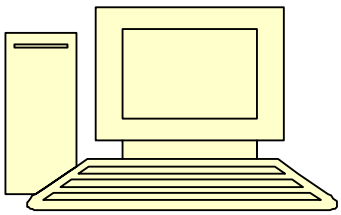
TRƯỜNG ĐẠI HỌC CẦN THƠ

Kỷ yếu:

HỘI NGHỊ TỔNG KẾT

5 NĂM NGHIÊN CỨU KHOA HỌC & ĐÀO TẠO

KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG



CẦN THƠ, 12/2011

MỤC LỤC

PHÂN TÍCH VÀ ĐÁNH GIÁ KẾT QUẢ ĐỔI MỚI TRONG ĐÀO TẠO VÀ NGHIÊN CỨU KHOA HỌC CỦA KHOA CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG GIAI ĐOẠN 2006 – 2011

Lê Quyết Thắng,
Khoa Công nghệ thông tin và Truyền thông, trường Đại học Cần Thơ

Tóm tắt

Được thành lập từ 1994, Khoa Công nghệ thông tin và Truyền thông (Khoa CNTT & TT) đã trải qua nhiều khó khăn và đã biết vượt qua và phát triển như ngày hôm nay. Nhìn lại thời gian qua, chúng ta có thể tổng kết lại thành 2 giai đoạn. Giai đoạn 1 (1991-2006): Xây dựng nguồn nhân lực và Ổn định chất lượng đào tạo, Giai đoạn 2 (2006-2011): Nâng cao chất lượng đào tạo và Ổn định định hướng Nghiên cứu Khoa học và Công nghệ. Phát triển tiếp theo của Khoa là Giai đoạn 3: Phát triển cân bằng Nghiên cứu khoa học và Đào tạo nguồn nhân lực, tiến tới đạt chuẩn Quốc gia ở mốc 2015 và chuẩn Khu vực ở mốc năm 2020. Bài viết này tổng kết những thành quả của Giai đoạn 2 nhằm xác định các thách thức và định hương các nhiệm vụ phải làm trong giai đoạn kế tiếp.

Bài viết sẽ tập trung trình bày các kết quả đổi mới cơ bản trong Đào tạo và Nghiên cứu khoa học của Khoa trong Giai đoạn 2. Đặc biệt bài viết có nêu bật sự chuyển mình trong nghiên cứu khoa học như là một sự hoàn thành "Bệ phóng", tạo điều kiện cho sự bay cao và nhanh của các nhóm nghiên cứu khoa học trong Khoa. Trong đó, nhóm nghiên cứu liên ngành DREAM cùng với phòng nghiên cứu Xử lý dữ liệu thông minh (LIDP) sẽ phải đóng vai trò chủ đạo. Một số kết quả nghiên cứu khoa học tiêu biểu được minh hoạ để chỉ rõ hơn các hướng nghiên cứu khác nhau. Các hướng này tuy khác nhau nhưng cùng mục tiêu: đưa công nghệ cao trong lĩnh vực Tin học đến gần với ứng dụng thực tế hơn, nhất là hỗ trợ cho công cuộc phát triển Kinh tế và Xã hội trên Vùng Đồng bằng sông Cửu long. Kết luận của bài viết đúc kết bốn đặc điểm chính tác động lên chất lượng đào tạo và ba tiêu chí đảm bảo sự bền vững của hoạt động nghiên cứu khoa học trong Khoa CNTT & TT.

Abstract

Established since 1994, the College of Information Technology and Communication (CICT) has gone through many difficulties, but has passed and grown today. Looking back the last five years, that can be summarized in two stages. The Phase 1 (1991-2006) is Building Human Resources and Stabilizing Training Quality. The Phase 2 (2006-2011): Improving Training Quality and Focusing Research on some important Fields. The next stage will be the Phase 3: Balancing Research and Human Resources Training, striving to meet national standards in the milestone 2015 and Regional (ASEAN) standards in the milestone 2020. This paper summarizes the results of Phase 2 to determine the challenges and tasks to be done in the next phase.

The content of this paper will focus on presenting the important innovations in Training and Scientific Research in the Phase 2. In particular, the growth of research in the last time has created favorable basis for the high and fast flying of the research teams of CICT. Among them, the DREAM team in collaboration with research Laboratory of Intelligent Data Processing (LIDP) will play a key role. Some typical results are shown to indicate more clearly the different research fields. Although different studies, but the same goal: bringing high technology in the Informatics closer to practical applications, especially to support the development of the economic and social problems on the Mekong Delta Region.

In the conclusions, this article summarized the four main characteristics affecting the quality of training and three criteria ensuring the sustainability of research activities in CICT.

1 TỔNG QUAN

Trường Đại học Cần thơ ngay từ khi ra đời cho đến nay vẫn được coi là trung tâm đào tạo và nghiên cứu khoa học về Nông nghiệp và Sinh học mạnh tầm Quốc gia. Mặc dầu vậy, đào tạo và nghiên cứu khoa học trong lĩnh vực Kỹ thuật-Công nghệ hỗ trợ cho Nông nghiệp và Sinh học còn yếu và trở thành sự bổ sung cấp thiết nhằm tạo sự đột phá trong phát triển Kinh tế và Khoa học Kỹ thuật trên toàn bộ Vùng Đồng bằng Sông Cửu long (ĐBSCL).

Nhằm đẩy nhanh tiến độ ứng dụng Công nghệ thông tin trên Vùng ĐBSCL, Trường Đại học Cần thơ đã quyết định xây dựng Khoa CNTT&TT. Được thành lập không chính thức từ năm 1990 thông qua Trung tâm Điện tử - Tin học và chính thức từ năm 1994, Khoa CNTT&TT đã trải qua 2 giai đoạn: Giai đoạn 1 (1991-2006): Xây dựng nguồn nhân lực và Ổn định đào tạo; Giai đoạn 2 (2006-2011): Nâng cao chất lượng đào tạo và Ổn định định hướng Nghiên cứu Khoa học và Công nghệ. Từ năm 2012 trở đi Khoa CNTT&TT sẽ bắt đầu một giai đoạn mới, Giai đoạn 3: Phát triển cân bằng Nghiên cứu khoa học và Đào tạo nhân lực, tiến tới đạt chuẩn Quốc gia ở mốc 2015 và chuẩn Khu vực ở mốc năm 2020.

Trong suốt Giai đoạn 2, KCNTT&TT đã thực hiện nhiều hoạt động mang tính đổi mới và đã về đích một cách xuất sắc để có thể chuyển sang Giai đoạn 3. Một số hoạt động nổi bật trong Giai đoạn 2 có thể được điểm qua như dưới đây.

i. Về Đào tạo: Nâng cao chất lượng đào tạo đại học là vấn đề đặt ra rất cấp thiết và cần phải được quy hoạch lâu dài và có lộ trình. Đổi mới chương trình đào tạo theo hệ thống tín chỉ là một trong những chính sách chiến lược. Nhận thức được tầm quan trọng này, Khoa đã nghiêm túc thực hiện các chỉ đạo của Trường theo lộ trình. Kết quả: Khoa đã là một trong một số khoa có thành tích luôn luôn thực hiện đúng hạn các kế hoạch đổi mới, như: Sửa đổi chương trình đào tạo, Viết lại các bài giảng, Đổi mới phương pháp giảng dạy và đánh giá sinh viên, Xây dựng chuẩn đầu ra, Áp dụng kiểm định và đánh giá đào tạo theo chuẩn AUN (ASEAN University Network).

ii. Về Đào tạo Sau đại học: Từ năm 2007, đã có tám tiến sĩ hoàn thành nghiên cứu ở nước ngoài và trở về tham gia công tác đào tạo Thạc sĩ Hệ thống thông tin (HTTT). Cùng lúc đó Khoa đã liên kết với trường Bách khoa Nantes (Pháp) để mở lớp Master chuyên ngành Khai khoáng dữ liệu cấp bằng Master Đại học Nantes. Chương trình liên kết này được triển khai theo hình thức đào tạo từ xa với sự hỗ trợ của các thiết bị chuyên dụng cho video-conference và phụ giảng tiếng Việt. Đến nay đã có 3 khoá tốt nghiệp chính quy (K14, K15, K16) và 2 khoá liên kết với tổng số với hơn 80 thạc sĩ HTTT và 9 thạc sĩ Nantes cùng với hơn 80 luận văn về các chuyên đề công nghệ tiên tiến cũng như khoa học máy tính. Kết quả này đã nâng cao uy tín của Khoa trong Trường cũng như trong toàn Vùng ĐBSCL.

iii. Về Nghiên cứu: trước năm 2006, các chủ đề nghiên cứu tản mạn và thường tập trung vào các đề tài ứng dụng xây dựng Hệ thống thông tin và E-learning. Những đề tài này phát sinh theo nhu cầu tin học hoá trong Trường, do đó chỉ thuần tuý là triển khai công nghệ mới trên khía cạnh quản lý mà chưa sâu. Bắt đầu từ 2006, nguồn nhân lực liên tục bổ sung bởi các giảng viên trẻ đã hoàn thành tiến sĩ và do vậy số lượng các bài báo chuyên ngành của Khoa tăng lên đáng kể (hàng năm trên dưới 30 bài báo được đăng hoặc báo cáo chính thức trên các tạp chí và hội nghị khoa học quốc tế và quốc gia). Ngoài ra một loạt các đề tài nghiên cứu cấp cơ sở, cấp tỉnh và cấp Nhà nước đã được thực hiện và nghiệm thu đúng hạn cho thấy Khoa CNTT&TT đã có thể tập trung nguồn lực vào ổn định định hướng nghiên cứu nhằm phát huy tối đa kết quả và thu hút được tài trợ. Đặc biệt đề tài nghiên cứu cấp Nhà nước KC.01.15/06-10 (Lê Quyết Thắng et al. [1]) đã tạo tiền đề cho việc hợp tác có hiệu quả với IRD (Pháp) và IFI (Hà nội). Đó là sự thành công

của việc thành lập được nhóm nghiên cứu liên ngành DREAM (Decision-support Research for Environmental Applications and Models) về Mô hình hóa và Mô phỏng các phương án đối phó với Biến đổi khí hậu trên Vùng ĐBSCL. Để tham gia hiệu quả các dự án nghiên cứu của nhóm DREAM, Khoa đã xác định hướng nghiên cứu chủ đạo: Mô hình hoá và Mô phỏng các vấn đề liên quan đến Môi trường, Kinh tế, Sinh học và Kỹ thuật; thông qua Xử lý thông minh các chùm dữ liệu tương ứng và qua đó chứng thực tính đúng đắn của Mô hình và Mô phỏng. Hướng nghiên cứu chủ đạo này là cơ sở để thành lập Phòng thí nghiệm Xử lý dữ liệu thông minh (Laboratory of Intelligent Data Processing - LIDP). LIDP đã tập trung được các kết quả nghiên cứu mới nhất về Mô hinh hoá và Mô phỏng trên một số tình huống của lan truyền dịch bệnh và biến đổi khí hậu trên Vùng ĐBSCL thông qua kết quả Xử lý dữ liệu thông minh.

iv. Về Hợp tác: Hợp tác để nâng cao chất lượng đào tạo và nghiên cứu là mục tiêu tiên quyết của Khoa. Khoa đã nhiều năm tìm kiếm đối tác trong nước cũng như quốc tế. Đa số đối tác dừng lại ở mức độ thăm dò, một số đối tác đòi hỏi chi phí cao và tất nhiên cũng có một số đối tác thật sự quan tâm đến hợp tác liên kết với một số ưu tiên dành cho Khoa. Các hợp tác hiệu quả cao vẫn là các hợp tác quốc tế.

Hợp tác quôc tế với trường Đại học Bách khoa Nantes (Pháp) từ 2007 là thành công đầu tiên. Hợp tác này cho phép chúng ta tổ chức lớp học Master song ngữ (tiếng Pháp) trực tuyến và dự kiến sẽ nâng dần lên mức độ Đào tạo liên kết cấp bằng đôi theo chuẩn châu Âu (Erasmus Mundus). Như vậy sinh viên theo học lớp này khi tốt nghiệp sẽ nhận bằng Master của trường Nantes. Nếu như chương trình này được tích hợp vào chương trình đào tạo thạc sĩ Hệ thống thông tin của Khoa thì chất lượng đào tạo thạc sĩ của Khoa sẽ được nâng cao đáng kể và trực tiếp được kiểm định bởi các trường lớn của châu Âu.

Hợp tác quốc tế với IRD (Institut for Research & Development, Pháp) từ 2010 đã giúp chúng ta xây dựng thành công nhóm nghiên cứu liên ngành DREAM. Nhóm nghiên cứu này cùng với LIDP đã thu nhận 3 nghiên cứu sinh và các thạc sĩ năm cuối để thực hiện một loạt các đề tài theo định hướng liên ngành: Môi trường, Mô hình toán và Mô phỏng. Hiện đã có 3 bài báo được báo cáo ở nước ngoài và nhiều bài báo khác được báo cáo trong nước.

Hợp tác quốc tế với Đại học Quebec à Trois Rivière (UQTR, Canada) bắt đầu từ 2011. Hợp tác này cho phép chúng ta nâng cấp chương trình đào tạo đại học song ngữ để cấp bằng đôi, đồng thời tạo điều kiện cho đội ngũ giảng viên và quản lý dần dần làm chủ được các hoạt động đào tạo trong môi trường tiên tiến.

Hợp tác quốc tế với Đại học Kemi-Tornio (Phần lan) thông qua một dự án của chính phủ Phần lan và được thực hiện trong 2 năm 2009-2010. Dự án này giúp chúng ta nhận một số sinh viên của Phần lan sang học chung với sinh viên Đại học Cần thơ và Đồng tháp trong 4 tháng dưới sự hướng dẫn của các giáo sư Phần lan. Sau đó chúng ta gửi một số sinh viên của Đại học Cần thơ và Đồng tháp sang Phần lan học một số học phần của trường đối tác có sự tham gia giảng dạy của 2 giảng viên của Đại học Cần thơ. Dự án này đã kết thúc và kết quả của nó giúp đối tác tiếp tục trình dự án mới để xây dựng chương trình đào tạo liên kết cấp bằng đôi trong vài năm tới.

v. Về Cơ sở vật chất và Môi trường: Cơ sở vật chất khang trang và môi trường sạch đẹp là yếu tố quan trọng hỗ trợ hiệu quả cho việc học tốt và dạy tốt. Bắt đầu từ những năm đầu 2000, Khoa đã tiếp nhận một số dự án quan trọng nhằm vào nâng cấp cơ sở vật chất, thiết bị đào tạo và cảnh quan môi trường. Từ năm 2003, hầu như tất cả các phòng học và làm việc trong Khoa đều được tu bổ và bố trí lại chức năng cho phù hợp với các nhiệm vụ đào tạo và nghiên cứu về Công nghệ thông tin. Đặc biệt, việc phủ sóng WIFI toàn bộ Khu 3, thiết kế và nâng cấp "sân cỏ trung tâm" thành công viên Khu 3, rất được sinh viên ưu

chuộng trong việc học nhóm và nghỉ ngơi sau những giờ học tập căng thẳng. Yếu tố này đã đóng góp không nhỏ vào sự thành công của các hoạt động đào tạo và nghiên cứu của Khoa.

Nhằm phát huy thế mạnh để tiếp tục phát triển trong thời gian tới, bài viết này sẽ phân tích và đánh giá các hoạt động tiêu biểu trong Đổi mới đào tạo và Nâng cao chất lượng Nghiên cứu khoa học của khoa CNTT&TT trong Giai đoạn: 2006-2011.

2 ĐỔI MỚI ĐÀO TẠO

2.1 Đào tạo CNTT theo hệ tín chỉ

Hệ thống đào tạo lấy sinh viên làm trung tâm đã chứng minh tính ưu việt khi tất cả các trường đại học đẳng cấp trên thế giới đều dựa trên nền tảng này. Vì vậy không thể nâng cao chất lượng đào tạo nếu như không thay đổi từ nhận thức đến việc thay đổi hệ thống đào tạo theo niên chế đồng thời với công tác kiểm định và đánh giá. Việc chuyển hệ thống đào tạo niên chế sang hệ thống tín chỉ là cấp thiết vì tất cả các đặc điểm của hệ thống này đều lấy tiêu chí Sinh viên là trung tâm.

Khoa CNTT & TT đã nhận thức rõ vấn đề và quyết tâm thực hiện thành công lộ trình chuyển đổi sang Hệ thống tín chỉ theo đúng thời gian quy định. Những công việc phức tạp trong quá trình chuyển đổi đã đòi hỏi toàn bộ giảng viên và cán bộ viên chức trong Khoa phải tham gia. Chúng ta đã hoàn thành hàng loạt các yêu cầu chuyển đổi:

i- Xác định lại tỷ lệ kiến thức: Cơ bản, Cơ sở và Chuyên ngành.

ii- Sắp xếp lại các môn học cơ sở và chuyên ngành sao cho phù hợp với lượng kiến thức quy định, trong đó có nhiều môn chuyên ngành mới đã được bổ sung.

iii- Điều chỉnh hoặc viết mới đề cương và bài giảng môn học.

iv- Điều động hầu như toàn bộ giảng viên làm Cố vấn học tập với nhiệm vụ: hướng dẫn sinh viên làm quen với vai trò Trung tâm của mình, gồm: lập kế hoạch học tập, đăng ký môn học sao cho vừa sức mình, đánh giá môn học.

v- Xây dựng chuẩn đầu ra có tham khảo chuẩn đầu ra của một số trường lớn về đào tạo Công nghệ - Kỹ thuật của Mỹ: Khi hoàn thành chuẩn đầu ra, chúng ta đã nhận thấy: cần tiếp tục điều chỉnh chương trình đào tạo và cải tiến quy trình đào tạo kỹ năng thực hành. Kỹ năng thực hành của sinh viên tốt nghiệp là một thách thức lớn cho các cố gắng nâng cao chất lượng đào tạo. Đây là vấn đề phức tạp đòi hỏi có sự liên kết và hợp tác chặt từ phía doanh nghiệp CNTT theo đúng nghĩa (Đầu tư để có Sản phẩm), đồng thời tiếp tục cải tiến đồng bộ quy trình quản lý chương trình đào tạo theo tín chỉ đối với các môn học có thực hành.

vi- Thực hiện kiểm định và đánh giá theo chuẩn AUN: Mặc dầu đây là công việc mới và khó, nhưng chúng ta đã hoàn tất công tác đánh giá chương trình đào tạo chuyên ngành Hệ thống thông tin (HTTT) và hiện đã chuyển sang đánh giá các chuyên ngành mới: Kỹ thuật phần mềm (KTPM), Mạng máy tính và Truyền thông (MTT & TT) và sang năm học mới sẽ bắt đầu đánh giá chuyên ngành còn lại: Khoa học máy tính (KHMT).

2.2 Kết quả đánh giá theo chuẩn AUN

Để có thể đánh giá, việc đầu tiên là phải thu thập thông tin. Phương pháp lấy thông tin như sau:

- Điều tra, thu thập ý kiến về chương trình đào tạo Hệ thống thông tin trên 3 nhóm đối tượng (Sinh viên, Cựu sinh viên, và Nhà tuyển dụng).
- Thu thập ý kiến phản hồi của SV dựa trên Phiếu nhận xét học phần do Trung tâm Đảm bảo chất lượng (TT ĐBCL) cung cấp. Trung bình mỗi HK có khoảng 23 học phần được đánh giá với khoảng 1200 phiếu nhận xét.

Từ các thông tin đã thu thập, chương trình đào tạo HTTT đã được coi là chương trình đạt điểm 5.11, xếp loại Trung bình khá. Theo đánh giá, chương trình HTTT có điểm mạnh trên các tiêu chí về Hỗ trợ sinh viên trong CSVC và thiết bị cũng như dịch vụ (đạt 6 điểm), nhưng điểm yếu là sự hợp tác với doanh nghiệp trong việc thiết kế chương trình cũng như điều chỉnh đề cương (đạt 4 điểm).

Như vậy có thể suy ra rằng các chương trình đào tạo của Khoa có chung điểm yếu, đó là sự hợp tác với các doanh nghiệp để điều chỉnh chương trình đào tạo hướng tới kỹ năng và gần gũi hơn với thực tế tuyển dụng.

2.3 Kết quả đổi mới Đào tạo

Kết quả nổi bật nhất trong công tác đổi mới: sinh viên năm 1 có kế hoạch mẫu và bắt buộc, sinh viên năm 2 và các năm sau đó đã quen với việc lập kế hoạch, điều chỉnh và đăng ký học phần theo năng lực của bản thân.

Giáo viên giảng dạy học phần có bài giảng, có tài liệu tham khảo có đánh giá giữa kỳ và cuối kỳ. Đặc biệt, hầu hết các giảng viên CNTT đều áp dụng phương pháp giảng dạy tích cực, lấy sinh viên làm trung tâm, tạo điều kiện cho sinh viên chủ động trong tự học và học nhóm.

Chương trình đào tạo được đổi mới hàng năm, nhất là các học phần cũng được điều chỉnh sau khi có những góp ý của sinh viên, cựu sinh viên và nhà tuyển dụng.

Kỹ năng thực hành đã được đổi mới nhiều nhưng vẫn là một thách thức lớn khi rất nhiều ý kiến của sinh viên, cựu sinh viên và đặc biệt các nhà tuyển dụng đánh giá không cao khả năng này của sinh viên CNTT khi ra trường.

2.4 Hướng phát triển

Rõ ràng thách thức cơ bản trong đào tạo là: Cân bằng giữa Kỹ năng thực hành và Khả năng tư duy. Khoa cần phải làm nhiều việc để vượt qua thách thức này. Những tiếp cận có thể giúp chúng ta đạt được mục tiêu:

- Tăng cường hợp tác với các doanh nghiệp lớn hoạt động trong lĩnh vực CNTT & TT: Nghiên cứu công nghệ mới theo yêu cầu của doanh nghiệp, đào tạo cấp chứng chỉ doanh nghiệp cùng với các loại chứng chỉ chuẩn quốc gia và quốc tế.
- Xây dựng các đề án đào tạo liên kết với trường uy tín trong Khu vực, tập trung và các chương trình đào tạo có trao đổi giáo viên, sinh viên, cấp 2 bằng.
- Chú ý quan tâm đến các chương trình đào tạo đã được kiểm định chuẩn quốc tế ABET. Nếu có tính khả thi và có cơ hội thì tiếp cận trao đổi và trình đề án kịp thời.

3 ĐÀO TẠO SAU ĐẠI HỌC

3.1 Quá trình chuẩn bị đội ngũ

Đội ngũ giảng dạy với học vị cao, từng trải nghiệm những thách thức trong học tập và nghiên cứu ở nước ngoài là chìa khoá để nâng cao chất lượng đào tạo và nghiên cứu. Trước năm 2006, Khoa tập trung gửi giảng viên trẻ đi học tập ở nước ngoài, song song với việc đó là ổn định tất cả các quy trình đào tạo kỹ sư ở các chuyên ngành của Công nghệ thông tin. Bắt đầu từ 2006, một số đông tiến sĩ hoàn thành nhiệm vụ học tập và nghiên cứu ở nước ngoài đã trở về và cũng là lúc Khoa đẩy mạnh các dự án đào tạo thạc sĩ và nghiên cứu khoa học. Đây cũng là chìa khoá chính tạo chỗ đứng ổn định cho các giảng viên đã được đào tạo trình độ cao ở nước ngoài trở về. Thực tế là Khoa đã bắt đầu đào tạo thạc sĩ từ năm 2007 với 2 hình thức: đào tạo chính quy thạc sĩ HTTT và đào tạo liên kết từ xa với Đại học Nantes; sau đó một loạt các đề tài Nhà nước cũng như dự án Quốc tế đã được duyệt và hoàn thành.

3.2 Đa dạng hóa phong cách giảng dạy

Đào tạo Sau đại học nói riêng và đào tạo nói chung luôn đặt ra bài toán về trao đổi phong cách giảng dạy với giảng viên quốc tế và so sánh. Qua đó tự các giảng viên của Khoa cũng như sinh viên phải sửa đổi phương pháp dạy và học của mình ngày càng chuẩn mực và phù hợp với mỗi người.

Ngay từ khoá đào tạo thạc sĩ đầu tiên (K14), Khoa đã chủ động mời một số giảng viên nước ngoài cũng như trong nước tham gia giảng dạy và hướng dẫn đề tài thạc sĩ. Chẳng hạn, GS Alain Boucher(1) giảng dạy về Thị giác máy tính (Computer Vision) đã cho thấy phương pháp dạy bằng tiếng Anh kết hợp với động tác hình thể để minh hoạ những khái niệm khó. Trong khi đó GS Alexis Drogoul(2) giảng dạy môn Mô hình hoá và Mô phỏng (Modelling and Simulation) đã cho phép sinh viên trực tiếp thực hành để hiểu rõ hơn trên phần mềm GAMA chuyên mô phỏng trên nền GIS.

3.3 Kết quả

Kết quả của công tác đào tạo thạc sĩ cho thấy rất nhiều luận văn thạc sĩ đã đạt được chuẩn mực nghiên cứu và có thể tiếp tục nghiên cứu cao hơn. Bằng chứng là đã có gần 10 bài báo được các thạc sĩ viết và công bố tại các hội nghị khoa học Quốc tế (RIVF'10(3), ICTACS'10(4), ...) và trong nước (@Hưng yên(5), ICTFIT'10(6), ICT.rda'10(7), @Cần thơ(8)...). Ngoài ra khoảng 95% các luận văn đều đạt loại khá trở lên và các sinh viên thực hiện các luận văn này đã được tin tưởng hơn trong công tác giảng dạy hoặc trở thành các CIO (Phụ trách Quản trị thông tin) tốt trong các công sở hoặc doanh nghiệp lớn.

Ghi chú

(1) Alain Boucher: Giáo sư giảng dạy cao học tại IFI (Hà nội) và hướng dẫn nghiên cứu về Nhận dạng ảnh tại Viện MICA (trường Đại học Bách khoa, Hà nội).

(2) Alexix Drogoul: Phụ trách nghiên cứu và phát triển tại phòng Nghiên cứu của IRD (Pháp) thuộc IFI (Hà nội). Năm 2012, GS sẽ trực tiếp là cố vấn cho nhóm nghiên cứu liên ngành DREAM, Đại học Cần thơ.

(3) RIVF'10: Viết tắt từ "Research, Innovation and Vision for the Future" tên của Hội thảo khoa học quốc tế năm 2010 về CNTT & TT tại Việt nam. Hội thảo này đã được chấp nhận đăng ký vào danh mục các hội thảo quốc tế của IEEE. RIVF'05 được tổ chức tại trường Đại học Cần thơ năm 2005.

(4) ICTACS'10: Viết tắt từ "International Conference on Theories and Applications of Computer Science" tên của Hội thảo khoa học năm 2010 về Khoa học máy tính. Hội thảo này được sáng lập bởi Khoa CNTT, trường Đại học Khoa học tự nhiên TP Hồ Chí Minh. Hội thảo này đã được tổ chức liên tục 4 năm và năm 2010 được tổ chức tại Trường Đại học Cần thơ. Các báo cáo của hội thảo được phản biện bởi các chuyên gia tin học quốc tế và được đăng trong số đặc biệt hàng năm về Khoa học máy tính của tạp chí Khoa học và Công nghệ, Viện Khoa học và Công nghệ Việt nam.

(5) @Hưng yên: Viết tắt từ "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông" tên của Hội thảo Quốc gia hàng năm về Công nghệ thông tin và Truyền thông do Viện Công nghệ thông tin tổ chức tại trường Đại học Hưng yên năm 2010. Các báo cáo sau khi được phản biện được đăng trong Kỷ yếu Hội thảo Quốc gia hàng năm và được xuất bản bởi nhà Xuất bản Khoa học và Kỹ thuật. @Cần thơ là Hội thảo Quốc gia được tổ chức tại trường Đại học Cần thơ năm 2011.

(6) ICTFIT'10: Viết tắt "Hội thảo Công nghệ thông tin" năm 2010 của Khoa Công nghệ thông tin, trường Đại học Khoa học tự nhiên, TP Hồ Chí Minh. Các bài báo của Hội thảo này sau khi được phản biện được in trong tạp chí "Tuyển tập công trình nghiên cứu Công nghệ thông tin và Truyền thông", nhà Xuất bản Khoa học và Kỹ thuật.

(7) ICT.rda'10: Viết tắt "Hội thảo Khoa học Quốc gia" tổng kết các kết quả nghiên cứu hàng năm của chương trình nghiên cứu Khoa học và Công nghệ Nhà nước trọng điểm KC.01/06-10. ICT.rda'10 là Hội thảo tổng kết Chương trình nghiên cứu trọng điểm này năm 2010. Các bài báo của Hội thảo sau khi được phản biện được đăng tại tạp chí Công nghệ thông tin và Truyền thông của Bộ thông tin và Truyền thông.

(8) @Cần thơ: Xem chú thích (5).

3.4 Hướng phát triển

Những ưu thế và khiếm khuyết của chương trình đào tạo thạc sĩ cho thấy cần phải từng bước chuẩn mực hoá chương trình đào tạo Thạc sĩ. Đầu ra của một thạc sĩ cần phải có kết quả là công trình được báo cáo trên một hội nghị Khoa học hoặc được đăng trên một tạp chí khoa học. Ngoài ra việc trình đề án Đào tạo tiến sĩ Khoa học máy tính để tiếp nhận

các thạc sĩ ưu tú là rất cần thiết và hỗ trợ rất lớn cho nguồn nhân lực nghiên cứu trình độ cao trên Vùng ĐBSCL.

4 NGHIÊN CỨU KHOA HỌC

4.1 Những thách thức trong nghiên cứu Khoa học

4.1.1 Môi trường nghiên cứu

Bất kỳ ai được đào tạo tiến sĩ ở nước ngoài trở về Việt nam làm việc đều cảm thấy rất khó tiếp tục nghiên cứu bởi 2 lý do (không đề cập lý do cơ bản về thu nhập):

- Thiếu nhóm làm việc chuyên sâu cùng một chủ đề.

- Thiếu chính sách chung về hỗ trợ tài chính đi báo cáo nước ngoài.

Vì vậy, nếu muốn tiếp tục nghiên cứu thì người nghiên cứu vẫn phải tìm học bổng đi nước ngoài hoặc tham gia nghiên cứu trong một dự án tầm quốc tế. Điều này dẫn đến các số liệu thống kê về tình hình nghiên cứu ở Việt nam có chỉ số về bài báo quôc tế rất thấp. Hậu quả của nó là sản phẩm công nghệ mới cũng như số lượng sản phẩm có đăng ký sở hữu trí tuệ cũng rất thấp trong khu vực Đông Nam Á.

Rõ ràng muốn xây dựng một môi trường nghiên cứu khoa học tốt, điều đầu tiên phải khắc phục 2 cái thiếu trên. Đây chính là thách thức quan trọng nhất đối với bài toán nâng cao năng lực nghiên cứu mặc dù đã có một đội ngũ đủ mạnh về NCKH.

4.1.2 Tiếp cận liên ngành

Nếu điểm qua tất cả các tạp chí khoa học trong nước về Khoa học cơ bản và cụ thể về lĩnh vực CNTT, có thể nhận xét sơ bộ: thiếu tiếp cận nghiên cứu liên ngành. Mặc dù trong các lĩnh vực Khoa học ứng dụng cũng có các bài báo viết về CNTT, nhưng phần lớn đó là những nghiên cứu nội bộ, không phản ánh được tiếp cận chuyên môn từ 2 lĩnh vực: Ứng dụng và CNTT. Có nghĩa là cùng một kết quả nghiên cứu, nhìn trên khía cạnh CNTT ta có một công trình mới (mô hình và giải thuật hiệu quả), nhưng nếu nhìn từ góc độ ứng dụng (được hỗ trợ phần mềm) thì ta cũng có một công trình mới (mô hình mới, ứng dụng hiệu quả cao).

Tiếp cận liên ngành tại Việt nam cũng không phải dễ xây dựng. Đối với NCKH về CNTT, mặc dầu xung quanh các nhóm nghiên cứu CNTT có rất nhiều yêu cầu nghiên cứu ứng dụng cần sự hỗ trợ sâu của CNTT. Thực tế cho thấy phần lớn các nghiên cứu ứng dụng chỉ sử dụng sản phẩm CNTT như sự hỗ trợ mà thiếu nghiên cứu chuyên sâu về CNTT.

Nghiên cứu CNTT thật sự hiệu quả nếu xuất phát từ tiếp cận liên ngành. Vì vậy cần phải có một cơ chế "bắt tay" hiệu quả giữa các nhà khoa học của các ngành khoa học ứng dụng và CNTT.

4.1.3 Kinh phí nghiên cứu và dự án khả thi

Để có được sự hỗ trợ tài chính cho nghiên cứu khoa học cần phải xây dựng dự án. Nhưng nghiên cứu khoa học thuần tuý lại có độ rủi ro cao. Vì vậy muốn có đề tài khả thi, các nhà khoa học phải có sẵn nhiều ý tưởng để một mặt có thể chọn được ý tưởng có độ rủi ro thấp, mặt khác phải nắm thời cơ và thời điểm thuận lợi để đưa ý tưởng có độ rủi ro thấp thành dự án thành công.

4.2 Sự chuyển mình trong Nghiên cứu Khoa học

4.2.1 Tạo môi trường nghiên cứu

Để có thể tạo ra không khí năng động trong nghiên cứu khoa học, việc đầu tiên phải tạo được môi trường thuận lợi, đồng thời khuyến khích các hoạt động NCKH của đội ngũ trẻ, những người luôn có hứng thú sáng tạo và đầy tiềm năng nghiên cứu. Khảo sát sơ bộ

các chủ đề nghiên cứu trên các bài báo công bố trên quốc tế và trong nước của một nhóm các tiến sĩ trẻ và nghiên cứu sinh là giảng viên của Khoa, chúng ta đã xác định được thế mạnh hiện tại của Khoa về NCKH trên các lĩnh vực: 1- Ứng dụng công nghệ E-learning, 2- Khai khoáng dữ liệu (Data Mining) và Hỗ trợ ra quyết định, 3- Tích hợp các hệ thống thông tin (Infomation System Integration) và Tương tác người máy, 4- Nhận dạng ảnh (Image Recognition), 5- Kiến trúc hướng dịch vụ (SOA) và đám mây (Cloud Computing), 6- Điều khiển thiết bị thông qua các tín hiệu sóng (Signal Processing for Device Controler).

Ngay từ năm 2003, Khoa đã được Trường giao phụ trách hướng nghiên cứu trọng điểm về Đào tạo từ xa. Sau một thời gian chuẩn bị, năm 2006, Nguyễn Văn Linh cùng nhóm nghiên cứu e-learning đã thực hiện đề tài cấp Bộ trọng điểm "Thiết kế và xây dựng chương trình e-learning trong đào tạo kỹ sư tin học liên thông từ cử nhân cao đẳng tin học", mã số B2006-16-44TD. Kết quả của đề tài là nhóm đã xây dựng được hệ thống e-learning trên nền Moodle, đồng thời bổ sung một số quy trình và modul. Chẳng hạn: Xác định cấu trúc một bài giảng tiền SCORM (Sharable Content Object Reuse Model) hay cấu trúc từng trang theo trang màn hình, Chuyển bài giảng có tiền cấu trúc SCORM có sang dạng chuẩn SCORM. Hệ thống này phục vụ rất hiệu quả cho chương trình đào tạo theo hệ thống tín chỉ.

Đòn bẩy cho việc tạo Môi trường nghiên cứu từ năm 2009 là việc tập trung được lực lượng nghiên cứu tham gia đề tài NCKH cấp Nhà nước: "**Nghiên cứu xây dựng các hệ thống thông tin hỗ trợ việc phòng chống dịch bệnh cây trồng và thuỷ sản cho vùng kinh tế trọng điểm** (mã số KC.01.15/06-10)". Tham gia đề tài còn có sự hỗ trợ của IRD (Pháp) và IFI (Hà nội). Sự có mặt của IRD đã tạo điều kiện cho chúng ta xây dựng thành công nhóm nghiên cứu liên ngành DREAM trong dự án quốc tế JEAI. Kết quả này thật sự đã tạo cho chúng ta môi trường thuận lợi, đưa các kết quả nghiên cứu của Khoa ra trường Quốc tế. Mặt khác việc hợp tác liên ngành cũng như hợp tác Vùng đã giúp chúng ta xây dựng được một số dự án quan trọng đến gần hơn với nhu cầu tin học hoá trong Vùng. Chẳng hạn, đề tài cấp Nhà nước về Cổng thông tin Khoa học công nghệ phục vụ nhu cầu ứng dụng Khoa học công nghệ trong Vùng ĐBSCL, đề tài Chế tạo thiết bị giảng dạy hỗ trợ giảng viên đứng bảng và giảng dạy trực quan ... là những đề tài phát sinh từ yêu cầu thực tế.

4.2.2 Định hướng nghiên cứu

Kết quả của đề tài KC.01.15/06-10 và sự thành lập nhóm nghiên cứu liên ngành DREAM đã chứng tỏ tính hiệu quả của định hướng nghiên cứu: Xử lý dữ liệu thông minh (Intelligent Data Processing) hướng tới xác nhận tính đúng đắn của các mô hình (Model Validity) trong Môi trường, Sinh học, Kinh tế, Giáo dục, Y tế và An ninh thông tin.

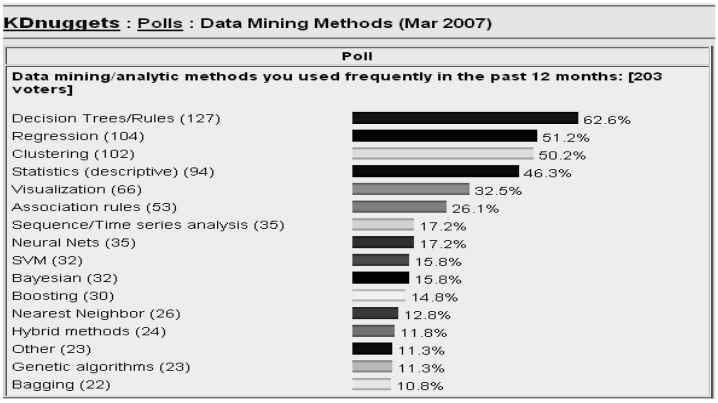
Thật vậy, trước đó hàng loạt các bài báo được công bố mang tính tản mạn, không tập trung và chỉ giải quyết các vấn đề đặt ra mang tính hàn lâm. Nhưng cũng với những kết quả đó, chẳng hạn như các cải tiến giải thuật trong Data Mining hoặc chỉ đơn giản sử dụng các giải thuật tiêu biểu, nếu đưa vào chứng thực các mô hình có tham số (Parametric Model Validity) hoặc phi tham số (Non Parametric Model Validity) với hệ thống các bộ luật trong Môi trường, Sinh học, Kinh tế, Giáo dục, Y tế và thậm chí trong An ninh thông tin, thì chúng ta đã có thể đi vào một không gian mênh mông đầy tiềm ẩn các sáng tạo kỳ thú.

Đề tài cấp Nhà nước KC.01.15/06-10 [1]: Xây dựng hệ thống hỗ trợ ra quyết định trong phòng chống dịch hại trên cây lúa và nuôi trồng tôm và cá là một minh hoạ tiểu biểu của định hướng trên. Chúng ta phải giải quyết bài toán đầu tiên: chứng thực hàng loạt các kinh nghiệm của các chuyên gia trong phòng chống dịch hại dưới dạng luật đồng thời bổ

sung các bộ luật mới (Mô hình phi tham số). Công việc này phải tiến hành đồng thời hàng loạt các công việc: Thu thập số liệu thực tế về dịch hại, Chuẩn mực số liệu, Lưu trữ số liệu, Thống kê số liệu, Khai thác dữ liệu (Data Mining) và cuối cùng kết xuất các bộ luật (kiểm chứng kiến thức đã có và bổ sung luật mới). Các bài toán tiếp theo là ứng dụng công nghệ mới để trình diễn các bộ luật dưới nhiều dạng khác nhau: công cụ Giao tiếp với người sử dụng (Cổng thông tin và Dịch vụ Web sử dụng Service Oriented Architecture - SOA), công cụ Hỗ trợ dự báo dịch hại (Dich vụ Web sử dụng Bayesien Network), công cụ Hiển thị bản đồ trực tuyến (WebGIS), công cụ Mô phỏng dịch hại (Dịch vụ Web sử dụng GIS-Based Multiagent Simulations), công cụ Hỗ trợ tìm kiếm thông tin theo ngữ nghĩa (Semantic Retrieval), công cụ Thống kê trực tuyến (OnLine Analytical Processing - OLAP), công cụ Tư vấn chẩn đoán và trị bệnh trong nông nghiệp (Decision Support System - DSS), công cụ hỗ trợ Viết báo cáo cộng tác với công thức toán động (WikiReport).

4.2.3 Giới thiệu các nghiên cứu tiêu biểu

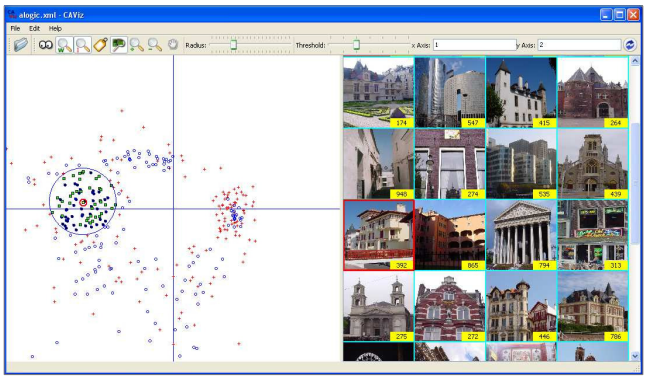
Hiện tại có rất nhiều các giải thuật khác nhau được sử dụng trong Khai thác dữ liệu, nên việc tổng hợp các giải thuật và xếp hạng theo nhu cầu sử dụng giúp một cái nhìn tổng quát và sau đó hỗ trợ lựa chọn một giải thuật phù hợp cũng rất cần thiết. Đỗ Thanh Nghị [2] đã liệt kê các giải thuật được ưu chuộng theo tỷ lệ ứng dụng *(Hình 1)*. Sau đó tác giả đã phân tích ý nghĩa của các giải thuật tiêu biểu để hướng dẫn người sử dụng lựa chọn giải thuật phù hợp với các tình huống thực tế.



Hình 1: Xếp hạng giải thuật theo tỷ lệ sử dụng

Các tác giả Nguyễn Thái Nghe và Đỗ Thanh Nghị [3] đã đề xuất phương pháp bổ sung cá thể hiếm xuất hiện khi thực hiện khám phá tri thức trên quần thể với nhiều lớp không đồng đều về lực lượng cá thể. Từ đó có thể ước lượng một cách tốt nhất ngưỡng phân lớp trên mạng Bayes.

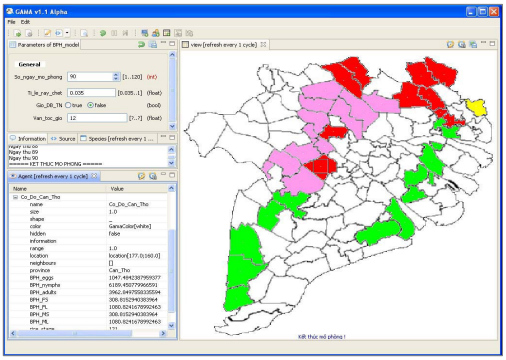
Phạm Nguyên Khang [4] đã sử dụng phương pháp Phân tích tương ứng (Correspondant Analysis - CA) đồng thời xây dựng giải thuật tích hợp nhiều môi trường khác nhau (CA cải tiến và khai thác Graphic Card) để tăng tốc tối đa quá trình khám phá tri thức, phân lớp, lấy chỉ mục theo nội dung ảnh trên một kho dữ liệu với hàng triệu ảnh. Kết quả phân lớp rất hiệu quả có thể thấy trong *Hình 2*.



Hình 2: Kết quả phân lớp ảnh theo CA

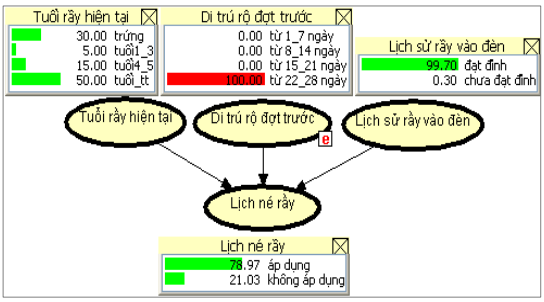
Từ khi đề tài KC.01.15/06-10 được triển khai đã có nhiều kết quả về Mô phỏng đa tác tử dựa trên nền GIS, nổi bật là kết quả của Phan Huy Cường [5]. Tác giả đã nghiên cứu khá kỹ tình hình diễn biến dịch hại rầy nâu khi chúng xuất hiện và phân tán trên Vùng ĐBSCL.

Dựa trên kết quả bắt rầy nâu từ bẫy đèn phân bố trên Vùng ĐBSCL, tác giả đã điều chỉnh lại một số thông số của mô hình phát tán rầy nâu. Với kết quả này, tác giả đã mô phỏng thành công sự lan truyền rầy nâu theo hướng gió, độ ẩm, ... và minh hoạ qua bản đồ phát tán rầy nâu *(Hình 3)*. Kết quả minh hoạ được giới chuyên môn về dịch hại rầy nâu đánh giá cao.



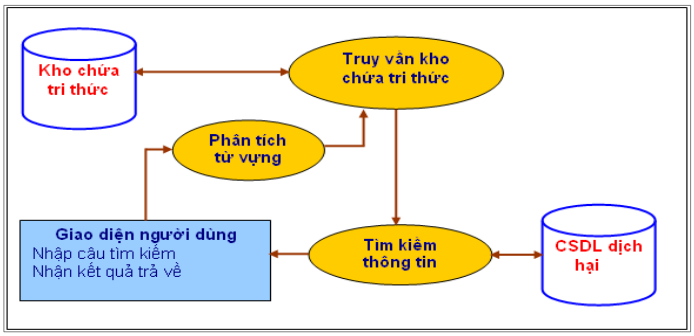
Hình 3: Mô phỏng sự phát tán rầy nâu trên Vùng ĐBSCL

Cũng từ bảng số liệu bẫy đèn bắt rầy nâu, Vũ Duy Linh [6] đã xây dựng thành công mô hình dự báo sự xuất hiện rầy nâu dựa trên mạng Bayesien *(Hình 4)* . Kết quả này có thể hỗ trợ người nông dân tiến hành các kế hoạch xuống giống theo phương pháp "né rầy" rất hiệu quả.



Hình 4: Minh hoạ mạng Bayes dự báo dịch rầy nâu

Các thông tin về các phương pháp phòng trừ dịch hại trong sản xuất nông nghiệp trên mạng có nhiều và rất hữu ích cho người nông dân. Nhưng đối với nông dân, việc nhớ và gõ từ khoá chuyên môn là rất khó khăn. Phan Thượng Cang [7] đã xây dựng dịch vụ Web ngữ nghĩa hỗ trợ người nông dân tìm kiếm thông tin về dịch hại bằng chính ngôn ngữ nói của mình theo sơ đồ trong *Hình 5*.



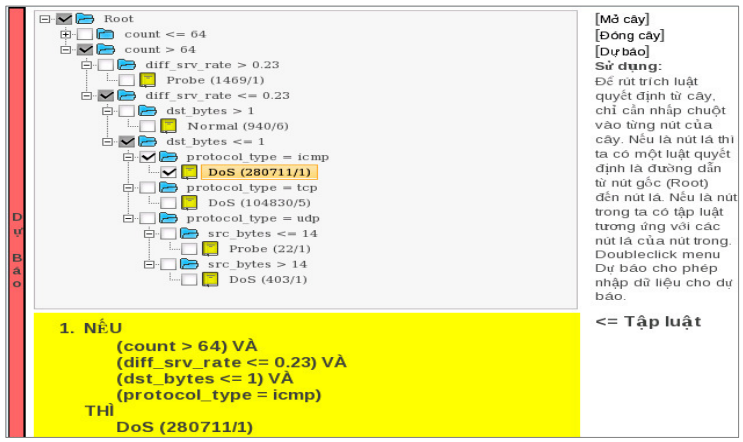
Hình 5: Kiến trúc mô hình tìm kiếm ngữ nghĩa

Vấn đề nhận dạng ký tự viết tay trực tuyến trên ô quy định đã được nghiên cứu bởi Trần Cao Đệ [8]. Tác giả sử dụng mô hình Markov ẩn để biểu diễn phân bố của ký tự viết tay, sau đó sử dụng giải thuật SVM (Support Vector Machine) để phân lớp và nhận dạng. Kết quả nhận dạng theo quy trình này cho thấy tăng độ chính xác so với các phương pháp trước đó.

Làm việc trên mạng và theo nhóm đã dần trở thành phương pháp làm việc hiệu quả. Người ta đã tạo ra các phần mềm Viết báo cáo cộng tác trên mạng để hỗ trợ các nhóm làm việc như vậy. Nhưng một dịch vụ Web hỗ trợ viết báo cáo và bổ sung các tính năng mới là cần thiết trong môi trường hợp tác trực tuyến như hiện nay. Lê Quyết Thắng cùng nhóm tác giả [9] đã nghiên cứu và tạo ra dịch vụ Web WikiReport dùng cho các nhóm viết báo cáo cộng tác theo nhiều cấu trúc khác nhau (cấu trúc Cây, cấu trúc Đẳng cấp hay cấu trúc Lai) và bổ sung khả năng viết công thức toán động (hiển thị công thức toán đồng thời với tính trực tiếp từ công thức). Dịch vụ này tạo khả năng viết báo cáo hoặc viết sách cho bất kỳ nhóm chuyên môn nào, kể cả các nhóm chuyên ngành toán lý.

Đoàn Hoà Minh và nhóm [10] đã nghiên cứu thành công mô hình kết nối sóng hồng ngoại giữa cây viết hồng ngoại và thiết bị thu sóng gắn với máy tính. Kết quả này cho phép nhóm tác giả sản xuất thành công thiết bị Hỗ trợ giáo viên dạy trên bảng tương tác với phần mềm giảng dạy và minh hoạ sống động các bài giảng trừu tượng.

Đối với vấn đề An ninh mạng, việc phát hiện Hacker và phân loại Hacker từ các hành vi tấn công mạng đã được rất nhiều tác giả nghiên cứu. Đã có nhiều kết quả về vấn đề này, nhưng vẫn còn có sai lầm không nhỏ khi gặp những tấn công mạng với tần số thấp chen lẫn những tấn công có tần số cao. Đỗ Thanh Nghị [11] đã cải tiến giải thuật học Cây quyết định với khoảng cách Kolmogorov-Smirnov và từ đó xây dựng được bộ luật nhận dạng các tấn công mạng tần số thấp cũng như tần số cao với độ chính xác tốt hơn. Tác giả đã biểu diễn kết quả này dưới dạng một mô hình trực quan của Cây quyết định như trong *Hình 6.* Đặc điểm này tạo khả năng quan sát trực tiếp, hỗ trợ nhanh việc chấp nhận hay bác bỏ bộ luật mới và tất nhiên là phát hiện nhanh sự thay đổi của bộ luật cũ.



Hình 6: Mô hình trực quan Cây quyết định nhận dạng tấn công mạng

5 KẾT LUẬN

Giai đoạn 2006-2011 chưa dài để chứng tỏ sự phát triển bền vững và cân bằng về Đào tạo và NCKH của Khoa CNTT & TT. Nhưng những kết quả vừa qua cũng cho thấy một "bệ phóng" tốt đã được xây dựng thành công, chuẩn bị cho thời kỳ mới "bay nhanh và cao" hơn nữa trong Đào tạo và NCKH về CNTT & TT chất lượng cao.

Qua phân tích các khía cạnh đổi mới trong Đào tạo và sự chuyển mình trong hoạt động NCKH thời gian qua, chúng ta có thể đúc kết lại thành 4 đặc điểm chính tác động lên chất lượng Đào tạo và 3 tiêu chí đảm bảo sự bền vững của hoạt động Nghiên cứu khoa học.

Bốn đặc điểm chính đối với Đào tạo

1- Hệ thống đào tạo theo tín chỉ và hướng tới sinh viên tạo cho sinh viên tính tự chủ và tự quyết định chất lượng học của chính mình.

2- Môi trường học tập và làm việc sạch, đẹp và thân thiện tạo tâm lý thoải mái cho việc dạy và học.

3- Hệ thống thiết bị đầy đủ đảm bảo hỗ trợ tốt cho sinh viên thực hành và nghiên cứu.

4- Kiên trì và kiên quyết Chuẩn mực hoá chương trình đào tạo theo hệ tín chỉ gồm: chuẩn đầu ra và chuẩn quy trình đào tạo; tiến tới thực hiện Kiểm định toàn bộ các chương trình đào tạo CNTT theo chuẩn Quốc tế phù hợp nhất.

Ba tiêu chí đối với NCKH

1- Môi trường hợp tác (Vùng và Quốc tế) và liên ngành luôn mang lại sự đa dạng của đề tài nghiên cứu.

2- Kết hợp Đề tài và Dự án đảm bảo phần thu tài chính hỗ trợ cho NCKH.

3- Kết hợp Đề tài, Luận văn Sau Đại học và Báo cáo khoa học đảm bảo gần như 100% sự thành công của đề tài NCKH.

Định hướng

Từ đây có thể thấy trong thời gian tới cần hoạch định chiến lược và kế hoạch theo một số định hướng phát triển như sau:

- Kiên trì xây dựng các chính sách đảm bảo duy trì được 4 yếu tố tác động chủ yếu lên chất lượng đào tạo và 3 tiêu chí thành công của NCKH.

- Xây dựng thành công một số chương trình đào tạo tiên tiến có kiểm định quốc tế nhằm đẩy nhanh tiến độ đưa các tiêu chí kiểm định quốc tế vào tất cả các chương trình đào tạo của Khoa.

- Đăng ký tổ chức hàng năm Hội thảo khoa học về CNTT & TT cấp Vùng.

- Xác định nguồn tài chính hỗ trợ cho xuất bản tạp chí khoa học Đại học Cần thơ số chuyên san hàng năm về CNTT & TT.

Tài liệu tham khảo

[1] Lê Quyết Thắng et al., Nghiên cứu xây dựng các hệ thống thông tin hỗ trợ việc phòng chống dịch bệnh cây trồng và thuỷ sản cho vùng kinh tế trọng điểm, Đề tài cấp Nhà nước mã số KC.01.15/06-10, Quyết định công nhận kết quả số 2934/Q-BKHCN, 2011.

[2] Đỗ Thanh Nghị et al., Xây dựng dịch vụ Web cho Khai mỏ dữ liệu, ICT.rda'10, sản phẩm của đề tài KC.02.15/06-10, 2011.

[3] Nguyễn Thái Nghe et al., Learning Optimal Threshold for Baysien Posterior Probabilities to Mitigate the Class Imbalance Problem, Tạp chí Khoa học và Công nghệ, Vol. 48, N^0. 4, 2010.

[4] Phạm Nguyên Khang et al., Intensive use of factorial correspondence analysis for large scale content-based image retrieval. In Advances in Knowledge Discovery and Management, AKDM'09, Springer-Verlag, 2009.

[5] Phan Huy Cường et al., An Agent-based Approach to the Simulation of Brown Plant Hopper (BPH) Invasions in the Mekong Delta, IEEE-RIVF Proceedings, 2010.

[6] Vũ Duy Linh et al., Dự báo né rầy theo thời gian, Kỷ yếu Hội thảo Quốc gia lần thứ XIII Hưng yên, Nhà xuất bản Khoa học và Kỹ thuật, 2011.

[7] Phan Thượng Cang et al., Dịch vụ hỗ trợ ngữ nghĩa cho nông dân tìm kiếm thông tin về dịch hại, Tạp chí Công nghệ thông tin và Truyền thông, tập V-1, số 6(26), 09-2011.

[8] Trần Cao Đệ, Bi-Character Model for On-line Cursive Handwriting Recognition, Tạp chí Khoa học và Công nghệ, Vol. 48, N^0. 4, 2010.

[9] Lê Quyết Thắng et al., Xây dựng dịch vụ báo cáo cộng tác hỗ trợ công thức toán động, Tuyển tập công trình nghiên cứu Công nghệ thông tin và Truyền thông 2010, Nhà Xuất bản Khoa học và Kỹ thuật, 2010.

[10] Đoàn Hoà Minh et al., Báo cáo tổng kết đề tài NCKH cấp cơ sở NGHIÊN CỨU THIẾT KẾ VÀ THỰC HIỆN BẢNG ĐIỆN TỬ TƯƠNG TÁC, mã số T2011-47, Trường ĐH Cần Thơ, 2011.

[11] Đỗ Thanh Nghị et al., Nhận dạng tấn công mạng với mô hình trực quan cây quyết định, Tạp chí Công nghệ thông tin và Truyền thông, tập V-1, số 6(26), 09-2011.

NGHIÊN CỨU XÂY DỰNG HỆ THỐNG E-LEARNING PHỤC VỤ CHƯƠNG TRÌNH ĐÀO TẠO KỸ SƯ TIN HỌC LIÊN THÔNG TỪ CỬ NHÂN CAO ĐẲNG TIN HỌC VÀ HỖ TRỢ TRONG ĐÀO TẠO THEO HỌC CHẾ TÍN CHỈ

Nguyễn Văn Linh, Phan Phương Lan, Phan Huy Cường,
Trần Ngân Bình , Trần Minh Tân
Khoa Công nghệ thông tin và Truyền thông, trường Đại học Cần thơ

Abstract

E-learning is one of the information technology (IT) application solutions in education. This paper presents the results of building an E-learning system for training IT engineers from the bachelor degree and for supporting training under the credit system at college of ICT, Can Tho university. These results include selecting the suitable solution for building an E-learning system, researching and choosing the standard and LMS, building some tools integrated into LMS Moodle, building curriculum and the process in the form of E-learning, building question bank and designing electronic lessons compliant SCORM standard.

Keywords: E-learning, Learning Management System (LMS), Moodle, Sharable Content Object Reference Model (SCORM)

Title: Building an E-learning system for training IT engineers from the bachelor degree and for supporting training under the credit system

Tóm tắt

E-learning là một trong các giải pháp ứng dụng CNTT trong giáo dục. Bài báo này trình bày kết quả nghiên cứu xây dựng hệ thống E-learning phục vụ chương trình đào tạo kỹ sư tin học liên thông từ cử nhân cao đẳng tin học và hỗ trợ trong đào tạo theo học chế tín chỉ tại Khoa CNTT&TT thuộc đại học Cần Thơ. Những kết quả nghiên cứu gồm chọn giải pháp xây dựng một hệ thống E-learning, nghiên cứu lựa chọn chuẩn và hệ nền cho E-learning, xây dựng một số công cụ tích hợp vào hệ nền cho E-learning, xây dựng chương trình và quy trình đào tạo liên thông theo hình thức E-learning, xây dựng ngân hàng câu hỏi và thiết kế bài giảng điện tử theo chuẩn SCORM.

Từ khóa: E-learning, hệ thống quản lý đào tạo, Moodle, chuẩn SCORM

1 ĐẶT VẤN ĐỀ

Trong giáo dục, đặc biệt là giáo dục bậc đại học và sau đại học, khi Việt Nam muốn rút ngắn khoảng cách về chất lượng đào tạo với các nước tiên tiến trên thế giới thì việc ứng dụng công nghệ thông tin (CNTT) là rất cần thiết. E-learning là một trong các giải pháp ứng dụng CNTT trong giáo dục. Với đội ngũ giảng viên và sinh viên đã quen ứng dụng CNTT, với cơ sở hạ tầng CNTT khá tốt, Khoa Công nghệ Thông tin và Truyền thông (CNTT&TT) thuộc Đại học Cần Thơ (ĐHCT) có rất nhiều thuận lợi để triển khai E-learning.

Bên cạnh đó, số lượng sinh viên tốt nghiệp từ các trường trung cấp, cao đẳng trong khu vực đồng bằng sông Cửu Long hàng năm là rất lớn. Đa số họ hiện đang làm công tác triển khai các ứng dụng CNTT trong các cơ quan xí nghiệp, trường học, doanh nghiệp. Họ đều có nguyện vọng được học liên thông lên cấp đào tạo cao hơn để nâng cao trình độ nhằm đáp ứng yêu cầu ngày càng tăng của xã hội. Tuy nhiên không phải ai cũng có điều kiện

học tập trung tại Cần Thơ. Vì vậy, E-learning là phù hợp đối với loại hình đào tạo liên thông này.

Vấn đề cuối cùng là áp lực của học chế tín chỉ. Một trong rất nhiều khó khăn của học chế tín chỉ là việc tự học của sinh viên. Với mỗi giờ lên lớp của giảng viên, sinh viên phải tự học 2 giờ. Thực hiện việc này là rất khó khăn vì sinh viên đã quen tâm lý học thụ động; thiếu tài liệu tham khảo và thiếu sự trợ giúp của giảng viên do thầy phải dạy quá nhiều và lớp quá đông. Một khó khăn khác đối giảng viên là việc đánh giá học phần. Theo quy chế, điểm tổng hợp đánh giá học phần được tính căn cứ vào một phần hoặc tất cả các điểm đánh giá bộ phận. Do sĩ số của các lớp quá lớn nên khâu tổ chức đánh giá bộ phận (do giảng viên thực hiện) rất khó khăn dẫn đến kết quả không khách quan, không có chất lượng. E-learning là một công cụ hỗ trợ tốt cho quá trình đào tạo theo học chế tín chỉ.

Từ các lý do vừa nêu trên, nhóm đề xuất đề tài nghiên cứu xây dựng hệ thống E-learning phục vụ chương trình đào tạo liên thông từ cử nhân lên kỹ sư tin học và hỗ trợ trong đào tạo theo học chế tín chỉ.

2 MỤC TIÊU

Xây dựng thành công một hệ thống E-learning chuẩn mực, bao gồm hệ thống quản lý đào tạo và các bài giảng điện tử phục vụ công tác đào tạo kĩ sư tin học liên thông từ cử nhân cao đẳng tin học nói riêng và đào tạo theo học chế tín chỉ nói chung. Việc ứng dụng E-learning trong dạy và học tại Khoa CNTT&TT sẽ góp phần đổi mới phương pháp giảng dạy, phương pháp học tập và phương pháp đánh giá trong đào tạo theo học chế tín chỉ.

Đối tượng mà hệ thống sẽ phục vụ là giảng viên, sinh viên liên thông từ cao đẳng tin học và sinh viên đại học của Khoa CNTT & TT.

3 PHƯƠNG PHÁP NGHIÊN CỨU

Do đây là một đề tài ứng dụng nên nhóm nghiên cứu không tập trung vào các lý thuyết khoa học chuyên sâu mà cố gắng tạo ra một tập hợp nội dung số, bổ sung thêm các công cụ hỗ trợ và áp dụng ngay vào trong thực tiễn đào tạo.

Nhóm nghiên cứu thực hiện:

- Phân tích các giải pháp xây dựng hệ thống E-learning để chọn ra giải pháp phù hợp.
- Thu thập và phân tích các tài liệu tham khảo từ các tổ chức chuyên nghiên cứu về E-learning để đề xuất chuẩn và hệ nền phù hợp cho E-learning.
- Nghiên cứu đến các cấp độ nhận thức của Benjamin S. Bloom và vận dụng vào việc xây dựng cấu trúc bài giảng điện tử theo chuẩn đã được lựa chọn.
- Nghiên cứu và xây dựng các công cụ hỗ trợ cho hệ nền đã chọn.
- Xây dựng nội dung số:
    - Tổ chức biên soạn các bài giảng thuộc chương trình đào tạo theo cấu trúc đã được xây dựng và đóng gói các bài giảng.
    - Tổ chức biên soạn ngân hàng câu hỏi phục vụ công tác đánh giá.
- Tổ chức tập huấn cho giảng viên sử dụng hệ thống và đưa hệ thống vận hành thực tế.

4 KẾT QUẢ

4.1 Chọn giải pháp xây dựng hệ thống E-learning

Các giải pháp xây dựng hệ thống E-learning tại Việt Nam có thể được nhóm lại theo ba dạng sau:

- *Kết hợp giữa công ty trong nước với đối tác nước ngoài.* Ở giải pháp dạng này, toàn bộ hệ thống E-learning đều do phía đối tác cung cấp. Trong một số trường hợp, công ty trong nước sử dụng bài giảng (phần quan trọng nhất của hệ thống) do đối tác cung cấp và đưa chúng lên một hệ thống LMS mã nguồn mở. Nhìn chung, giải pháp này phù hợp với những công ty kinh doanh Việt Nam thực hiện đào tạo trong thời gian ngắn.
- *Tự xây dựng hệ thống.* Tự xây dựng hệ thống cho riêng mình là một giải pháp rất tốn kém cả về mặt thời gian, tiền bạc cũng như công sức. Nó phù hợp với những công ty hoặc các tổ chức đào tạo lớn với khả năng mạnh về tài chính cũng như nhân lực phát triển phần mềm.
- *Xây dựng hệ thống dựa trên phần mềm nguồn mở.* Giải pháp dạng này không những giúp các đơn vị triển khai khá hiệu quả và phù hợp với yêu cầu thực tiễn mà vẫn có thể phát triển, nâng cấp hệ thống.

Với các nguồn lực của Khoa CNTT & TT (nhân lực, vật lực và tài lực), chúng tôi đề nghị xây dựng hệ thống E-learning theo giải pháp thứ 3.

4.2 Nghiên cứu lựa chọn chuẩn và hệ nền cho E-learning

Trong rất nhiều chuẩn và nhiều hệ quản lý đào tạo mã nguồn mở như hiện nay, việc chọn được một chuẩn và một hệ quản lý đào tạo phù hợp trước khi xây dựng hệ thống E-learning theo giải pháp thứ ba là một trong những công việc phải được thực hiện sớm nhất.

Về việc đánh giá các chuẩn và các hệ nền, nhóm không xây dựng bộ tiêu chí đánh giá mà dựa vào kết quả đánh giá của các tổ chức trên thế giới để lựa chọn chuẩn và hệ nền cho ứng dụng của mình. Trên kết quả phân tích các tài liệu đó, nhóm đề xuất chọn chuẩn SCORM (Sharable Content Object Reference Model) và hệ thống quản lí đào tạo (Learning Management System - LMS) mã nguồn mở Moodle.

Nhóm nghiên cứu đã tổ chức tập huấn về sử dụng hệ thống Moodle cho giảng viên trong khoa (bao gồm cả bộ môn Viễn thông và Kỹ thuật điều khiển). Nhờ công tác này mà hiện nay hầu hết các giảng viên đều sử dụng hệ thống E-learning của khoa để hỗ trợ giảng dạy, học tập và thi, kiểm tra cho cả sinh viên chính quy của trường và sinh viên ngoài trường.

4.3 Xây dựng chương trình và quy trình đào tạo liên thông theo hình thức E-learning

Trong quá trình xây dựng chương trình đào tạo liên thông, nhóm nghiên cứu đã tham khảo chương trình đào tạo kỹ sư CNTT và chương trình đào tạo cử nhân cao đẳng Tin học của Khoa CNTT&TT, trường ĐHCT, tham khảo khung chương trình đào tạo của ACM (Association for Computing Machinery). Từ đó, nhóm đề xuất cấu trúc chương trình đào tạo liên thông và quy trình tổ chức đào tạo.

Mỗi giảng viên và sinh viên đều được cấp một tài khoản để truy cập vào hệ thống. Sinh viên sử dụng hai phần ba thời gian được sắp trong thời khóa biểu và cả thời gian trống khác để tham gia các hoạt động trên hệ thống E-learning như: học bài, tham khảo tài liệu, làm bài tập lớn, tham gia diễn đàn, thảo luận trực tuyến, làm kiểm tra,v.v. Một phần ba thời gian còn lại dành cho giảng viên và sinh viên gặp nhau trực tiếp để giải đáp các thắc mắc còn tồn đọng. Ngoài ra, đối với các phần thực hành trên máy thì sinh viên phải tham gia đầy đủ, bình thường như các sinh viên khác.

Chương trình và quy trình đào tạo này được thực hiện cho hai khóa sinh viên liên thông của khoa.

4.4 Thiết kế bài giảng điện tử

Để có cơ sở khoa học sư phạm cho việc xây dựng cấu trúc bài giảng điện tử, nhóm nghiên cứu quan tâm đến các cấp độ nhận thức của Benjamin S. Bloom (1913-1999). Nhóm sử dụng các cấp độ nhận thức để đặt mục tiêu học tập và chuẩn về kiến thức, kỹ năng để kiểm tra, đánh giá. Sáu cấp độ nhận thức từ thấp đến cao gồm: biết, hiểu, vận dụng, phân tích, tổng hợp và đánh giá.

Nhóm nghiên cứu sử dụng chuẩn đóng gói nội dung SCORM để đóng gói bài giảng điện tử. Mỗi học phần sẽ được chia thành nhiều bài giảng, từng bài giảng được đóng gói thành SCO (Sharable Content Object) hoàn chỉnh. Thực tế, đa số các bài giảng được soạn bằng MS Words nên nhóm đã nghiên cứu và cài đặt thành công phần mềm Word2SCO cho phép tách, chuyển đổi, đóng gói hoàn toàn tự động, kết quả sinh ra sẽ là các SCO ở dạng nén có thể sử dụng trong các LMS dùng chuẩn SCORM.

Phục vụ đào tạo liên thông, các giảng viên đã biên soạn 30 bài giảng điện tử. Các bài giảng này đã được đóng gói nhờ công cụ Word2SCO và đưa lên hệ thống E-learning của khoa.

4.5 Xây dựng ngân hàng câu hỏi

Trong công tác đào tạo, khâu đánh giá là rất quan trọng, nó góp phần quyết định chất lượng đào tạo. Việc đánh giá theo quá trình và sử dụng hình thức trắc nghiệm khách quan luôn được khuyến khích.

Để đảm bảo xây dựng được ngân hàng câu hỏi có chất lượng, nhóm nghiên cứu thực hiện quy trình gồm 5 bước sau: (1) Thiết lập ma trận kiến thức đáp ứng yêu cầu; (2) xây dựng ngân hàng câu hỏi, (3) tổ chức thi, kiểm tra, (4) phân tích câu hỏi và (5) điều chỉnh câu hỏi.

Thông qua hệ quản lý đào tạo Moodle, giảng viên tạo các câu hỏi trắc nghiệm với nhiều hình thức khác nhau như câu hỏi tính toán, câu hỏi so khớp, câu hỏi đúng sai, câu hỏi trả lời ngắn, câu hỏi đa lựa chọn,v.v. Cũng qua hệ thống này, giảng viên thiết lập các lựa chọn cho đề thi, xem và lưu trữ kết quả chấm thi tự động. Ngoài ra, hệ Moodle còn giúp giảng viên thực hiện việc phân tích các câu hỏi trên cơ sở thống kê và dựa vào các chỉ số: độ khó, độ lệch chuẩn, chỉ số phân biệt và hệ số phân biệt. Từ đó, giảng viên có thể chỉnh lý, bổ sung hay loại bỏ câu hỏi.

Các giảng viên trong khoa đã xây dựng được ngân hàng có khoảng 2630 câu hỏi (tính đến tháng 1/2010) và vẫn đang tiếp tục bổ sung, cập nhật. Nhìn chung các giảng viên thường sử dụng loại câu hỏi đa lựa chọn. Ngân hàng câu hỏi này đã và đang được các giảng viên sử dụng thường xuyên để tổ chức thi, kiểm tra.

4.6 Xây dựng một số công cụ tích hợp vào hệ nền cho E-learning

Moodle là phần mềm của cộng đồng, phần mềm mang tính chất phục vụ chung. Vì vậy, để phù hợp với yêu cầu ứng dụng E-learning tại Khoa CNTT&TT, các công cụ hỗ trợ hệ thống Moodle cần phải được xây dựng. Nhóm nghiên cứu đã thiết kế và cài đặt các công cụ sau:

- *Điều khiển tiến trình học của học viên trong một course.* Công cụ này cho phép giảng viên có thể tạo ra tiến trình cho một course. Cụ thể, khi thiết lập từng chủ đề trong một course, giảng viên phải qui định các chủ đề tiên quyết mà sinh viên phải đạt được trước khi học chủ đề mà giảng viên đang thiết lập. Ngoài ra, giảng viên cũng phải thiết lập tỉ lệ phần trăm số điểm cho mỗi hoạt động trong mỗi chủ đề. Như vậy, khi tham gia vào một course nào đó, học viên phải học theo tiến trình mà giảng viên đặt ra.
- *Điều khiển tiến trình học của học viên trong chương trình học.* Mục đích của công cụ này là này cho phép giảng viên bắt buộc học viên phải học theo tiến trình mà chương trình đào tạo đã quy định. Tương tự như công cụ được đề cập ở trên, khi thiết lập một course,

giảng viên phải qui định các course tiên quyết mà sinh viên phải đạt được trước khi học course giảng viên đang thiết lập.

- *Quản lí ngân hàng câu hỏi và ra đề tự động*. Công cụ này giúp giảng viên quản lí ngân hàng câu hỏi trong course theo từng chủ đề, loại câu hỏi và độ khó. Ngoài ra, giảng viên có thể ra đề tự động theo tiêu chí đặt ra (chủ đề, loại câu hỏi và độ khó) không phải chọn từng câu hỏi và thêm vào đề thi như chức năng hiện tại của hệ thống Moodle.

5 KẾT LUẬN VÀ KIẾN NGHỊ

5.1 Kết luận

Kết quả quan trọng nhất mà nhóm nghiên cứu đã làm được đó là đưa hệ thống E-learning vào hoạt động tại Khoa CNTT&TT thuộc trường ĐHCT, tạo ra một kênh học tập khác góp phần nâng cao chất lượng đào tạo. Ngày nay việc sử dụng hệ thống E-learning đã trở thành tự giác đối với hầu hết giảng viên trong khoa vì những lợi ích thiết thực mà hệ thống mang lại.

5.2 Kiến nghị

Hiện nay chương trình đào tạo liên thông thay đổi theo hệ thống tín chỉ nên không còn lớp học dành riêng cho sinh viên liên thông. Mặc dù hệ thống E-learning của khoa không còn phục vụ cho các lớp học đó nhưng nó đã trở thành một hệ thống thân quen với các giảng viên và sinh viên trong khoa. Vì thế, nhóm nghiên cứu đề nghị hệ thống này vẫn được duy trì và chuyển qua phục vụ đào tạo theo hệ thống tín chỉ. Để phục vụ tốt hơn, hệ thống cần được nâng cấp như tăng tính bảo mật, tạo ra các bài giảng điện tử dạng video cho các học phần trong chương trình đào tạo mới, v.v.

TÀI LIỆU THAM KHẢO

B. S. Bloom (Ed.), "Taxonomy of Educational Objectives: The Classification of Educational Goal". Susan Fauer Company, 1956, pp. 201-207.

Đỗ Trung Tá, "Ứng dụng CNTT-TT để đổi mới giáo dục Đại học ở Việt Nam: Bốn cột trụ lớn". Báo Bưu điện Việt Nam, 2005.

Graf, S. & List, B. "An Evaluation of Open Source E-Learning Platforms Stressing Adaptation Issues. Proceedings of the International Conference on Advanced Learning Technologies. Kaohsiung", Taiwan, 2005, pp. 163-165.

Huỳnh Ngọc Phiên, Trần Đại Dũng, Huỳnh Ngọc Chương, Võ Quốc Bảo, "Distance Education". Bangkok, Thailand, tháng 12-1997.

Lâm Quang Thiệp, "Cơ sở của các phương pháp trắc nghiệm". Tài liệu đào tạo.

Lê Quyết Thắng, Nguyễn Văn Linh và Phan Huy Cường, "Cấu trúc cơ bản của một giáo trình điện tử dành cho tự học và công cụ cài đặt nó". Tham luận tại hội thảo quốc gia về công nghệ thông tin (ICT .rda '03). Hà Nội, tháng 4-2003.

Lorin W. Anderson et al, "Taxonomy for Learning, Teaching, and Assessing — A Revision of Bloom's Taxonomy of Educational Objectives". Addison Wesley Longman, 2001.

Ngô Trung Việt, "Mô hình tổ chức cơ sở học tập e-learning". Tài liệu đào tạo.

Ngô Trung Việt, "E-learning, một hình thức học tập mới". Tài liệu đào tạo.

Nguyễn Ngọc Bình, Nguyễn Thúc Hải và Đỗ Văn Uy; "Kiến trúc nền cho e-learning và hệ đào tạo trên mạng BKVIEWS". Tham luận tại hội thảo quốc gia về công nghệ thông tin (ICT .rda '03). Hà Nội, tháng 4-2003.

Nguyễn Ngọc Đệ, Dương Ngọc Thành, Võ Thị Thanh Lộc, Nguyễn Phú Son; "Thực trạng, nhu cầu và giải pháp cung cấp thông tin khoa học công nghệ khu vực đồng bằng sông Cửu Long"; Kỷ yếu hội thảo báo cáo một số kết quả của đề tài nghiên cứu cơ sở khoa học xây dựng mạng thông tin KH&CN khu vực đồng bằng sông Cửu Long; Cần thơ, 8/2009.

Nguyễn Văn Linh, "Sử dụng bài giảng ghi trên đĩa CD để thay thế một phần công tác giảng dạy trực tiếp của giảng viên và tăng cường tính tự học của sinh viên". Tham luận tại hội thảo Tổng kết 5 năm đổi mới phương pháp giảng dạy của trường Đại Học Cần Thơ. Cần thơ, tháng 2 năm 2003

Paul MacEke, "Directions in e-learning". IBM Corp 2000.Van den Berg, K. "Finding Open options: An Open Source software evaluation model with a case study on Course Management Systems". Master Thesis. Tilburg University, Netherland, 2005.

Website về chuẩn SCORM, ADL, http://www.adlnet.org

Website về hệ quản lý đào tạo Moodle, http://moodle.org

Website về hệ quản lý đào tạo ILIAS, http://www.ilias.uni-koeln.de

Website về hệ quản lý đào tạo ATutor, http://www.atutor.ca

MỘT GIẢI PHÁP THÊM CHỨC NĂNG CỦA BẢNG TƯƠNG TÁC VÀO HỆ THỐNG PROJECTOR-COMPUTER HOẶC LCD-COMPUTER.

Đoàn Hòa Minh[1], Nguyễn Khắc Nguyên[2], Bùi Minh Quân[1]

Abstract

Interactive whiteboards (IWB) have been produced and widely used in the world. In Vietnam, IWBs are rarely used because their price is still high. With the development trend in education in Vietnam recently, more and more teachers use Powerpoint with projectors as the main teaching method. However, this method also has some weaknesses, especially when the lecturer wants to present his or her ideas by handwritten method. Therefore, it is necessary to find out a solution to add IWB's functionalities to a projector-computer system or a LCD-computer. That is the reason why we have carried out this project.

Key words: Interactive whiteBoard, IWB, IWB using Wii Remote, infrared pen, detecting location unit, calibration, toolbar.

Title: A solution of adding the functions of interactive whiteboard into the system of projector-computer or the system of LCD-Computer.

Tóm tắt

Bảng tương tác (interactive whiteboard, IWB) đã được sản xuất và được sử dụng khá phổ biến trên thế giới. Riêng ở Việt Nam, số IWB được sử dụng còn rất hiếm do giá thành của nó còn khá cao. Với sự phát triển của giáo dục và đào tạo ở nước ta, hiện nay phương thức giảng dạy trình chiếu Powerpoint dùng projector đã trở nên thịnh hành, đặc biệt là trong các trường đại học và cao đẳng. Phương pháp này cũng có những nhược điểm riêng của nó, chẳng hạn như trong các trường hợp cần giải thích thêm hoặc làm bài tập, khi đó thầy phải kết hợp trình chiếu và viết bảng. Để kết hợp các hình thức này với nhau ta cần IWB. Từ nhu cầu thực tế trên, chúng tôi đã nghiên cứu giải pháp thêm các tiện ích của IWB vào hệ thống projector-computer hoặc LCD-computer.

Từ khóa: Bảng tương tác, IWB, IWB sử dụng Wii Remote, bút hồng goại, bộ phận lấy tọa độ, định khung tác động, thanh công cụ.

1 MỞ ĐẦU

Bảng điện tử tương tác (interactive whiteboards, IWB) là một màn hiển thị tương tác (interactive display) lớn được kết nối với bộ xử lý hoặc máy vi tính mà trên đó người dùng có thể viết, vẽ, xóa và sử dụng các phần mềm máy tính bằng ngón tay, bút trâm (stylus) hoặc bút cảm ứng (sau đây sẽ được gọi chung là bút). Nó không chỉ thực hiện các chức năng của bảng phấn (nhưng không có bụi phấn) mà còn cho phép thực hiện các ứng dụng của công nghệ thông tin. Các thao tác trực tiếp trên IWB cho phép thực hiện các chức năng như sau:

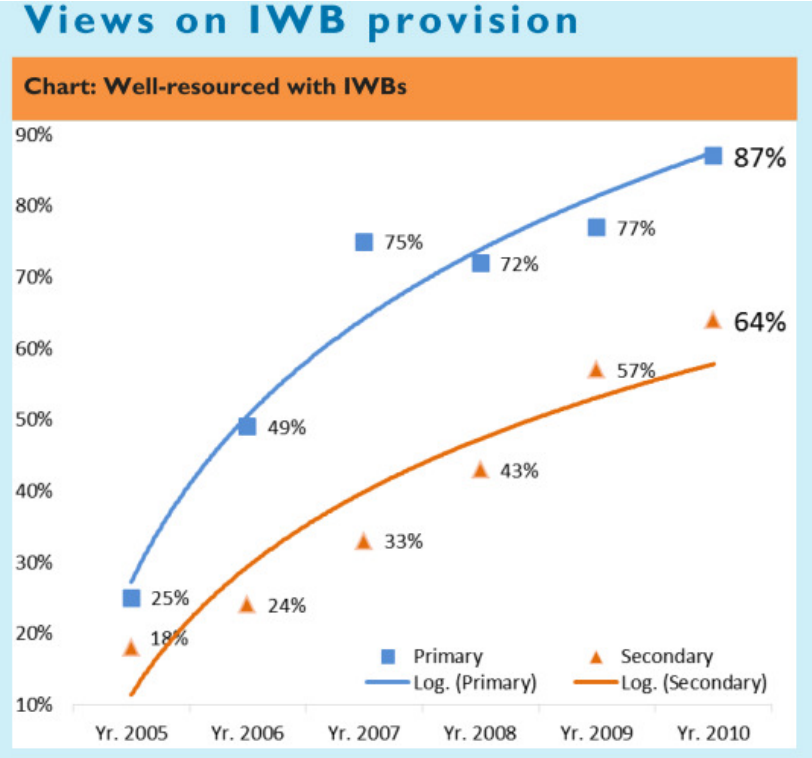
- Viết, vẽ, xóa, đánh dấu và lưu các trang dữ liệu này vào bộ nhớ máy tính. Nếu có cài đặt phần mềm OCR (optical character recognition) trên máy tính thì các chữ viết thảo (cursive writing) có thể chuyển đổi thành văn bản (text).

[1] Khoa Công nghệ thông tin & Truyền thông, Trường Đại học Cần Thơ.
[2] Khoa Công nghệ, Trường Đại học Cần Thơ.

- Điều khiển trình chiếu Powerpoint.
- Điều khiển chạy các phần mềm trên máy (duyệt, mở, lưu và xóa tập tin; kết nối mạng và duyệt web, thực hiện các ứng dụng đa phương tiện,…).

Nói chung, với một bút chúng ta có thể thực hiện tất cả các chức năng trên như dùng chuột máy tính điều khiển trên giao diện người dùng.

IWB đã được sản xuất và được sử dụng khá phổ biến trên thế giới. BESA (British Educational Suppliers Association) đã thực hiện sự khảo sát hằng năm về công nghệ thông tin trong các trường học ở Vương quốc Anh, theo tường trình năm 2010, có 87% lớp bậc tiểu học và 64% lớp bậc trung học được trang bị IWB (Hình 1). Cũng ở Anh Quốc, công ty Decision Tree Consulting chuyên nghiên cứu thị trường ở London đã điều tra nhu cầu về IWB ở 66 quốc gia trên thế giới và tiên đoán rằng trung bình cứ 7 lớp học sẽ được trang bị 1 IWB vào năm 2011. Các tính năng ưu việt của IWB rất thuận lợi cho việc trình bày báo cáo và giảng dạy trên lớp.



Hình 1: Sơ đồ tăng trưởng hằng năm của IWB (% số lớp có IWB)
(http://www.besa.org.uk/besa/documents/index.jsp?mode=search&topic=0&r=true&b=true&o=true&keyword=ICT+in+UK+State+Schools&x=41&y=13)

Riêng ở Việt Nam, số IWB được sử dụng còn rất hiếm do giá thành của nó khá cao (từ 1.000 USD đến 10.000USD tùy loại và chất lượng, chưa tính thuế và phí chuyên chở, phân phối). Bảng điện tử tương tác đang là miềm mơ ước của nhà giáo.

Với sự phát triển của giáo dục và đào tạo ở nước ta, hiện nay phương thức giảng dạy trình chiếu dùng Powerpoint dùng projector đã trở nên thịnh hành, đặc biệt là trong các trường đại học và cao đẳng. Phương pháp này cũng có những nhược điểm riêng của nó, chẳng hạn như trong các trường hợp cần giải thích thêm hoặc làm bài tập, khi đó thầy phải kết hợp trình chiếu và viết bảng. Để kết hợp các hình thức này vào một ta cần IWB.

Từ nhu cầu thực tế trên, chúng tôi đã nảy sinh ý tưởng thêm các tiện ích của IWB vào hệ thống Projector-Computer (hoặc LCD-Computer). Điều này có nghĩa là từ bộ projector và máy vi tính có sẵn, chúng ta thiết kế thêm "bộ thiết bị hỗ trợ" để sử dụng hệ thống này tương tự như một IWB. Một vấn đề quan trọng khác cần phải tính đến, đó là bộ thiết bị hỗ trợ phải dễ cài đặt và có giá phù hợp với khả năng của một giáo viên bình thường. Khi đó công trình này mới có hiệu quả thật sự. Bộ thiết bị hỗ trợ gồm bộ phận giao tiếp với máy tính (lấy tọa độ) và một bút. Bộ thiết bị này có thể kết hợp với máy tính và projector sẵn có để tạo thành bảng điện tử tương tác. Thiết bị hỗ trợ có thể kết nối với máy tính thông qua đường truyền vô tuyến, bluetooth hoặc dây dẫn gắn vào cổng USB, người giáo viên hoặc thuyết trình viên có thể đứng trên bảng dùng bút để thực hiện các tính năng của IWB.

2 PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Tìm hiểu nguyên lý hoạt động của IWB

Sau khi tìm hiểu nguyên lý hoạt động của nhiều loại IWB khác nhau, chúng tôi có thể tổng hợp thành một nguyên lý chung. Trước tiên chúng ta cần định nghĩa các khái niệm sau:

- **Bút:** là công cụ dùng để viết, vẽ, đánh dấu, xóa và điều khiển phần mềm máy tính bằng các tương tác trên giao diện đồ họa người dùng. Bút có thể là ngón tay, bút trâm, bút cảm ứng,… Có thể phân làm hai loại: bút thụ động là loại không có nguồn cấp điện và chỉ tương tác cơ học như ngón tay, bút trâm (stylus); bút tích cực là loại bút cần có nguồn cấp điện, như bút cảm ứng điện từ, bút hồng ngoại.
- **Bảng hay màn hình:** là thiết bị để hiển thị các nội dung cần trình bày và giao diện người dùng. Bảng có thể là màn hiển thị thụ động (màn vải, vách tường, bảng thông thường,… có màu nền thích hợp để chiếu hình ảnh lên đó) hoặc là màn hiển thị tích cực (như màn hình LCD, màn hình cảm ứng).
- **Bộ xử lý:** có thể lắp đặt thành một hệ thống với màn hình hoặc là một máy tính cá nhân (PC) được kết nối với projector và/hoặc bảng.

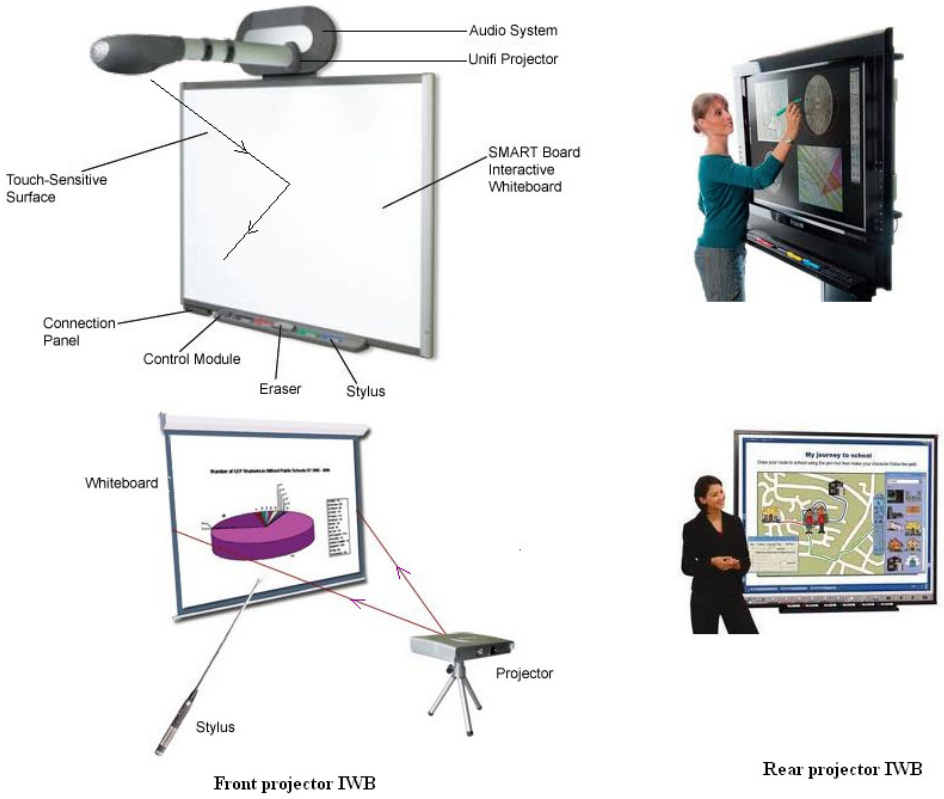
Từ các khái niệm đã được định nghĩa, chúng ta có thể phát biểu một cách ngắn gọn nguyên lý hoạt động của IWB như sau: *Bút được dùng để xác định tọa độ của một hoặc nhiều điểm trên bảng bằng thao tác của người dùng, các tọa độ này được tách và truyền về bộ xử lý và được xem như là kiểu dữ liệu con trỏ. Phần mềm sẽ xử lý các điểm tọa độ để thực hiện các chức năng của IWB.*

2.2 Phân loại IWB

Chúng ta có các cách phân loại khác nhau dựa vào các tiêu chí khác nhau:

2.2.1. Phân loại dựa vào vị trí nguồn sáng

Theo cách phân loại này thì có 2 loại IWB. Đó là IWB chiếu trước (Front projector IWB) và IWB chiếu sau (Rear projector IWB).



Hình 2: IWB loại chiếu trước và chiếu sau

- **IWB chiếu trước** thường dùng màn hiển thị thụ động và hình ảnh được chiếu từ một projector đặt ở phía trước. Bút được dùng là loại cảm ứng hồng ngoại hoặc bút camera kỹ thuật số. Nhược điểm của IWB chiếu trước là in bóng cánh tay của người trình bày lên màn ảnh ngay cả khi đứng ở tư thế né sang một bên.
- **IWB chiếu sau** có nguồn sáng chiếu từ phía sau của màn hình. Màn hình thuộc loại tinh thể lỏng (LCD) hoặc plasma có tính năng màn cảm ứng (touchscreen). Bút được dùng là loại bút thụ động. IWB chiếu sau không bị che bóng bởi người thuyết trình, nhưng giá thành thường đắc hơn nhiều so với IWB chiếu trước.

2.2.2. Phân loại dựa vào công nghệ chế tạo hoặc nguyên lý hoạt động

Theo tiêu chí này thì IWB được chia thành nhiều loại như sau:

(1) IWB màn hình cảm ứng điện trở (resistive touchscreen), bút được dùng là loại thụ động.

(2) IWB màn hình cảm ứng điện từ (electromagnetic touchscreen), bút được dùng là loại tích cực cảm ứng điện từ (cuộn dây gắn trên đầu bút tương tác với hệ thống dây dẫn trong suốt trên màn hình hoặc đặt phía sau màn hình).

(3) IWB màn hình cảm ứng điện dung (capacitive touchscreen), bút được dùng là loại thụ động.

(4) IWB áp dụng công nghệ màn sáng hồng ngoại (infrared light curtain) hoặc màn sáng laser (laser light curtain), bút được dùng là loại thụ động.

(5) IWB áp dụng công nghệ "phát hủy sự phản xạ toàn phần bên trong" (frustrated total internal reflection). Bút được dùng là loại thụ động. Nguyên lý hoạt

động của IWB loại này được tóm tắt như sau: Ánh sáng hồng ngoại bị phản xạ toàn phần bởi một lớp trong suốt và mềm dẻo trên mặt bảng nên không thể truyền ra ngoài. Khi ấn đầu bút vào mặt bảng, sự phản xạ toàn phần bị phá vỡ, ánh sáng hồng ngoại thoát ra ngoài mặt bảng và nó được thu bởi một camera. Một phần mềm xử lý sẽ chuyển các điểm sáng thu được từ camera thành sự di chuyển con trỏ chuột.

(6) IWB áp dụng công nghệ "bút camera" và "điểm mẫu" (dot pattern). Trên màn hình được tích hợp các điểm mẫu vô cùng nhỏ (không thể thấy được bằng mắt thường). Bút được sử dụng là loại tích cực, nó là bút kỹ thuật số kết nối vô tuyến với máy tính, có một camera nhỏ được gắn ở đầu bút để đọc các điểm mẫu, nhờ đó nó xác định một cách chính xác các điểm tọa độ trên màn hình, tách sự thay đổi tọa độ và gởi về máy tính. Đây là công nghệ bản quyền của tập đoàn Anoto, Thụy Điển.

(7) IWB dựa trên kỹ thuật siêu âm: Bảng loại này có hai nguồn phát siêu âm được đặt ở hai gốc của màn hình và hai cảm biến thu siêu âm đặt ở hai góc còn lại. Sóng siêu âm được truyền trên mặt bảng. Một số điểm (mặt) phản xạ sóng được đặt ở biên của màn hình tạo ra sóng phản xạ cho mỗi nguồn sóng ở các vị trí khác nhau và truyền đến cảm biến với các khoảng cách xác định. Bút được dùng là loại thụ động. Khi đầu bút ấn vào màn hình, sóng siêu âm truyền qua điểm này sẽ bị chặn và cảm biến thu sẽ truyền thông tin sự kiện đến bộ điều khiển.

(8) IWB kết hợp cả hai kỹ thuật siêu âm và hồng ngoại.

(9) Ngoại ra người ta còn phân loại dựa vào tiêu chí di động và IWB được chia làm hai loại: xách tay và cố định.

2.3 Tìm hiểu các giải pháp và kết quả đã được nghiên cứu và thực hiện

Với mục tiêu đưa thêm các chức năng của IWB vào hệ thống projector-computer (hoặc LCD-Computer) có sẵn, trong quá trình tìm kiếm các giải pháp và kết quả đã được nghiên cứu và thực hiện. Chúng tôi đặc biệt chú ý đến IWB sử dụng Wii Remote controller do Johnny Chung Lee, người Đài Loan đề xuất vào năm 2007, báo cáo tại TED năm 2008 và đã đưa các videoclip giới thiệu về hệ thống này trên YouTube (http://johnnylee.net/projects/wii/). Kỹ thuật này đã được các thành viên trong nhóm nghiên cứu của chúng tôi là Lý Phát Hải Linh và Lê Hữu Kỳ Quang, Sở GD-ĐT Hậu Giang, dùng thử và phổ biến trên trang web http://haugiang.edu.vn/toan/.

IWB sử dụng Wii Remote (tay bấm game của bộ chơi game Nintendo) là một cách sử dụng hệ thống projector-computer (hoặc LCD-Computer) sẵn có để thực hiện chức năng của IWB. Thiết bị Wii Remote controller được kết nối với máy tính thông qua kỹ thuật Bluetooth. Bút được dùng thuộc loại tích cực, ở đầu bút có gắn một LED phát ánh sáng hồng ngoại. LED được cấp điện bằng nguồn pin và đóng ngắt cấp điện nhờ một "công tắc thường hở", LED chỉ được cấp điện và phát tia hồng ngoại khi công tắc được ấn. Bảng là màn chiếu của projector (màn hình thụ động) và cũng có thể là màn hình LCD. Khi đặt đầu bút tại một điểm trên bảng và ấn công tắc, chùm tia hồng ngoại phát ra bị phản xạ bởi mặt bảng truyền ngược trở về camera được đặt trong thiết bị Wii Remote. Camera "đọc" các điểm tọa độ trên bảng thông qua các điểm ảnh hồng ngoại mà nó thu được và chuyển đổi thành các cặp giá trị tọa độ, các tín hiệu này được truyền về máy tính thông qua kết nối Bluetooth. Các phần mềm được hỗ trợ để thực hiện chức năng của bảng tương tác xử lý sự thay đổi tọa độ trên màn hình như là các sự kiện chuột, mỗi lần ấn thả công tắc trên bút tương đương với

thao tác nhắp chuột trái (click), hai lần ấn liên tục tương đương với nhắp đôi chuột (double click), ấn – giữ - kéo tương đương với rê chuột (drag).

Để thực hiện hệ thống này chúng ta cần trang bị: projector hoặc màn hình LCD; Wii Remote Controller; bút hồng ngoại (tự làm); máy tính cá nhân (PC) có thể kết nối Bluetooth với thiết bị Wii Remote và được cài đặt các phần mềm sau: NET Framework 3.5; WiimoteWhiteboard (miễn phí, tải về tại địa chỉ http://www.softpedia.com/progDownload/Wiimote-Whiteboard-Download-163702.html) kết hợp với AnnotatePro (mua bản quyền với giá 19,95 USD, tải về tại http://annotatepro.com/) hoặc chỉ dùng SmoothBoard (mua bản quyền với giá 29,95 USD, tải về tại http://www.smoothboard.net/). Tổng chi phí thực hiện khoảng 1.500.000 VNĐ (không tính Projector, PC có hỗ trợ Bluetooth, giả sử đã có sẵn).

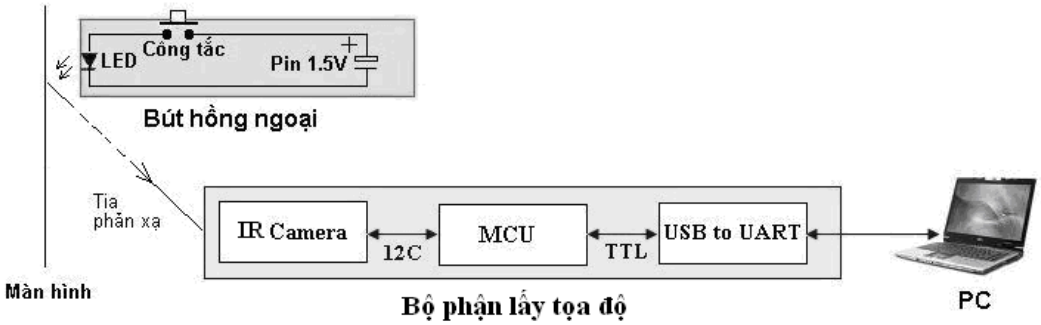
Đây là một công nghệ được coi là ít tốn tiền. Tuy nhiên, với tổng chi phí như trên, chúng tôi cho rằng vẫn còn cao và một giáo viên bình thường vẫn còn phải đắn đo khi muốn trang bị. Chúng tôi cố gắng tìm cách để giảm chi phí xuống thấp hơn nữa, chi phí trọn gói thấp hơn 1.000.000 NĐ. Thêm vào đó còn có nhiều khó khăn cho người dùng, đó là: máy tính phải có khả năng kết nối Bluetooth với thiết bị Wii Remote Controller và phải cài đặt các phần mềm trên máy tính. Việc cài đặt các phần mềm và thực hiện kết nối Bluetooth cũng không phải dễ dàng đối với một người dùng bình thường và cũng là vấn đề phiền phức. Hơn nữa, nguồn cấp điện dùng pin cũng là một hạn chế với công suất tiêu thụ đáng kể của Wii Remote Controller.

3 KẾT QUẢ VÀ THẢO LUẬN

Để có được một sản phẩm của riêng mình, giảm chi phí và khắc phục các hạn chế như đã nêu, chúng tôi đã nghiên cứu và đề ra một số giải pháp. Sau đây là giải pháp mà chúng tôi chọn để thiết kế và thử nghiệm.

3.1 Mô tả phần cứng

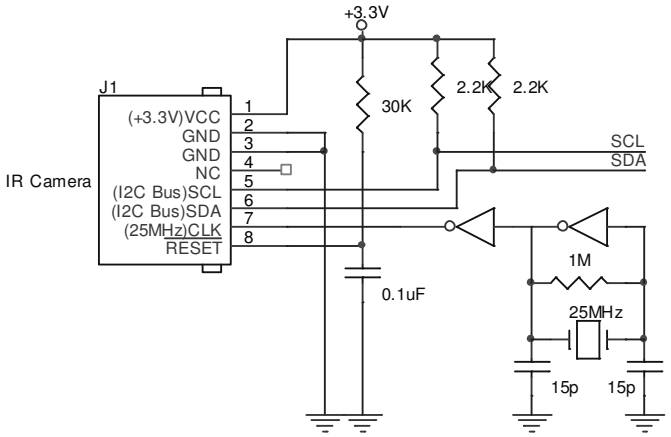
Phần cứng của hệ thống gồm có *bút hồng ngoại* và *bộ phận lấy tọa độ* (hình 3). Bút hồng ngoại khá đơn giản và có thể tự làm một cách dễ dàng. Nó bao gồm một LED phát hồng ngoại, một công tắc thường hở và nguồn cấp điện (pin 1.5V). Bộ phận lấy tọa độ bao gồm camera hồng ngoại (IR camera), mạch điều khiển trung tâm (micro controller unit, MCU) và mạch chuyển đổi USB-UART. Bộ lấy tọa độ có chức năng thu nhận tín hiệu từ camera hồng ngoại, xử lý dữ liệu và truyền về máy tính thông qua chuẩn USB. Camera hồng ngoại thu nhận điểm sáng hồng ngoại và trả về kết quả là tọa độ của điểm sáng (có khả năng nhận được đồng thời nhiều điểm sáng), từ kết quả này mạch điều khiển sẽ tiếp nhận, xử lý và truyền về máy tính để tiếp tục xử lý.



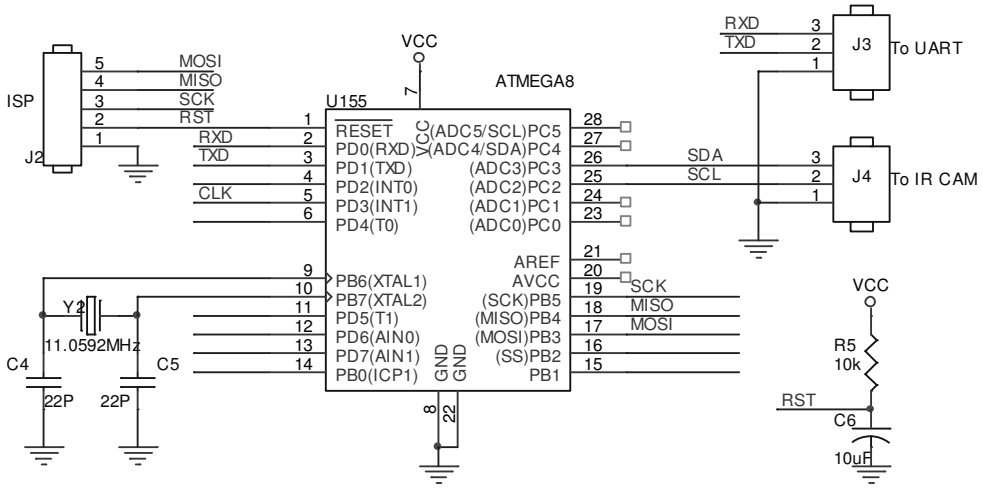
Hình 3: Phần cứng của hệ thống.

Chúng tôi thử nghiệm với loại IR camera được dùng trong thiết bị chơi game Wii Remote Controller. Camera này được tích hợp một bộ xử lý, bộ lọc hồng ngoại và có khả năng nhận đồng thời bốn điểm sáng, vì vậy nó có khả năng xử lý đồng thời bốn đối tượng chuyển động. Dữ liệu ngõ ra của camera là tọa độ của những điểm sáng hồng ngoại, do đó ta không thể dùng camera này để chụp ảnh. Camera có độ phân giải 1024 x 768 điểm, tần số hoạt động là 25MHz, góc nhìn hiệu quả là 33^0 theo chiều ngang và 23^0 theo chiều dọc. Nguồn sáng hồng ngoại nhạy nhất là ở tần số 940nm. Hình 4 trình bày sơ đồ kết nối IR camera với MCU.

MCU có chức năng thiết lập các thông số cho IR camera, thu nhận dữ liệu từ camera (theo chuẩn I2C), xử lý dữ liệu đã nhận được từ camera và truyền về PC theo đúng định dạng đã được quy ước. Hình 5 trình bày sơ đồ nguyên lý của mạch MCU.



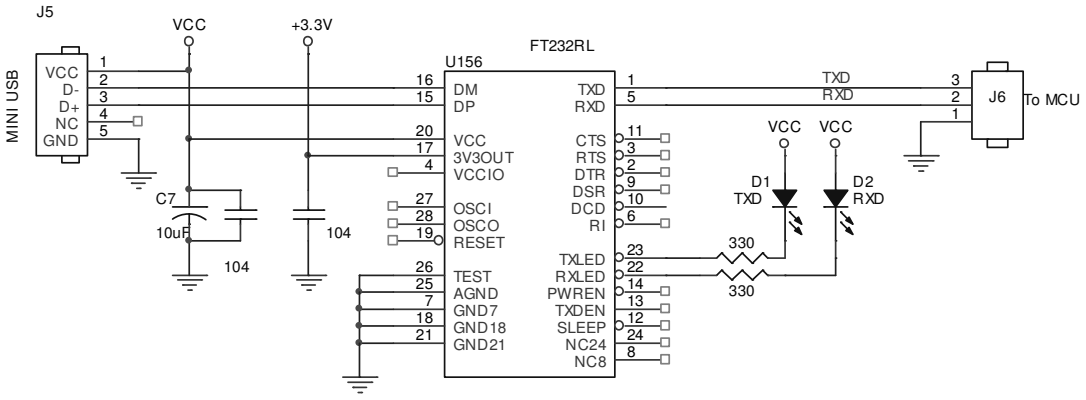
Hình 4: Sơ đồ kết nối IR camera với MCU



Hình 5: Sơ đồ nguyên lý của mạch MCU.

Mạch chuyển đổi USB-UART có chức năng chuyển đổi tín hiệu UART ↔ USB, tạo ra một kết nối giữa máy tính với hệ thống mạch điều khiển camera bên ngoài. Hình 6 trình bày sơ đồ nguyên lý của mạch này.

Vì bộ tiền xử lý kết nối với máy tính qua cổng USB nên nó được cấp điện từ cổng này.



Hình 6: Mạch chuyển đổi USB-UART

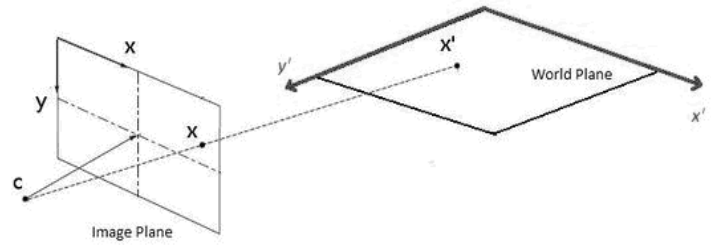
3.2 Phát triển phần mềm

Có hai gói phần mềm được thực hiện cho hệ thống: phần mềm nhúng trên bộ lấy tọa độ kết nối với máy tính và phần mềm trên máy tính.

Phần mềm trên máy tính có nhiệm vụ tiếp nhận và xử lý các cặp tọa độ từ bộ tiền xử lý (bộ phận lấy tọa độ) chuyển về để thực hiện các chức năng của IWB. Chúng tôi đã phát triển gói phần mềm này bằng ngôn ngữ Java (lý do đơn giản là vì chúng tôi đã quen thuộc với ngôn ngữ này). Ngoài các gói trong thư viện chuẩn của Java chúng tôi đã sử dụng thêm các gói mở rộng như: Robot (điều khiển con trỏ chuột), javax.comm (nhận dữ liệu từ cổng com), Jama.matrix (tính toán trên ma trận).

Các vấn đề cần giải quyết khi thiết kế phần mềm bao gồm:

Định khung tác động (Calibration): Khung tác động là phạm vi mà trong đó camera chụp được điểm sáng. Đó là mặt phẳng toạ độ thật (world plane). Khi chọn vị trí và hướng của *bộ lấy tọa độ* thì khung tác động có thể bị lệch so với mặt bảng. Vì vậy chúng ta cần xác định vị trí và kích thước khung tác động đúng với vị trí và kích thước của bảng. Việc làm này được gọi là định khung tác động. Chú ý rằng, các cặp toạ độ điểm sáng truyền về máy tính lại thuộc về mặt phẳng toạ độ ảnh (image plane). Vì vậy, chúng ta cần một giải thuật chuyển đổi tọa độ giữa mặt phẳng tọa độ thật và mặt phẳng toạ độ ảnh. Chúng tôi đã sử dụng giải thuật 2D Homography. Đây là giải thuật khá phổ biến và chúng ta dễ dàng tìm thấy trên mạng. Homography thực chất là một phép biến đổi xạ ảnh (projective transformation), nó thực hiện sự chuyển đổi một điểm trong không gian này sang một điểm trong không gian khác và ngược lại. Hình 7 minh hoạ việc chuyển đổi một điểm x' trong mặt phẳng thật tương ứng thành điểm x trong mặt phẳng ảnh. Phần mềm định khung tác động sẽ hiển thị lần lượt ký hiệu (chọn là ⌖) trên bốn góc của khung tác động, bắt đầu là góc trên phía trái màn hình, có tọa độ là (0,0). Ta chỉnh bộ lấy tọa độ cho ký hiệu ⌖ đầu tiên này nằm ở sát góc, sau đó đặt đầu bút vào đó và ấn công tắc (ta gọi thao tác này là nhắp, tương ứng như nhắp chuột trái), ký hiệu đầu tiên sẽ lặn đi và ký hiệu góc



Hình 7: Minh họa biến đổi xạ ảnh

phải trên xuất hiện. Ta tiếp tục nhắp vào ký hiệu này và nó sẽ lặn đi, ký hiệu góc phải dưới xuất hiện, tiếp tục cho đến cuối cùng là góc trái dưới.

Tạo thanh công cụ (Toolbar): Thanh công cụ là một giao diện đồ họa người dùng hiện lên màn hình khi chương trình được khởi động, trên đó là các nút lệnh được biểu diễn bằng các biểu tượng thể hiện chức năng của nút lệnh đó (Hình 8: Một ví dụ minh họa thanh công cụ). Khi muốn chọn một chức năng, ta nhắp bút hồng ngoại vào biểu tượng tương ứng.

Hình 8: Minh họa một thanh công cụ

Để bút hồng ngoại có vai trò như chuột máy tính, ta phải lập trình định nghĩa các sự kiện chuột (nhắp, thả, rê và nhắp đôi). Ta cần phải khai báo một luồng (stream) để đọc các điểm ảnh trả về, mỗi điểm có ba thông tin: tọa độ ngang, tọa độ dọc và thời gian. Có một số vấn đề cần giải quyết khi phát triển phần mềm như sau: Mỗi thao tác nhắp bút không diễn ra tức thời mà có thời gian bắt đầu (ấn) và kết thúc (thả), chùm tia sáng hồng ngoại phát ra từ bút có bán kính lớn (độ hội tụ kém), công tắc có thể bị rung hoặc dội khi ấn, bị nhiễu do các nguồn sáng khác. Vì vậy sẽ có nhiều điểm (tọa độ) trả về tương ứng với một thao tác nhắp – thả và số lượng điểm cũng khác nhau cho những lần nhắp khác nhau. Số điểm trả về khác nhau về không gian (tọa độ) và thời gian (thời điểm). Vì vậy, vấn đề là phải chọn vùng tác động, điểm tác động và thời gian cho mỗi thao tác nhắp bút. Qua thực nghiệm, chúng tôi nhận thấy vùng tác động có kích thước lớn nhất là 30x30 điểm và xác định được khoảng thời gian lớn nhất giữa 2 điểm liền kề trong cùng một thao tác nhắp (ký hiệu là t_{max} = 0.005giây), từ đó quyết định là thao tác nhắp chuột nếu tổng số tọa độ đếm theo chiều ngang hoặc chiều dọc nhỏ hơn 30 và chọn điểm tác động là điểm thứ 2 trong luồng dữ liệu thu được, và thời điểm xuất hiện điểm cuối trong luồng là thời điểm thả chuột, nếu thời gian xuất hiện hai điểm liền kề lớn hơn t_{max} thì chuyển sang thao tác nhắp kế tiếp, nếu số điểm xuất hiện theo chiều ngang hoặc chiều dọc quá 30 thì quyết định là thao tác rê và nếu khoảng thời gian giữa hai thao tác nhắp nhỏ hơn 0.05 giây thì quyết định là nhắp đôi . Giải thuật chi tiết cho phần mềm thực hiện chức năng của IWB tương đối dài dòng, chúng tôi sẽ trình bày trong các số tiếp theo.

4 KẾT LUẬN VÀ ĐỀ NGHỊ

Giải pháp gia tăng tiện ích của IWB cho hệ thống trình chiếu Projector/LCD kết hợp với PC là giải pháp nhằm nâng cao hiệu quả giảng dạy với chi phí thấp, giáo viên có thể tự trang bị bộ công cụ hỗ trợ này, trong điều kiện các trường đã trang bị các bộ Projector/LCD và PC. Tất nhiên, chúng ta cũng có thể sử dụng kỹ thuật này trong báo cáo khoa học, tiếp thị,… Trong trường hợp giảng dạy hoặc trình bày cho một nhóm nhỏ, chúng ta cũng có thể sử dụng màn hình của máy tính để làm bảng tương tác. Khi đó không cần projector hoặc màn hình LCD lớn.

Với hiệu quả mà IWB mang lại, nhu cầu về IWB trên thế giới sẽ gia tăng, công nghiệp sản xuất IWB sẽ phát triển nhanh chóng về số lượng và giá thành cũng sẽ giảm dần. Trong tương lai gần, IWB sẽ được sử dụng đại trà trong trường học ở nước ta. Tuy nhiên, chúng tôi nghĩ rằng giải pháp gia tăng tiện ích cho các thiết bị có sẵn cũng là giải pháp có hiệu quả kinh tế cao vì vậy nó không chỉ đáp ứng cho nhu cầu trước mắt.

Khi có sản phẩm của chính mình (phần cứng lẫn phần mềm), chúng tôi sẽ phổ biến kỹ thuật này trong các trường học, từ phổ thông cho đến đại học. Chúng tôi cũng sẽ tìm cách giảm chi phí đến mức thấp nhất có thể để một giáo viên bình thường có thể tự trang bị.

TÀI LIỆU THAM KHẢO

1 Michelle R. Davis, 2007. Whiteboard Inc Interactive features fuel demand for modern chalkboards. http://www.edweek.org/dd/articles/2007/09/12/02board.h01.html.

2 Moss, G., Armstrong, V., Jewitt,…, 2010. The Interactive Whiteboards, Pedagogy and Pupil Performance Evaluation: An Evaluation of the Schools Whiteboard Expansion (SWE) Project: London Challenge. Institute of Education (2007).

3 Lý Phát Hải Linh, 2010. Giới thiệu bảng tương tác giá rẻ. Chuyên san khoa học của tỉnh Hậu Giang.

4 http://www.besa.org.uk/besa/documents/index.jsp?mode=search&topic=0&r=true&b=true&o=true&keyword=ICT+in+UK+State+Schools&x=41&y=13

5 http://johnnylee.net/projects/wii/

6 http://davidlongman.com/documents/Interactive%20Whiteboards/IWB%20Research%20London%20Challenge%20RR816.pdf

7 http://en.wikipedia.org/wiki/Interactive_whiteboard

8 http://www.ted.com/pages/about

9 http://mmlab.disi.unitn.it/wiki/index.php/2D_Homography:_Algorithm_and_Tool

10 Đoàn Hòa Minh, 2011. Báo cáo tổng kết đề tài NCKH cấp cơ sở NGHIÊN CỨU THIẾT KẾ VÀ THỰC HIỆN BẢNG ĐIỆN TỬ TƯƠNG TÁC, mã số T2011-47, Trường ĐH Cần Thơ.

PHÁT TRIỂN PHẦN MỀM THÊM CHỨC NĂNG CỦA BẢNG TƯƠNG TÁC CHO HỆ THỐNG PROJECTOR-COMPUTER HOẶC LCD-COMPUTER

Đoàn Hòa Minh[1], Nguyễn Khắc Nguyên[2], Bùi Minh Quân[1]

Abstract

To add the functionalities of interactive whiteboards (IWB) to a projector-computer system or an LCD-computer, we established the connection of hardware devices including an infrared pen, a location detecting unit and a computer. There are two software packages synchronized with this hardware system, those are: the software embedded on the location detecting unit and the software setup in the computer. In this article, we will present the algorithms for the development of these software packages

Key words: software packages, embedded software, setup software, calibration, 2D Homography algorithm, toolbar.

Title: Development softwares for adding the functions of interactive whiteboard into the system of projector-computer or LCD-Computer.

Tóm tắt

Để thêm vào chức năng bảng tương tác (interactive whiteboard, IWB) cho hệ thống computer (PC)-projector hoặc PC-LCD, chúng tôi đã thực hiện mô hình kết nối gồm các thiết bị phần cứng là bút hồng ngoại (infrared pen), bộ phận lấy tọa độ (detecting location unit) và computer. Kèm theo đó là hai gói phần mềm, phần mềm nhúng trên bộ phận lấy tọa độ và phần mềm cài đặt trên trên computer. Trong bài này, chúng tôi trình bày chi tiết về giải thuật để phát triển các gói phần mềm.

Từ khóa: Gói phần mềm, phần mềm nhúng, phần mềm cài đặt, sự định khung, giải thuật 2D Homography

1 GIỚI THIỆU

Với hiệu quả mà IWB mang lại trong đào tạo, nhu cầu về IWB trên thế giới sẽ gia tăng, công nghiệp sản xuất IWB sẽ phát triển nhanh chóng về số lượng và giá thành cũng sẽ giảm dần. Trong tương lai gần, IWB sẽ được sử dụng phổ biến trong trường học. Tuy nhiên, chúng tôi nghĩ rằng giải pháp gia tăng tiện ích cho các thiết bị có sẵn cũng là giải pháp có hiệu quả kinh tế cao vì vậy nó không chỉ đáp ứng cho nhu cầu trước mắt. Trong đó, giải pháp gia tăng tiện ích của IWB cho hệ thống trình chiếu Projector/LCD kết hợp với PC là giải pháp nhằm nâng cao hiệu quả giảng dạy với chi phí thấp, trong điều kiện các trường đã trang bị các bộ Projector/LCD và PC. Tất nhiên, chúng ta cũng có thể sử dụng kỹ thuật này trong báo cáo khoa học, tiếp thị,… Trong trường hợp giảng dạy hoặc trình bày cho một nhóm nhỏ, chúng ta cũng có thể sử dụng màn hình của máy tính để làm bảng tương tác. Khi đó không cần projector hoặc màn hình LCD lớn.

Hiện nay, có một vài nhóm nghiên cứu theo hướng này. Trong đó, giải pháp dùng tay chơi game Wii Remote controller của Nintendo do Johnny Chung Lee, người Đài

[1] Khoa Công nghệ thông tin & Truyền thông, Trường Đại học Cần Thơ.
[2] Khoa Công nghệ, Trường Đại học Cần Thơ.

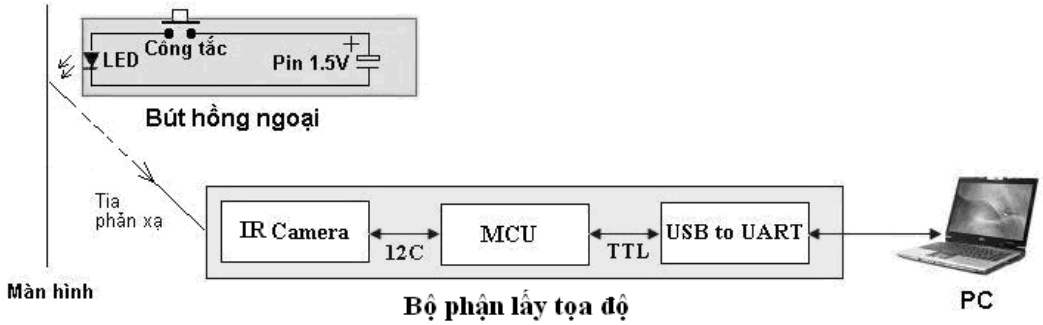
Loan, đề xuất vào năm 2007 là một nghiên cứu tiêu biểu (http://johnnylee.net/projects/wii/). Để thực hiện hệ thống này chúng ta cần trang bị: projector hoặc màn hình LCD; Wii Remote Controller; bút hồng ngoại (tự làm); PC có thể kết nối Bluetooth với thiết bị Wii Remote và được cài đặt các phần mềm sau: NET Framework 3.5; WiimoteWhiteboard (version,URL http://www.softpedia.com/, miễn phí) kết hợp với AnnotatePro (URL: http://annotatepro.com/, giá 19,95 USD) hoặc chỉ dùng SmoothBoard của Creative Commons (version 2.0, phát hành năm 2010, URL: http://www.smoothboard.net/, giá 29,95 USD). Tổng chi phí thực hiện khoảng 1.500.000 VNĐ (không tính Projector, PC có hỗ trợ Bluetooth, giả sử đã có sẵn).

Mục tiêu của chúng tôi là tạo ra sản phẩm của chính mình (phần cứng lẫn phần mềm) với giá thành thấp nhất, và sẽ phổ biến kỹ thuật này trong các trường học, từ phổ thông cho đến đại học.

Trong bài viết này, chúng tôi trình bài giải thuật để phát triển các gói phần mềm thêm chức năng của IWB vào hệ thống projector-PC hoặc LCD-PC.

2 NỘI DUNG NGHIÊN CỨU

Hệ thống được thiết kế với sơ đồ phần cứng như hình 1.



Hình 1: Sơ đồ phần cứng

Bút hồng ngoại là dụng cụ dùng để viết, vẽ, xóa,… và kích hoạt các phần mềm ứng dụng trên máy tính. Khi bấm công tắc LED sẽ phát ra tia sáng hồng ngoại chiếu vào mặt bảng và phản xạ về bộ phận lấy tọa độ. *Bộ phận lấy tọa độ* gồm IR camera, mạch điều khiển trung tâm (Micro controller unit, MCU) và mạch chuyển đổi USB-UART. Bộ phận lấy tọa độ có nhiệm vụ thu các điểm sáng hồng ngoại, mã hóa và chuyển đổi thành dữ liệu tọa độ thích hợp và truyền về PC qua cổng USB.

Để hệ thống có thể hoạt cần phải có hai gói phần mềm: gói *phần mềm nhúng trên chip vi điều khiển của bộ phận lấy tọa độ* (sẽ gọi là "phần mềm nhúng") và gói *phần mềm cài đặt trên computer* (sẽ gọi là "phần mềm cài đặt").

2.1 Phần mềm nhúng (Embedded software)

Phần mềm nhúng trên chíp vi điều khiển của mạch điều khiển trung tâm thực hiện các chức năng như sau:

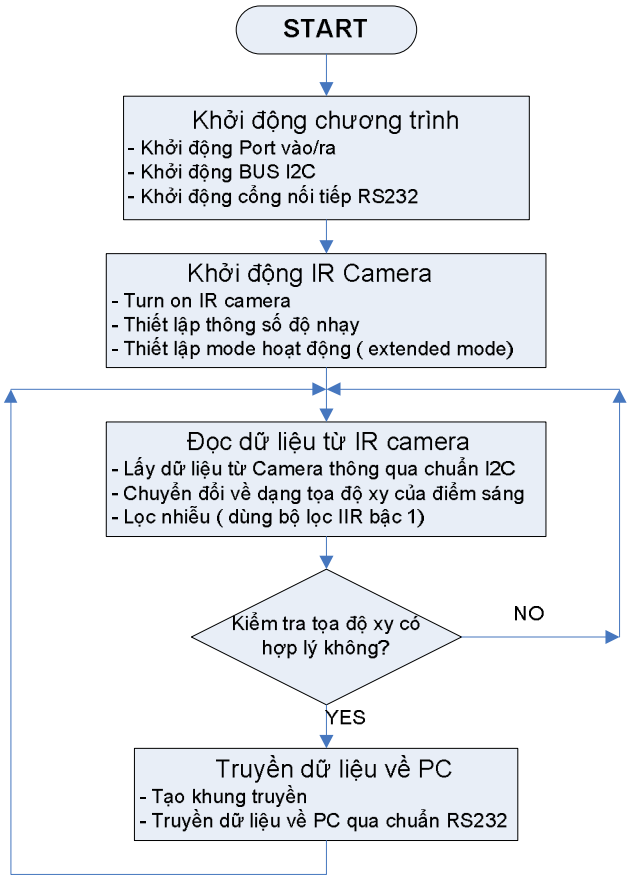
- Kết nối với IR camera nhằm mục đích: Thiết lập các thông số hoạt động cho camera như độ nhạy, độ phân giải, định dạng dữ liệu ra và mode hoạt động, . . . ; thu thập tín hiệu tọa độ điểm sáng hồng ngoại từ IR camera.
- Xử lý tín hiệu nhận được từ IR camera: lọc thông thấp nhằm loại bỏ một phần nhiễu, chuyển đổi tín hiệu tọa độ cho phù hợp với mục đích sử dụng.
- Định khung dữ liệu tọa độ của điểm sáng hồng ngoại mà IR nhận được gởi về PC thông qua chuẩn RS232 (dùng chip chuyển đổi USB ↔ RS232).

Phần mềm nhúng trên chíp vi điều khiển được lập trình bằng ngôn ngữ C, dùng trình biên dịch CodevisionAVR để biên dịch và nạp trình vào chíp.

Khung dữ liệu mà MCU truyền về PC có dạng như sau:

X (2byte)	Y (2byte)	\r (1byte)	\n(1byte)

Lưu đồ chương trình trên chip vi điều khiển ATmega8:



Hình 2: Lưu đồ chương trình trên chip vi điều khiển ATmega8

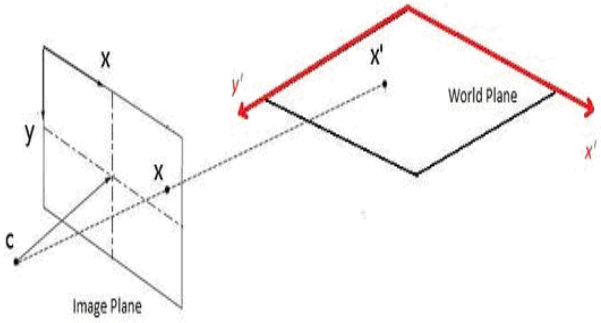
2.2 Phần mềm cài đặt (setup software)

Phần mềm cài đặt trên máy tính có nhiệm vụ tiếp nhận và xử lý các cặp tọa độ từ *bộ phận lấy tọa độ* chuyển về để thực hiện các chức năng của IWB. Chúng tôi đã phát triển gói phần mềm này bằng ngôn ngữ Java (lý do đơn giản là vì chúng tôi đã quen thuộc với ngôn ngữ này). Ngoài các gói trong thư viện chuẩn của Java chúng tôi đã sử dụng thêm

các gói mở rộng như: javax.comm (nhận dữ liệu từ cổng com), Jama.matrix (tính toán trên ma trận).

2.2.1 Định khung tác động với giải thuật 2D Homography

IR camera ghi nhận các tọa độ của điểm sáng hồng ngoại thay đổi theo hành động của người dùng từ lúc "nhấp bút" (bấm công tắc của bút hồng ngoại) đến lúc "thả bút" (thả công tắc của bút hồng ngoại) trong không không gian làm việc của nó, không gian thật (world plane). Sau đó PC xử lý tín hiệu lưu trữ (dãy tọa độ) và trình bài lại các hành động đó trong một không gian khác, không gian ảnh (image plane). Giải thuật Homography được dùng để thực hiện việc chuyển đổi này. Một hành động của người dùng được tính từ lúc nhấp bút đến lúc thả bút (một tác vụ). Vấn đề là làm thế nào để PC biết đây là một click chuột (mouse click), một rê chuột (mouse drap) hay một click đôi (double click).



Hình 3: Minh học biế đổi xạ ảnh

Homography thực chất là một phép biến đổi xạ ảnh (projective transformation). Đây là giải thuật khá phổ biến và chúng ta dễ dàng tìm thấy trên mạng. Một điểm x_i trong không gian thứ nhất tương ứng với điểm x_i' trong không gian thứ hai.

Việc chuyển đổi giữa các điểm được tính thông qua một ma trận Homography (H) được định nghĩa như sau:

$$\begin{bmatrix} x_i' \\ y_i' \\ w_i' \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ w \end{bmatrix}$$

Ngược lại, tọa độ điểm x_i được tính từ x_i' thông qua ma trận H^{-1} bởi phương trình: $\mathbf{x_i = H^{-1} x'_i}$

Để tính được ma trân H cần có 8 điểm tọa độ, 4 điểm đầu vào và 4 điểm đầu ra. Các điểm đầu vào được chọn tại 4 góc của không gian thật, 4 điểm đầu ra được chọn tại 4 góc của không gian ảnh (màn hình của PC).

2.2.2 Định nghĩa các sự kiện chuột

Để bút hồng ngoại thực hiện được vai trò như chuột máy tính, ta phải lập trình định nghĩa các sự kiện chuột (nhắp, thả, rê và nhắp đôi). Mỗi hành động được định nghĩa bởi một phương phương thức (method) trong lớp (class) Robot của Java, phương thức mousePress định nghĩa hành động bấm công tắc của bút hồng ngoại; mouseRelease định nghĩa hành động thả công tắc và mouseMove định nghĩa hành động di chuyển bút. Từ các phương thức này ta định nghĩa các sự kiện chuột:

- mouseClick = mousePress + mouseRelease (nhắp chuột)
- mouseDrap = mousePress + mouseMove +… + movemove + mouseRelease (rê chuột)
- doubleClick = mouseClick + mouseClick (nhắp đôi)

a. Cấu trúc lưu trữ

Ta cần phải khai báo một dòng dữ liệu (stream) để đọc các điểm ảnh trả về, mỗi điểm có ba thông tin: tọa độ ngang, tọa độ dọc và thời gian. Khi *bộ phận lấy tọa độ* thu được tin hiệu từ bút hồng ngoại trong không gian làm việc, sẽ lưu thành một điểm (Point) vào danh sách liên kết (một mảng Point được đặt tên là **pe**). Cấu trúc của Point bao gồm ba thành phần:

X : tọa độ trục x (hoành độ), loại biến int 4 byte.
Y: tọa độ trục y (tung độ), loại biến int 4 byte.
Time: thời gian hệ thống (system time) của PC, loại long 64 bit.

b. Các vấn đề cần giải quyết

Mỗi thao tác nhắp bút không diễn ra tức thời mà có thời gian bắt đầu (ấn) và kết thúc (thả), chùm tia sáng hồng ngoại phát ra từ bút có bán kính lớn (độ hội tụ kém), công tắc có thể bị rung hoặc dội khi ấn. Do đó sẽ có nhiều điểm (tọa độ) trả về tương ứng với một thao tác "nhắp thả" và số lượng điểm cũng khác nhau cho những lần nhắp thả khác nhau. Số điểm trả về khác nhau về không gian (tọa độ) và thời gian (thời điểm). Có quá nhiều điểm trả được lưu vào mảng **pe** chờ xử lý, có thể tràn bộ nhớ hoặc làm cho tốc độ xử lý các sự kiện sẽ chậm.Vì vậy, có một số vấn đề cần phải quyết định như sau: giải pháp xác định điểm bắt đầu và điểm kết thúc tác vụ của người dùng; giải pháp lọc bỏ bớt số điểm trong một thao tác; xác định một tác vụ của người dùng là mouseClick, mouseDrap, doubleClick; làm thế nào để khắc phục được nhiễu.

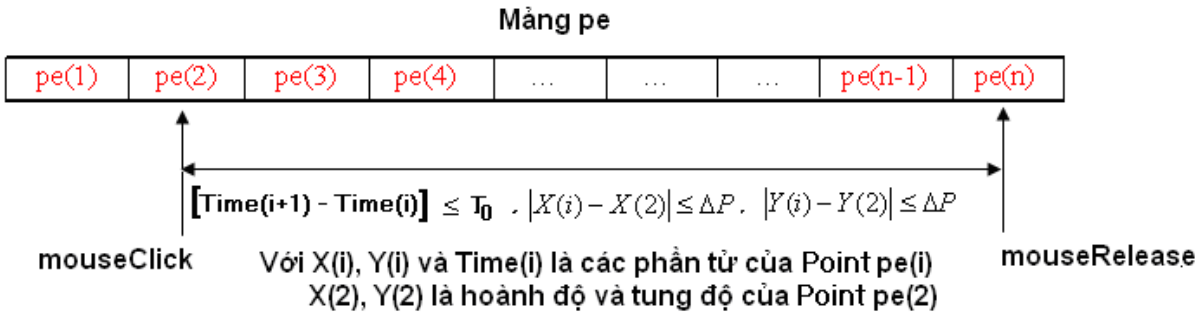
c. Các giải pháp được đề xuất

- *Nguyên tắc thời gian:* nguyên tắc này dựa vào thành phần Time của các Point. Nó được dùng để quyết định thời gian kết thúc của một tác vụ (mouseClick hay mouseDrap). Ta chọn thời gian hệ thống (system time) của máy tính làm chuẩn.

- *Nguyên tắc không gian:* nguyên tắc này dựa vào khoảng cách tọa độ giữa các điểm trong cùng tác vụ nhắp thả. Nguyên tắc này được dùng để quyết định một tác vụ của người dùng là mouseClick hay mouseDrap.

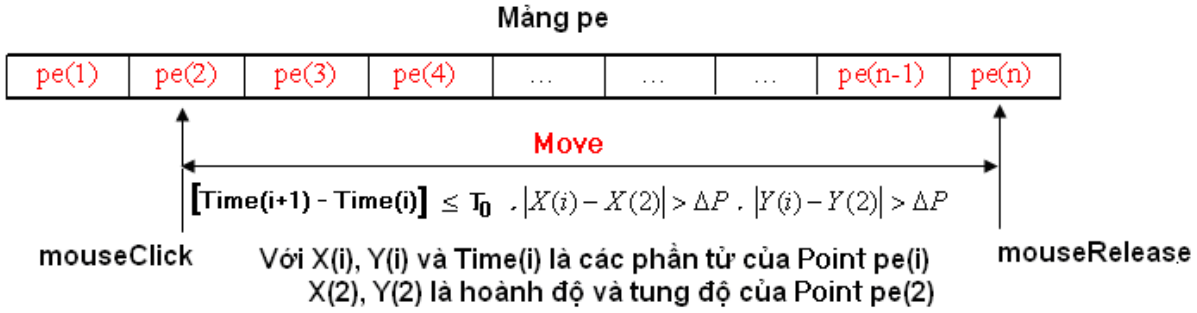
- *Chọn điểm nhắp chuột (pointClick):* ta cần chọn một điểm pe(i) trong mảng **pe** là điểm nhắp chuột, để từ đó so sánh với các điểm khác trong mảng nhằm ra quyết định thao tác của người dùng là mouseClick, mouseDrap hay doubleClick, chúng tôi chọn điểm p(2) làm pointClick mà không chọn pe(1) vì điểm pe(1) được cho là chưa ổn định.

- *Quyết định một tác vụ là mouseClick:* khi nhận được một mảng (qe) các Point, phần mềm sẽ kiểm tra các thành phần các điểm pe(i), nếu có hiệu [Time(i+1)- Time(i)] của hai điểm liên tiếp lớn hơn giá trị qui định T_0 nào đó hoặc thời gian chờ điểm kế tiếp lớn hơn T_0 thì quyết định thả chuột (mouseRelease) và chờ tác vụ mới, đồng thời nếu trị tuyệt đối của hiệu hoành độ $|X(i) - X(2)| \le \Delta P$ (với ΔP là một giá trị qui định có đơn vị là pixel) hoặc trị tuyệt đối của hiệu tung độ $|Y(i) - Y(2)| \le \Delta P$ thì quyết định là mouseClick.



Hình 4: Minh họa mảng pe trong tác vụ nhắp chuột.

- *Quyết định một tác vụ là mouseDrap:* khi nhận được một mạng (qe) các Point, chương trình kiểm tra tọa độ các điểm kế tiếp trong cùng một tác vụ, nếu trị tuyệt đối của hiệu hoành độ $|X(i)-X(2)|>\Delta P$ hoặc trị tuyệt đối của hiệu tung độ $|Y(i)-Y(2)|>\Delta P$ thì quyết định là mouseDrap, và sẽ để con trỏ di chuyển qua các điểm mới. Khi có một điểm thỏa điều kiện trên thì các điểm sau trong cùng tác vụ không cần kiểm tra nữa. Chương trình kiểm tra các hiệu [Time(i+1)- Time(i)] của hai điểm liên tiếp, khi giá trị này lớn hơn giá trị qui định T_0 nào đó hoặc thời gian chờ điểm kế tiếp lớn hơn T_0 thì quyết định thả chuột (mouseRelease) và kết thúc một tác vụ mouseDrap.



Hình 5: Minh họa mảng pe trong tác vụ mouseDrap

- *Quyết định một tác vụ là doubleClick:* tác vụ doubleClick gồm hai tác vụ mouseClick liên tiếp với khoảng cách thời gian nhỏ hơn giá trị qui định ΔT.

- *Chọn* T_0, ΔT, ΔP: dựa vào quan sát thực nghiệm chúng tôi đã chọn các giá trị này như

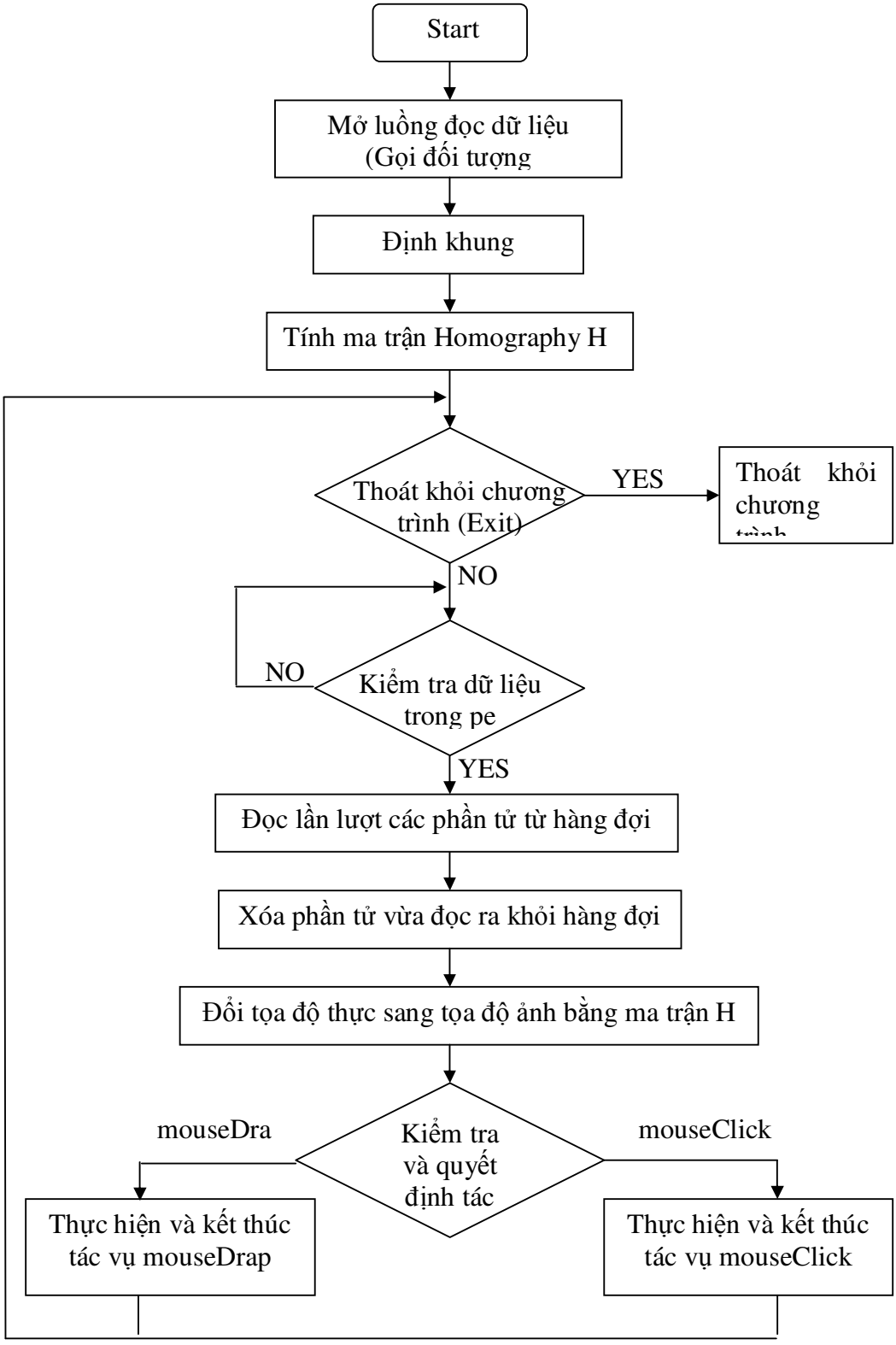
sau: $T_0 = 0.005s$, $\Delta T = 0.05$s, $\Delta P = 30$ pixel.

Gói phần mềm cài đặt được thiết kế gồm 12 lớp:

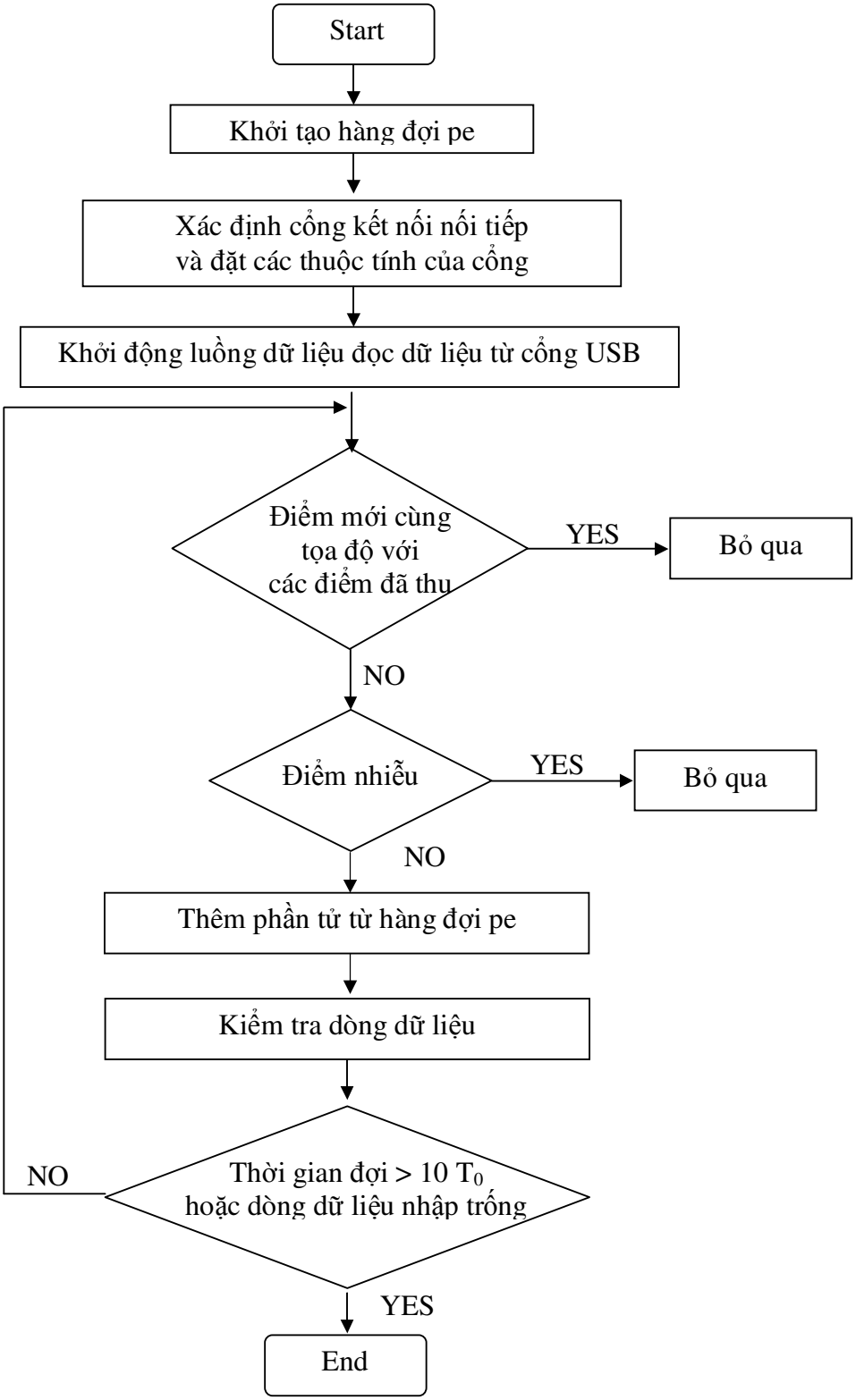
1. *MyApp.java :* khởi động kết nối thiết bị nhận dữ liệu, bắt đầu định khung, quyết định tác vụ là click, drap, doubleclick. Khởi động thanh công cụ điều khiển thanh công cụ vẽ.

2. *SimpleWrite.java:* kiểm tra cổng kết nối thiết bị, đọc dữ liệu từ thiết bị, kiểm tra dữ liệu và đưa vào danh sách chờ xử lý.
3. *Homography.java:* thực hiện chuyển đổi giữa mặt phẳng thực và mặt phẳng ảnh, định khung cho mặt phẳng thực.
4. *ScreenCapture.java:* chụp hình desktop để xác định độ phân giải, lấy 4 tọa độ trên 4 góc của màn hình PC làm các điểm đầu ra tính ma trận H.
5. *WildcardFileFilter.java:* cho phép tìm kiếm đường dẫn phần mềm thực thi (powerpoint, explorer,…) linh động vơi các phiên bản khác nhau.
6. *CanvasPanel.java:* lắng nghe và xử lý các sự kiện chuột.
7. *MainTool.java:* quản lí và điều khiển các thanh công cụ.
8. *ColorBuutonPanel.java:* khởi động bảng chọn màu.
9. *StarFrame.java:* khởi tạo khung cho phép chọn cổng COM kết nối thiết bị, xác nhận việc định khung màn hình làm việc thành công.
10. *DrawShape.java:* tạo thanh công cụ trình chiếu powerpoint, khỏi tạo chức năng tạo bảng trắng, cho phép vẽ hình cơ bản trên bảng trắng và desktop đang trình chiếu.
11. *MyPoint.java:* định nghĩa cấu trúc điểm (Point).
12. *WiiBoardJ.java:* hiển thị khung hiệu chỉnh khung làm việc, xác định 4 gốc tọa độ điểm đầu vào xây dựng ma trận Homography.

Sau là lưu đồ thuật toán của hai lớp quan trọng, đó là SimpleWrite và MyApp.



Hình 6: Lưu đồ thuật toán lớp MyApp.java



Hình 7: Lưu đồ thuật toán lớp SimpleWrite.java

3 KẾT QUẢ ĐẠT ĐƯỢC

3.1 Kết quả đạt được

- Thiết kế và lắp ráp phần cứng gồm: bộ phận lấy tọa độ kết nối với PC qua cổng USB và bút hồng ngoại.
- Thiết kế và thực hiện hai gói phần mềm: phần mềm nhúng trên chip vi điều khiển của thiết bị lấy tọa độ và phần mềm cài đặt trên PC.
- Hệ thống thực thi được đầy đủ các chức năng đã đề ra:
  + Duyệt các thư mục, tập tin và mở tập tin;
  + Điều khiển trình chiếu các slide PowerPoint;
  + Vẽ, viết, đánh dấu và xóa trên bảng (màn chiếu của projector);
  + Điều khiển truy cập internet;
  + Lưu các dữ liệu đã viết và vẽ.

Các mặt còn hạn chế:

- Còn một số chức năng của bảng tương tác đã được hỗ trợ nhưng chưa hoàn thiện như: thao tác vẽ tự do còn khó khăn; chưa thực hiện được đoạn cong nhỏ ở điểm khởi đầu, vì vậy chưa viết chữ nhỏ được; chưa xử lý thay đổi bề rộng nét vẽ; chưa có thể di chuyển/quay các đối tượng hình vẽ trên màn hình; chưa thay đổi độ lớn của gôm xóa;…
- Còn một số chức năng chưa được thực hiện: nhấp chuột phải, thay đổi nét vẽ; highligh; tạo lưới (grid) trên màn hình; snapshot; …
- Trường tương tác nhỏ và việc định hướng IR camera còn khó khăn.

4 KẾT LUẬN VÀ ĐỀ NGHỊ

Cho đến thời điểm này, chúng tôi đã cơ bản hoàn thành các mục tiêu đã đề ra. Chúng tôi mong muốn được nhiều người đóng góp ý kiến về nhu cầu ứng dụng kết quả của đề tài này, về ưu và nhược điểm của hệ thống, các chức năng cần thêm vào, ... và các ý kiến khác.

Đề nghị nhóm nghiên cứu cố gắng hoàn chỉnh sản phẩm trong thời gian sớm nhất và giảm giá thành đến mức thấp nhất có thể được, nhằm hỗ trợ quí thầy, cô có được công cụ tốt để nâng cao chất lượng dạy và học.

TÀI LIỆU THAM KHẢO

1 Đoàn Hòa Minh, 2011. Báo cáo tổng kết đề tài NCKH cấp cơ sở NGHIÊN CỨU THIẾT KẾ VÀ THỰC HIỆN BẢNG ĐIỆN TỬ TƯƠNG TÁC, mã số T2011-47, Trường ĐH Cần Thơ.

2 David Kriegman, 2007. Homography Estimation. Computer Vision I CSE 252A.

3 http://johnnylee.net/projects/wii/

4 http://en.wikipedia.org/wiki/Interactive_whiteboard

5 http://www.ted.com/pages/about

6 http://mmlab.disi.unitn.it/wiki/index.php/2D_Homography:_Algorithm_and_Tool

7 http://cseweb.ucsd.edu/classes/wi07/cse252a/homography_estimation/homography_estimation.pdf

8 http://www.csse.uwa.edu.au/~pk/Research/MatlabFns/Projective/homography2d.m

10 NĂM NGHIÊN CỨU KHAI MỎ DỮ LIỆU

Đỗ Thanh Nghị[1], *Phạm Nguyên Khang*[2]

Abstract: *Knowledge discovery in databases (KDD) and data mining is one of top 10 emerging technologies of the 21st century (Technology Review of MIT, February 2001). Therefore, KDD and data mining are well-known as an important domain in computer sciences. In this paper, we present the KDD process and data mining technology. And then, we survey top 10 challenges problems in data mining, our activities research on the data mining and knowledge discovery field over the last 10 years and future works.*

Keywords: Knowledge discovery in databases, data mining, top 10 data mining algorithms, top 10 challenges problems.

Title: 10 Years Research on Data Mining

Tóm tắt: *Khám phá tri thức và khai mỏ dữ liệu được xem là một trong top 10 công nghệ nổi bật của thế kỷ XXI (Tạp chí công nghệ số ra tháng 2/2001 trường MIT). Nghiên cứu, học tập, làm chủ được công nghệ khám phá tri thức và khai mỏ dữ liệu có ý nghĩa quan trọng, quyết định cho sự phát triển của thầy trò Khoa Công Nghệ Thông Tin Trường Đại học Cần Thơ. Bài viết trình bày quá trình công nghệ khám phá tri thức và khai mỏ dữ liệu. Chúng tôi cũng điểm qua các vấn đề nan giải của lãnh vực khai mỏ dữ liệu, các thành tựu nghiên cứu của Khoa trong 10 năm qua và hướng phát triển sắp tới của lãnh vực khám phá tri thức và khai mỏ dữ liệu.*

Từ khóa: Khám phá tri thức, khai mỏ dữ liệu, top 10 giải thuật khai mỏ dữ liệu, top 10 vấn đề nan giải.

1. GIỚI THIỆU

Trong những năm 1990, cuộc cách mạng kỹ thuật số cho phép số hóa thông tin dễ dàng và chi phí thấp, thêm vào đó là sự phát triển của công nghệ thông tin bao gồm cả phần cứng lẫn phần mềm, công nghệ truyền thông, web, internet đã góp phần đưa máy tính vào các sinh hoạt thường nhật của con người. Tất cả các hoạt động kinh doanh, vui chơi giải trí, nghiên cứu khoa học, giáo dục, truyền thông đều có sự hỗ trợ của máy tính. Hệ quả kéo theo khối lượng lớn dữ liệu được sinh ra và lưu trữ trong các cơ sở dữ liệu, thiết bị lưu trữ như băng từ, đĩa từ. Từ năm 1999, Giáo sư P. Lyman và các cộng sự của ông ở Đại học Berkeley đã tiến hành thống kê dữ liệu được sinh ra hằng năm trên toàn cầu. Kết quả chỉ trong năm 2002-2003 (tham khảo ở địa chỉ http://www.sims.berkeley.-edu/research/projects/how-much-info-2003), dữ liệu toàn cầu tăng 5 Exabytes.

Vấn đề đặt ra là làm sao chúng ta có thể rút trích tri thức quan trọng từ các kho dữ liệu khổng lồ. Các tri thức phục vụ cho các tổ chức, cơ quan, công ty bao gồm việc phát hiện quan trọng trong khoa học, các dự báo chính xác về thời tiết và các thảm họa tự nhiên, những tri thức cho phép ta xác định được nguyên nhân và phương pháp điều trị các căn bệnh hiểm nghèo,v.v. Sự ra đời của công nghệ khám phá tri thức và khai mỏ dữ liệu [Fayyad et al., 1996] trong những năm gần đây nhằm đáp ứng nhu cầu cần thiết của các tổ chức, cơ quan, công ty về phát hiện tri thức từ các kho dữ liệu khổng lồ.

[1] Bộ môn Khoa Học Máy Tính, Khoa CNTT&TT, Đại học Cần Thơ, Email: dtnghi@cit.ctu.edu.vn
[2] Bộ môn Khoa Học Máy Tính, Khoa CNTT&TT, Đại học Cần Thơ, Email: pnkhang@cit.ctu.edu.vn

Các ứng dụng thành công của công nghệ khai mỏ dữ liệu có thể tìm thấy trong rất nhiều lãnh vực như: tiếp thị, ngân hàng, bảo hiểm, y tế, sinh học, phát hiện gian lận, tìm kiếm thông tin, lọc thư rác, phân loại văn bản.

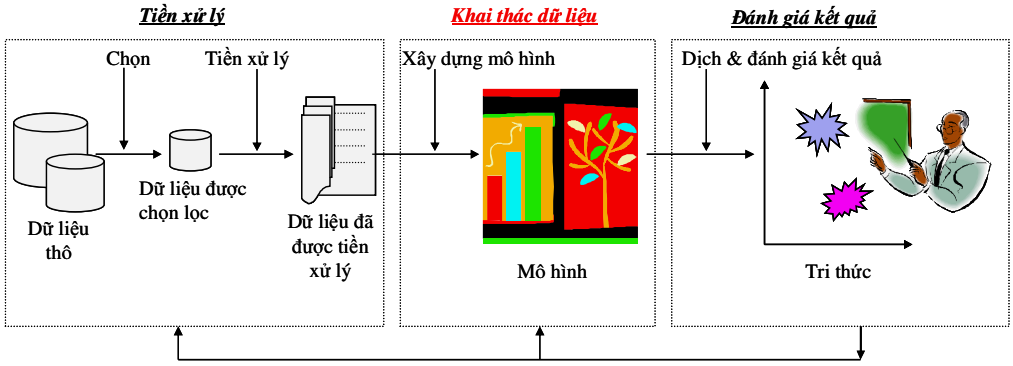
Tạp chí về công nghệ của trường MIT số ra tháng 1-2 năm 2001 cho rằng khai mỏ dữ liệu là một trong 10 công nghệ nổi bật nhất của thế kỷ XXI.

Nắm bắt được xu thế công nghệ khám phá tri thức và khai mỏ dữ liệu, Khoa Công Nghệ Thông Tin Trường Đại Học Cần Thơ đã đầu tư rất nhiều công sức trong nghiên cứu, triển khai ứng dụng, đào tạo nguồn nhân lực trong 10 năm qua. Kết quả đạt được rất đáng khích lệ với 5 tiến sĩ, 10 thạc sĩ, xuất bản hơn 100 công trình được đăng tải trong và ngoài nước chuyên ngành khám phá tri thức và khai mỏ dữ liệu. Bài viết từng bước trình bày quá trình nghiên cứu của Khoa trong chuyên ngành trong 10 năm qua và hướng phát triển trong tương lai.

Trong phần 2, chúng tôi sẽ trình bày tóm tắt công nghệ khám phá tri thức và khai mỏ dữ liệu. Phần 3 điểm qua các vấn đề nan giải trong khai mỏ dữ liệu. Các thành tựu nghiên cứu của Khoa được trình bày trong phần 4 trước khi kết luận và hướng phát triển trong phần 5.

2. KHÁM PHÁ TRI THỨC VÀ KHAI MỎ DỮ LIỆU

Theo [Fayyad et al., 1996], công nghệ khám phá tri thức từ dữ liệu được định nghĩa là sự trích xuất từ dữ liệu những thông tin hữu ích nhưng tiềm ẩn và chưa được biết đến. Khai mỏ dữ liệu là một bước quan trọng trong quá trình khám phá tri thức từ dữ liệu. Khai mỏ dữ liệu thực hiện việc khảo sát, phân tích tỉ mỉ một lượng lớn dữ liệu nhằm phát hiện ra các mẫu hoặc các luật có ý nghĩa.



Hình 1: Quá trình khám phá tri thức

Quá trình khám phá tri thức như mô tả trong hình 1 là một quá trình lặp phức tạp, sử dụng nhiều kỹ thuật như cơ sở dữ liệu, máy học, phương pháp thống kê trong phân tích dữ liệu, hiển thị dữ liệu, trí tuệ nhân tạo, nhằm tìm ra những tri thức từ kho dữ liệu lớn.

Hình 2: Các lãnh vực liên quan đến khám phá tri thức và khai mỏ dữ liệu

Quá trình khám phá tri thức bao gồm 3 bước chính: tiền xử lý, khai mỏ dữ liệu và đánh giá kết quả. Từ mục tiêu đề ra của ứng dụng, ở bước tiền xử lý chúng ta cần thực hiện:

- Tập hợp dữ liệu từ nguồn dữ liệu khác nhau,
- Chọn dữ liệu cần thiết cho mục tiêu đề ra, mẫu tin, trường dữ liệu,
- Biểu diễn dữ liệu, chuyển đổi kiểu sao cho phù hợp với giải thuật khai mỏ dữ liệu mà bước tiếp theo sử dụng,
- Làm sạch dữ liệu, khắc phục đối với trường dữ liệu rỗng, dư thừa, hoặc dữ liệu không hợp lệ, có thể tinh giảm dữ liệu hơn.

Sau khi đã tiền xử lý dữ liệu xong, đến bước khai mỏ dữ liệu tiến hành xây dựng các mô hình với sự hỗ trợ của:

- Máy học,
- Trí tuệ nhân tạo,
- Phân tích dữ liệu nhiều chiều bằng phương pháp thống kê,
- Hoặc bằng phương pháp trực quan hiển thị dữ liệu.

Các giải thuật khai mỏ dữ liệu được sử dụng nhiều trong cộng đồng khám phá tri thức bao gồm: luật kết hợp, *k* láng giềng, phân lớp Bayes thơ ngây, học bằng cây quyết định, Bagging, Boosting, máy học véctơ hỗ trợ, rừng ngẫu nhiên, gom cụm *k*-Means [Wu & Kumar, 2009].

Bước khai mỏ dữ liệu được xem là trung tâm của quá trình khám phá tri thức. Công việc rất phức tạp, lặp đi lặp lại các công việc như: xây dựng mô hình, tạo tri thức về dữ liệu, kiểm định lại mô hình, nếu chưa đạt thì phải xây dựng mô hình khác.

Khai mỏ dữ liệu tập trung giải quyết các vấn đề cơ bản như phân lớp (classification, supervised classification), hồi quy (regression), gom nhóm (clustering, unsupervised classification) và luật kết hợp (association rules).

3. TOP 10 VẤN ĐỀ NAN GIẢI CỦA KHAI MỎ DỮ LIỆU

Trong thực tế, công nghệ phám phá tri thức và khai mỏ dữ liệu đã được ứng dụng với nhiều thành công trong hầu hết các lãnh vực như: ngân hàng, bảo hiểm, marketing, quan hệ khách hàng, dự báo bệnh, bào chế thuốc, sinh-tin học, .v.v. Tuy nhiên, vẫn còn nhiều vấn đề phức tạp đặt ra cho cộng đồng khám phá tri thức và khai mỏ dữ liệu cần giải quyết. Cụ thể là trong nghiên cứu mới đây về khám phá tri thức và khai mỏ dữ liệu.

[Yang & Wu, 2006] đã phối hợp cùng với các chuyên gia đầu ngành về khai mỏ dữ liệu như P. Domingos, C. Elkan, J. Gehrke, J. Han, D. Herkerman, D. Keim, J. Liu, D. Madigan, G. Piatetsky-Shapiro, V. Raghavan, A. Tuzhilin và B. Wah đã chỉ ra top 10 vấn đề lớn cần giải quyết trong khai mỏ dữ liệu như sau.

1. Phát triển lý thuyết hợp nhất cho khai mỏ dữ liệu: do các giải thuật hiện nay được phát triển theo đặc thù của ứng dụng, cụ thể như giải thuật cho phân lớp, giải thuật cho bài toán hồi quy, luật kết hợp, .v.v. Cần phải có lý thuyết hợp nhất để làm việc hiệu quả hơn.
2. Xử lý dữ liệu lớn, nhiều chiều, tốc độ nhanh cho dòng dữ liệu (data stream): đa số các giải thuật không có khả năng hoặc không đáp ứng được yêu cầu ràng buộc về thời gian, chất lượng mô hình khi xử lý dữ liệu lớn hàng TB hay số chiều có thể lên đến hàng tỉ.
3. Khai mỏ dữ liệu tuần tự (sequence data), dữ liệu thời gian (time series data): dữ liệu tuần tự và dữ liệu thời gian thường bị nhiễu, mặc dù có những công cụ cho phép xử lý nhiễu, nhưng không thật sự hiệu quả. Câu hỏi đặt ra là làm thế nào có thể xử lý hiệu quả các vấn đề phân lớp, dự báo, gom cụm trên dữ liệu này.
4. Khai mỏ dữ liệu phức tạp (complex data): dữ liệu trong thực tế thường phức tạp và phi cấu trúc, chẳng hạn như dữ liệu văn bản (text), hình ảnh, âm thanh, vidéo, do đó khai mỏ dữ liệu phức tạp vẫn là chủ đề khó.
5. Khai mỏ dữ liệu trong mạng (network): khám phá mối quan hệ trong mạng xã hội cũng rất cần các kỹ thuật khai mỏ dữ liệu, chẳng hạn như khám phá quy luật của nhóm người dùng trong cộng đồng thường liên lạc với nhau trên diễn đàn, hay đánh giá độ tin cậy của các bài trên mạng wikipedia. Ngoài ra, khám phá hay phát hiện tấn công mạng từ xa như từ chối dịch vụ cũng là vấn đề phức tạp cần đến kỹ thuật khai mỏ dữ liệu.
6. Khai mỏ dữ liệu phân tán và đa tác tử (multi-agent): trong các ứng dụng, dữ liệu có thể phân tán trên nhiều nguồn khác nhau, làm thế nào để xử lý dữ liệu phân tán. Cộng đồng cũng muốn kết hợp kỹ thuật khai mỏ dữ liệu với lý thuyết trò chơi cho phép mở rộng ứng dụng.
7. Khai mỏ dữ liệu cho vấn đề của công nghệ sinh học và môi trường: có rất nhiều vấn đề của công nghệ sinh học và môi trường cần đến các công cụ khám phá tri thức và khai mỏ dữ liệu, chẳng hạn như dự đoán cấu trúc phân tử, quản lý nguồn nước, đất, .v.v.
8. Vấn đề liên quan đến quá trình khai mỏ dữ liệu: phát triển môi trường khám phá tri thức một cách trực quan, tích hợp các công cụ cho cả tiền xử lý, khai mỏ và đánh giá kết quả, giúp người sử dụng tránh được các lỗi khi khai mỏ dữ liệu.
9. Đảm bảo an ninh, tính riêng tư và tích hợp dữ liệu: giải thuật khai mỏ cần đảm bảo tính riêng tư và an toàn của dữ liệu riêng của người dùng, chẳng hạn như khai mỏ dữ liệu thư điện tử hay đầu tư chứng khoán. Trong khi tích hợp tri thức vào hệ thống, người ta cần phát triển những độ đo để đánh giá việc tích hợp tri thức của bộ sưu tập dữ liệu (collection of data) và mẫu (pattern).
10. Xử lý dữ liệu không cân bằng (unbalanced), có giá khác nhau (cost-sensitive) và dữ liệu biến động (non-static): trong thực tế, dữ liệu thay đổi theo thời gian do đó mô hình khai mỏ dữ liệu cần theo sát dữ liệu. Dữ liệu có giá khác nhau, chẳng hạn như nhận dạng bệnh ung thư hay không phải ung thư thì giá phải trả cho sai lầm của việc để lọt bệnh ung thư (bệnh bị dự đoán là không bệnh) sẽ cao hơn nhận

dạng lầm là bệnh (không bệnh bị dự đoán là bệnh). Ngoài ra, dữ liệu phân lớp trong thực tế rất lệch nhau (1 lớp ta quan tâm rất bé so với lớp còn lại), giải thuật thường thiết kế để cho độ chính xác toàn cục cao nhưng dễ bị mất dự đoán lớp nhỏ.

4. THÀNH QUẢ CỦA 10 NĂM NGHIÊN CỨU KHAI MỎ DỮ LIỆU

Nguồn nhân lực của chuyên ngành khai mỏ dữ liệu của Khoa Công Nghệ Thông Tin và Truyền Thông, Trường Đại học Cần Thơ, được đào tạo từ các trường đại học, các viện nghiên cứu nổi tiếng ở châu âu như Đại học Bách khoa Toulouse, Nantes, Đại học Rennes, Đại học Paris-11, Đại học Lyon, Trường Viễn thông Bretagne, Viện Nghiên cứu Tin học và Tự động INRIA, Cộng hòa Pháp, Đại học Leuven, Vương quốc Bỉ, Đại học Québec Montréal, Canada, Đại học Hildesheim, Cộng hòa Liên bang Đức, Đại học Wellington, Tân Tây Lan. Nhóm nghiên cứu năng động, sáng tạo, đã tạo ra nền tảng nghiên cứu vững chắc, với hàng trăm công trình nghiên cứu khoa học được đăng tải, tham gia tổ chức các sự kiện, ban chương trình hội thảo trong nước và quốc tế.

Nhóm nghiên cứu do TS. Đỗ Thanh Nghị[1], TS. Phạm Nguyên Khang[2], TS. Nguyễn Văn Hòa[3], Nguyễn Thanh Bình[4], tập trung vào đề xuất:

Lớp giải thuật máy học véc-tơ hỗ trợ cho phân lớp dữ liệu có dung lượng khổng lồ. Các giải thuật mới cải tiến thời gian thực thi, yêu cầu về bộ nhớ, dựa trên chiến lược học tăng trưởng, song song, phân tán, tận dụng khả năng xử lý của đồng xử lý toán học của card đồ họa và điện toán đám mây.

Lớp giải thuật rừng ngẫu nhiên của cây quyết định xiên phân (sử dụng đa thuộc tính khi phân hoạch) giúp phân lớp hiệu quả các tập dữ liệu có số chiều lớn, nhiễu như dữ liệu gien, dữ liệu ảnh.

Lớp giải thuật học cây quyết định cho phân lớp dữ liệu không cân bằng (lớp quan tâm có số mẫu tin quá nhỏ chiếm ít hơn 10% của lớp còn lại. Các giải pháp tập trung vào phương pháp lấy mẫu, thay thế chiến lược phân hoạch sử dụng các hàm entropy bất đối xứng, Kolmogorov-Smirnov, chiến lược tập hộ mô hình, và luật gán nhãn.

Khai thác dữ liệu phức tạp kiểu Interval, Taxonomy, dữ liệu phi cấu trúc như văn bản, hình ảnh bằng cách xây dựng hàm nhân phi tuyến của máy học véc-tơ hỗ trợ, phương pháp hiển thị dữ liệu nhiều chiều, phép giảm chiều như phân tích thành phần chính, phân tích biệt lập tuyến tính, phân tích tương ứng, phân tích ngữ nghĩa tiềm ẩn.

Hiển thị và khai thác tri thức từ mạng xã hội từ phép giảm chiều, mô hình phân cấp, phương pháp hiển thị đồ thị, tương tác, trực quan.

[1] http://www.cit.ctu.edu.vn/~dtnghi
[2] http://www.cit.ctu.edu.vn/~pnkhang
[3] http://staff.agu.edu.vn/nvhoa
[4] http://www.cit.ctu.edu.vn/~ntbinh

Phương pháp hiển thị dữ liệu hỗ trợ cho quá trình khám phá tri thức và khai mỏ dữ liệu, từ tiền xử lý, xây dựng mô hình phân lớp, diễn dịch kết quả thu được từ máy học véc-tơ hỗ trợ, cây quyết định, môi trường lập trình trực quan.

Triển khai ứng dụng như nhận dạng dấu vân tay, lọc ảnh, lọc thư rác, phân tích dữ liệu văn bản, tìm kiếm ảnh, học xếp hạng trong tìm kiếm chuyên gia. Các chương trình, thư viện ứng dụng được tìm thấy tại địa chỉ[1].

Nhóm của TS. Huỳnh Xuân Hiệp[2], Lâm Chí Nguyện, tập trung vào khai thác các độ đo hấp dẫn nhằm đánh giá chất lượng của tri thức (luật kết hợp). Nghiên cứu đóng một vai trò rất quan trọng trong giai đoạn hậu xử lý các luật kết hợp của tiến trình khai phá tri thức từ dữ liệu. Ngoài ra, TS. Huỳnh Xuân Hiệp còn nghiên cứu mạng Bayes cho dự báo, đánh giá chất lượng mẫu tuần tự.

Nhóm của TS. Trần Cao Đệ[8] chủ yếu tập trung vào xử lý dữ liệu văn bản, phân lớp văn bản, phát hiện sao chép và nhận dạng chữ viết tay. Các kỹ thuật bao gồm trích chọn đặc trưng, giảm chiều với chỉ mục ngữ nghĩa tiềm ẩn, xây dựng mô hình máy học véc-tơ hỗ trợ, cây quyết định cho phân lớp.

TS. Trương Quốc Định[8] tập trung xây dựng mô hình tìm kiếm thông tin dựa trên lý thuyết đồ thị. Nhóm cũng nghiên cứu tóm tắt văn bản, phát hiện sao chép, nhận dạng chữ viết tay.

TS. Lê Thanh Vân[3] nghiên cứu và đề xuất giải thuật mới cho gom cụm dữ liệu dựa trên pretopology. Giải thuật đã được triển khai ứng dụng cho cơ sở y tế.

Phạm Thế Phi[4] đề xuất giải pháp khai thác dữ liệu đa phương tiện. Các đề xuất bao gồm học tổng quát và biệt lập, mô hình gom cụm, mô hình xác suất, canh lề đa ngữ.

Lê Văn Lâm[5] nghiên cứu về nhận dạng tấn công mạng từ tiếp cận trích chọn đặc trưng và máy học.

Nguyễn Thái Nghe[6] đề xuất cải tiến mạng Bayes, tiêu chí đánh giá hiệu quả của mô hình phân lớp dữ liệu không cân bằng. Ngoài ra, Nguyễn Thái Nghe cũng ứng dụng khai mỏ dữ liệu trong E-learning.

Đội ngũ nghiên cứu khám phá tri thức và khai mỏ dữ liệu của Khoa Công Nghệ Thông Tin và Truyền Thông, Đại Học Cần Thơ đã năng động, sáng tạo, đăng tải được

[1] http://www.cit.ctu.edu.vn/~dtnghi/v4miner
[2] http://www.cit.ctu.edu.vn
[3] TS. Lê Thanh Vân đã chuyển về ĐHBK TP.HCM từ tháng 10/2011
[4] http://people.cs.kuleuven.be/~phithe.pham
[5] http://ecs.victoria.ac.nz/Main/GradVanLamLe
[6] http://www.ismll.uni-hildesheim.de/personen/nguyen_en.html

hàng trăm công trình nghiên cứu khoa học chủ yếu là bài báo và tạp chí quốc tế[1], sách khai mỏ dữ liệu [Đỗ Thanh Nghị, 2011], thành lập phòng nghiên cứu xử lý dữ liệu thông minh[2] tập trung nghiên cứu, đào tạo về khai mỏ dữ liệu, dự báo và mô phỏng. Tham gia vào tổ chức các hội thảo, workshop, ban chương trình của hội thảo như: AVEC2008-2009, DMIN2008-2010, ACIIDS2010-2012, QDC2008-2011, AKDM2010, AusDM2004, MCO2008-2010, RNTI2005-2011, I3, Journal of EA2009, Pattern Recognition Elsevier 2008, v.v. Chúng ta cũng thường xuyên tổ chức hơn 20 chuyên đề, tạo cơ hội để gặp gỡ, trao đổi, chia sẻ kinh nghiệm trong nghiên cứu và triển khai ứng dụng. Nhóm cũng đã đào tạo được đội ngũ sinh viên, giảng viên kế thừa nghiên cứu và phục vụ cho đào tạo. Nhóm cũng phát triển được thư viện, chương trình phục vụ cho nghiên cứu và triển khai ứng dụng.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Qua trình bày thành tựu đạt được của nhóm nghiên cứu tại Khoa, chúng ta có thể thấy được nghiên cứu tập trung vào 5 trong top 10 vấn đề khó của khai mỏ dữ liệu:

Kết quả nghiên cứu hiện nay cũng là thành tựu cá nhân, riêng lẻ. Chưa có sự kết hợp của một nhóm lớn. Trong tương lai, nên tập hợp thành nhóm nghiên cứu lớn, chia sẻ, hợp tác trong các chủ đề lớn. Các nghiên cứu của nhóm vẫn cần tập trung vào vấn đề khai thác tri thức từ các mạng xã hội, chẳng hạn như facebook. Tìm kiếm thông tin vẫn là lãnh vực cần nhiều nghiên cứu của máy học xếp hạng. Hầu hết, nhóm nghiên cứu của chúng ta vẫn chỉ khai thác dữ liệu kiểu số, văn bản, hình ảnh, nhưng chưa có nghiên cứu nào liên quan đến âm thanh, chuỗi thời gian, vidéo. Kết nối với sinh-tin học để thực hiện nghiên cứu liên ngành.

Tuy giải thuật khai mỏ dữ liệu hiện có rất phong phú, đa dạng, nhưng không có phương pháp nào là tốt trong tất cả các trường hợp. Điều này được biết như *«no free lunch theorem»,* có nghĩa là một giải thuật khai mỏ dữ liệu được biết là tốt cho một trường hợp cụ thể này thì trong trường hợp khác lại xử lý kém hiệu quả so với phương pháp khác. Hầu hết các giải thuật khai mỏ dữ liệu hiện nay được phát triển rất đặc thù cho trường hợp cụ thể nào đó (ad-hoc). Vấn đề của khai mỏ dữ liệu vẫn còn rất được quan tâm để cải thiện quá trình khám phá tri thức từ dữ liệu.

Triển khai ứng dụng vẫn là vấn đề cần quan tâm. Trước tiên vẫn là nguồn dữ liệu chúng ta chưa có nhiều, gây khó khăn cho quá trình triển khai ứng dụng. Các phương pháp được phát triển hiện nay cần hiệu chỉnh để đáp ứng được nhu cầu thực tế như: nhận dạng vân tay, đọc bản số xe, lọc ảnh, lọc thư rác, lọc web rác, phân lớp văn bản, tìm kiếm chuyên gia, nhận dạng tấn công mạng, kiểm thử phần mềm, dự báo dịch bệnh cây trồng, dự báo kinh tế, v.v. Các thư viện chương trình vẫn được phát triển nhỏ lẻ theo dạng "prototype", rất khó tái sử dụng.

[1] http://www.cit.ctu.edu.vn/index.php?option=com_content&task=view&id=596&Itemid=372&lang=vi
[2] http://www.cit.ctu.edu.vn/~t2i

Tài liệu tham khảo

Đỗ Thanh Nghị. *Khai mỏ dữ liệu, minh họa bằng ngôn ngữ R*. Nhà xuất bản Đại học Cần Thơ, 2011.

U. Fayyad, G. Piatetsky-Shapiro and P. Smyth. From Data Mining to Knowledge Discovery in Databases. in *AI Magazine*, vol. 17, n° 3, 1996, p. 37-54.

X. Wu and V. Kumar. *Top 10 Algorithms in Data Mining*. Chapman & Hall/CRC, 2009.

Q. Yang and X. Wu. 10 Challenging Problems in Data Mining Research. *Journal of Information Technology & Decision Making 5(4)*:597-604, 2006.

CHỈ MỤC NGỮ NGHĨA TIỀM ẨN VÀ ỨNG DỤNG

Trần Cao Đệ
Khoa Công nghệ Thông tin & Truyền thông
Đại học Cần Thơ

Tóm tắt

Kỹ thuật Chỉ mục ngữ nghĩa tiềm ẩn (Latent Semantic Indexing-LSI) dựa trên cơ sở toán học là phân tích giá trị đơn và thường được dùng trong xử lí văn bản để tìm mối liên hệ ngữ nghĩa giữa các thuật ngữ và các khái niệm trong một tập hợp các văn bản phi cấu trúc. Nhiều nghiên cứu chỉ ra rằng các từ được dùng trong cùng một ngữ cảnh thường có ngữ nghĩa tương tự nhau. LSI có khả năng trích xuất khái niệm bằng cách thiết lập mối quan hệ giữa các thuật ngữ đồng xuất hiện trong cùng ngữ cảnh. Kỹ thuật này thường được dùng trong khai khoáng dữ liệu văn bản, phân loại văn bản. Nó có ưu điểm so với một số kỹ thuật khác được dùng trong xử lí ngữ nghĩa văn bản nhờ vào việc phân tích tương quan, đồng xuất hiện của các thuật ngữ hơn là dựa vào nghĩa của thuật ngữ. Các nghiên cứu khác nhau đã chỉ ra rằng LSI vượt qua được các khó khăn chính trong xử lí ngôn ngữ đó là từ đồng nghĩa (synonymy) và đa nghĩa (polysemy).

Bài viết này giới thiệu kỹ thuật LSI và ứng dụng của kỹ thuật này trong phân loại văn bản từ đó đề xuất ý tưởng áp dụng kỹ thuật này trong nhận dạng thông qua việc áp dụng vào nhận dạng chữ viết tay.

Từ khóa: ***Chỉ mục ngữ nghĩa tiềm ẩn, Nhận dạng chữ viết tay, Phân loại văn bản, Phân tích giá trị đơn, Xử lí văn bản.***

LATENT SEMANTIC INDEXING AND ITS APPLICATIONS

Abstract

Latent semantic indexing (LSI), based on mathematic basic known as singular value decomposition (SVD), has been used in text processing to detect (latent) semantic links between terms in unstructured documents. Different research has illustrated that words that are used in the same context may have similar semantic meaning. LSI can extract concepts by establish the relationship between terms that co-appeared in a context. It has been used in text mining and text classification. It has been known better than a few techniques thank for analyzing of the correlation, the co-occurrence between terms rather than using the semantic of term. Many research show that LSI overcomes two principle difficulties in language processing which are synonymy and polysemy.

This paper introduces LSI and its application in text classification from which we propose the idea to apply it into recognition problem in general and in handwriting character as particular.

Keywords: ***Latent semantic indexing, handwriting recognition, text classification, singular decopostion value, text processing.***

1. GIỚI THIỆU

Chỉ mục ngữ nghĩa tiềm ẩn (LSI) xuất hiện từ năm 1990 bởi Deerwester và các đồng sự [1]. Nó đầu tiên được nghĩ ra để lập chỉ mục văn bản phục vụ cho tìm kiếm và thu hồi thông tin văn bản. Có nhiều đồn đoán rằng nó được dùng trong cỗ máy tìm kiếm của Google. LSI về bản chất là dựa trên cơ sở toán học là phân tích giá trị đơn (Singular Value Decomposition). LSI có khả năng trích xuất khái niệm bằng cách thiết lập mối quan hệ giữa các thuật ngữ đồng xuất hiện trong cùng ngữ cảnh. Ví dụ "Paris Hilton" được kết hợp với khái niệm là "người mẫu", "diễn viên", "người thừa kế tập đoàn Hilton" hơn là kết hợp với khái niệm "thủ đô nước Pháp", "Khách sạn Hilton" hay "Tháp Eiffel". "Eiffel Paris" thì kết hợp với "Tháp Eiffel" hơn là với "Eiffel Software" hay "Paris Hilton".

Kỹ thuật này thường được dùng trong khai khoáng dữ liệu văn bản, phân loại văn bản [1]. Nó có ưu điểm so với một số kỹ thuật khác được dùng trong xử lí ngữ nghĩa văn bản nhờ vào việc phân tích tương quan, đồng xuất hiện trong một ngữ cảnh của các thuật ngữ hơn là dựa vào nghĩa của thuật ngữ. Các nghiên cứu khác nhau đã chỉ ra rằng LSI vượt qua được các khó khăn chính trong xử lí ngôn ngữ đó là từ đồng nghĩa (synonymy) và đa nghĩa (polysemy) [2].

Ngày nay, LSI được ứng dụng trong nhiều bài toán khác nhau, chẳng hạn:

- Khám phá tri thức [3]
- Tự động phân loại văn bản [4]
- Tóm tắt văn bản [5]
- Gán bài viết và người phản biện [6]
- Tự động gán từ khóa diễn giải ảnh [7]
- Phát hiện sao chép mã nguồn phần mềm [8]
- Lọc thư rác [9]
- Chấm bài luận tự động [10]

Trong các nghiên cứu của chúng tôi, LSI được áp dụng vào một số bài toán như: tóm tắt văn bản [11], phát hiện sao chép văn bản [12], tìm kiếm đa ngữ [13], phân loại văn bản [14]. Từ nghiên cứu phân loại văn bản với máy học SVM kết hợp với việc tách giá trị đơn SVD trong [14], chúng tôi nhận thấy rằng có thể áp dụng cùng ý tưởng vào một bài toán phân loại bất kỳ chẳng hạn như nhận dạng mặt người, nhận dạng chữ viết,…

Phần tiếp theo chúng tôi giới thiệu kỹ thuật LSI và áp dụng kỹ thuật này vào phân loại văn bản cùng với máy học vector hỗ trợ. Kế đến, áp dụng ý tưởng giải quyết bài toán đó vào bài toán nhận dạng chữ viết tay.

2. PHÂN LOẠI VĂN BẢN VỚI MÁY HỌC VECTOR HỖ TRỢ VÀ KẾT HỢP VỚI PHÂN TÍCH GIÁ TRỊ ĐƠN

2.1 Bài toán phân loại văn bản

Việc phân loại văn bản tự động có thể xem là một bài toán phân lớp dữ liệu, đó là việc gán các nhãn phân loại lên một văn bản mới dựa trên mức độ tương tự của văn bản đó so với các văn bản đã được gán nhãn trong tập huấn luyện. Một cách hình thức, cho một tập $\{S_i\}$ hữu hạn các chủ đề mỗi chủ đề gồm một số văn bản $\{D_{ij}\}$ đã được xác định trước là thuộc về chủ đề S_i (bằng chuyên gia con người chẳng hạn). Khi cần xem xét một văn bản T để phân loại nó thuộc về chủ đề nào, người ta có thể dùng kỹ thuật phân lớp dữ liệu để gán nhãn S_k nào đó cho T.

Để thực hiện phân loại văn bản với máy học vector hỗ trợ, trước hết văn bản cần được mô hình hóa theo mô hình không gian vector. Ở đây chúng tôi xem xét các văn bản ngắn như một bài viết trên báo điện tử và mô hình hóa chúng thành một vector chứa trọng số các từ. Một cách hình thức, một văn bản D có thể xem như là một vector $D(x_1,\ldots,x_m)$

trong đó x_i là trọng số của từ thứ i trong bộ từ vựng đang xét. Ví dụ xét các văn bản tiếng Việt, thì bộ từ vựng bao gồm "tất cả" các từ có thể có của tiếng việt. Tuy nhiên, việc sử dụng bộ từ vựng như vậy là quá lớn, nhiều dư thừa. Cách tiếp cận thực tế là xét một tập hợp đủ lớn các văn bản, gọi là tập ngữ liệu (corpus), tập từ vựng là tập tất cả các từ có mặt trong tập ngữ liệu đó. Chẳng hạn chúng tôi đã thu thập được 2000 văn bản thuộc 10 chủ đề để làm thực nghiệm trong bài viết này. Tập ngữ liệu này chứa 19749 từ tiếng việt được tách theo giải thuật MMSEG [1.1.1.15].

Có nhiều các định trọng số x_i cho mỗi từ trong văn bản $D(x_1,\ldots,x_m)$. Cách đơn giản nhất là gán $x_i = 0$ nếu từ không có trong D và $x_i=1$ nếu từ i có trong D. Trong nghiên cứu của chúng tôi, trọng số của từ được gán bằng chỉ số TF*IDF (TF- tần suất xuất hiện của từ trong văn bản; IDF: tần suất xuất hiện ngược trong văn bản):

$$x_i = TF_i * \log\left(\frac{N}{DF_i}\right) \quad (1)$$

- TF_i là số lần xuất hiện của từ thứ i trong văn bản D.
- DF_i là tổng số văn bản có chứa từ thứ i trong tập ngữ liệu.
- N là tổng số văn bản trong tập ngữ liệu.

Như vậy tập ngữ liệu sẽ có thể được mô hình hóa như là một ma trận $A= (a_{ij})^T$ trong đó mỗi cột của ma trận là một vector biểu diễn cho một văn bản, cột i biểu diễn cho văn bản thứ i. Nói chung, A sẽ là một ma trận thưa và có số chiều lớn, trong trường hợp tập ngữ như trên thì A có kích thước 19749 x 2000.

2.2 Phân tích giá trị đơn (Single Value Decomposition-SVD)

Phân tích giá trị đơn là phân tích toán học nền tảng trong kỹ thuật chỉ mục ngữ nghĩa tiềm ẩn (LSI) đã được dùng rộng rãi trong tìm kiếm và thu hồi thông tin dạng văn bản. Ý tưởng chính của giải thuật [1.1.1.1,1.1.1.2] như sau:

Cho ma trận $\mathbf{A}_{mxn}$, ma trận **A** có thể được phân tích thành tích của ba ma trận theo dạng:

$\mathbf{A = USV^T}$, trong đó:

- U là ma trận trực giao mxm có các cột là các vector đơn bên trái của A.
- **S** là ma trận mxn có đường chéo chứa các giá trị đơn, không âm có thứ tự giảm dần: $\boldsymbol{\delta_1 \geq \delta_2 \geq \ldots \geq \delta_{min(m,n)} \geq 0}$.
- **V** là ma trận trực giao nxn có các cột là các vector đơn bên phải của **A**.

Hạng của ma trận **A** là số các số dương trên đường chéo chính của ma trận S**.** Giả sử hạng của **A** là r, tức là rank(**A**)=r, thì số Frobenius (Frobenius norm) của **A** là :

$$\|A\|_F = \sqrt{\sum_{i=1}^{r} \delta_i^2} \quad (2)$$

Trong phân tích này, ma trận **A** thường là một ma trận thưa có kích thước lớn. Để giảm số chiều của ma trận người ta thường tìm cách xấp xỉ ma trận **A** (có hạng r) bằng một ma trận $\mathbf{A_k}$ có hạng là k nhỏ hơn r rất nhiều. Ma trận xấp xỉ của **A** theo kỹ thuật LSI chính là: $\mathbf{A_k = U_k S_k V_k^T}$**,** trong đó

- $\mathbf{U_k}$ là ma trận trực giao mxk có các cột là k cột đầu của ma trận U.
- $\mathbf{S_k}$ là ma trận đường chéo kxk chứa các phần tử đầu tiên $\boldsymbol{\delta_1, \delta_2, \ldots, \delta_k}$ trên đường chéo chính.
- $\mathbf{V_k}$ là ma trận trực giao nxk có các cột là k cột đầu của ma trận **V**.

Người ta đã chứng minh rằng $\mathbf{A_k}$ là một xấp xỉ của **A** với sự thay đổi Frobenius nhỏ nhất. Việc xấp xỉ này có thể xem như chuyển không gian đang xét (r chiều) về không gian k chiều, với k nhỏ hơn rất nhiều so với r. Về mặt thực hành việc cắt ma trận A về số

chiều k còn loại bỏ nhiễu và tăng cường các mối liên kết ngữ nghĩa tiềm ẩn giữa các từ trong tập văn bản. Chúng tôi sẽ áp dụng kỹ thuật xấp xỉ này để rút ngắn số chiều của không gian đặc trưng. Khởi đầu, mỗi văn bản được mô hình hóa thành một vector cột trong không gian xác định bởi $\mathbf{A}_{mxn}$. Sau khi cắt $\mathbf{A}_{mxn}$ về $\mathbf{A}_k$, các tất cả các vector đang xét đều được chiếu lên không gian $\mathbf{A}_k$ để có số chiều k theo công thức:

$$\text{Proj}(x) = x^T \mathbf{U}_k \sum_k^{-1} \quad (3)$$

2.3 Thực nghiệm phân loại văn bản với máy học vector hỗ trợ (SVM)

Trong thực nghiệm chúng tôi dùng tập ngữ liệu gồm 2000 tài liệu thuộc 10 chủ đề: Công nghệ thông tin (CNTT), Thế giới, Thể thao, Kinh doanh, Giáo dục, Tình yêu, Giải trí, Sức khỏe, Nấu ăn, Pháp luật. Các tài liệu này được sưu tập từ nhiều báo điện tử như dantri.com.vn, vnexpress.net, 123suckhoe.com, thuongthucgiadinh.com, 24h.com.vn. Tập ngữ liệu chứa 19749 từ tiếng việt được tách theo giải thuật MMSEG [1.1.1.15]. Đây là số lượng từ còn lại sau khi loại các từ không quan trọng (stopwords). Tập ngữ liệu được mô hình hóa như là một ma trận trọng số tính bằng chỉ số TF*IDF của các từ. Nó có kích thước 19749 x 2000 phần tử.

Bảng 1: Ma trận kiểm chứng (confusion matrix) trên tập dữ liệu huấn luyện

Tên lớp	Mã lớp	1	2	3	4	5	6	7	8	9	10	Tổng số mẫu	Độ chính xác
CNTT	1	**141**	1	2	2	1	0	0	0	3	0	150	94%
Thế giới	2	0	**133**	1	5	1	3	0	1	4	2	150	88.67%
Thể thao	3	0	3	**144**	0	0	0	0	0	3	0	150	96%
Kinh doanh	4	2	9	1	**128**	0	1	2	0	2	5	150	85.33%
Giáo dục	5	4	0	2	1	**134**	3	0	1	2	3	150	89.33%
Tình yêu	6	0	0	0	0	2	**140**	1	2	4	1	150	93.33%
Giải trí	7	0	0	1	0	0	4	**144**	0	0	1	150	96%
Sức khỏe	8	0	1	0	0	0	0	0	**147**	2	0	150	98%
Nấu ăn	9	0	0	0	0	0	2	0	1	**147**	0	150	98%
Pháp luật	10	0	0	0	4	0	2	0	0	0	**144**	150	96%
												1500	**93.47%**

Nói cách khác mỗi văn bản được mô hình hóa như một vector 19749 đặc trưng. Số đặc trưng này là quá lớn để SVM có thể đạt hiệu quả cao. Do đó bước tiếp theo là dùng SVD phân tích giá trị đơn và xấp xỉ ma trận trên về kích thước nhỏ hơn nhằm rút gọn số chiều không gian đang xét.

Để rút gọn số chiều không gian đặc trưng, chúng tôi đã áp dụng phương pháp phân tích giá trị đơn và rút gọn số chiều về k=50 (đã thử nghiệm một số giá trị khác nhau của k). Sau đó chúng tôi thực hiện tính toán để đưa tất cả vector đặc trưng về các vector đặc trưng mới bằng phép chiếu (2) với kích thước không gian rút gọn là 50.

Trong giai đoạn phân lớp thì văn bản phân lớp cũng được vector hóa theo TF*IDF các từ và chiếu lên không gian A_{50} này để có kích thước 50 đặc trưng.

Để thực hiện huấn luyện máy học, 150 văn bản trên mỗi chủ đề được lấy ra ngẫu nhiên từ tập ngữ liệu làm tập huấn luyện. 50 văn bản cho mỗi chủ đề còn lại trong tập ngữ liệu sẽ làm tập kiểm chứng độc lập. Bảng 1 trình bày kết qủa thực hiện kiểm tra 20-folds trên tập huấn luyện có dùng kết hợp SVD. Kết quả này cho thấy việc huấn luyện cho kết quả rất cao trên không gian rút gọn A_{50}.

Ở đây máy học SVM là tuyến tính với C bằng 19000 và epxilon bằng 0.001. Kết quả kiểm chứng trên tập kiểm chứng độc lập được cho trong bảng 2 (với cùng các tham số máy học SVM).

Bảng 2: Ma trận kiểm chứng (confusion matrix) trên tập dữ liệu kiểm chứng độc lập

Tên lớp	Mã lớp	1	2	3	4	5	6	7	8	9	10	Tổng số mẫu	Độ chính xác
CNTT	1	**47**	1	0	2	0	0	0	0	0	0	50	94%
Thế giới	2	0	**39**	0	7	1	1	1	0	1	0	50	78%
Thể thao	3	0	0	**50**	0	0	0	0	0	0	0	50	100%
Kinh doanh	4	0	0	0	**47**	0	0	1	0	2	0	50	94%
Giáo dục	5	1	0	1	0	**45**	1	2	0	0	0	50	90%
Tình yêu	6	0	0	0	0	1	**47**	0	0	2	0	50	94%
Giải trí	7	0	1	1	1	0	2	**45**	0	0	0	50	90%
Sức khỏe	8	0	0	0	0	0	0	0	**49**	1	0	50	98%
Nấu ăn	9	0	0	0	0	0	0	0	0	**50**	0	50	100%
Pháp luật	10	0	0	0	0	0	0	0	0	0	**50**	50	100%
												500	**93.8%**

Kết quả so sánh giữa việc dùng bộ phân lớp SVM trên tập dữ liệu rút gọn số chiều (dùng SVD) và tập dữ liệu thô (chưa dùng SVD) được cho trong bảng 3.

Bảng 3: So sánh hiệu quả giữa phân loại trên tập kiểm chứng độc lập có dùng và không dùng SVD

Tên lớp	Áp dụng SVD	
	Có	Không
CNTT	94%	78%
Thế giới	78%	72%
Thể thao	100%	100%
Kinh doanh	94%	92%
Giáo dục	90%	82%
Tình yêu	94%	84%
Giải trí	90%	82%
Sức khỏe	98%	86%
Nấu ăn	100%	98%
Pháp luật	100%	78%
Hiệu quả Trung bình	**93.8%**	**85.2%**

Thực nghiệm này cho thấy việc dùng SVD trong lựa chọn đặc trưng, rút ngắn số chiều không gian làm tăng độ chính xác của phân lớp từ 85.2% lên 93.8%.

3. ÁP DỤNG SVD VÀO TRÍCH CHỌN ĐẶC TRƯNG NHẬN DẠNG CHỮ VIẾT TAY

Phương pháp dùng phân tích giá trị đơn để rút ngắn số chiều không gian đặc trưng trong bài toán phân loại văn bản đã trình bày có thể tổng quát hóa để dùng trong việc trích chọn đặc trưng cho các bài toán nhận dạng khác nhau. Phần tiếp theo sẽ cụ thể hóa ý tưởng này và kiểm chứng bằng cách nhận dạng chữ số viết tay.

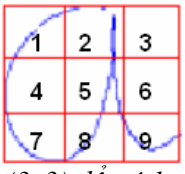
Bài toán nhận dạng chữ viết tay cũng là một bài toán phân lớp, và theo cách tiếp cận đó, nó cũng bao gồm các bước cơ bản giống như bài toán phân loại văn bản:

- Trích xuất đặc trưng
- Lựa chọn đặc trưng (rút gọn số đặc trưng)
- Áp dụng giải thuật máy học để huấn luyện

- Dùng máy học đã xây dựng để nhận dạng (dự đoán lớp).

3.1 Trích xuất đặc trưng

Trong nghiên cứu này, chúng tôi dùng bộ dữ liệu chuẩn về chữ viết tay UNIPEN. Đây là bộ dữ liệu chữ viết tay trực tuyến thu được từ tablet. Vì vậy, chúng tôi mong muốn trích cả các đặc trưng trực tuyến (online) và không trực tuyến (offline). Dữ liệu thu được từ tablet là dữ liệu trực tuyến, đó là tập hợp các nét viết, mỗi nét viết là một dãy có thứ tự theo thời gian các điểm từ lúc đặt viết xuống đến lúc nhấc bút lên. Các đặc trưng liên quan đến thời gian có thể xem là các đặc trưng trực tuyến (ví dụ điểm bắt đầu, điểm kết thúc một nét viết). Dữ liệu trực tuyến thu được từ tablet có thể xem như một ảnh động (ảnh vector), từ đó có thể tạo ra một ảnh tĩnh (ảnh raster) và trích xuất các đặc trưng từ ảnh này như là các đặc trưng không trực tuyến. Để xây dựng các đặc trưng cho máy học và nhận dạng kí tự, 7 nhóm đặc trưng cho dữ liệu không trực tuyến (offline) đã được đề nghị bởi L. Heute [1.1.1.16] được dùng, bao gồm: 7 bất biến Hu (Moment Hu); các bản đồ chiếu ngang, dọc (horizontal and vertical projections); các profiles trên, dưới, trái phải; giao các đường thẳng ngang, dọc; lỗ trống, các cung lòng chão(concave arcs); các điểm cực trái, phải, trên, dưới; các điểm kết thúc và giao điểm. Ảnh của một kí tự thường được chia bởi lưới cơ sở 3x3 để trích các đặc trưng trong một ô lưới. Hình 1 cho ví dụ về một ô lưới 3x3 của kí tự "a". Hình 2a cho ví dụ về bản đồ chiếu và hình 2b cho ví dụ về profiles.



Hình 1: Lưới cơ bản (3x3) để trích các đặc trưng theo ô.

2a. 2b.

Hình 2: Bản đồ chiếu (projections) và profiles

Các bất biến Radon [1.1.1.17] và Zernike 1.1.1.18] cũng được trích trong danh sách các đặc trưng. Bây giờ danh sách các đặc trưng được bổ sung thêm các đặc trưng từ dữ liệu trực tuyến (online) như: điểm bắt đầu, điểm kết thúc, số nét viết, hướng viết tại nơi bắt đầu, hướng viết tại nơi kết thúc. Tổng cộng, mỗi kí tự được trích 254 đặc trưng trực tuyến và không trực tuyến.

Kế đến, 254 đặc trưng vừa nêu sẽ được chọn lọc lại nhằm tăng tốc độ xử lí và loại bỏ dư thừa. Việc lựa chọn đặc trưng là vấn đề lớn, chưa có cách giải quyết thỏa đáng. Thông thường người ta dùng heuristic theo kiểu tham ăn (greedy) ví dụ giải thuật Best-First [1.1.1.19] để lựa chọn đặc trưng. Giải thuật này thực hiện rất chậm, nhất là trong trường hợp số đặc trưng là lớn. Trong nghiên cứu này chúng tôi thực hiện SVD để cắt giảm số chiều của không gian đặc trưng.

Cụ thể chúng tôi dùng tập dữ liệu huấn luyện gồm 4086 mẫu của 10 lớp dữ chữ số (0..9) viết tay trong UNIPEN. Mỗi lớp trung bình khoảng 400 mẫu. Thực hiện rút trích 254 đặc trưng đã trình bày. Như vậy tập dữ liệu huấn luyện được mô hình hóa như là một ma trận 254 x 4086, mỗi cột tương ứng với một mẫu trong tập dữ liệu học. Thực hiện SVD cắt ma trận này còn lại 100 chiều sau đó thực hiện phân lớp với SVM. Kết quả

được so sánh với trường hợp thực hiện phân lớp trên tập dữ liệu thô trên toàn bộ 254 đặc trưng.

Bảng 4 : So Sánh giữa việc dùng SVD trong lựa chọn đặc trưng và không dùng SVD trong lựa chọn đặc trưng.

	Số mẫu huấn luyện	Độ dài vector đặc trưng	Độ chính xác (10-folds)
Có dùng SVD	4086	90	99.2%
Không dùng SVD	4086	254	98.4%

Kết quả này cho thấy, SVD có thể được dùng trong lựa chọn đặc trưng. Về mặt phương pháp, nó có thể được dùng để làm rút ngắn số chiều không gian, loại bỏ dưa thừa và nhiễu. Vì nó làm rút ngắn số chiều của KG đặc trưng nên giúp cải thiện tốc độ học và phân lớp. Nó cũng có thể làm tăng độ chính xác.

4. KẾT LUẬN

Trong bài viết này chúng tôi đã trình bày phương pháp phân loại văn bản dựa trên máy học SVM. Đóng góp của chúng tôi là đã đề xuất dùng kỹ thuật phân tích giá trị đơn (SVD) để rút ngắn số chiều của không gian đặc trưng. Chúng tôi đã kiểm chứng đề xuất này trên 2000 tập tin độc lập tập huấn luyện thuộc 10 chủ đề với máy học SVM. Kết quả cho thấy rằng việc dùng SVD để phân tích và rút gọn số chiều của không gian đặc trưng đã nâng cao hiệu quả phân lớp SVM.

Các kiểm chứng thực nghiệm dựa trên tập hợp các mẫu độc lập với các mẫu dùng để xây dựng máy học cho thấy rằng hiệu quả của máy học SVM trong bài toán phân loại văn bản là ổn định, không phải là học vẹt. Việc phân tích giá trị đơn để rút gọn số chiều của không gian đặc trưng là hoàn toàn thích hợp cho bài toán phân loại văn bản, một bài toán mà không gian đặc trưng lớn, có nhiều nhiễu.

Kết quả nghiên cứu này có thể áp dụng vào các bài toán phân lớp khác nhau. Ở dây chúng tôi đã thực nghiệm trên 10 lớp dữ liệu chữ số viết tay, thực hiện rút số đặc trưng từ 254 về 90. Kết quả cho thấy độ chính xác nâng lên một ít (98.4% 99.2%). Độ tăng này không nhiều do kết quả nhận dạng 10 lớp số đã khá cao rồi. Tuy nhiên nó có ý nghĩa phương pháp luận: có thể rút ngắn số chiều không gian đặc trưng, tránh dự thừa và loại bỏ nhiễu. Chúng tôi sẽ tiếp tục nghiên cứu việc lựa chọn đặc trưng bằng phân tích giá trị đơn SVD và hi vọng sẽ cải tiến hiệu quả nhận dạng ảnh nói chung, nhận dạng chữ viết tay nói riêng.

Tài liệu tham khảo

1. Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R., Indexing by latent semantic analysis, Journal of the American Society for Information Science, 41(6): 390-407,1990.
2. Dumais, S., Latent Semantic Analysis, ARIST Review of Information Science and Technology, vol. 38, 2004, Chapter 4.
3. Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery, the Sedona Conference, 2007, pp. 189–223.
4. Foltz, P. W. and Dumais, S. T. Personalized Information Delivery: An analysis of information filtering methods, Communications of the ACM, 1992, 34(12), 51-60.
5. Gong, Y., and Liu, X., Creating Generic Text Summaries, Proceedings, Sixth International Conference on Document Analysis and Recognition, 2001, pp. 903–907.
6. Yarowsky, D., and Florian, R., Taking the Load off the Conference Chairs: Towards a Digital Paper-routing Assistant, Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very-Large Corpora, 1999, pp. 220–230.

7. Monay, F., and Gatica-Perez, D., On Image Auto-annotation with Latent Space Models, Proceedings of the 11th ACM international conference on Multimedia, Berkeley, CA, 2003, pp. 275–278.
8. Maletic, J., and Marcus, A., Using Latent Semantic Analysis to Identify Similarities in Source Code to Support Program Understanding, Proceedings of 12th IEEE International Conference on Tools with Artificial Intelligence, Vancouver, British Columbia, November 13–15, 2000, pp. 46–53.
9. Gee, K., Using Latent Semantic Indexing to Filter Spam, in: Proceedings, 2003 ACM Symposium on Applied Computing, Melbourne, Florida, pp. 460–464.
10. Foltz, Peter W., Laham, Darrell, and Landauer, Thomas K., Automated Essay Scoring: Applications to Educational Technology, Proceedings of EdMedia, 1999.
11. TRAN Cao De, ONO Kinji, Content-Based Image Retrieval: Object representation by the Density of Feature Points, RIVF'03 Hanoi, Vietnam, 10-13 February 2003.
12. De Cao Tran, Tri Cao Tran, Copy Detection Using Latent Semantic Similarity, IEEE International Conference on Research, Innovation & Vision for the Future, RIVF 2008, July 13-17, 2008, Ho Chi Minh City, Vietnam.
13. Trần Cao Đệ,Tìm kiếm tài liệu học tập đa ngôn ngữ với kỹ thuật chỉ mục ngữ nghĩa tiềm ẩn (latent semantic indexing), Hội thảo khoa học về CNPM và phần mềm nhóm, công nghệ tri thức và giải pháp mã nguồn mở cho hệ thống E-Learning, Huế, tháng 9 năm 2006
14. Trần Cao Đệ, Phân loại văn bản với máy học vector hỗ trợ (SVM) kết hợp với phân tích giá trị đơn, Hội thảo CNTT Quốc Gia @ 2011, Cần Thơ, 2011.
15. Chih-Hao Tsai, MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm. http://technology.chtsai.org/MMSEG/, 2000.
16. L. Heutte, T. Paquet, J.V. Moreau, Y. Lecourtier, C. Olivier, A structural/ statistical feature based vector for handwritten character recognition, Pattern Recognition Letters, Vol 19, pp. 629–641, 1998.
17. Dattatray V. Jadhao, Raghunath S. Holambe, Feature Extraction and Dimensionality Reduction Using Radon and Fourier Transforms with Application to Face Recognition, Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007) - Volume 02, Pages 254-260 , 2007.
18. Miao Zhenjiang, Zernike moment-based image shape analysis and its application, Pattern Recognition Letters 21 (2000), pp. 169 – 177.
19. Pudil P, Novovicova J, Kittler J. Floating search methods in feature selection, Pattern, Recognition Letters 1994, 15: 1119-1125.

TÌM KIẾM ẢNH THEO NỘI DUNG SỬ DỤNG PHÂN TÍCH NGỮ NGHĨA TIỀM ẨN THEO MÔ HÌNH XÁC SUẤT

Phạm Nguyên Khang[1], *Võ Trí Thức*[1]

TÓM TẮT

*Trong bài báo này, chúng tôi trình bày kỹ thuật Phân tích ngữ nghĩa tìm ẩn theo mô hình xác suất (Probabilistic latent semantic analysis, PLSA) và ứng dụng của nó vào tìm kiếm ảnh theo nội dung. PLSA là một kỹ thuật được phát triển để phân tích, xử lý dữ liệu văn bản theo mô hình túi từ (bag of words). Áp dụng PLSA trên dữ liệu văn bản cho phép phát hiện tự động các chủ đề của văn bản và biểu diễn lại văn bản trong không gian các chủ đề, có số chiều nhỏ hơn rất nhiều so với không gian các từ. Để có thể áp dụng được PLSA trên dữ liệu ảnh, ảnh phải được biểu diễn theo mô hình túi từ. Chúng tôi đề xuất sử dụng mô hình túi "từ của ảnh" (visual words) để biểu diễn nội dung ảnh. Tập các "từ của ảnh" được xây dựng bằng cách áp dụng một giải thuật gom nhóm (ví dụ như k-means) lên các đặc trưng cục bộ SIFT. Với cách biểu diễn này, ảnh được mô hình như văn bản và như thế ta có thể áp dụng PLSA lên dữ liệu ảnh. Kết quả của PLSA được sử dụng để tạo chỉ mục cho ảnh phục vụ cho việc tìm kiếm ảnh theo nội dung. Thực nghiệm trên các tập ảnh Caltech4, Caltech101 và Unbench, PLSA cho kết quả tốt hơn rất nhiều so với phương pháp TF*IDF trên mô hình túi từ gốc.*

Từ khóa: Tìm kiếm ảnh theo nội dung, SIFT, k-means, PLSA.

Title: Content-based image retrieval using Probabilistic latent Semantic Analysis.

ABSTRACT

*In this paper, we present Probabilistic Latent Semantic Analysis (PLSA) and its application in content based image retrieval. PLSA is originally developed for textual data processing using the bag of words model. The model represents a textual corpus by a contingency table crossing documents and terms/words. PLSA allows to discover latent topics and to represent documents in the topic space. In order to apply PLSA on images, we have to represent them in a form of a contingency table. We have proposed to use the bag of visual words model to describe the content of images. The visual vocabulary is constructed from a set of local features, SIFT, on images. Numerical results on Caltech4, Caltech101 and Unbench dataset show that PLSA outperform TF*IDF method on the original bag of words model.*

Keywords: Content based image retrieval, SIFT, k-means, Probalilistic Latent Semantic Analysis.

1. GIỚI THIỆU

Cùng với nhu cầu tìm kiếm văn bản, nhu cầu tìm kiếm ảnh cũng nhận được nhiều quan tâm của người sử dụng. Tuy nhiên, với một số lượng ảnh lớn lưu trữ thì công việc tìm kiếm trở nên vô cùng khó khăn. Để giải quyết vấn đề này, các hệ thống tìm kiếm ảnh đã ra đời như: Yahoo, Google Image Search,... Các hệ thống này cho phép người sử dụng nhập truy vấn về các ảnh cần quan tâm. Thông qua việc phân tích các văn bản đi kèm ảnh, hệ thống gửi trả các ảnh tương ứng với truy vấn của người dùng. Đây là một cách tiếp cận mà nhận được nhiều sự quan tâm của nhiều công trình khoa học trên thế giới. Tìm kiếm ảnh theo nội dung (Content Based Images Retrieval CBIR) hay truy vấn theo nội dung ảnh (Query Based Image Content QBIC) là một ứng dụng của thị giác máy tính đối với bài toán tìm kiếm ảnh [J. Eakins, 1999]. Dựa vào nội dung ảnh (Content-Based)

[1] Bộ môn Khoa Học Máy Tính – Khoa CNTT&TT – Trường Đại học Cần Thơ, Khu III, Số 1, Lý Tự Trọng, Quận Ninh Kiều, Tp.Cần Thơ. pnkhang@cit.ctu.edu.vn

[1] Bộ môn Khoa Học Máy Tính – Khoa CNTT&TT – Trường Đại học Cần Thơ, Khu III, Số 1, Lý Tự Trọng, Quận Ninh Kiều, Tp.Cần Thơ. vtthuc@cit.ctu.edu.vn

nghĩa là việc tìm kiếm sẽ phân tích nội dung thực sự của các bức ảnh. Nội dung ảnh ở đây được thể hiện bằng màu sắc, hình dạng, kết cấu (texture), các đặc trưng cục bộ (local features) [T.C. Siew, 2008],… hay bất cứ thông tin nào có từ chính nội dung ảnh. Đây là một tác vụ khó vì sự thay đổi về màu sắc, ánh sáng, thay đổi góc chụp ảnh, vật thể bị che khuất, ảnh hưởng đến phông nền lên vật thể. Gần đây, việc sử dụng các đặc trưng cục bộ SIFT (scale-invariant feature transform) đã mang lại nhiều thành tựu đáng kể trong phân tích ảnh. Trước hết, người ta tìm các điểm đặc biệt (interest points) trên ảnh. Các điểm này thường là những điểm nằm trong các vùng có kết cấu (texture) đặc biệt. Sau đó từng điểm đặc biệt, ta sẽ mô tả bằng một vector đặc trưng được trích lọc từ vùng xung quanh của điểm đặc biệt này. Một vector đặc trưng là một vector 128 chiều, trung bình một ảnh sẽ có khoảng 1000 vector đặc trưng. Để so sánh sự tương tác của 2 ảnh, ta đếm số cặp vector đặc trưng "khớp" với nhau giữa 2 ảnh. Phương pháp này cho kết quả rất tốt trong trường hợp ảnh thay đổi về độ sáng, màu sắc, góc chụp và ngay cả trường hợp vật thể được quan tâm bị che khuất một phần. Tuy nhiên nhược điểm chính của phương pháp này là độ phức tạp về không gian lưu trữ là rất lớn. Và thời gian tìm kiếm cũng tăng lên, do dữ liệu không được lưu trữ trong bộ nhớ trong. Hơn nữa với mỗi ảnh, ta phải thực hiện khoảng 1000 truy vấn (mỗi ảnh có khoảng 1000 vector đăc trưng). Sivic và các cộng sự đã trình bày một phương pháp để biểu diễn ảnh nhằm giải quyết vấn đề không gian lưu trữ và tốc độ truy vấn [J. Sivic & A. Zisserman, 2003]. Theo đó, các vector đặc trưng sẽ được gom nhóm bằng một giải thuật gom nhóm (clustering) nào đó (ví dụ như *k*-means). Mỗi nhóm được định nghĩa như là một *từ của ảnh* (visual word). Kế đến, các vector đặc trưng sẽ được gán vào các nhóm tương ứng. Thứ tự xuất hiện của các visual words không quan trọng và thường được bỏ qua. Đếm số vector của một ảnh được gán vào từng visual words ta sẽ thu được phân phối của các visual words trong ảnh này. Cách biểu diễn này cho phép biểu diễn một tập các vector đặc trưng SIFT (khoảng 1000 vector) thành một vector duy nhất, tương tự như cách biểu diễn văn bản bằng mô hình "túi từ" (bag of words) trong phân tích dữ liệu văn bản. Ta có thể áp dụng PLSA trên dữ liệu văn bản cho phép phát hiện tự động các chủ đề của văn bản và biểu diễn lại văn bản trong không gian các chủ đề, có số chiều nhỏ hơn rất nhiều so với không gian các từ. Để có thể áp dụng được PLSA trên dữ liệu ảnh thì ảnh cũng phải được biểu diễn theo mô hình túi từ. Chúng tôi đề xuất sử dụng mô hình túi "*từ của ảnh*" (visual words) để biểu diễn nội dung ảnh. Vậy với tập ảnh đầu vào ta sẽ biểu diễn được bằng một bảng tần số (contingency table) P(d, w) với các hàng (*d*) tương ứng với ảnh và các cột (*w*) tương ứng với các visual words. Phần tử P(d, w) mô tả số lần xuất hiện của từ *w* trong văn bản *d*. Sau đó áp dụng PLSA để phân tích bảng ngẫu nhiên tạo chỉ mục cho ảnh phục vụ cho việc tìm kiếm ảnh theo nội dung và kết quả thu được là 2 bảng giá trị P(z|d) và P(w|z). Với z là số biến tiềm ẩn.

Phần tiếp theo của bài viết này được trình bày như sau: phần 2 trình bày ngắn gọn về giải thuật SIFT, *k*-means, PLSA mà chúng tôi sử dụng vào việc tìm kiếm ảnh theo nội dung. Phần 3 trình bày các kết quả thực nghiệm tiếp theo sau đó là kết luận và hướng phát triển.

2. PHƯƠNG PHÁP

2.1 Đặc trưng SIFT

Phần này trình bày phương pháp trích rút các đặc trưng cục bộ bất biến SIFT của ảnh. Các đặc trưng này bất biến với việc thay đổi tỉ lệ ảnh, quay ảnh, đôi khi là thay đổi điểm nhìn và thêm nhiễu ảnh hay thay đổi cường độ chiếu sáng của ảnh. Phương pháp được lựa chọn có tên là Scale-Invariant Feature Transform (SIFT) và đặc trưng trích rút được gọi là đặc trưng SIFT (SIFT Feature). Các đặc trưng SIFT này được rút trích từ các

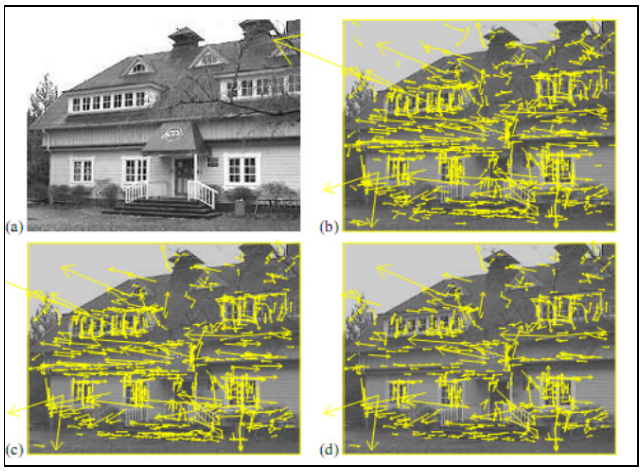
điểm đặc biệt cục bộ (Local Interest Point). Phương pháp này được phát triển bởi David G. Lowe [D.G. Lowe, 1998] [D.G. Lowe, 2004].

Phương pháp trích rút các đặc trưng bất biến SIFT được tiếp cận theo phương pháp thác lọc, theo đó phương pháp được thực hiện lần lượt theo các bước sau:

- *Bước 1:* Phát hiện các điểm cực trị Scale-Space (Scale-Space extrema detection): Bước đầu tiên này tiến hành tìm kiếm các điểm đặc biệt trên tất cả các tỉ lệ và vị trí của ảnh. Nó sử dụng hàm different-of-Gaussian để xác định tất cả các điểm đặc biệt tiềm năng mà bất biến với quy mô và hướng của ảnh.

- *Bước 2:* Định vị các điểm đặc biệt (keypoint localization): Một hàm kiểm tra sẽ được đưa ra để quyết định xem các điểm đặc biệt tiềm năng có được lựa chọn hay không?

- Loại bỏ các điểm đặc biệt có độ tương phản thấp.
- Một số điểm đặc biệt dọc theo các cạnh không giữ được tính ổn định khi ảnh bị nhiễu cũng bị loại bỏ.



Hình 1: Quá trình lựa chọn các điểm đặc biệt

(a) Ảnh gốc, (b) Các điểm đặt biệt được phát hiện,

(c) Ảnh sau khi loại bỏ các điểm đặc biệt có độ tương phản thấp,

(d) Ảnh sau loại bỏ các điểm đặc biệt dọc theo cạnh.

- *Bước 3:* Xác định hướng cho các điểm đặc biệt (Orientation assignment): Xác định hướng cho các điểm đặc biệt được chọn.

- ***Bước 4:*** **Mô tả các điểm đặc biệt** (Keypoint descriptor): Các điểm đặc biệt sau khi được xác định hướng sẽ được mô tả dưới dạng các vector đặc trưng nhiều chiều. Điểm đặc biệt sau khi được xác định hướng sẽ được biểu diễn dưới dạng các vector 4x4x8=128 chiều.

Chúng tôi áp dụng giải thuật SIFT để trích chọn các điểm đặc biệt của ảnh truy vấn và ảnh trong cơ sở dữ liệu (vector đặc trưng SIFT). Như vậy, với mỗi điểm đặc biệt trên ảnh ta biểu diễn nó bằng một vector 128 chiều. Kết quả thu được là mỗi ảnh trung bình chúng ta có khoảng 1000 vector đặc trưng 128 chiều.

2.2 Giải thuật *k*-means:

k-means là một thuật toán được áp dụng khá nhiều trong gom nhóm dữ liệu vì hiệu năng và tính hiện thực khá tốt. Tuy nhiên ngoài việc cần cho trước số cụm, *k*-means

còn đòi hỏi phải chọn trước k điểm làm trọng tâm, việc chọn ngẫu nhiên này có thể cho ra các kết quả khác nhau.

Mỗi ảnh sau khi SIFT ta có khoảng 1000 vector đặc trưng 128 chiều thì được biểu diễn tương ứng với một vector duy nhất. Sau khi áp dụng giải thuật SIFT, để so sánh sự tương tác của 2 ảnh, ta đếm số cặp vector đặc trưng "khớp" với nhau giữa 2 ảnh. Mỗi ảnh có khoảng 1000 vector đặc trưng thì thực hiện việc so sánh này thì độ phức tạp về không gian lưu trữ là rất lớn. Để giải quyết vấn đề trên chúng tôi sử dụng mô hình túi "từ của ảnh" (visual words) để biểu diễn nội dung ảnh thông qua giải thuật gom nhóm *k*-means. Như vậy, với mỗi ảnh SIFT (khoảng 1000 vector đặc trưng) sau khi áp dụng *k*-means ta sẽ biểu diễn nó bằng 1 vector duy nhất. Với tập ảnh đầu vào ta sẽ biểu diễn bằng một bảng ngẫu nhiên P(d, w) với *d* tương ứng với ảnh và *w* tương ứng với các visual words. Giải thuật được sử dụng đầu tiên bởi James MacQueen vào năm 1967[1]. Nội dung giải thuật *k*-means như sau:

Tiền xử lý: số hóa dữ liệu thành dạng điểm trong không gian nhiều chiều, tọa độ mỗi điểm là (x, y, z, t, ...)

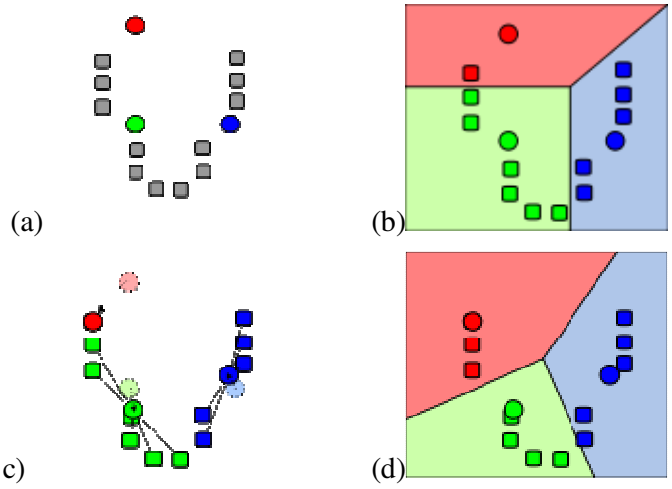
- ***Bước 1:*** Tạo ngẫu nhiên k điểm để làm điểm trọng tâm, mỗi trọng tâm sẽ xác định một tập con.

- ***Bước 2:*** Lần lượt tính khoảng cách giữa các điểm trong tập dữ liệu đang xét với từng điểm trọng tâm, điểm nào gần trọng tâm thứ i nhất sẽ thuộc về tập hợp i. Khoảng cách có thể tính bằng công thức Euclidean hoặc Mahattan.

- ***Bước 3:*** Trong mỗi tập con, ta lần lượt tính lại trọng tâm mới bằng cách lấy trung bình cộng của tất cả các điểm.

⇨ **Lặp lại bước 2 và 3 cho đến khi nào tọa độ của tất cả các điểm trọng tâm không còn thay đổi nữa.**

Giải thuật có thể được mô tả bằng những hình sau:



Hình 2: *Quá trình gom nhóm dữ liệu*

(a) Tạo k điểm trọng tâm. (b) Tính khoảng cách giữa các điểm trọng tâm

(c) Tính lại trọng tâm mới. (d) Kết quả sau khi lập bước 2 và 3

Độ phức tạp: O (nkt)

[1] http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

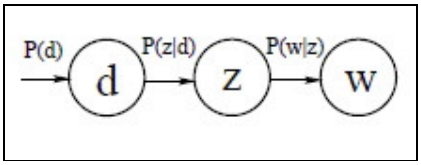
- ✓ n là số đối tượng
- ✓ k là số nhóm
- ✓ t là số lần lặp (k << n, t << n)

2.3 Phân tích ngữ nghĩa tiềm ẩn theo mô hình xác suất (PLSA)

Sau khi áp dụng *k*-means để gom cụm các vector đặc trưng lại thì với tập ảnh đầu vào ta sẽ có được một bảng tần số với các hàng tương ứng với ảnh và các cột tương ứng với các visual words. Tiếp theo, ta sử dụng PLSA để phân tích bảng ngẫu nhiên trên tạo chỉ mục cho ảnh và phục vụ cho việc tìm kiếm ảnh theo nội dung. Nội dung của giải thuật PLSA như sau:

Mô hình tổng quát: [T. Hofmann, 1999] Ý tưởng của PLSA dựa trên mô hình xác suất hay còn gọi là mô hình biến ẩn cho tập dữ liệu kép (d, w), trong đó mỗi xuất hiện (d, w) đều tương tác với một lớp biến tiềm ẩn $z \in Z = \{z_1, z_2, \ldots, z_K\}$. Mô hình được mô tả như sau:

- ✓ Chọn một tài liệu (ảnh) d với xác suất P(d),
- ✓ Chọn một chủ đề ẩn z với xác suất P(z|d),
- ✓ Sinh một từ w (visual words) với xác suất P(w|z).



Hình 3: Mô hình PLSA

$$P(d, w) = P(d)P(w \mid d) \quad (1)$$

Trong đó:
$$P(w \mid d) = \sum_{z \in Z} P(w \mid z)P(z \mid d) \quad (2)$$

Thay (2) vào (1)
$$P(d, w) = P(d) \sum_{z \in Z} P(w|z)P(z|d) \quad (3)$$

Áp dụng công thức Bayes với P(z|d)= P(z)P(d|z) vào (3) ta có được:

$$P(d, w) = \sum_{z \in Z} P(z)P(d|z)P(w|z) \quad (4)$$

Trong mô hình này ta đã giả sử *d* và *w* độc lập có điều kiện theo *z*. Chẳng hạn, các điều kiện về tài liệu (ảnh) *d* hay các từ *w* hoàn toàn độc lập với trạng thái của lớp tiềm ẩn *z*. Số lượng biến tiềm ẩn *z* nhỏ hơn rất nhiều so với số lượng từ hay tài liệu trong hệ thống. Ở đây, *z* đóng vai trò như một biến chủ đề.

Từ mô hình xác suất, ta có hàm khả năng như sau:

$$L = \sum_{d \in D} \sum_{w \in W} n(d, w) \log P(d \mid w), \quad (5)$$

với *n(d, w)* là tần suất xuất hiện của từ w trong tài liệu d.

Vấn đề được đặt ra là cần cực đại hóa khả năng trên (Maximum Likelihood) thông qua thuật toán EM (Expectation Maximization). Thuật toán EM thực hiện chuyển đổi giữa hai bước E (Expectation) và M (Maximization) như sau:

Bước E: Tính xác suất cho các biến tiềm ẩn với các giá trị tham số hiện tại. Đó là khả năng xuất hiện của từ *w* trong tài liệu *d* trình bày chủ đề *z* với công thức sau:

$$P(z \mid d,w) = \frac{P(z)P(d \mid z)P(w \mid z)}{\sum_{z'} P(z')P(d \mid z')P(w \mid z')}, \qquad (6)$$

Bước M: Cập nhật lại các tham số cho việc tính xác suất ở bước E để tìm ra các tham số làm cho cực đại hóa hàm khả năng:

$$P(w \mid z) = \frac{\sum_d n(d,w)P(z \mid d,w)}{\sum_{d,w'} n(d,w')P(z \mid d,w')} \quad (7) \qquad P(w \mid z) = \frac{\sum_d n(d,w)P(z \mid d,w)}{\sum_{d,w'} n(d,w')P(z \mid d,w')}$$

$$P(d \mid z) = \frac{\sum_w n(d,w)P(z \mid d,w)}{\sum_{d',w} n(d',w)P(z \mid d',w)} \quad (8) \qquad P(z \mid d) = \frac{\sum_w n(d,w)P(z \mid d,w)}{\sum_{w,z'} n(d,w)P(z' \mid d,w)} \quad (8')$$

$$P(z) = \frac{1}{R}\sum_{d,w} n(d,w)P(z \mid d,w) \qquad (9), \quad R \equiv \sum_{d,w} n(d,w)$$

Sau m bước lặp tính toán giữa bước M và E thì kết quả P(w|z) và P(z|d) sẽ hội tụ.

Độ phức tạp: (O(d*w)

- ✓ d: số tài liệu (số ảnh).
- ✓ w: số visual words.

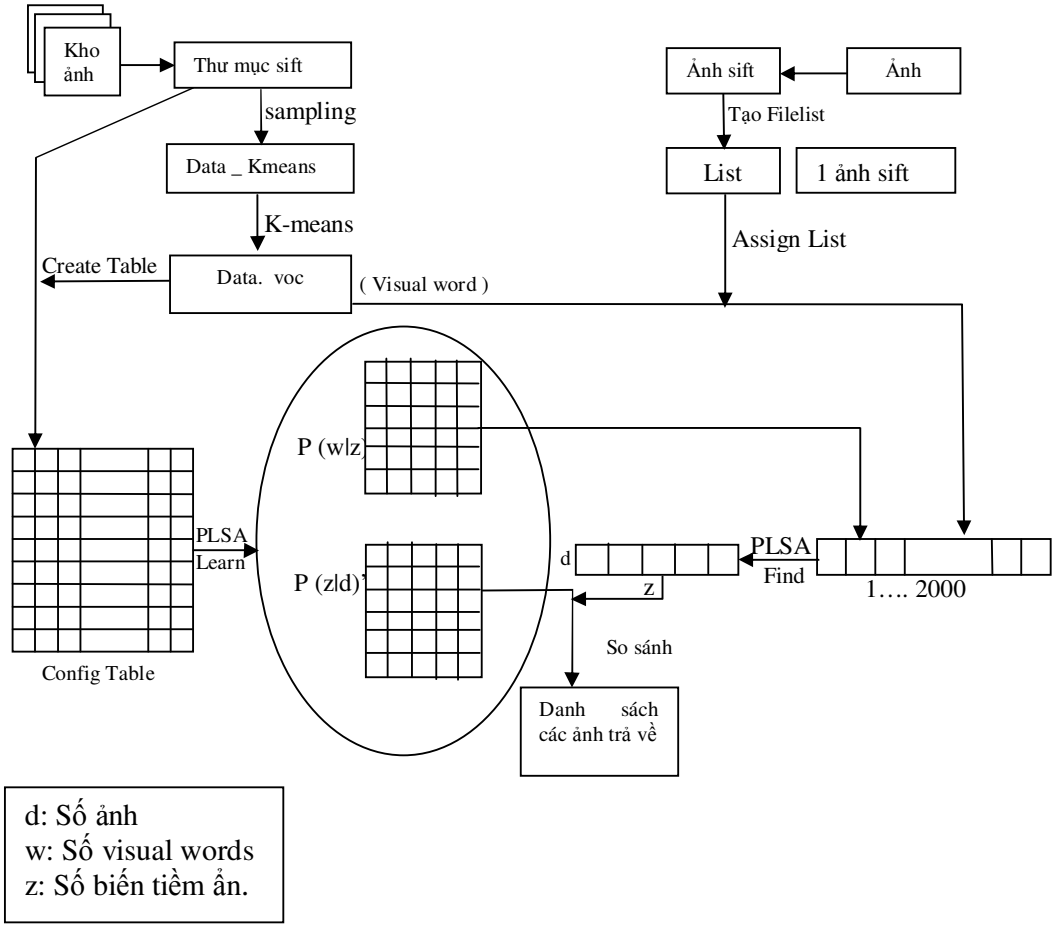
Sau khi gom nhóm tập dữ liệu ảnh ta có được bảng ngẫu nhiên P(d, w) với dòng tương ứng là các ảnh và cột là các visual words. Ta sẽ áp dụng bảng ngẫu nhiên này trên PLSA. PLSA gồm có 2 giai đoạn:

- ✓ **Giai đoạn 1: Tạo chỉ mục (học)**

Đầu vào là bảng giá trị ngẫu nhiên kho dữ liệu ảnh P(d, w) với d là số ảnh và w là số visual words. Và lần lượt ta khởi tạo ngẫu nhiên 2 bảng giá trị là P(z|d) và P(w|z) có giá trị thuộc [0; 1]. Và ta thực hiện theo các bước của giải thuật PLSA, kết quả thu được là 2 bảng giá trị P(z|d) và P(w|z). Với z là các biến tiềm ẩn.

- ✓ **Giai đoạn 2: Tìm ảnh**

Cũng như trên đầu vào là bảng giá trị ngẫu nhiên P(d, w) của ảnh truy vấn (d=1) và khởi tạo bảng P(z|d). Tiếp đến là bảng giá trị P(w|z) trong giai đoạn 1 được đưa vào để tính toán. Kết quả thu được là bảng giá trị tương đương với ảnh truy vấn là P(z|d). Tiếp đến ta sẽ áp dụng công thức tính khoảng cách của vector của ảnh truy vấn với từng vector trong bảng giá trị của kho dữ liệu để có được kết quả tìm kiếm. Giá trị vector có khoảng cách càng nhỏ với vector truy vấn thì ảnh tương ứng với vector đó càng giống với ảnh truy vấn. Mô hình có thể được mô tả thông qua hình sau:



Hình 4: Mô hình tổng thể của hệ thống

3. KẾT QUẢ THỰC NGHIỆM

Để có thể đánh giá hiệu quả của giải thuật, chúng tôi cài đặt giải thuật PLSA bằng ngôn ngữ lập trình C/C++. Bên cạnh đó, để tăng khả năng tính toán thì chúng tôi đã sử dụng Cblas[1] [2]để hỗ trợ cho việc tính toán các phép tính về ma trận.

Chúng tôi tiến hành thực nghiệm trên tập ảnh Caltech4, Caltech101 và Unbench. Với những tập dữ liệu có sẵn tập học và tập thử, chúng tôi dùng tập học để huấn luyện. Sau đó, dùng tập thử để kiểm tra. Ở đây, chúng tôi chia tập dữ liệu ra là 3 và sử dụng 2 tập để huấn luyện, 1 tập để kiểm tra. Tập ảnh kết quả là k ảnh gần giống nhất với ảnh truy vấn.

Chúng tôi sử dụng độ chính xác trung bình (Average Precision) để đánh giá kết quả thực nghiệm của giải thuật:

Độ chính xác trung bình được định nghĩa như sau [Mu Zhu, 2004]:

[1] http://math-atlas.sourceforge.net/

[2] http://www.gnu.org/software/gsl/manual/html_node/GSL-CBLAS-Library.html

$$AP = \frac{\sum_{k=1}^{n} P@K \times I(K)}{\sum_{j=1}^{n} I(j)}$$

Trong đó:

- ✓ n là số phần tử trong cơ sở dữ liệu.
- ✓ $P@K = \frac{Match@K}{K}$

  (Match@K = số các đối tượng phù hợp ở K vị trí đầu tiên)
- ✓ I(K) = 1 nếu đối tượng ở vị trí K phù hợp với câu truy vấn, ngược lại I(K) = 0

Giá trị trung bình của m truy vấn:

$$MAP = \frac{\sum_{i=1}^{m} AP_i}{m}$$

Thử nghiệm hệ thống với 3 tập dữ liệu Caltech4, Caltech101, Unbench cho kết quả là bảng 1, 2 và 3:

Bảng 1: Kết quả thực nghiệm trên tập Caltech4

Phương pháp	P@10	P@20	P@50	P@100	P@200	MAP
tf*idf	0.879	0.868	0.849	0.828	0.795	0.663
PLSA, z=5	0.862	0.862	0.858	0.850	0.84	0.800
PLSA, z=7	0.938	0.936	0.931	0.925	0.916	0.863
PLSA, z=15	0.964	0.958	0.950	0.942	0.928	0.824
PLSA, z=30	0.968	0.962	0.952	0.942	0.924	0.773
PLSA, z=50	0.967	0.959	0.946	0.935	0.914	0.751

Bảng 2: Kết quả thực nghiệm trên tập Caltech101

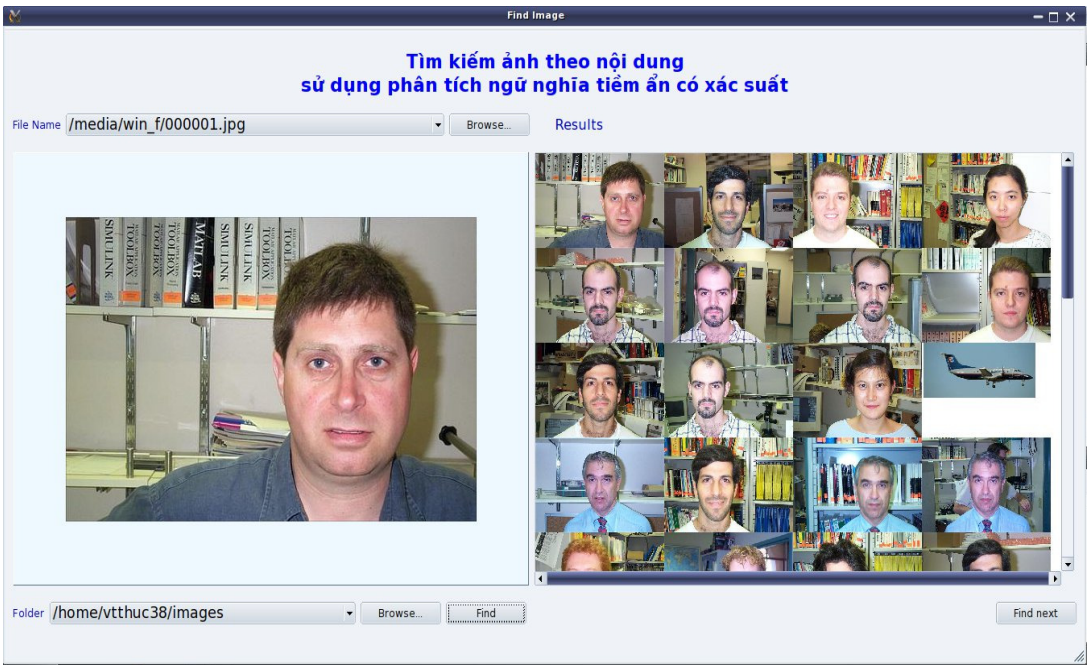
Phương pháp	P@10	P@20	P@50	P@100	P@200	MAP
tf*idf	0.258	0.244	0.217	0.192	0.166	0.136
PLSA, z=30	0.294	0.277	0.252	0.231	0.210	0.174
PLSA, z=50	0.326	0.307	0.279	0.252	0.277	0.195
PLSA, z=75	0.333	0.312	0.283	0.255	0.255	0.186
PLSA, z=100	0.333	0.312	0.28	0.251	0.221	0.181
PLSA, z=200	0.337	0.317	0.286	0.255	0.223	0.185

Bảng 3: Kết quả thực nghiệm trên tập Unbench

Phương pháp	P@3	MAP
tf*idf	0.258	0.295
PLSA, z=30	0.294	0.174
PLSA, z=50	0.326	0.195
PLSA, z=75	0.333	0.186
PLSA, z=100	0.333	0.181
PLSA, z=200	0.337	0.185

Từ các kết quả thống kê trên, ta thấy rằng độ chính xác trung bình của hệ thống là khá cao. Với tập dữ liệu Caltech4, khi cho z=30 thì ta có kết quả P@10=0.968 và

MAP=0.773 cao hơn so với phương pháp tf*idf thì P@10=0.879 và MAP=0.663. Hay với tập dữ liệu Caltech101 với z=50 thì P@50=0.279 và MAP=0.195 cũng cao hơn so với phương pháp tf*idf là P@50=0.217 và MAP=0.136. Đặc biệt, theo khảo sát của thực nghiệm, hệ thống cho kết quả rất chính xác với kết quả đầu tiên trả về. Đối với tập dữ liệu có chứa ảnh giống hệt so với ảnh truy vấn, thì khả năng ảnh thứ nhất được trả về giống hệt với ảnh truy vấn là tuyệt đối. Sau đây là hình ảnh minh họa tìm kiếm ảnh trên tập Caltech4:



Hình 5: Ví dụ minh họa tìm kiếm ảnh

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi đã trình bày mô hình tìm kiếm ảnh theo nội dung thông qua giải thuật SIFT, *k*-means và PLSA. Với giải thuật SIFT được áp dụng để trích chọn các điểm đặc biệt trong ảnh và giải thuật gom nhóm *k*-means để gom nhóm các điểm đặc biệt (vector đặc trưng) trên ảnh này. Với tập ảnh đầu vào ta sẽ biểu diễn nó bằng một bảng ngẫu nhiên với các hàng tương ứng với ảnh và các cột tương ứng với các visual words. Sau cùng là áp dụng giải thuật PLSA để phân tích bảng ngẫu nhiên trên tạo chỉ mục cho ảnh và phục vụ cho việc tìm kiếm ảnh theo nội dung.

Đề tài đã tiến hành thử nghiệm mô hình trên 3 tập dữ liệu là Caltech4, Caltech101 và Unbench. Kết quả kiểm tra trên tập dữ liệu Caltech4, khi cho z=30 thì ta có kết quả P@10=0.968 và MAP=0.773, tập dữ liệu Caltech101 với z=50 thì P@10=0.326 và MAP=0.195 và tập dữ liệu Unbench với z=200 thì P@3=0.337 và MAP là 0.185. Từ những kết quả bước đầu cho thấy tính khả quan và đúng đắn của mô hình.

Trong thời gian tới, chúng tôi sẽ tiến hành thử nghiệm mô hình với nhiều tập dữ liệu khác nhau nữa. Bên cạnh đó, chúng tôi sẽ phát triển hệ thống để tìm kiếm ảnh trên Internet. Và sẽ phát triển thêm cũng tìm ảnh ngữ nghĩa theo nội dung nhưng sử dụng phân phối Dirichlet và sử dụng phân tích tương ứng. Để có sự đánh giá tổng quát hơn về mức độ hiệu quả giữa các giải thuật.

TÀI LIỆU THAM KHẢO

T. Hofmann**,** *Probabilistic Latent Semantic Indexing*, International Computer Science Institute, Berkeley, CA & EECS Department, CS Division, UC Berkeley, 1999.

T. Hofmann*, Probabilistic Latent Semantic Analysis*, Uncertainity in Artificial Intelligence, UAI'99, Stockholm, 1999.

J. Eakins, Margaret Graham*, Content-based Image Retrieval*, University of Northumbria at Newcastle, 10-1999.

T.C. Siew, *Feature selection for content-based image retrieval using statistical discriminant*, Faculty of Computer Science and Information System Universiti Teknologi Malaysia, 10-2008.

D.G. Lowe, *Distinctive Image Features from Scale-Invariant Keypoints,* International Journal of Computer Vision, 2004.

D.G. Lowe , *Object Recognition from Local Scale-Invariant Features*, Computer Science Department University of British Columbia Vancouver, B. C., V6T 1Z4, Canada, 1999.

J. Sivic and A. Zisserman, *Video Google: A Text Retrieval Approach to Object Matching in Video*, Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003.
M.Zhu, *Recall, Precision and Average Precision*, Department of Statistics & Actuarial Science of Waterloo University, 2004.

XÂY DỰNG DỊCH VỤ BẢN ĐỒ TƯƠNG TÁC VỚI CÁC WEBSERVICES DỰA TRÊN KIẾN TRÚC SOA

Nguyễn Văn Kiệt, Trương Xuân Việt*, Lê Quyết Thắng***

**Trung tâm Công nghệ Phần mềm Đại học Cần Thơ (CUSC)*

*** Khoa Công nghệ Thông tin & Truyền thông – Đại học Cần Thơ*

Abstract:
From the recent years, belong to the developing of the WEB 2.0, the WebGIS application have a great innovation not only in the applied technologies but also in the back-end applications. The stages of WebGIS can be divided into the ways to develop the different applications: in the first days, the developers must create by them-self all necessary components for the application; and soon after, the architecture of WebGIS application is gradually completed with the appears of a various server-side applications (such as MapServer, ArcGIS Services, GeoServer, ...), and also client-side applications (KaMap, OpenLayers, …). Specially, the OPENGIS standard appeared with the great agreement of a lot companies and the organizations promotes truly this kind of application.
Actually, this research is not aimed to describe the history of the WebGIS applications, we have to go to the depth of the inside services. In which, two following main purposes are proposed and resolved: *(1) develop a kernel of server-side application for WebGIS named MicroGIS (version 1.1), complying with the OPENGIS standards* and *(2) create the interfaces to connect with the others WebServices*. This research is already applied to develop the WebGIS application in the SOA architecture in the national project KC.01.06/05-10 of Vietnam.

1. GIỚI THIỆU

Trong những năm gần đây, sự phát triển của WEB 2.0, ứng dụng WebGIS có sự thay đổi lớn không chỉ trong ứng dụng công nghệ mà còn trong các ứng dụng đầu cuối. Các giai đoạn của WebGIS có thể được chia thành hai hướng phát triển khác nhau: giai đoạn đầu tiên, người phát triển phải tạo ra tất cả các thành phần cho ứng dụng; giai đoạn tiếp theo, với kiến trúc của ứng dụng WebGIS đang dần được hoàn thiện với những ứng dụng server-side (MapServer, ArcGIS Services, GeoServer, ...), cũng như các ứng dụng phía client (KaMap, OpenLayers, …). Đặc biệt, chuẩn OPENGIS xuất hiện với sự đồng thuận của nhiều công ty và các tổ chức, các công ty này đã và đang xây dựng các ứng dụng WebGIS dựa trên chuẩn này.
Trong bài báo này, chúng tôi không mô tả lịch sử phát triển của các ứng dụng WebGIS, chúng tôi sẽ nghiên cứu sâu bên trong của dịch vụ. Trong đó, hai mục chính được đề xuất và hướng giải quyết: (1) *xây dựng nhân của ứng dụng server-side cho WebGIS (MicroGIS) theo chuẩn OpenGIS của OGC và (2) tạo giao diện (interface) kết nối với các dịch vụ web*. Nghiên cứu này được ứng dụng để xây dựng ứng dụng WebGIS theo kiến trúc SOA trong đề tài nghiên cứu khoa học cấp nhà nước "Nghiên cứu xây dựng các hệ thống thông tin hỗ trợ việc phòng chống dịch bệnh cây trồng và thuỷ sản cho vùng kinh tế trọng điểm" (KC.01.06/05-10).

2. ĐẶT VẤN ĐỀ

Hiện nay, công nghệ GIS được nghiên cứu và ứng dụng đa dạng, đa tầng trong nhiều lĩnh vực khác nhau (phân tích các sự kiện, dự đoán tác động và hoạch định chiến lược). Chính vì thế, việc tìm hiểu các dịch vụ web cho ứng dụng GIS (WebGIS) và xây dựng ứng dụng khai thác dịch vụ nay không chỉ là một yêu cầu mang tính khoa học của ngành công nghệ thông tin mà còn là một đòi hỏi của chính thực tiễn đời sống trong bối cảnh toàn cầu hoá.

Tuy nhiên các ứng dụng WebGIS hiện nay chủ yếu tập trung giải quyết các bài toán xử lý dữ liệu địa lý nội tại và hiển thị dữ liệu đã xử lý trên nền Web, chưa thật sự chú trọng giải quyết các bài toán về kết hợp giữa dữ liệu địa lý tập trung và dữ liệu phi địa lý phân tán.

Vấn đề đặt ra là làm thế nào để có thể kết hợp hai nguồn dữ liệu này với nhau theo một cách thức chung nhưng vẫn đảm bảo được khả năng giao tiếp và truyền tải thông tin theo một chuẩn chung thống nhất (chuẩn OpenGIS) giữa Server và Client. Tạo điều kiện thuận lợi cho quá trình khai thác và trình bày dữ liệu phi địa lý phân tán khắp nơi lên bản đồ địa lý tập trung.

3. CƠ SỞ LÝ THUYẾT

3.1. Kiến trúc hướng dịch vụ

Kiến trúc hướng dịch vụ là một hướng tiếp cận mới trong kiến trúc phần mềm ứng dụng. Trong đó, một ứng dụng được cấu thành từ một tập các thành phần độc lập, phân tán, phối hợp "mềm dẻo" với nhau được gọi là các dịch vụ (services).

Bốn nguyên tắc chính của một hệ thống SOA: Sự phân định ranh giới rạch ròi giữa các dịch vụ, các dịch vụ hoạt động tự động, các dịch vụ chia sẻ lược đồ, tính tương thích của dịch vụ dựa trên chính sách

Các tính chất của một hệ thống SOA: Loose coupling, sử dụng lại dịch vụ, sử dụng dịch vụ bất đồng bộ, quản lý các chính sách, coarse granularity, khả năng cộng tác, tự động dò tìm và ràng buộc động, tự hồi phục.

Lợi ích của SOA: Sử dụng lại những thành phần có sẵn, giải pháp ứng dụng tổng hợp cho doanh nghiệp, tính loose coupling giúp tăng tính linh hoạt và khả năng triển khai cài đặt, hỗ trợ đa thiết bị và đa nền tản, tăng khả năng mở rộng và khả năng sẵn sàng cung cấp.

3.2. Chuẩn OpenGIS

OGC (Open Geospatial Consortium) là một tổ chức xây dựng các chuẩn với tính chất đồng tâm, tự nguyện, có tính toàn cầu và phi lợi nhuận. OGC dẫn dắt việc phát triển các chuẩn cho các dịch vụ trên cơ sở vị trí và không gian địa lý

Ngày nay, OGC là một tổ chức quốc tế của 401 công ty (số liệu ngày 30/08/2010), các tổ chức chính phủ và các trường đại học tham gia vào quá trình tìm tiếng nói chung để phát triển các đặc tả giao tiếp cho cộng đồng. Chúng ta thường gọi đó là các đặc tả OPENGIS (OpenGIS Specifications). Các đặc tả OpenGIS hỗ trợ các giải pháp đồng vận hành, tích hợp làm cho dữ liệu địa lý luôn sẵn sàng phục vụ cho Web, các dịch vụ trên nền tảng định vị, các dịch vụ không dây và phù hợp với các xu hướng chính của công nghệ thông tin. Các đặc tả sẽ tăng cường sức manh cho các nhà phát triển công nghệ nhằm biến các dịch vụ và thông tin không gian phức tạp trở nên dễ dàng truy cập và hữu ích bới hầu hết các loại ứng dụng.

3.3. Web Map Service (WMS)

WMS là một dịch vụ giúp tạo ra các bản đồ dựa trên các dữ liệu địa lý. Bản đồ ở đây được hiểu như một cách thể hiện trực quan của dữ liệu địa lý, còn bản thân bản đồ không được xem là dữ liệu. Các bản đồ này được hiển thị dưới định dạng ảnh như PNG, GIF, JPEG hoặc các định dạng thành phần đồ họa vector như SVG (Scalable Vector Graphics), WebCGM (Web Computer Graphics Metafile). Một WMS sẽ hỗ trợ ba operation, trong đó có hai operation đầu là bắt buộc cho mọi WMS.

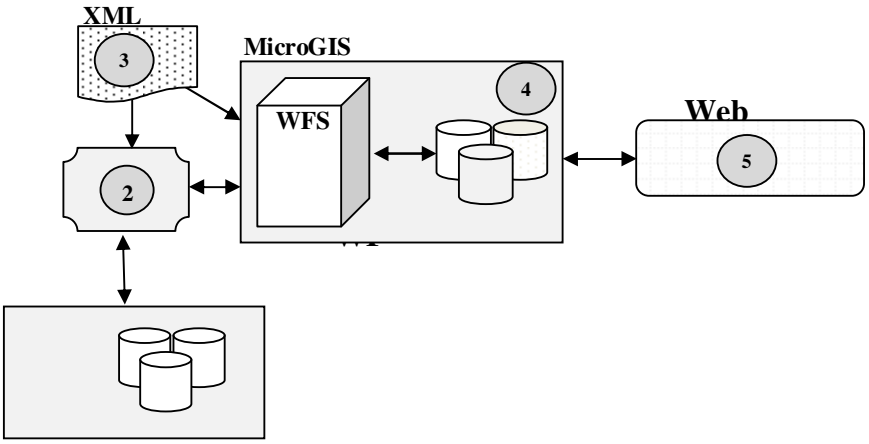
- GetCapabilities: cung cấp các thông tin metadata ở mức dịch vụ, đó là đặc tả cho các thông tin của dịch vụ WMS và các tham số cần thiết cho các yêu cầu.
- GetMap: cung cấp ảnh bản đồ khi nhận được các tham số về chiều và thông tin không gian địa lý hợp lệ.
- GetFeatureInfo: truy vấn thông tin của các feature trên bản đồ.

3.4. Web Feature Service (WFS)

WFS cho phép một client nhận và cập nhật dữ liệu không gian được mã hóa trong GML từ nhiều WFS khác nhau. WFS hỗ trợ các thao tác INSERT, UPDATE, DELETE, LOCK, QUERY và DISCOVERY trên các feature địa lý và phi địa lý sử dụng giao thức HTTP.

4. NỘI DUNG NGHIÊN CỨU

4.1. Mô hình dịch vụ MiroGIS



Hình 1: Mô hình tổng thể dịch vụ MicroGIS

Hình 1 mô tả kiến trúc tổng thể dịch vụ MicroGIS, mô hình gồm 5 thành phần đảm nhiệm các công việc sau:

(1) Xây dựng dịch vụ WFS theo chuẩn OpenGIS. Dữ liệu của gói này chủ yếu tương tác với cơ sở dữ liệu không gian.

(2) Chủ yếu tập trung vào phân tích các yêu cầu tương tác giữa dịch vụ WFS và các phân hệ khác (Lúa, mô phỏng, thủy sản,…).

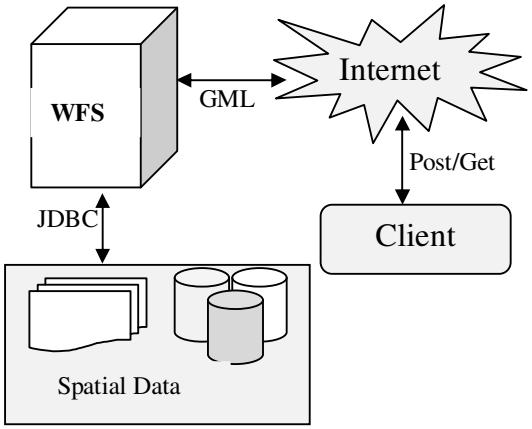
(3) Tập trung vào phân tích dữ liệu tương tác giữa phân hệ WFS với các phân hệ khác (Lúa, mô phỏng, thủy sản,…).

(4) Tích hợp nội dung từ 1 đến 3 nhằm tạo ra hệ thống hoàn thiện (Microgis) có khả năng tương tác với các phân hệ khác (Lúa, mô phỏng, thủy sản,…).

(5) Xây dựng công cụ tìm kiếm và hiển thị bản đồ trên web cho phép người dùng có thể xem bản đồ và các hiện trạng dịch bệnh từ phân hệ Lúa, mô phỏng, thủy sản,… theo không gian và thời gian.

4.2. Mô hình tổng thể dịch vụ WFS

Hình 2 mô tả chi tiết các thành phần tham gia và nối kết của dịch vụ WFS. Trong đó client kết nối và lấy dữ liệu thông qua HTTP Get/Post, kết quả đáp ứng yêu cầu của dịch vụ cho một client là tập GML, tập này chứa các thông tin liên quan đến vùng dữ liệu thỏa yêu cầu của client. Các kết nối từ WFS đến các database không phân biệt khoản cách địa lý. Điều lưu ý trong mô hình là client phải có khả năng phân tích tập GML dựa trên đặc tả GML Schema và hỗ trợ hiển thị dữ liệu vector.



Hình 2: Mô hình dịch vụ WFS

4.2.1. GetCapabilities

Quá trình xử lý GetCapabilities (hình 3) được thực hiện theo thứ tự sau:

(1) Parser: nhận và phân tích yêu cầu từ Client. Sau khi nhận và phân tích yêu cầu, nếu yêu cầu được xác định là GetCapabilities thì lớp Parser chuyển yêu cầu này đến lớp Processor.

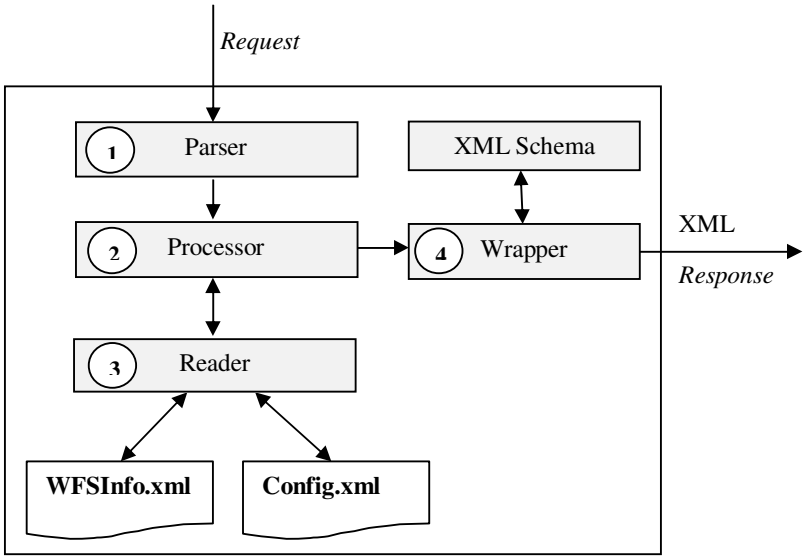
(2) Processor: nhận yêu cầu từ Parser, do yêu cầu GetCapabilities liên quan đến việc mô tả thông tin về dịch vụ WFS nên lớp Processor thực hiện gọi lớp Reader để đọc thông tin.

(3) Reader: nhận yêu cầu từ Processor, thực hiện đọc thông tin từ hai tập tin WFSInfo.xml và Config.xml theo yêu cầu của Processor. Cấu trúc WFSInfo.xml và Config.xml được mô tả phụ lục 2.

(2) Processor: nhận kết quả từ lớp Reader và chuyển thông tin này sang lớp Wrapper.

(4) Wrapper: nhận thông tin từ Processor và thực hiện đóng gói dữ liệu theo định dạng XML dựa trên XML Schema đã được đặc tả, sau đó trả kết quả về Client.

- Mô hình xử lý

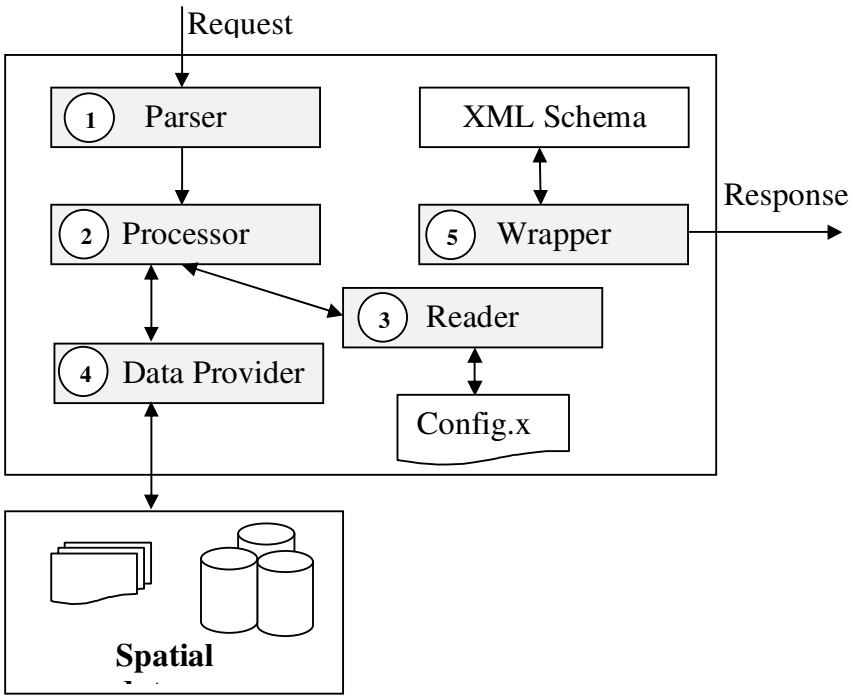


Hình 3: Mô hình xử lý GetCapabilities

4.2.2. GetDescribeFeatureType

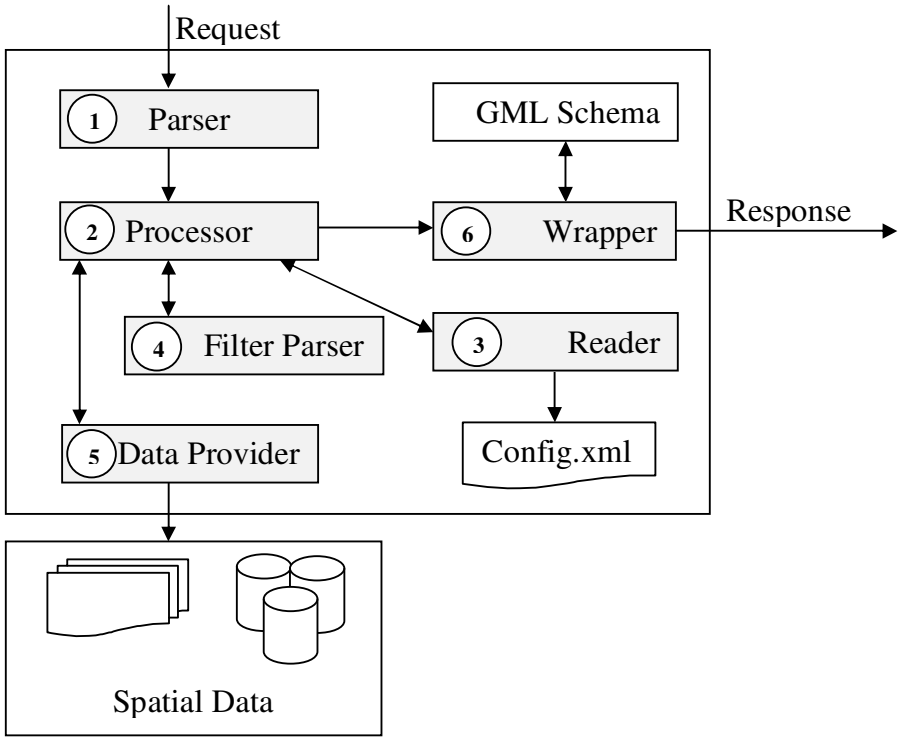
Quá trình xử lý GetDescribeFeatureType (hình 4) được thực hiện như sau:
(1) Parser: nhận và phân tích yêu cầu từ Client, nếu yêu cầu được xác định là GetDescribeFeatureType thì lớp Parser chuyển yêu cầu này đến lớp Processor.
(2) Processor: nhận yêu cầu từ lớp Parser, nếu yêu cầu là GetDescribeFeatureType thì lớp Processor sẽ thực hiện gọi lớp Reader để đọc các thông tin về FeatureType.
(3) Reader: nhận yêu cầu từ lớp Processor, đọc các thông tin theo yêu cầu của lớp Processor. Sau khi đọc xong gởi lại thông tin cho lớp Processor.
(2) Processor: nhận kết quả từ lớp Reader sau đó thực hiện kết nối đến cơ sở dữ liệu không gian thông qua DataProvider. Sau khi nhận được thông tin về cấu trúc của FeatureType lớp Processor sẽ chuyển thông tin đó qua lớp Wrapper.
(5) Wrapper: nhận thông tin từ lớp Processor sau đó đóng gói theo đặc tả và gởi kết quả về Client.

- *Mô hình xử lý*



Hình 4: Mô hình xử lý GetDescribeFeatureType

4.2.3. GetFeature



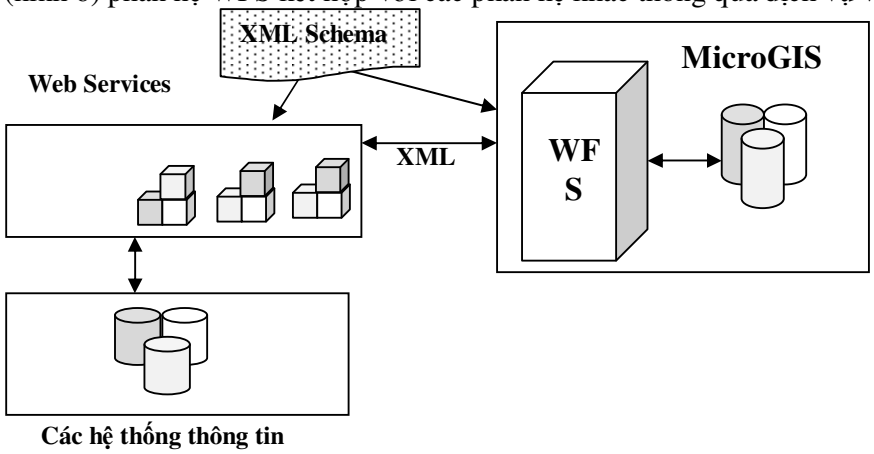
Hình 5: Mô hình xử lý GetFeature

(1) Parser: nhận và phân tích yêu cầu từ Client, nếu yêu cầu được xác định là GetFeature thì lớp Parser chuyển yêu cầu này đến lớp Processor.

(2) Processor: nhận yêu cầu từ lớp Parser, nếu yêu cầu được xác định là GetFeature lớp Processor sẽ thực hiện gọi lớp Reader để kiểm tra thông tin về FeatureType.
(3) Reader: nhận thông tin từ lớp Processor và thực hiện kiểm tra thông tin do lớp Processor yêu cầu. Sau khi kiểm tra lớp Reader sẽ trả thông tin về cho lớp Processor.
(2) Processor: khi nhận được thông tin từ lớp Reader, lớp Processor chuyển yêu cầu từ client sang lớp Filter Parser.
(4) Filter Parser: nhận yêu cầu từ lớp Processor, chuyển các yêu cầu từ Client sang SQL Query và chuyển kết quả về cho lớp Processor.
(2) Processor: tổng hợp kết quả trả về từ 2 lớp Reader và Filter Parser gởi đến lớp Data Provider.
(5) Data Provider: nhận yêu cầu từ lớp Processor và thực hiện lấy thông tin từ cơ sở dữ liệu không gian theo yêu cầu đồng thời trả kết quả về lớp Processor.
(2) Processor: nhận kết quả từ lớp Data Provider và gởi kết quả đó đến lớp Wrapper.
(6) Wrapper: nhận thông tin từ lớp Processor và thực hiện đóng gói dữ liệu theo GML Schema và trả kết quả về Client.

4.3. Mô hình kết hợp

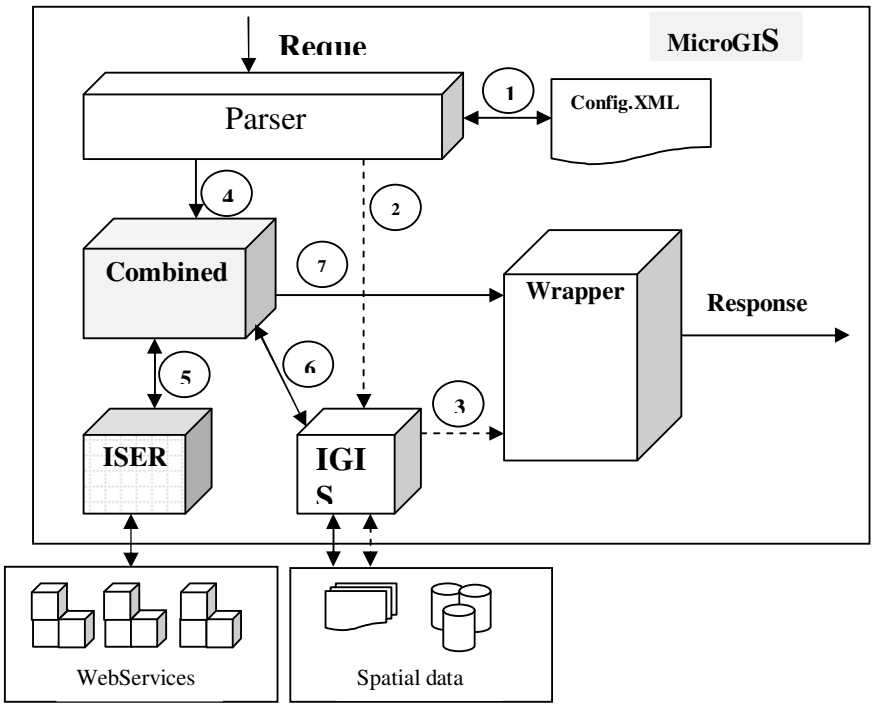
Mô hình (hình 6) phân hệ WFS kết hợp với các phân hệ khác thông qua dịch vụ web.



Hình 6: Mô hình kết hợp

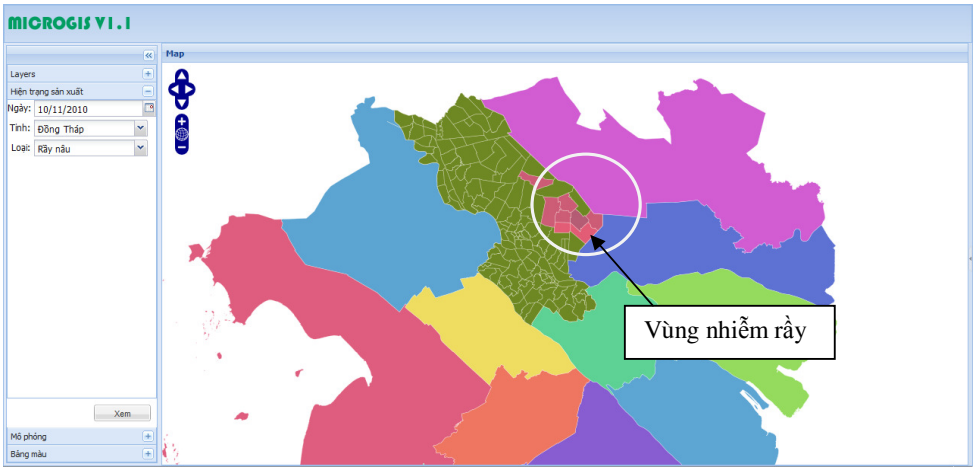
- Chi tiết mô hình kết hợp

Các thành phần trong mô hình kết hợp
Parser: phân tích các yêu cầu để điều hướng việc xử lý.
IGIS: nhận yêu cầu lấy dữ liệu (dữ liệu không gian) từ gói Combined hoặc Parser.
Combined: nhận yêu cầu từ Parser, điều hướng xử lý cho hai gói ISER và IGIS.
ISER: nhận yêu cầu từ gói Combined liên quan đến việc giao tiếp với dịch vụ web.
Wrapper: nhận dữ liệu từ IGIS hoặc Combined đóng gói dữ liệu theo chuẩn OpenGIS.



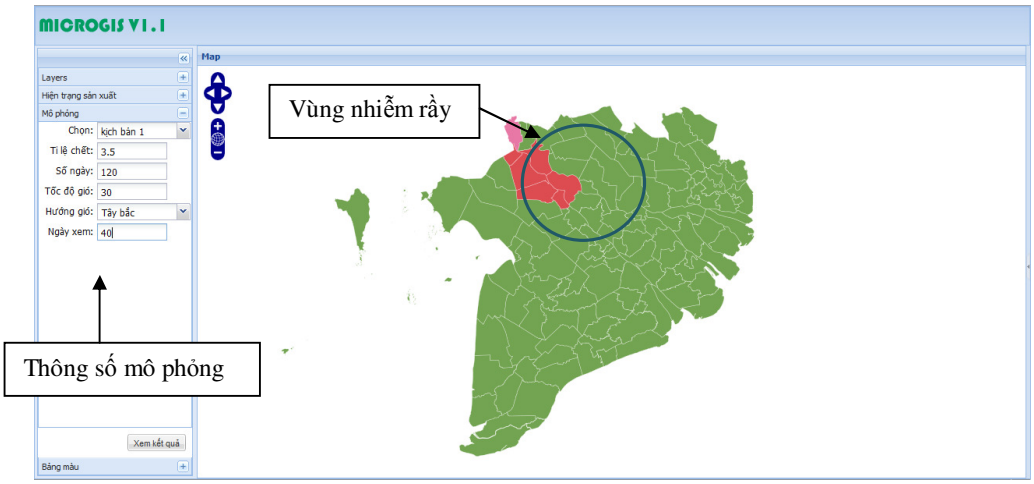
Hình 7: Chi tiết mô hình kết hợp

4.4. Kết quả nghiên cứu



Hình 8: Kết nối với phân hệ Hiện trạng sản xuất

Hình 8 thể hiện dữ liệu về tình hình nhiễm Rầy tại tỉnh Đồng Tháp theo thời gian Rầy (mức màu thể hiện mức độ nhiễm), kết quả trên là quá trình nối kết dữ liệu thông qua dịch vụ đã xây dựng với phân hệ Hiện trạng sản xuất.



Hình 9: Kết nối với phân hệ mô phỏng

Hình 9 thể hiện dữ liệu về vùng nhiễm Rầy (mức màu thể hiện mức độ nhiễm) dựa trên các thông số do người dùng nhập vào, kết quả trên là quá trình giao tiếp giữa hệ thống đã xây dựng với phân hệ Mô phỏng.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Trong nghiên cứu này, chúng tôi đã xây dựng dịch vụ WFS theo chuẩn OpenGIS.
Phân tích, xác định giải pháp bao gồm về kỹ thuật và nội dung cho dịch vụ web đã xây dựng tương tác với các hệ thống thông tin khác.
Xây dựng công cụ trên nền web cho phép tương tác, xem và tìm kiếm thông tin trên bản đồ (MicroGIS Client).
Định nghĩa được chuẩn tương tác cho phép xây dựng các WebServices theo kiến trúc quy định nhằm tăng cương khả năng nối kết với dịch vụ WebGIS.

5.2. Hướng phát triển

MicroGIS chỉ hỗ trợ Basic WFS cần xây dựng thêm Xlink WFS và Transaction WFS.
Đối với tập GML quá lớn làm hiệu xuất thực thi rất chậm.Cần cải thiện tốc độ thực thi cả hai phía client và server. Có thể giải quyết theo đề xuất sau: Loại bỏ các đường biên trùng nhau (TS. Lê Quyết Thắng) và Tách dữ liệu từ cấu trúc và gom cụm dữ liệu theo ngữ nghĩa (Qingting Wei; Jihong Guan).
Xây dựng cơ chế thiết lập hàng chờ đối với trường hợp số lượt truy cập vào hệ thống quá lớn.
Cần cải tiến giải pháp nối kết dữ liệu phi địa lý mềm dẻo hơn.
Cần nghiên cứu cách thức lưu trữ kết quả mô phỏng để người dùng có thể xem kết quả mô phỏng trên bản đồ một cách liên tục theo chuỗi ngày mô phỏng.

Tài liệu tham khảo

Tiếng Việt

1. Đặng Văn Đức (2001), *Hệ thống thông tin địa lý*, NXB Khoa học và Kỹ thuật, Hà Nội.
2. Võ Quang Minh (1998), *Bài giảng môn học Hệ thống thông tin địa lý*, Khoa Nông nghiệp, Đại học Cần Thơ.
3. Bộ Thông tin truyền thông (2008), *Danh mục tiêu chuẩn về ứng dụng công nghệ thông tin trong cơ quan nhà nước*, Hà Nội.

Tiếng Anh

4. Scott Davis (2007), *GIS for Web Developers*, The Pragmatic Programmers LLC.
5. Mark Endrei (2004), *Service-Oriented Architecture and Web Services*, IBM.
6. Elliotte Rusty Harold (2002), *Processing XML with Java: A Guide to SAX, DOM, JDOM, JAXP, and TrAX*, Addison Wesley.
7. John R. Herring (2010), *Simple feature access - Part 1: Common architecture*, Open Geospatial Consortium Inc.
8. Brinkhoff, T., Kriegel, H.-P., and Seeger, B. (1994), *Efficient Multi-step Processing of Spatial Joins Using R-trees*, MN, pp. 237-246.
9. William Lalonde (2002), *Styled Layer Descriptor Implementation Specification*, Open Geospatial Consortium Inc.
10. Sun Microsystems, The Java™ Web Services Tutorial, pp.25-86.
11. Dr. Markus Müller (2006), *Symbology Encoding Implementation Specification*, Open Geospatial Consortium Inc.
12. Clemens Portele (2007), *Geography Markup Language (GML) Encoding Standard 3.2.1*, Open Geospatial Consortium Inc.
13. Keith Ryden (2005), *Simple feature access - Part 2: SQL option*, Open Geospatial Consortium Inc.
14. Ahmet Sayar, Marlon Pierce and Geoffrey Fox, *OGC Compatible Geographical Information Systems Web Services*, Indiana University Computer Science Department.
15. Panagiotis A. Vretanos (2005), *Web Feature Service Implementation Specification 1.1.0*, Open Geospatial Consortium Inc.
16. Panagiotis A. Vretanos (2005), *Filter Encoding Implementation Specification 1.1.0*, Open Geospatial Consortium Inc.
17. Arliss Whiteside, Jim Greenwood (2010), *OGC Web Services Common Standard*, Open Geospatial Consortium Inc.
18. Arliss Whiteside, John D. Evans (2008), *Web Coverage Service (WCS) Implementation Standard*, Open Geospatial Consortium Inc.
19. Albert K.W. Yeung, *Spatial Database Systems*, Springer. pp.93-125.

XÂY DỰNG HỆ THỐNG TRUYỀN DỮ LIÊU VÀ ĐIỀU KHIỂN HIỂN THỊ TỪ XA QUA HỆ THỐNG MẠNG ĐIỆN THOẠI DI ĐỘNG

Lâm Chí Nguyện, Thái Minh Tuấn, Đoàn Hoà Minh
Khoa Công nghệ thông tin và Truyền thông, trường Đại học Cần thơ

ABSTRACT

The model wireless connection (WIFI, GPRS, Internet, GSM, ...) have demonstrated the technological superiority of the control device or remote computer. In particular, the transfer protocol can be used TCP / IP or other protocols of each system. The contents of this paper we present a model for remote control a computer or a device via SMS of telecommunications services. We have applied this model to the data transmission system and remote displays media data.

Key words:Wifi, SMS, GRRS, TCP/IP, Remote control, computer

TÓM TẮT

Các mô hình nối kết không dây (WIFI, GPRS, Internet,GMS,…) đã chứng tỏ được ưu thế trong công nghệ điều khiển thiết bị hoặc máy tính từ xa. Trong đó, giao thức truyền tín hiệu được sử dụng có thể là TCP/IP hoặc các giao thức riêng biệt của từng hệ thống. Trong nội dung bài báo này chúng tôi trình bày một mô hình điều khiển máy tính từ xa thông qua hệ thống tin nhắn SMS của nhà cung cấp dịch vụ viễn thông. Chúng tôi đã áp dụng mô hình này cho hệ thống truyền dữ liệu và hiển thị thông tin từ xa.

1 MỞ ĐẦU:

Trong bối cảnh điện thoại di động thông minh (smart phone), máy tính bảng (tablet computer), như một máy vi tính thu nhỏ, có xu hướng chiếm lĩnh thị trường với giá thành thấp dần và có nhiều ứng dụng ngày càng phong phú. Cùng với việc xuất hiện ngày càng nhiều thiết bị có tính di động và truyền dẫn không dây, thu phát theo công nghệ hồng ngoại, bluetooth, wifi,… đã tạo nên sự đa dạng về các dạng thức kết nối hệ thống mạng.

Để giải quyết nhu cầu dung lượng nhớ của thiết bị di động, các hãng sản xuất thực hiện công nghệ bộ nhớ rời, như thẻ nhớ, và công nghệ này đã phát triển một cách mạnh mẽ. Với khả năng lưu trữ của các thiết bị di động ngày càng cao và băng thông đường truyền không dây tăng đáng kể đã cho phép truyền tải dữ liệu đa phương tiện (âm thanh và hình ảnh) với chất lượng cao.

Từ đó nảy sinh các yêu cầu sau:

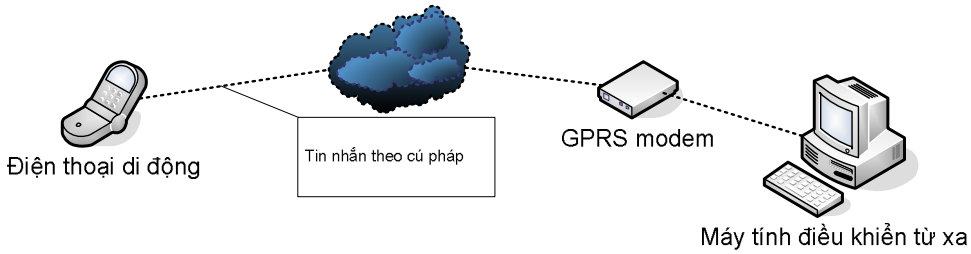
- Nghiên cứu một cách toàn diện các kỹ thuật kết nối mạng không dây và di động cũng như các giải pháp truyền tải dữ liệu đa phương tiện nhằm xây dựng các mô hình kết nối tốt nhất trong các điều kiện cụ thể.
- Nghiên cứu các mô hình truyền tải dữ liệu trên các hệ thống mạng không dây và di động nhằm giảm thiểu lưu lượng truyền tải dữ liệu, nhằm làm tăng tốc độ truy cập cũng như tăng hiệu quả kinh tế trong triển khai sử dụng các thiết bị có tính di động.

Bài viết này trình bày nội dung và kết quả nghiên cứu các mô hình nối kết giữa điện thoại di động và máy tính nhằm điều khiển hiển thị dữ liệu đa phương tiện từ xa. Đây cũng là một vấn đề mang tình thực tế trong công nghệ quảng cáo và giám sát từ xa.

2 NỘI DUNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Mô hình chung

Điều khiển máy tính từ xa bằng điện thoại di động thông qua mạng máy tính và sử dụng các giao thức TCP/IP là một giải pháp được sử dụng phổ biến hiện. Tuy nhiên quá trình kết nối máy tính sẽ phụ thuộc vào hạ tầng mạng giữa máy điều khiển và máy bị điều khiển. Vì vậy, chúng tôi đề xuất một hệ thống cho phép người dùng có thể điều khiển thiết bị hoặc máy tính từ xa thông qua dịch vụ SMS của mạng viễn thông. Hệ thống hoạt động theo mô hình Client-Server, trong đó Client (điện thoại di động) và Server (máy tính + GSM modem) trao đổi thông tin qua các tin nhắn SMS.



Hình 1: Mô hình hệ thống kết nối

Theo mô hình này, các thiết bị tối thiểu cần có:

- Điện thoại di động
- GPRS modem
- Máy tính điều khiển từ xa
- Sim điện thoại của nhà cung cấp dịch vụ viễn thông

Máy tính được điều khiển từ xa thông qua một tập hợp các cú pháp tin nhắn cho trước, có thể gọi là một bộ giao thức được quy đinh cho hệ thống này.

2.2 Giao thức làm việc giữa Client và Server

Để yêu cầu server thực hiện công việc nào đó, client sẽ gởi đến server các lệnh (tin nhắn SMS), đồng thời server khi tiếp nhận tin nhắn và thực hiện yêu cầu tương ứng sẽ có tin nhắn phản hồi cho client. Vì nối kết giữa client và server không duy trì liên tục cho nên các tác vụ yêu cầu giữa hai bên cần đơn giản, ngắn gọn tránh duy trì trạng thái nối kết quá lâu. Chúng tôi đề xuất một số quy tắc tạo ra cú pháp câu lệnh điều khiển bằng SMS như sau :

2.2.1 Cú pháp gởi tin nhắn từ client :

Các tin nhắn điều khiển từ client có thể chứa cùng lúc nhiều câu lệnh. Mỗi câu lệnh trên một dòng. Cú pháp câu lệnh được mô tả như sau:

<Câu lệnh>	01 Khoảng trắng	Tham số cho câu lệnh nếu có

Người có thể soạn tin nhắn trực tiếp bằng điện thoại, hoặc gởi tin nhắn thông quan một giao diện đồ họa trên điện thoại của mình.

2.2.2 Cú pháp tin nhắn từ server:

Tin nhắn trả lời của server có 02 loại: *tin nhắn phản hồi kết quả thực hiện* và *tin nhắn cung cấp thông tin* cho client.

- Tin nhắn phản hồi gồm có 03 trường hợp:

- Yêu cầu được chấp nhận
- Yêu cầu bị từ chối
- Yêu cầu sai cú pháp

Cú pháp câu lệnh được mô tả như sau:

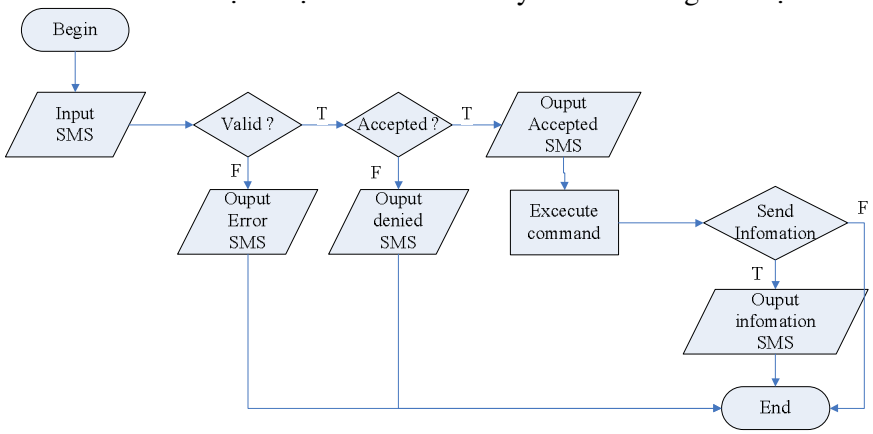
<Phản hồi>	01 Khoảng trắng	Ghi chú

- Tin nhắn cung cấp thông tin,
  Cú pháp câu lệnh được mô tả như sau:

<Loại thông tin>	01 Khoảng trắng	Các thông tin đính kèm, cách nhau bằng dấu |

2.3 Tiến trình thực hiện một yêu cầu :

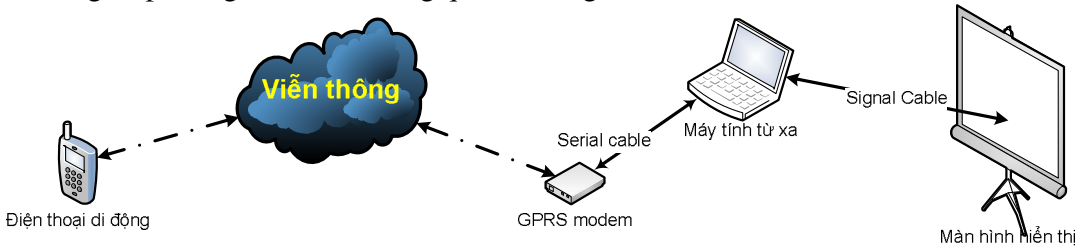
- Phía client : soạn tin nhắn theo cú pháp gởi đến số điện thoại của Server
- Phía Server khi nhận được tin nhắn sẽ xử lý theo lưu đồ giải thuật sau:



Hình 2: Lưu đồ giải thuật xử lý tin nhắn ở phía Server

3 KẾT QUẢ

Từ mô hình được đề xuất ở phần trên, chúng tôi đã thiết kế một mô hình điều khiển hiển thị thông đa phương tiện từ xa thông qua hệ thống SMS.



Hình 3: Mô hình hệ thống điều khiển hiển thị thông tin từ xa

3.1.1 Các chức năng của hệ thống:

Client: (điện thoại di động)

- Xem danh sách các tập tin có thể hiển thị trên server.
- Xem thông tin tập tin đang được hiển thị trên server.
- Yêu cầu server hiển thị một tập tin có trong danh sách.

Server: (máy tính + GSM modem)

- Nhận các yêu cầu từ client.

- Trả lời các yêu cầu của client.
- Thực thi các yêu từ client nếu yêu cầu hợp lệ.

3.1.2 Giao thức làm việc giữa Client và Server:

Client:

Để yêu cầu server thực hiện công việc nào đó, client sẽ gởi đến server các lệnh (tin nhắn SMS) sau:

+ Yêu cầu server gởi danh sách các file có thể thực thi.

LIST

+ Yêu cầu server gởi thông tin file đang thực thi

INFO

+ Yêu cầu server tạm dừng file đang thực thi.

STOP

+ Yêu cầu server thực thi file có tên <filename>, cú pháp câu lệnh được mô tả như sau:

PLAY	01 Khoảng trắng	filename

Server:

Khi nhận được yêu cầu của Client, Server sẽ trả lời bằng các lệnh sau:

+ Yêu cầu bị từ chối, , cú pháp câu lệnh được mô tả như sau:

DE	01 Khoảng trắng	Ghi chú

+ Yêu cầu sai cú pháp, , cú pháp câu lệnh được mô tả như sau:

ER	01 Khoảng trắng	Ghi chú

+ Yêu cầu được chấp thuận, , cú pháp câu lệnh được mô tả như sau:

OK	01 Khoảng trắng	Ghi chú

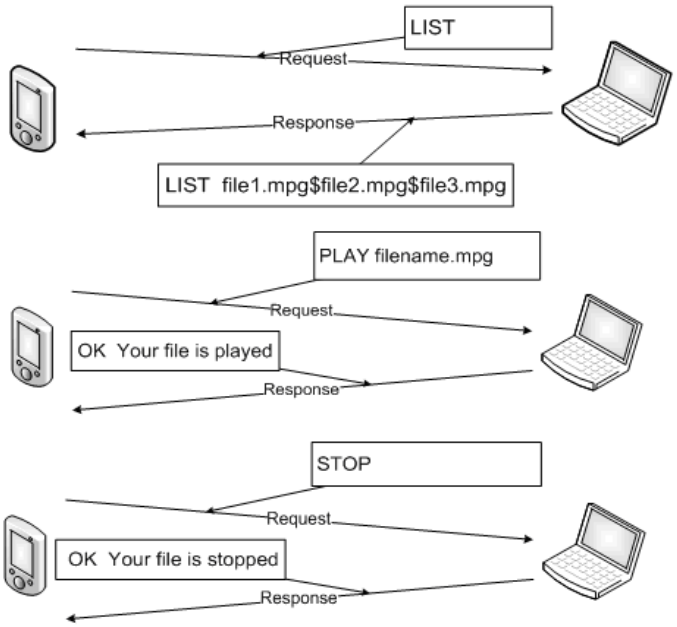
+ Thông tin tập tin đang được thực thi, , cú pháp câu lệnh được mô tả như sau:

FI	01 Ký tự |	Tên tập tin	01 Ký tự |	Dung lượng tập tin

+ Danh sách tập tin đang được server quản lý, , cú pháp câu lệnh được mô tả như sau:

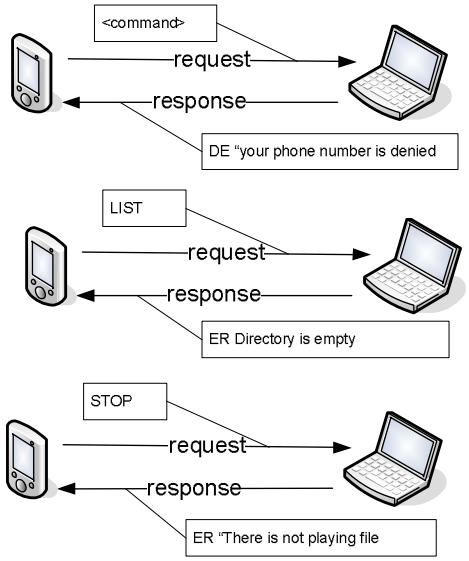
LI	01 Khoảng trắng	01 Ký tự $	Tên tập tin 1	01 Ký tự $	Tên tập tin 2	….

Mô hình một phiên giao dịch thành công



Hình 4: Giao thức truyền phiên giao dịch thành công

Mô hình một phiên giao dịch không thành công.



Hình 5: Giao thức truyền phiên giao dịch không thành công

3.1.3 Phần mềm phía máy tính từ xa

Gởi nhận tin nhắn thông qua GPRS/GSM modem:

GPRS/GSM modem là một thiết bị truyền dẫn không dây có thể làm việc với mạng viễn thông. Nó có chức năng giống như một modem quay số (dial-up modem). GPRS/GSM modem gởi nhận thông tin thông qua sóng vô tuyến của hệ thống viễn thông. GPRS/GSM modem được lắp đặt tại máy tính từ xa như là một thết bị ngoại vi thông qua các cổng giao tiếp: PC Card, PCMCIA Card, COM, USB. Giống như điện thoại di động GPRS/GSM modem cần có thẻ SIM của nhà cung cấp dịch vụ viễn thông để có thể sử dụng các dịch vụ truyền dẫn thông tin trên các tần số của công ty viễn thông.

Trên máy tính, chúng ta sử dụng tập hợp các câu lệnh AT commands mở rộng để điều khiển GPRS/GSM Modem. Với tập lệnh AT mở rộng này ta có thể thực hiện các chức năng:

- Đọc, viết, xoá các tin nhắn SMS.
- Gởi tin nhắn SMS.
- Quan sát trạng thái tín hiệu giao tiếp (signal strength).
- Xem, viết, tìm kiếm danh bạ được lưu trên SIM.

GPRS Modem có thể dùng để truyền nhận tin nhắn SMS với tốc tộ trung bình khoảng 30 tin nhắn trong 01 phút, trong khi đó GSM modem có thể xử lý trung bình 6 đến 10 tin nhắn trong 01 phút. Vì vậy, trong đề tài này chúng tôi sử dụng GPRS modem để truyền nhận tin nhắn điều khiển SMS.

Kết nối với GPRS modem

Thiết bị GPRS modem được gắn kết với máy tính thông qua cổng USB, để quản lý và làm việc với thiết bị này chúng tôi sử dụng gói phần mềm Java communication API của Java cung cấp.

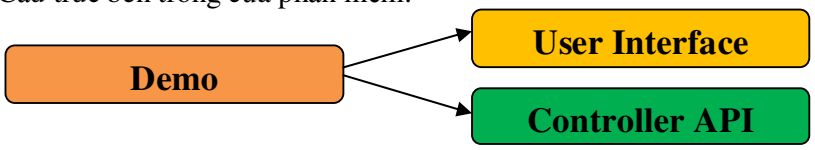
Để thực hiện điều khiển, giao tiếp với modem chúng tôi sử dụng thông qua tập lệnh AT mở rộng. AT commands là một tập lệnh dùng để điều khiển các modem quay số được mở rộng thêm các câu lệnh để có thể điều khiển các GPRS/GSM modem. Chương trình giao tiếp được với GPRS/GSM Modem sẽ điều khiển modem thông qua gởi nhận từng byte dữ liệu tại hai luồng thông tin (InputStream và OutputStream) của nối kết Serial_Connection đến thiết bị.

Kiến trúc gói phần mềm DEMO trên máy tính từ xa

Trong phần mềm chúng tôi sử dụng các gói thư viện hàm của Java có sẵn:

- Java communication API: là gói thư viện phục vụ cho các nối kết với các cổng giao tiếp ngoại vi: PCI, COM, LPT, USB…
- SMSLib: là gói thư viện phần mềm mã nguồn mở được phát triển bằng ngôn ngữ java .
- Các gói thư viện chuẩn của JAVA: là các gói thư viện hàm chuẩn được hỗ trợ bởi J2SE.

Cấu trúc bên trong của phần mềm:

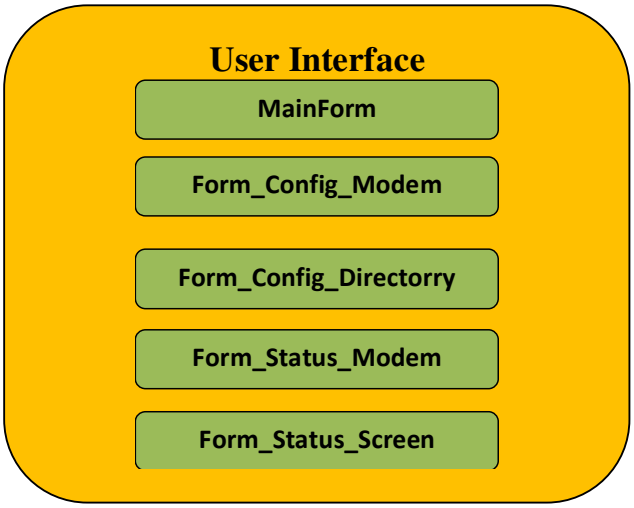


Hình 6: Cấu trúc bên trong của phần mềm

User Interface: Gói giao diện UserInterface bao gồm các lớp phục vụ việc giao tiếp với người dùng, giao tiếp với GPRS modem, giao tiếp với phần mềm hiển thị thông tin. Trong gói UserInterface đó các lớp giao diện lần lượt :

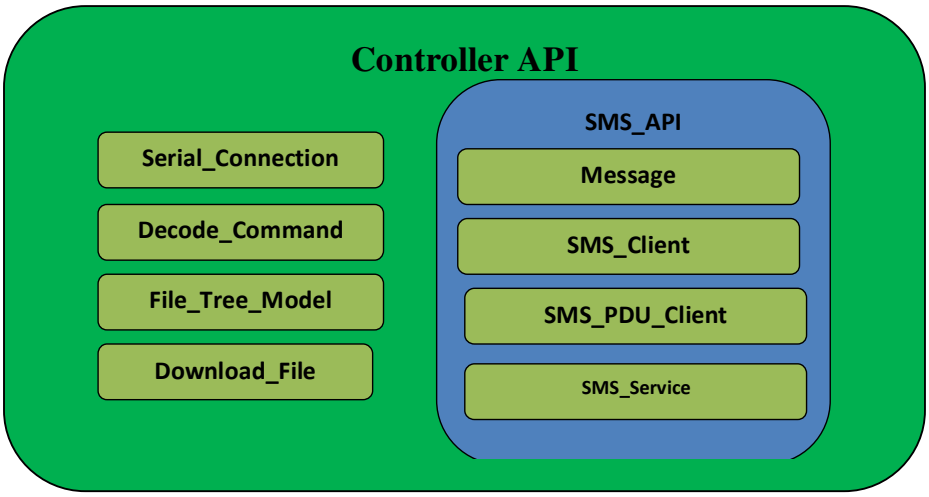
- Mainform: là giao diện chính của chương trình
- Form_Config_Modem: là giao diện cho phép thức hiện cấu hình nối kết đến modem
- Form_Config_Directory: là giao diện cho phép cấu hình các thông tin lưu trữ cục bộ và lưu trữ từ xa.
- Form_Status_Modem: là giao diện hiển thị trạng thái gởi nhận thông tin của GPRS Modem, và các điều khiển cơ bản của Modem.

- Form_Status_Screen: là giao diện hiển thị trang thái của chương trình hiển thị, và các điều khiển cơ bản liên quan đên hiển thị thông tin.



Hình 7: Các class tạo giao diện người dùng

ControllerAPI: là tập hợp các lớp đối tượng phục vụ cho các chức năng gởi nhận tin nhắn SMS. Các lớp được thiết kế trong gói Controller API:



Hình 8: Các class của controller API

- Gói SMS_API:
  - ✓ Message: là lớp đối tượng mô tả về một tin nhắn điều khiển
  - ✓ SMS_Client : là lớp đối tượng gởi nhận thông tin bằng tin nhắn SMS đơn giản.
  - ✓ SMS_PDU_Client : là lớp đối tượng giúp giải mã và mã hóa tin nhắn theo định dạng PDU (protocol description unit).
  - ✓ SMS_Service : là lớp đối tượng thường trực dùng để quản lý GPSR modem, luôn lắng nghe trên modem để tiếp nhận các tin nhắn điều khiển và gởi các tin nhắn trả lời.
- Serial_Connection: Lớp đối tượng quản lý kết nối với GPRS modem

- Download_File: là lớp đối tượng phục vụ quản lý file trên thư mục từ xa
- File_Tree_Model: là lớp đối tượng cho mô hình thư mục các file cục bộ theo dạng cây
- Decode_Command: là lớp đối tượng thực hiện việc ánh xạ câu lệnh nhận được từ tin nhắn thành chức năng thực hiện trêm máy tính và ngược lại gởi tin nhắn kết quả về điện thoại điều khiển.

Phần mềm phía Client / thiết bị di động

Trong mô hình hoạt động của hệ thống, người dung có thẻ tự soạn thảo tin nhắn điều khiển gởi đến máy tính từ xa. Tuy nhie để đơn giản hóa thao tác điều khiển, cũng như tránh sai sót trong quá trình soạn thảo tin nhắn, chúng tôi đã phát triển một phần mềm điều khiển từ xa trên điện thoại di động. Phần ứng dụng trên điện thoại di động đóng vai trò client trong hệ thống phần mềm đã được nêu ở mục trên Chức năng của ứng dụng là cho phép người dùng đầu cuối yêu cầu thực hiện các yêu cầu điều khiển hệ thống thông qua giao diện đồ họa trên điện thoại di động.

Một số lớp (class) của giao diện người dùng:

Lớp Screen: Là lớp cha của các lớp TextBox, Form, List, và Alert. Lớp Screen định nghĩa các thuộc tính và phương thức tổng quát cho các lớp trên. Thông thường ta ít khi điều khiển sắp xếp các thành phần của giao diện người dùng trên màn hình của điện thoại di động. Việc sắp xếp thật sự phụ thuộc vào nhà sản xuất.

Lớp TextBox: Cho phép người dùng nhập và soạn thảo văn bản. Lập trình viên có thể định nghĩa số ký tự tối đa, giới hạn loại dữ liệu nhập (số học, mật khẩu, email,…) và hiệu chỉnh nội dung của textbox.

Lớp Alert: Alert hiển thị một màn hình pop-up trong một khoảng thời gian. Thông thường Alert được dùng để cảnh báo hay báo lỗi.Thời gian hiển thị có thể được thiết lập bởi ứng dụng. Alert có thể được gán các kiểu khác nhau (alarm, confirmation, error, info, warning), các âm thanh tương ứng sẽ được phát ra.

Lớp List: Lớp List là một Screen chứa danh sách các lựa chọn chẳng hạn như các radio button, StringItem,… Người dùng có thể tương tác với list

Lớp Form: Lớp Form là một thành phần hiển thị cho phép chứa nhiều item trên cùng một màn hình (giống như frame, panel trong lập trình J2SE). Các item có thể là DateField, TextField, ImageItem, StringItem, ChoiceGroup, …

Trong ứng dụng chúng ta sẽ tạo lớp MainList thừa kế lớp List, chứa danh sách các tập tin có thể thao tác. Các lớp OptionForm, DownloadForm thừa kế lớp Form để cho phép người dùng nhập các thông tin cần thiết. Sử dụng lớp Alert để hiện các thông báo của ứng dụng cho người dùng.

Hình 11: Một số hình ảnh về phần mềm điều khiển trên di động

Thư viện hàm API nhắn tin (Wireless measaging API- WMA)

Bộ API nhắn tin không dây của J2ME được đặc tả trong JSR 120 (Java Specification Request), cho phép người dùng gởi và nhận các tin nhắn văn bản hoặc nhị phân ngắn trên kết nối không dây. WMA dựa trên khung kết nối mạng tổng quát (GCF), được định nghĩa trong đặc tả CLDC (Connected Limited Device Configuration) của J2ME. Các tin nhắn được gởi và nhận với WMA được gởi trên các mạng không dây của điện thoại di động và các thiết bị tương tự khác, có thể là GPRS/GSM hay CDMA. WMA hỗ trợ Short Message Service (SMS), Multimedia Message Service (MMS) và Cell Broadcast Short Message Service (CBS).

Để gởi hoặc nhận tin nhắn, ứng dụng trước hết phải tạo một instance của giao diện MessageConnection, sử dụng GCF connection factory. Một địa chỉ URL chuyển cho phương thức java.microedition.io.Connector.open() chỉ định giao thức sử dụng (SMS, MMS hoặc CBS), và số điện thoại đích, cổng, hoặc cả hai.

Ví dụ chúng ta có một số URL như sau:

```
sms://+840989570010
sms://+840989570010:1234
sms://:1234
cbs://:1234
```

URL trong hai dạng đầu tiên mở kết nối client, ứng dụng kết nối đến một server với địa chỉ thiết bị và cổng chỉ định. Nếu cổng không chỉ định, sẽ dùng cổng nhắn tin mặc định của ứng dụng. Dạng URL thứ ba mở một kết nối server trên thiết bị, cho phép ứng dụng đợi và hồi đáp tin nhắn đến từ các thiết bị khác. Dạng cuối cùng cho phép ứng dụng lắng nghe tin nhắn broadcast từ người điều hành mạng.

Đoạn code minh họa việc tạo nối kết và gởi một SMS.

```
String address = "sms://" + mReceiver;
MessageConnection       conn    =    (MessageConnection)
                        Connector.open(address);
TextMessage txtmessage = (TextMessage)conn
      .newMessage(MessageConnection.TEXT_MESSAGE);
txtmessage.setAddress(address);
txtmessage.setPayloadText(msgString);
conn.send(txtmessage);
```

Như đã miêu tả ở phần 2.3.1, giao tiếp giữa phần ứng dụng client chạy trên điện thoại di động và phần server được thực hiện qua các tin nhắn SMS. Để dễ dàng quản lý chương trình, chúng ta sẽ tạo ra 3 Thread: một dùng để gởi các tin nhắn tới server, một dùng để

nhận các tin nhắn từ server gởi về và thread chính dùng để xử lý các sự kiện thực hiện các yêu cầu của người dùng. Trong ứng dụng chúng ta sẽ tạo lớp **Sender** dùng để sinh ra Thread gởi tin nhắn và công việc nhận tin nhắn sẽ được thực thi bởi 1 Thread sinh ra bởi lớp **Main**.

4 KẾT LUẬN VÀ ĐỀ NGHỊ

Qua quá trình thực hiện đề tài chúng tôi đã kiểm chứng tính khả thi của mô hình điều khiển máy tính từ xa.

Xây dựng thành công phần mềm ứng dụng hỗ trợ việc truyền, nhận và hiển thị thông tin đa phương tiện giữa các thiết bị có tính di động với nhau và giữa các thiết bị có tính di động với thiết bị hoặc màn hình cố định theo mô hình kết nối mạng không dây và di động đã chọn. Phần mềm này có các tính năng sau: truyền lệnh và dữ liệu đa phương tiện giữa điện thoại di động với PC; quản lý, duyệt hệ thống tập tin trên PC, điều khiển mở tập tin và hiển thị từ xa bằng điện thoại di động.

Căn cứ vào các kết quả đạt được chúng tôi đề nghị mô hình điều khiển máy tính từ xa có thể áp dụng trong thực tế như trong nhiều lĩnh vực, sau đây là mộ số ứng dụng tiêu biểu:

- ✓ Trong thông tin tuyên truyền và quảng cáo: ghép các màn hình nhỏ thành một màn hình lớn theo yêu cầu, theo dõi và thay đổi thông tin được hiển thị trên các màn hình lớn từ xa bằng thiết bị di động (điện thoại di động).
- ✓ Điều khiển giờ từ xa cho các "đồng hồ tháp" bằng thiết bị di động.
- ✓ Trong giám sát và cảnh báo: giám sát tình hình ở một văn phòng, một phân xưởng, một công trình từ xa bằng điện thoại di động; theo dõi từ xa dữ liệu thu thập từ cảm biến; báo động, báo trộm;…

Trong đào tạo:

- ✓ Xây dựng mô hình thí nghiệm và các bài thực hành cho về mạng không dây và di động, về lập trình cho thiết bị di động cho sinh viên các ngành Tin học và Điện tử -Viễn thông.
- ✓ Làm cơ sở cho nhiều đề tài luận văn tốt nghiệp cho sinh viên các ngành Tin học và Điện tử -Viễn thông.
- ✓ Làm cơ sở cho đề tài nghiên cứu cấp trường do sinh viên thực hiện, đề tài **"Xây dựng mô hình Đăng ký môn học, tra cứu thông tin học tập bằng tin nhắn SMS".**

TÀI LIỆU THAM KHẢO

[1]. Frank H.P. Fitzek & Frank Reichert, **Mobile Phone Programming and its Application to Wireless Networking**, Springer Science, 2007, ISBN 978-1-4020-5969-8 (e-book).
[2]. Ngô Bá Hùng, Nguyễn Công Huy, **Lập trình truyền thông**, NXB giao thông vận tải, 2008.
[3]. Đoàn Hòa Minh, Báo cáo tổng kết đề tài NCKH cấp cơ sở **"NGHIÊN CỨU XÂY DỰNG HỆ THỐNG PHẦN MỀM TRUYỀN DỮ LIỆU VÀ HIỂN THỊ TỪ XA TRÊN HỆ THỐNG MẠNG KHÔNG DÂY VÀ DI ĐỘNG", mã số T2010-31, 2010.**
[4] http://vietnamnet.vn/cntt/201010/Mot-may-tinh-nhieu-man-hinh-939157/
[5] http://vietbao.vn/Vi-tinh-Vien-thong/TV-plasma-150-inch-chieu-hinh-anh-to-bang-nguoi-that/11039423/217/
[6] http://www.cinemassivedisplays.com/
[7] http://en.wikipedia.org/wiki/Multi-monitor

XỬ LÝ DỮ LIỆU KHÔNG CÂN BẰNG: TIẾP CẬN LẤY MẪU GIẢM KÍCH THƯỚC DỮ LIỆU

Phạm Xuân Hiền[1], *Bùi Minh Quân*[2], *Huỳnh Xuân Hiệp*[3]

Tóm tắt

Trong bài viết này, chúng tôi trình bày hai giải thuật lấy mẫu đơn giản(EasyEnsemble) và lấy mẫu cân bằng(BalanceCascade) để phân loại dữ liệu không cân bằng. Lấy mẫu đơn giản là lấy mẫu từ một số tập con của lớp lớn kết hợp với tất cả các mẫu của lớp nhỏ và sử dụng giải thuật AdaBoost để huấn luyện phân lớp. Giải thuật lấy mẫu cân bằng là huấn luyện học một các tuần tự. Nghĩa là lấy mẫu giống như giải thuật lấy mẫu đơn giản. Tuy nhiên, ở mỗi bước lấy mẫu,các mẫu của lớp lớn đã được phân loại đúng sẽ được loại bỏ ở các bước sau. Cả hai phương pháp lấy mẫu thực hiện đơn giản và có mức cân bằng giảm trong khi xây dựng mô hình. Thực nghiệm đánh giá trên nhiều tập dữ liệu có mức cân bằng khác nhau. Kết quả thực nghiệm trên hai phương pháp cho kết quả cải thiện đối với lớp nhỏ. Khi tập dữ liệu không cân bằng ở mức cao, kích thước dữ liệu của lớp nhỏ rất ít thì phương pháp lấy mẫu cân bằng hiệu quả hơn phương pháp lấy mẫu đơn giản.

Từ khóa: lấy mẫu rút gọn, phương pháp lấy mẫu đơn giản, phương pháp lấy mẫu không cân bằng, Giải thuật Adaboost.

Title: Approach Under-Sampling for Imbalanced Data

Abstract

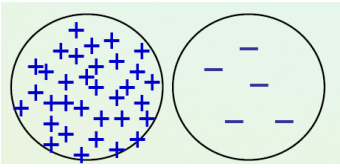
In this paper, we present two algorithms, called EasyEnsemble and BalanceCascade to classify imbalanced data. EasyEnsemble samples several subsets from the majority class, add to minority class and combines AdaBoost algorithm to train the classifier. BalanceCascade trains the learners sequentially. Its sampling likes EasyEnsemble. Howerver, each step the majority class examples which are correctly classified by the current trained learners are removed from further consideration. Both sampling methods are simple to implement and balance while reducing modeling. Experimental evaluation of multiple data sets have different balance. Experimental results show that both methods improved predictive accuracy of positive minority. When a dataset is highly imbalanced and the size of the minority class is very small, BalanceCascade is more effective than EasyEnsemble.

Keywords: under-sampling, EasyEnsemble, BalanceCascade, Adaboost.

[1] Bộ môn Khoa Học Máy Tính, khoa CNTT&TT, Trường ĐH Cần Thơ, pxhien@cit.ctu.edu.vn
[2] Tổ Kỹ Thuật Máy Tính, CNTT&TT, Trường ĐH Cần Thơ, bmquan@cit.ctu.edu.vn
[3] Bộ môn Công Nghệ Phần Mềm, khoa CNTT&TT, Trường ĐH Cần Thơ, hxhiep@cit.ctu.edu.vn

1. GIỚI THIỆU

Cùng với sự hỗ trợ của máy tính, khai thác dữ liệu có thể cho ta giải pháp thích hợp để ra quyết định. Trong khai thác dữ liệu, có hàng loạt tiếp cận máy học hỗ trợ để ra quyết định: C4.5[Quinlan, 1993], mạng Bayes[Fridman & Goldszmidt, 1996], máy học vector hỗ trợ[Vapnil, 1995],… Các tiếp cận này được ứng dụng thành công trong hầu hết các lĩnh vực về phân tích dữ liệu, phân loại văn bản, phân loại gien, phân loại bệnh,…Tuy nhiên, các ứng dụng thực tế dữ liệu thu được là các tập dữ liệu không cân bằng. Tập dữ liệu không cân bằng là tập có sự phân phối không đồng đều nhau. Tập dữ liệu có chứa nhiều thể hiện (instance) ở một số lớp, và có ít thể hiện ở các lớp còn lại.



Hình 1: Dữ liệu không cân bằng

Dữ liệu không cân bằng có các lớp xấp xỉ không bằng nhau, tỉ lệ dữ liệu không cân bằng giữa các lớp có thể lên đến 1/100000 [Chawla et al., 2002]. Vấn đề không cân bằng dữ liệu ảnh hưởng rất lớn đến việc phân loại dữ liệu, nhưng các giải thuật học cơ bản không quan tâm đến vấn đề này, kết quả phân loại thường hướng về lớp lớn. Nghĩa là các tiếp cận học truyền thống thường phân loại sai với lớp nhỏ, nhưng trong thực tế lớp dữ liệu cần quan tâm thường có rất ít thể hiện. Ngoài ra trong khai phá dữ liệu, việc đánh giá thuật toán học theo cách truyền thống dựa vào độ chính xác phân loại (predictive accuracy). Độ chính xác phân loại là tỉ lệ các mẫu trong tập kiểm tra được phân loại đúng trên mô hình nào đó. Độ chính xác phân loại đánh giá trên toàn tập dữ liệu, thường dự báo không tốt đối với lớp nhỏ. Các giải thuật này có xu hướng gán nhãn cho mẫu chưa biết nhãn thuộc lớp có số phần tử lớn và bỏ qua lớp nhỏ. Vì các giải thuật phân loại dữ liệu này thường dựa vào nguyên tắc số đông để gán nhãn cho các mẫu chưa xác định. Lớp nhỏ có ít thể hiện, cho nên khó gán nhãn. Điều này không thích hợp với tập dữ liệu không cân bằng mà chúng ta quan tâm đến lớp nhỏ. Giá phải trả cho phân loại sai ở lớp nhỏ là rất cao. Ví dụ trong tập dữ liệu không cân bằng ở mức 99, giải thuật học cơ bản phân loại chính xác đến 99%, nghĩa là hầu như các mẫu mới sẽ được phân loại đúng ở lớp đa số, và tỉ lệ lỗi chỉ 1%, có thể các mẫu mới ở lớp nhỏ luôn bị phân loại sai, làm mất tính dự đoán của lớp nhỏ[Liu et al., 2008]. Việc phân loại sai ở các lớp nhỏ thì tốn nhiều chi phí như phân loại phát hiện các bệnh hiểm, bệnh ung thư, phát hiện gian lận thẻ tín dụng, spam mail, …Vì vậy, ta cần quan tâm đến việc xây dựng mô hình trên tập dữ liệu không cân bằng sao cho phân loại tốt lớp có ít thể hiện. Mô hình phân loại được xây dựng nhằm cân bằng lại tập dữ liệu, loại bỏ các thể hiện xem là nhiễu, bất thường. Trong quá trình phân loại làm giảm bớt ảnh hưởng của lớp không cân bằng, giảm chi phí. Kết quả đạt được là phân loại tốt lớp nhỏ, lớp được quan tâm. Điều này cũng được thực hiện bởi các tác giả [Liu et al., 2008] với phương pháp lấy mẫu đơn giản(EasyEnsemble) và lấy mẫu cân bằng(BalanceCascade) được viết bằng Matlab.

Phần tiếp theo của bài viết này được trình bày như sau: phần 2 trình bày ngắn gọn về cách lấy mẫu, giải thuật Adaboost, giải thuật lấy mẫu đơn giản, giải thuật lấy mẫu cân bằng. Phần 3 trình bày các kết quả thực nghiệm. Phần cuối trình bày kết luận và hướng phát triển.

2. CÁCH LẤY MẪU

Lấy mẫu giảm[Kubat & Matwin, 1997] là lấy mẫu ở lớp lớn, số mẫu được lấy ra có thể bằng với số mẫu của lớp nhỏ hay theo tỉ lệ 1:2, 1:4,… Phương pháp lấy mẫu làm giảm mức không cân bằng giữa lớp lớn và lớp nhỏ. Tập dữ liệu huấn luyện để xây dựng mô hình giảm mức không cân bằng, phân loại tốt hơn các phần tử thuộc lớp nhỏ do phân loại dựa trên nguyên tắc số đông bị phá vỡ. Ngoài ra, việc lấy mẫu này còn loại bỏ một số mẫu thuộc lớp lớn nên quá trình xây dựng mô hình phân loại có thời gian ngắn.

Một số tiếp cận lấy mẫu giảm[Chawla et al., 2002] như: lấy mẫu ngẫu nhiên trên lớp lớn cho đến khi số mẫu được lấy bằng với số mẫu thuộc lớp nhỏ. Kết hợp số mẫu này với số mẫu lớp nhỏ xây dựng mô hình phân loại. Tiếp cận khác là sử dụng kết hợp lấy mẫu giảm và lấy mẫu tăng thêm. Ví dụ: phân loại ảnh vết dầu loang từ tập dữ liệu không cân bằng với phân bố 42(2%) ảnh có vết dầu loang và 2.471(98%) ảnh giống như vậy nhưng không có vết dầu loang, 100 mẫu được lấy từ phần ảnh có vết dầu loang, và lấy 100 mẫu ngẫu nhiên từ tập còn lại, sau đó học trên tập cân bằng này với giải thuật cây quyết định. Kết quả tỉ lệ lỗi ở lớp nhỏ 14% và 4% đối với lớp lớn.

Lấy mẫu giảm là phương pháp phổ biến trong tiếp cận dữ liệu không cân bằng. Nhiều thực nghiệm nghiên cứu trước đây chỉ lấy tập con trong tập dữ liệu lớn, một số mẫu trong tập dữ liệu lớn bị bỏ qua trong quá trình huấn luyện, dữ liệu trở nên cân bằng hơn và thời gian huấn luyện tương đối nhanh. Tuy nhiên, việc lấy mẫu giảm sẽ làm mất dữ liệu có chứa thông tin hữu ích để phân loại tốt cho tập dữ liệu không cân bằng. Vì thế, nhóm tác giả đã đưa ra đề xuất nhằm khai thác hiệu quả tập dữ liệu cho kết quả phân loại tốt và thời gian thực hiện giải thuật tương đương với việc lấy mẫu giảm là phương pháp lấy mẫu đơn giản và phương pháp lấy mẫu cân bằng[Liu et al., 2008].

Trong quá trình xây dựng giải thuật, phương pháp lấy mẫu đã đề cập ở trên có sử dụng phương pháp tập hợp phân loại: AdaBoost

Giải thuật Adaboost (Adaptive Boosting) [Freund & Schapire, 1995] sinh ra tập phân lớp và bầu chọn chúng một cách tuần tự. Giải thuật Adaboost duy trì tập trọng số trên tập dữ liệu huấn luyện S và điều chỉnh trọng số sau mỗi lần học phân lớp trên giải thuật học cơ bản. Có hai cách mà Adaboost sử dụng trọng số để xây dựng tập huấn luyện mới S' đối với giải thuật học cơ bản. Boosting bằng cách lấy mẫu, các ví dụ được lấy ra từ tập S với xác suất tương ứng với trọng số. Cách thứ hai là Boosting bằng trọng số, các giải thuật học cơ bản trực tiếp đảm nhận trọng số của tập huấn luyện. Toàn bộ tập huấn luyện S được gán trọng số. Mục đích là nhằm cực tiểu lỗi trên các phân phối đầu vào khác nhau để xây dựng mô hình phân lớp có độ chính xác cao.

Phương pháp lấy mẫu giảm được đề cập sử dụng Adaboost cách thứ 2. Cho T lần thử, tập huấn luyện được gán trọng số S_1, S_2, …, S_T được sinh ra tuần tự và xây dựng T phân lớp C_1, C_2, …, C_T, sử dụng trọng số bầu chọn để phân lớp C* cuối cùng. Trọng số của mỗi phân lớp phụ thuộc vào lần thực hiện trên tập huấn luyện dùng xây dựng nó. Lưu ý, nếu tỉ lệ lỗi của mô hình học không tốt hơn đoán ngẫu nhiên thì AdaBoost không làm việc.

Giải thuật AdaBoost cho phân lớp nhị phân:

Đầu vào:

1. Tập dữ liệu: D= {(x_1,y_1), (x_2,y_2),… (x_m,y_m)},
2. Giải thuật học cơ sở *L* ,
3. T số lần học

Xử lý:

- $D_1(i)=1/m$ % khởi tạo phân phối trọng số
- *For t=1 to T:*

- $h_t = L(\mathrm{D}, Dt)$; % huấn luyện lớp yếu phân phối *Dt*
- $\varepsilon_t = \Pr_{x \sim D_t, y} I[h_t(x) \neq y]$; % độ đo lỗi của h_t
- Nếu $\varepsilon_t > 0.5$ then break
- $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$ % xác định trọng số của h_t
- $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \left\{ \begin{array}{l} \exp(-\alpha_t)\ if\ h_t(x_i) = y_i \\ \exp(\alpha_t)\ if\ h_t(x_i) \neq y_i \end{array} \right\}$ % cập nhật phân phối

  $= \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$ % với Z_t là hệ số chuẩn hóa

- end

Đầu ra: $H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

Bảng 1: Ý nghĩa các ký hiệu trong giải thuật Adaboost

Ký hiệu	**Ý nghĩa**
X	Không gian các thể hiện, X={(x_1,y_1), (x_2,y_2),… (x_m,y_m)}
Y	Tập các nhãn, Y= {-1,+1}
T	Số lần lặp trong giải thuật
D	Tập dữ liệu học, D=(x_i,y_i) (i=1, 2,… m)
D_t	Phân phối mà tập học tuân theo
L	Giải thuật học cơ sở
h	Mô hình sau khi huấn luyện tập D bằng giải thuật *L*
I[.]	Hàm chỉ định, có giá trị là 1 nếu biểu thức trong đó là đúng, ngược lại nhận giá trị 0.

Ý tưởng chính của giải thuật là duy trì phân phối tập trọng số trên tập huấn luyện[Schapire, 1999]. Dt(i) là trọng số của mẫu huấn luyện i ở lần lặp t. Giá trị khởi tạo của mỗi mẫu bằng nhau, nhưng ở mỗi lần lặp trọng số sẽ tăng cho lớp yếu phân lớp sai, nhằm lưu ý đến các mẫu khó nhận dạng trong tập huấn luyện. Việc học lớp yếu là tìm giả thuyết ht: X→{-1,+1}. Giả thuyết của lớp yếu tốt được đo bởi lỗi:

$$\varepsilon_t = \Pr_{x \sim D_t, y} I[h_t(x) \neq y] = \sum_{i:h_t(x_i) \neq y_i} D_t(i)$$

Giả thuyết H cuối cùng là bầu chọn theo số đông trọng số. Phân lớp nhị phân, tỉ lệ lỗi dự đoán phải nhỏ hơn dự đoán ngẫu nhiên là 0.5, nghĩa là xác suất đoán đúng phải từ 0.5 trở lên.

Adaboost dễ xây dựng, giải thuật không có tham số (ngoại trừ vòng lặp T), nó không đòi hỏi phải biết trước giới hạn lỗi vì thế nó có thể kết kợp với bất kỳ phương thức để học lớp yếu. Giải thuật cho kết quả chính xác hơn, cao hơn dự đoán ngẫu nhiên ở lớp yếu thay vì giải thuật học được thiết kế có độ chính xác trên toàn tập.

Giải thuật lấy mẫu đơn giản: Từ tập dữ liệu huấn luyện tạo ra T tập con lấy từ lớp chứa nhiều mẫu (lớp lớn, lớp negative, lớp âm). Chọn ngẫu nhiên một tập con kết hợp với các mẫu thuộc lớp có ít mẫu mà ta cần quan tâm (lớp nhỏ, lớp positive, lớp dương),

huấn luyện với phương pháp tập hợp có số lần lặp là s_i. Thực hiện các bước trên T lần và tổng hợp các mô hình cho ra kết quả phân loại.

Giải thuật lấy mẫu đơn giản được trình bày như sau:

Đầu vào:

1. Tập huấn luyện với lớp nhỏ (lớp positive, lớp dương) có P mẫu, lớp lớn (lớp negative, lớp âm) có N mẫu (|P|<|N|)
2. Số lượng tập con T lấy mẫu từ lớp lớn N
3. Số lần lặp s_i để huấn luyện tập H_i trong giải thuật AdaBoost.

Xử lý:

1. i = 0
2. repeat
3. i = i + 1
4. Lấy mẫu ngẫu nhiên tập con N_i từ tập N, ($|N_i|=|P|$)
5. Sử dụng P và N_i để học mô hình H_i, H_i là giả thuyết học từ giải thuật AdaBoost với số lần lặp s_i, phân lớp yếu $h_{i,j}$ và trọng số tương đương $\alpha_{i,j}$, ngưỡng của phương pháp phân loại tập hợp là θ_i

$$H_i(x) = \text{sgn}\left(\sum_{j=1}^{s_i} \alpha_{i,j} h_{i,j}(x) - \theta_i\right)$$

6. Until i = T

Đầu ra:

$$H(x) = \text{sgn}\left(\sum_{i=1}^{T} \sum_{j=1}^{s_i} \alpha_{i,j} h_{i,j}(x) - \sum_{i=1}^{T} \theta_i\right)$$

Phương pháp này sinh ra T tập con. Đầu ra của tập con thứ i là phân lớp H_i với phân lớp yếu $h_{i,j}$, h_{ij} trích xuất từ các đặc trưng của phương pháp học tập hợp, α_{ij} là trọng số của mô hình, θ_i là ngưỡng của mô hình tập hợp. N_i chứa những thông tin khác nhau từ tập huấn luyện của lớp lớn N. Cuối cùng tổng hợp $h_{i,j}$ và tạo thành phân lớp tập hợp. Đầu ra của giải thuật lấy mẫu đơn giản là tập hợp các mô hình đơn giản, trông giống như "tập hợp của tập hợp". Kết hợp Boosting với chiến lược giống như bagging đối với phân phối cân bằng. Nhiều nghiên cứu đã chứng minh rằng Boosting giảm bias trong khi bagging giảm variance. Phương pháp này sử dụng mẫu phân loại để nâng tập Boosting và lấy mẫu không dựa trên xác suất. Các mẫu thuộc lớp nhỏ đều được đưa vào mô hình huấn luyện. Vì thế, khi tập có sự không cân bằng cao hay số mẫu của lớp nhỏ cực hiếm thì những mẫu này vẫn được đưa vào tập huấn luyện để xây dựng mô hình.

Giải thuật lấy mẫu cân bằng: Cấu trúc của phương pháp lấy mẫu cân bằng gần giống như phương pháp lấy mẫu đơn giản. Sự khác biệt là phương pháp lấy mẫu này có sự điều chỉnh ngưỡng sao cho đạt đến tỉ lệ lỗi và loại bỏ các mẫu phân loại đúng thuộc lớp lớn trong mỗi lần thực hiện. Tập dữ liệu huấn luyện ở bước sau sẽ chứa tất cả các mẫu thuộc lớp nhỏ và các mẫu trong tập lớn lớn đã loại bỏ các mẫu được phân loại đúng của mô hình trước. Cuối cùng, phân loại mẫu mới thuộc lớp nhỏ khi chúng thuộc lớp nhỏ trong tất cả mô hình H_i. Giải thuật lấy mẫu cân bằng được trình bày như sau:

Đầu vào:

1. Tập huấn luyện với lớp nhỏ (lớp positive) có P mẫu, lớp lớn (lớp negative) có N mẫu (|P|<|N|)
2. Số lượng tập con T lấy mẫu từ lớp lớn N
3. Số lần lặp s_i để huấn luyện tập H_i trong giải thuật AdaBoost.

Xử lý:

1. i = 0
2. $f = \sqrt[T-1]{\frac{|P|}{|N|}}$, f là tỉ của các mẫu thuộc lớp lớn phân vào lớp nhỏ (fpr)
3. repeat
4. i = i + 1
5. Lấy mẫu ngẫu nhiên tập con N_i từ tập N, ($|N_i|=|P|$)
6. Sử dụng P và N_i để học, xây dựng được mô hình H_i, H_i là giả thuyết học từ giải thuật AdaBoost với số lần lặp s_i, phân lớp yếu $h_{i,j}$ và trọng số của mô hình là $\alpha_{i,j}$, ngưỡng của phương pháp phân loại tập hợp là θ_i
7. $$H_i(x) = \text{sgn}\left(\sum_{j=1}^{s_i} \alpha_{i,j} h_{i,j}(x) - \theta_i\right)$$
8. Điều chỉnh θ_i để $H_i(x)$ đạt tỉ lệ lỗi là f
9. Loại bỏ các mẫu phân lớp đúng của lớp lớn.
10. Until i = T

Đầu ra:

$$H(x) = \text{sgn}\left(\sum_{i=1}^{T} \sum_{j=1}^{s_i} \alpha_{i,j} h_{i,j}(x) - \sum_{i=1}^{T} \theta_i\right)$$

Khi xây dựng mô hình lấy mẫu cân bằng đầu vào là tập dữ liệu không cân bằng chứa |P| mẫu thuộc lớp nhỏ, |N| mẫu thuộc lớp lớn. Từ lớp lớn tạo ra tập con N_i có số phần tử bằng với số phần tử của lớp nhỏ. Kết hợp tập con N_i với các mẫu thuộc lớp nhỏ để xây dựng mô hình phân loại tập hợp và thực hiện T lần lấy mẫu từ lớp lớn. Sau đó thực hiện điều chỉnh ngưỡng sao cho phù hợp với mức không cân bằng và loại bỏ các mẫu phân loại đúng của lớp lớn dựa trên ngưỡng. Cuối cùng phân loại dữ liệu dựa trên tập giá trị ngưỡng của các mô hình.

3. KẾT QUẢ THỰC NGHIỆM

Chúng tôi đã cài đặt thực nghiệm bằng ngôn ngữ Java để xây dựng mô hình phân loại trên tập dữ liệu không cân bằng dựa vào tiếp cận lấy mẫu rút gọn với phương pháp lấy mẫu đơn giản, lấy mẫu cân bằng. Các chức năng được cài đặt để hỗ trợ xây dựng mô hình như sau:

- EasyEnsemble: chức năng lấy mẫu đơn giản
- BalanceCascade: chức năng lấy mẫu cân bằng
- Adaboost: chức năng xây dựng tập mô hình
- AdjustFPRate: chức năng điều chỉnh ngưỡng
- Predict: chức năng dự đoán mẫu

- BoostData: chức năng sinh mẫu

Trong quá trình cài đặt để thực nghiệm chúng tôi đã sử dụng một số lớp có sẵn trong thư viện weka: Instances, Classifier. Chúng tôi sử dụng giải thuật học cơ sở J48 (C4.5 trong weka là J48) với 40 cây được xây dựng. Chúng tôi cũng chú ý đến nghi thức kiểm tra dùng tập học và tập thử theo tỉ lệ 50:50. Tập dữ liệu không cân bằng nhị phân được chia làm 2 phần bằng nhau. Một tập học dùng để xây dựng mô hình, tập kiểm thử dùng để phân loại dựa trên mô hình đã xây dựng.

Thực nghiệm sử dụng tập dữ liệu đầu vào là các tập dữ liệu không cân bằng thu thập được từ [Blake et al.] có dạng .ARFF.

Chúng tôi tiến hành thực nghiệm trên 8 tập dữ liệu được chọn để thực nghiệm. Các tập dữ liệu này có tỉ lệ không cân bằng khác nhau. Các tập dữ liệu được trình bày theo trình tự mức không cân bằng tăng.

Tên tập dữ liệu	Mức không cân bằng	Số thể hiện		Ý nghĩa
		Lớp dương	Lớp âm	
Haberman	2.8	81	225	Kiểm tra sự sống còn của bệnh nhân ung thư vú
Cmc	3.4	333	1140	Điều tra sử dụng loại tránh thai nào?
Pendigits	9.4	1055	9937	Nhận dạng chữ viết tay
Balance	11.8	49	576	Điều tra tâm lý
Car	25.6	65	1663	Đánh giá xe
Letter	26.2	734	19266	Nhận dạng phân loại ký tự
Ann	42.4	166	7034	Bệnh tuyến giáp
Page-blocks	194	28	5445	Phân loại khối văn bản

Bảng 2: Mô tả các tập dữ liệu

Thực nghiệm chia các tập dữ liệu thành 2 kịch bản để phân tích đánh giá tiếp cận mới dựa trên mức không cân bằng và số mẫu của lớp nhỏ trong tập dữ liệu. Lựa chọn phân tích 2 kịch bản này nhằm kiểm chứng ảnh hưởng của phương pháp lấy mẫu đơn giản, lấy mẫu cân bằng đối với tập dữ liệu có mức không cân bằng khác nhau từ thấp đến cao (mức không cân bằng tối đa để thực nghiệm là 194) và số phần tử của lớp quan tâm rất nhỏ (khoảng 14 mẫu huấn luyện). Đồng thời kiểm tra khả năng đáp ứng của tiếp cận mới đối với tập dữ liệu có mức không cân bằng cao và có số mẫu của lớp quan tâm rất nhỏ.

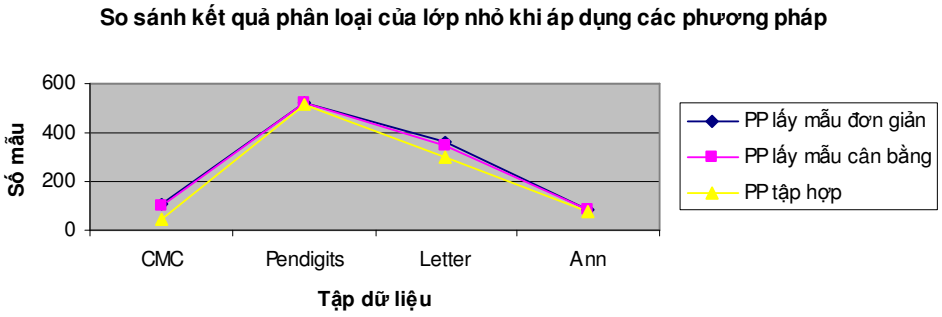
- Kịch bản 1: thực nghiệm trên các tập dữ liệu có mức không cân bằng khác nhau. Thực nghiệm này nhằm kiểm tra khả năng ảnh hưởng của tiếp cận đối với tập dữ liệu có mức không cân bằng cao.
- Kịch bản 2: thực nghiệm trên các tập dữ liệu không cân bằng có số mẫu ở lớp nhỏ rất ít. Thực nghiệm này nhằm kiểm tra khả năng đáp ứng của tiếp cận đối tập không cân bằng cao và số mẫu lớp nhỏ rất ít.

Các bước thực hiện:

1. Chọn các tập có mức không cân bằng khác nhau.

2. Mỗi tập dữ liệu gốc chia thành 2 phần bằng nhau dùng để huấn luyện và kiểm thử.
3. Sử dụng tập huấn luyện để xây dựng mô hình phân loại theo phương pháp lấy mẫu đơn giản, lấy mẫu cân bằng.
4. Sử dụng tập kiểm thử để phân tích đánh giá trên mô hình đã xây dựng.
5. Lấy giá trị trung bình trên 30 lần thực nghiệm và ghi nhận kết quả đạt được, thống kê số mẫu phân loại.
6. Cuối kịch bản, vẽ đồ thị so sánh đánh giá hai phương pháp trên các tập dữ liệu có mức không cân bằng khác nhau với giải thuật AdaBoost thông thường.

Kết quả kịch bản 1:

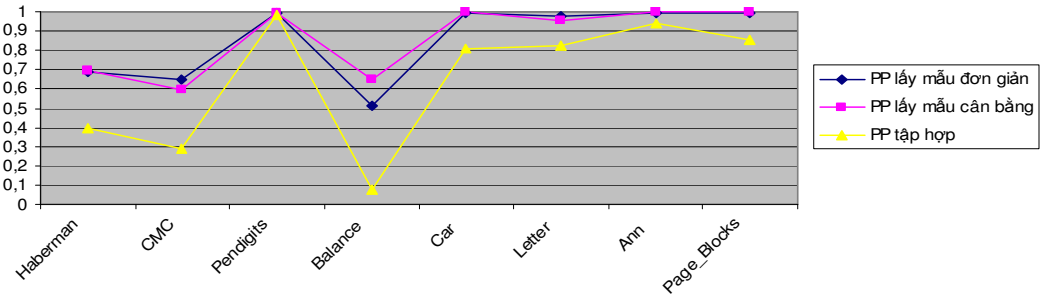


Đồ thị 1: Kết quả phân loại trên các tập dữ liệu dựa trên mức không cân bằng

Kết quả kịch bản 2:

Bảng 3: Kết quả phân loại của lớp nhỏ

Tập dữ liệu	PP lấy mẫu đơn giản	PP lấy mẫu cân bằng	PP tập hợp
Haberman	0,685	**0,693**	0,4000
CMC	**0,649**	0,594	0,2874
Pendigits	**0,994**	0,992	0,9810
Balance	0,512	**0,652**	0,0800
Car	0,994	**0,998**	0,8125
Letter	**0,98**	0,955	0,8229
Ann	0,993	**0,998**	0,9398
Page_Blocks	0,996	**0,999**	0,8571



Đồ thị 2: So sánh độ chính xác của lớp nhỏ trên các phương pháp

Thực nghiệm cho thấy tiếp cận mới với phương pháp lấy mẫu đơn giản, lấy mẫu cân bằng luôn ảnh hưởng đến tỉ lệ phân bố các mẫu ở các lớp khác nhau. Độ chính xác của tất cả các tập luôn được cải thiện ở tất cả các trường hợp. Mức độ chính xác của lớp nhỏ được cải thiện hầu như là như nhau ở tất cả các tập. Điều này được lý giải là trong tiếp cận mới đã đưa dữ liệu cân bằng để xây dựng mô hình đơn, và nhiều mô hình được xây dựng để hạn chế việc mất thông tin của lớp lớn khi thực hiện lấy mẫu rút gọn.

Tóm lại, đối với tập dữ liệu không cân bằng áp dụng tiếp cận lấy mẫu đơn giản và lấy mẫu cân bằng thì ảnh hưởng đến phân loại dữ liệu, chúng cho kết quả tốt trên các tập thực nghiệm. Nghĩa là số mẫu được phân loại đúng tăng lên đáng kể đối với lớp có ít mẫu đồng thời cũng hạn chế phân loại sai các mẫu ở lớp lớn. Các tập có số mẫu lớn thường phương pháp lấy mẫu đơn giản hiệu quả hơn hơn phương pháp lấy mẫu cân bằng và có số mẫu dự đoán đúng ở lớp quan tâm cao hơn. Tuy nhiên, đối với các tập có rất ít mẫu thuộc lớp quan tâm thì phương pháp lấy mẫu cân bằng cải thiện các mẫu được phân loại đúng thuộc lớp nhỏ. Đối với một số tập không cân bằng, khi áp dụng phương pháp lấy mẫu đơn giản, phương pháp lấy mẫu cân bằng không những cải thiện độ chính xác của lớp nhỏ mà còn cải thiện độ chính xác của toàn tập. Tất cả điều này cho thấy tiếp cận mới mang lại hiệu quả tích cực khi phân loại. Cả hai phương pháp luôn cải thiện độ chính xác của lớp nhỏ trên tập dữ liệu không cân bằng và cho kết quả phân loại ổn định. Đặc biệt, phương pháp lấy mẫu cân bằng có thể giải quyết vấn đề dữ liệu không cân bằng ở mức cao và số mẫu của lớp nhỏ cực hiếm.

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong bài báo này, chúng tôi đã trình bày và thực nghiệm phương pháp lấy mẫu đơn giản và phương pháp lấy mẫu cân bằng trên tập dữ liệu không cân bằng. Lấy mẫu giảm làm thay đổi sự phân phối các mẫu giữa các lớp. Sự thay đổi có xu hướng di chuyển các mẫu bị phân loại sai ở các mô hình trước về lớp đúng của nó. Kết quả làm giảm tỉ lệ lỗi và tăng tỉ lệ phân loại đúng của lớp nhỏ.

Các tập dữ liệu không cân bằng trong tiếp cận này được xem như tập dữ liệu nhị phân (Lớp quan tâm có ít mẫu được xem như lớp dương, các lớp còn lại là lớp âm). Tuy nhiên trong thực tế, người ta quan tâm đến nhiều lớp có ít mẫu thì tiếp cận này bị hạn chế. Để áp dụng phương pháp lấy mẫu rút gọn này rộng rãi chúng ta cần cải thiện tiếp cận lấy mẫu rút gọn trên tập dữ liệu đa lớp.

Ngoài ra, phương pháp lấy mẫu này sử dụng trọng số để phân loại các mẫu là như nhau nhưng mỗi mẫu mang thông tin khác nhau và thể hiện các đặc trưng khác nhau, việc gán trọng số bằng nhau là không phù hợp, không nói lên được mức độ quan tâm đối với các mẫu. Vì vậy, chúng tôi đề xuất một phương pháp mới đối với dữ liệu không cân bằng là kết hợp lấy mẫu rút gọn với trọng số khác nhau giữa các mẫu.

TÀI LIỆU THAM KHẢO

Blake, E.Keogh, and C. J. Merz, *UCI repository of machine learning databases* [http://www.ics.uci.edu/~mlearn/MLRepository.html], Department of Information and Computer Science, University of California, Irvine, CA.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *SMOTE: Synthetic minority over-sampling technique*, Journal of Artificial Intelligence Research, vol. 16, 2002, pp. 321–357.

N. Friedman, M.Goldszmidt, *Learning Beyesian Networks with Local Structure*, Proc. UAI-96, 1996, pp 252-262.

Y. Freund, R. Schapire, *A decision-theoretic generalization of online learning and an application to boosting*, Journal of Computer and System Sciences, to appear, 1995.

M. Kubat and S. Matwin, *Addressing the curse of imbalanced training sets: One-sided selection*, Proceedings of the 14th International Conference on Machine Learning, Nashville, TN, 1997, pp. 179–186.

X.-Y. Liu, J. Wu, and Z.-H. Zhou, *Exploratory under-sampling for class-imbalance learning*, Proceedings of the 6th IEEE International Conference on Data Mining, Hong Kong, 2008, pp. 965–969.

J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

R. E. Schapire, *A brief introduction to boosting*, Proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 1999, pp. 1401–1406.

V. Vapnil, *The Nature of Statistical Learning Theory*, Spinger-Verlag, 1995.

ĐÁNH GIÁ CHẤT LƯỢNG MẪU TUẦN TỰ TƯƠNG TÁC DỰA TRÊN HƯỚNG TIẾP CẬN MẠNG BAYES

Trần Minh Tân[1], Huỳnh Xuân Hiệp[1], Julien Blanchard[2]

ABSTRACT

Event sequences data have been involved in almost every aspect of social activities such as data about calling log, stock market, telecommunication network, ect. The mining of sequences data aiming at filtering the essential knowledge to support experts in forecasting, decision support systems is of great significance. The process of mining serial episode can result in expected samples as well as redundant valueless ones. In this paper, we suggest a new approach to interactive serial episode and use interestingness/quality measures on Bayesian Networks to determine quality of serial episode to extract the useful knowledge. Experiments on environmental data has proved initial positive results from this approach.

Keywords: serial episode, serial episode rule, interestingness/quality measures, Bayesian Networks.

TÓM TẮT

Dữ liệu tuần tự thường xuất hiện rất nhiều trong đời sống xã hội như dữ liệu về các cuộc gọi điện thoại, thị trường chứng khoán, hệ thống viễn thông,…Khai thác dữ liệu tuần tự với mục đích trích lọc tri thức để hỗ trợ cho các chuyên gia ứng dụng trong các hệ thống dự báo, hỗ trợ ra quyết định,…là điều rất quan trọng. Quá trình khai thác mẫu tuần tự có thể đạt được các mẫu mong đợi, nhưng cũng có những mẫu không cần thiết. Trong bài viết này chúng tôi đề xuất một tiếp cận mới trên cơ sở tiếp cận mẫu tuần tự tương tác và sử dụng các độ đo chất lượng/lợi ích dựa trên mạng bayesian để xác định chất lượng của mẫu tuần tự nhằm trích xuất các tri thức có ích. Các thực nghiệm trên dữ liệu về môi trường đã cho thấy những kết quả khả quan ban đầu từ tiếp cận này.

Từ khóa: mẫu tuần tự, luật tuần tự, độ đo lợi ích/chất lượng, mạng Bayes.

1 GIỚI THIỆU

Trong xã hội ngày nay, có rất nhiều thông tin được lưu trữ theo dạng chuỗi các sự kiện xảy ra theo thứ tự thời gian gọi là dữ liệu tuần tự *(H. Mannila, et al., 1997)*. Tùy theo các lĩnh vực khác nhau mà dữ liệu sẽ được lưu trữ khác nhau. Trong cuộc sống hằng ngày có nhiều vấn đề liên quan với nhau, cũng như công việc này lại ảnh hưởng tới công việc khác, hay sự kiện đã xảy ra sẽ kéo theo sự kiện khác xuất hiện. Việc tiên đoán một sự việc khác sẽ xảy ra nếu như biết được một sự việc đã xảy ra ở hiện tại, đó là điều mong muốn của con người. Một bác sĩ nếu biết rõ tiền sử bệnh trọng của bệnh nhân thì sẽ dễ dàng chuẩn đoán bệnh và đưa ra cách điều trị cũng như giải pháp phòng bệnh thích hợp hơn cho bệnh nhân của mình. Hiện nay, nếu ta nắm bắt được qui luật thay đổi của khí hậu theo từng vùng miền khác nhau, thì ta cũng dễ dàng đưa ra những chiến lược để ứng phó và giúp cho các vụ mùa cũng tốt hơn…

Hầu hết những vấn đề được trình bày ở trên đều hướng đến mục tiêu muốn tìm ra một qui luật mang tính phổ biến hiện nay. Việc xác định được tri thức tốt nhất từ

[1] Khoa Công Nghệ Thông Tin & Truyền Thông – Trường Đại học Cần Thơ, Việt Nam
[2] Đại học bách khoa Nantes, Cộng hòa Pháp

dữ liệu tuần tự là vấn đề quan trọng và được quan tâm nhiều trong các hệ thống hỗ trợ ra quyết định và dự báo. Mục tiêu của khai thác dữ liệu tuần tự là trích lọc được các mẫu tuần tự thật sự tiềm năng và mang tính thực tiễn ứng dụng cao. Vấn đề cần được giải quyết là làm sao để loại bỏ đi những mẫu tuần tự không cần thiết (kém chất lượng), nên cần có giải pháp để đo chất lượng của mẫu tuần tự.

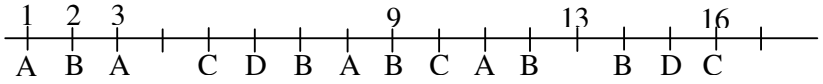
Trong bài viết này chúng tôi đề xuất một hướng nghiên cứu mới với việc sử dụng các độ đo chất lượng trên các mẫu tuần tự dựa vào cơ sở tiếp cận mạng Bayes với các mẫu được hình thành với kích thước biến động trong quá trình tương tác với người sử dụng. Các mẫu tuần tự được đánh giá chất lượng sẽ là những tri thức tốt hỗ trợ quá trình ra quyết định.

Bài viết được tổ chức thành 6 phần. Phần thứ nhất giới thiệu tính cấp thiết của việc nghiên cứu việc đánh giá chất lượng các mẫu. Phần thứ hai trình bày tiếp cận về dữ liệu tuần tự dạng tương tác và mạng Bayes. Phần thứ ba mô tả các phương pháp đánh giá chất lượng mẫu tuần tự với độ đo chất lượng hình thành trên cơ sở tiếp cận mạng Bayes. Phần thứ tư trình bày khái quát hệ thống đã được cài đặt để đánh giá chất lượng mẫu tuần tự do chúng tôi xây dựng. Phần thứ năm trình bày các kết quả thực nghiệm về đánh giá chất lượng mẫu tuần tự. Phần cuối cùng đưa ra các kết luận và hướng nghiên cứu tiếp theo.

2 DỮ LIỆU TUẦN TỰ TƯƠNG TÁC VÀ MẠNG BAYES

2.1 Chuỗi sự kiện

Cho một chuỗi s trên tập các sự kiện E, chuỗi s gồm: (s, T_s, T_e), với $s = \langle (A_1, t_1), (A_2, t_2), ... (A_n, t_n) \rangle$ là một chuỗi các sự kiện tuần tự hay là dữ liệu tuần tự *(H. Mannila, et al., 1997)* được sắp xếp theo thời gian, $A_i \in E \quad \forall$ i = 1..n và $t_i \leq t_{i+1}$ $\forall$ i=1..n-1. Cặp (A_i, t_i) là sự kiện A xảy ra ở thời điểm i, T_s là thời điểm bắt đầu của chuỗi tuần tự và T_e là thời điểm kết thúc chuỗi tuần tự; với $T_s \leq t_i < T_e$ (là thời điểm sau thời điểm của sự kiện cuối cùng trong chuỗi). Hình sau đây mô tả chuỗi dữ liệu tuần tự s= (s, 1, 17).



Hình 7: Chuỗi dữ liệu tuần tự

2.2 Mẫu tuần tự

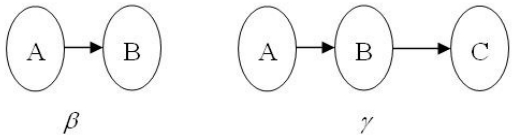
Một mẫu episode *(H. Mannila, et al., 1997)* là một bộ ba (V, ≤, g) với V là tập các nút chứa các sự kiện, ≤ là một quan hệ trên V và ánh xạ g: V $\rightarrow$ E là một ánh xạ kết hợp giữa các nút với các loại sự kiện. Kích thước của episode α là số sự kiện xảy ra trong mẫu α. Episode α là tuần tự khi quan hệ ≤ có thứ tự (ví dụ $x \leq y$ hay $y \leq x$, $\forall x, y \in V$).

2.3 Luật tuần tự

Bên cạnh việc khai thác các mẫu tuần tự phổ biến từ chuỗi dữ liệu tuần tự, thì luật tuần tự cũng rất quan trọng, luật tuần tự cũng tương đối giống mẫu tuần tự, nó hỗ trợ rất mạnh cho ứng dụng để dự báo về một lĩnh vực nào đó.

Luật tuần tự mà ta quan tâm là luật episode *(H. Mannila, et al., 1997)* : có dạng là $\beta \Rightarrow \gamma$ với β, γ là các mẫu tuần tự (serial episode) và β là mẫu con của mẫu γ ($\beta \preceq \gamma$).

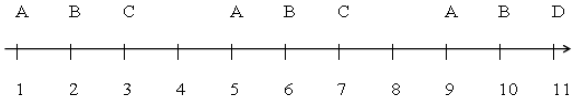
Xét về mặt đồ thị, thì biểu diễn đồ thị của β là đồ thị con của đồ thị γ



Hình 8: Ví dụ về luật tuần tự

2.4 Mẫu tuần tự tương tác

Trong quá trình khai thác mẫu episode, bên cạnh những mẫu tiềm năng thì có thể thu được những mẫu ít phổ biến. Chẳng hạn như một giám đốc công ty thường quan tâm đến thói quen của các khách hàng sau khi mua sữa thì họ sẽ mua thêm đường, cà phê hay là mua dầu lửa. Như vậy hai mẫu tuần tự sữa|đường|cà phê; sữa| dầu lửa thì mẫu tuần tự nào phổ biến hơn để từ đó công ty có chiến lược kinh doanh hàng hóa cho phù hợp ? (sữa tương tác với đường, cà phê, dầu lửa thế nào ?) Kỹ thuật khai thác mẫu episode theo hướng tiếp cận WINEPI hay MINEPI *(H. Mannila, et al., 1997)* sẽ sinh ra những mẫu phổ biến, nhưng đồng thời cũng sinh ra những mẫu bất thường (ví dụ như mẫu bánh mì| xà bông). Do đó, giải pháp sinh mẫu tuần tự mà các sự kiện trong mẫu có liên quan hay tương tác với nhau là cần thiết, nhằm mục đích sinh ra những mẫu đáng quan tâm, với thời gian khai thác được rút ngắn hơn. Kỹ thuật khai thác mẫu theo mô hình tương tác được đề xuất như sau. Xét chuỗi dữ liệu tuần tự như Hình 9



Hình 9: Chuỗi tuần tự

Cho tham số w=1, giả sử cần quan tâm đến mẫu có chứa sự kiện B. Như vậy việc khai phá mẫu sẽ tìm những mẫu tương tác với B, cụ thể như 2 loại mẫu (*, B) và (B, *). Theo Hình 9 , những mẫu tuần tự liên quan đến sự kiện B là: AB, với $t_B - t_A \leq w$; BC, với $t_C - t_B \leq w$; BD,với $t_D - t_B \leq w$

Với w = 1, thì mẫu AB xuất hiện ở ba lần trong chuỗi ở các vị trí (1,2); (5,6) và (9,10). Tương ứng với w=5, thì mẫu AB sẽ xuất hiện thêm hai vị trí nữa là (1,6) và (5,10). Nếu tiếp tục xét các mẫu tương tác với mẫu AB trong cửa sổ w =1, thì kết quả các mẫu sẽ có dạng là (*, A, B) hoặc (A, B, *), cụ thể theo Hình 9 sẽ có các mẫu tương tác với mẫu AB là : ABC, với $t_C - t_A \leq 2w$; ABD, với $t_D - t_A \leq 2w$

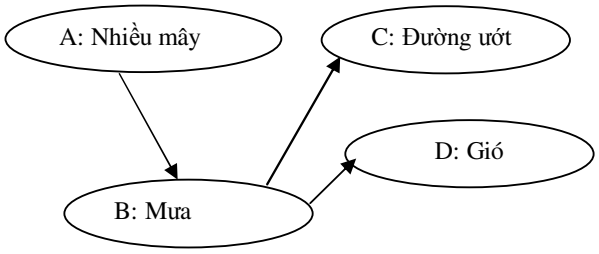
Việc tìm mẫu tuần tự tương tác là quá trình khai thác các sự kiện xảy ra (các sự kiện tương tác) trước hoặc sau các sự kiện đã biết (mẫu tuần tự đã biết). Mô hình tổng quát cho quá trình khai thác mẫu tuần tự tương tác như sau:

1. Nếu mẫu ứng viên có 1 sự kiện X , thì tìm mẫu tương tác có 2 sự kiện và có chứa sự kiện X, sao cho $t_e - t_s \leq w$; t_s: thời gian ở sự kiện bắt đầu của mẫu; t_e: thời gian ở sự kiện kết thúc mẫu.

2. Trường hợp mẫu ứng viên có n sự kiện ($n \geq 2$), thì mẫu tương tác cần khai phá sẽ có n+1 sự kiện, sao cho $t_e - t_s \leq nw$

2.5 Mạng Bayes

Mạng Bayes là một đồ thị biểu diễn phân phối xác suất có điều kiện trên tập các nút, thường được dùng để biểu diễn tri thức của chuyên gia theo các lĩnh vực khác nhau. Mạng Bayes còn được gọi là mạng ý niệm (belief network), hay mạng nhân quả (causal network) *(Ohm Sornil and Sunatashee Poonvutthikul, 2006)*. Mạng Bayes là một mô hình đồ thị có hướng không chu trình (DAG) bao gồm tập các nút (nodes) và các cung (edges). Các nút trong mạng là tập các biến (variable), các biến thể hiện các sự kiện (events) hay tập các thuộc tính. Các cung diễn giải mối quan hệ phụ thuộc giữa các biến (mối quan hệ phụ thuộc giữa các sự kiện) và được xem như là mối quan hệ nhân quả (Hình 10)



Hình 10: Quan hệ trong mạng Bayes

Theo Hình 10, thì nguyên nhân dẫn đến mưa là do có mây, sau khi mưa thì dẫn đến đường ướt và có gió. Xét về mẫu tuần tự tương tác, *các mẫu tương tác với mẫu mưa như sau:*

Mây | **Mưa**

Mưa | *Đường ướt*

Mưa | *Gió*

Luật tuần tự được sinh ra từ mẫu tuần tự như sau:

Mây → Mưa;

Mưa → gió;

Mưa → đường ướt.

Trường hợp tổng quát, cho một đồ thị mạng Bayes $G = (V, E)$, với V là tập các nút và E là tập các cung. Giả sử tập V có n nút từ $A_1 \ldots A_n$, khi đó ta có phân phối xác suất như sau:

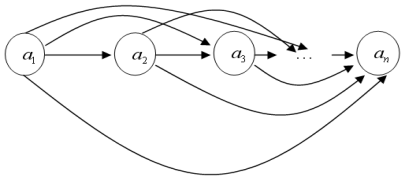
$$P(A_1, A_2, ..., A_n) = \prod_{i=1}^{n} P(A_i \mid parents(A_i)) \quad (1.1)$$

3 ĐÁNH GIÁ CHẤT LƯỢNG MẪU TUẦN TỰ

Hiện nay, một trong những mục đích chính cần quan tâm khi khai thác mẫu tuần tự là vấn đề về chất lượng hay độ hấp dẫn (quality/interestingness) của mẫu. Việc khai thác mẫu có thể sử dụng nhiều hướng tiếp cận khác nhau *(R. Agrawal, 1996; H. Mannila, 1997)*, nhưng để trích lọc được những mẫu/luật thật sự tiềm năng hay có ích (interesting), nhằm loại bỏ bớt đi những mẫu không cần thiết thì cần phải có phương pháp để đánh giá chất lượng của chúng. Để đánh giá mẫu, thì sự cần thiết phải thực hiện là đo chất lượng mẫu. Do đó, chúng tôi sử dụng hướng tiếp cận mạng Bayes để tính toán độ đo chất lượng (interestingness measures) của mẫu tuần tự *(Kuralmani Vellaisamy and Jinyan Li, 2006)*. Theo hướng tiếp cận mạng Bayes, ta xem các phần tử (sự kiện) của mẫu tuần tự như là tập các biến (nút), các cung thể hiện mối quan hệ phụ thuộc giữa các biến với nhau (sự phụ thuộc các sự kiện với nhau). Độ đo chất lượng của mẫu tuần tự được xác định dựa trên phân phối xác suất có điều kiện của mạng Bayes. Sau khi có được độ đo chất lượng của mẫu, ta sẽ thu được tri thức có ích qua các độ đo, nhằm hỗ trợ ra quyết định trong thực tiễn hiện nay. Có hai hướng tiếp cận để xác định độ đo phụ thuộc của các sự kiện trong mẫu *(Kuralmani Vellaisamy and Jinyan Li, 2006):* hướng tiếp cận dựa vào sự phụ thuộc mạnh (strong dependency) và hướng tiếp cận dựa vào sự phụ thuộc yếu (weak dependency).

3.1 Phương pháp đo mẫu theo sự phụ thuộc mạnh

Khi xét một sự kiện a_i xảy ra ở thời điểm t, thì phải tính đến sự ảnh hưởng của tất cả các sự kiện xảy ra trước sự kiện a_i, mô hình này là hướng tiếp cận dựa vào sự phụ thuộc mạnh (strong dependency). Cho một mẫu episode α có n sự kiện, $\alpha = a_1 a_2 a_3 ... a_n$, ta có mô hình sau:



Hình 11: Mô hình hướng tiếp cận theo sự phụ thuộc mạnh

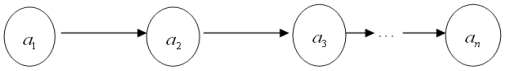
Khi đó các độ đo của mẫu (có n sự kiện, $n \geq 2$) theo sự phụ thuộc mạnh giữa các sự được xác định bởi độ đo BSD (Bayesian Strong Dependency) và BSC (Bayesian confidence for Strong Dependency).

$$BSD = \prod_{j=2}^{n} P(a_j / a_1..a_{j-1}) / P(a_j)$$
$$BSC = \prod_{j=2}^{n} P(a_1..a_j) / P(a_1..a_{j-1}) \quad (1.2)$$

3.2 Phương pháp đo mẫu theo sự phụ thuộc yếu

Hướng tiếp cận dựa vào sự phụ thuộc yếu (weak dependency) là ta xem xét một sự kiện chịu sự ảnh hưởng bởi sự kiện trước nó như thế nào, không quan tâm đến tất cả các sự kiện trong quá khứ. Ví dụ a_n phụ thuộc vào a_{n-1}, a_{n-1} phụ thuộc

vào a_{n-2}…. Một mẫu tuần tự (serial episode) α có n sự kiện tuần tự ($a_1a_2…a_n$) được thể hiện qua hình sau:



Hình 12: Mô hình hướng tiếp cận theo sự phụ thuộc yếu

Khi đó các độ đo BWD (Bayesian Weak Dependency) và BWC (Bayesian confidence for weak dependency) được xác định như sau:

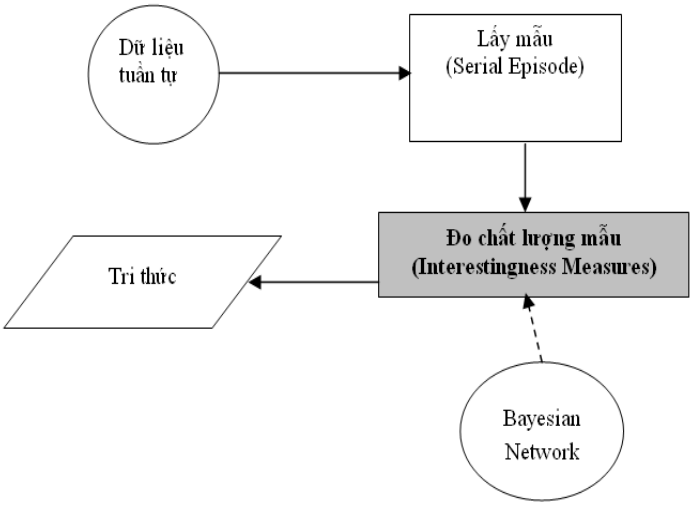
$$BWD(\alpha) = \prod_{j=2}^{n} P(a_j / a_{j-1}) / P(a_j)$$

$$BWC(\alpha) = \prod_{j=2}^{n} P(a_{j-1}a_j) / P(a_{j-1}) \quad (1.3)$$

Đối với một mẫu tuần tự, việc thực hiện đo chất lượng của mẫu để đánh giá tri thức (mẫu/luật), ta cần quan tâm áp dụng cho cả hai trường hợp phụ thuộc mạnh và yếu để tính các độ đo theo sự phụ thuộc của mạng Bayes. Sau khi có được giá trị các phép đo của mẫu, ta sẽ trích lọc được tri thức từ dữ liệu.

4 MÔ HÌNH HỆ THỐNG

Trong phần này chúng tôi xây dựng một công cụ có chức năng khai thác, đánh giá và trích lọc tri thức có ích từ dữ liệu tuần tự (mẫu tuần tự). Mô hình tổng thể của hệ thống như sau:



Hình 13: Mô hình hệ thống

Mô hình hệ thống được thực hiện như sau:

- Đọc vào một tập tin dữ liệu tuần tự, cấu trúc file csv như sau:

time, event
1,A
2,B
3,C
...

A, B, C là các sự kiện xảy ra, tùy vào dữ liệu theo từng lĩnh vực.

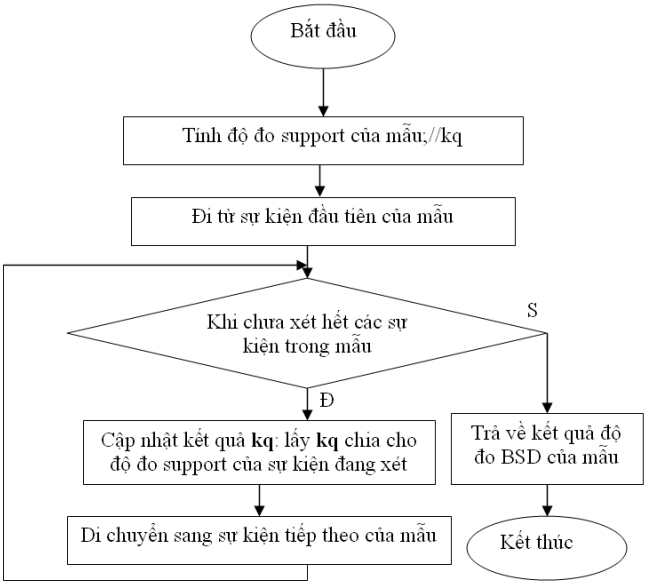
- ✓ Ví dụ: dữ liệu về thị trường chứng khoán, lưu trữ các chỉ số tăng hay giảm của một công ty; A: chỉ số tăng cao; B: không tăng; C: chỉ số giảm…

- ✓ Hoặc dữ liệu về khí tượng thủy văn, lưu trữ các sự kiện thay đổi khí hậu của một vùng; A: trời không mưa; B: mưa rào nhẹ; C: mưa rất to;…

- Bước tiếp theo: hệ thống sẽ tiến hành khai thác mẫu tương tác theo mô hình được trình bày ở mục 2.4
- Sau khi khai thác mẫu tuần tự, ta có thể tiến hành thực hiện các độ đo mẫu dựa trên hướng tiếp cận mạng Bayes (thể hiện phương pháp đo như phần 3)
- Công cụ dễ dàng cho kết quả các tri thức có ích (mẫu tuần tự chất lượng) từ các độ độ chất lượng của mẫu

4.1 Lưu đồ xác định độ đo BSD

Việc tính toán độ đo chất lượng BSD (Bayesian strong dependency) của một mẫu tuần tự theo công thức (1.2) được thể hiện như sau:



Từ công thức (1.2) được triển khai như sau:

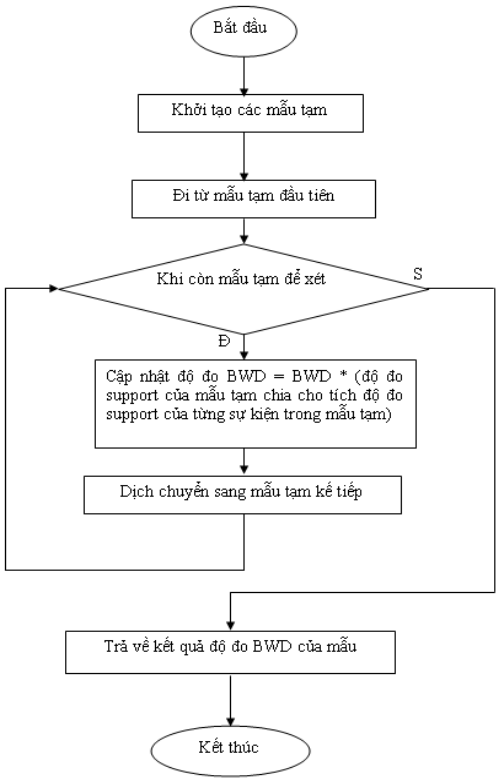
$$BSD = \frac{P(a_1 a_2 a_3 ... a_n)}{P(a_1)P(a_2)P(a_3)...P(a_n)} = \frac{Sup(a_1 a_2 a_3 ... a_n)}{Sup(a_1)Sup(a_2)Sup(a_3)...Sup(a_n)}$$

Do đó độ độ BSD của mẫu được xác định bằng độ đo support của mẫu chia cho tích các độ đo support của các sự kiện trong mẫu.

Hình 14: Lưu đồ tính độ đo BSD

4.2 Lưu đồ xác định độ đo BWD

Việc tính toán độ đo chất lượng BWD (Bayesian Weak Dependency) của một mẫu tuần tự theo công thức (1.3) được thể hiện như sau:



Một mẫu tuần tự có n sự kiện, $\beta = a_1 a_2 a_3 ... a_n$, độ đo BWD của mẫu β được triển khai theo công thức (1.3) như sau:

$$BWD = \prod_{j=2}^{n} P(a_j / a_{j-1}) / P(a_j) \quad -$$

$$= \frac{P(a_1 a_2)}{P(a_1)P(a_2)} x \frac{P(a_2 a_3)}{P(a_2)P(a_3)} x \frac{P(a_3 a_4)}{P(a_3)P(a_4)} x ... x \frac{P(a_{j-1} a_j)}{P(a_{j-1})P(a_j)}$$

Ta tạo một mảng episode (mảng tạm) chứa các cặp sự kiện:

$tam_0 = a_1a_2$; $tam_1 = a_2a_3$; $tam_2 = a_3a_4$; ……; $tam_{n-2} = a_{n-1}a_n$

mỗi mẫu tam_i có 2 sự kiện (kích thước = 2)

- Tạo vòng lặp để tính BWD theo luật xích trên từng mẫu tạm, độ đo BWD của mẫu sẽ là tích của:

* Ứng với từng mẫu tạm: lấy độ đo support của mẫu tạm chia cho tích độ đo support của từng sự kiện trong mẫu tạm

Hình 15: Lưu đồ tính độ đo BWD

5 THỰC NGHIỆM

5.1 Giới thiệu dữ liệu thực nghiệm

Việc nắm bắt qui luật biến đổi các sự kiện về khí hậu của vùng Đồng bằng Sông Cửu Long, để giúp ích cho việc ra quyết định xây dựng các chiến lược kinh tế-xã hội ứng với sự biến đổi khí hậu của vùng hiện nay là điều quan trọng hàng đầu.

Ứng với mỗi vùng sẽ có các sự kiện khí hậu biến đổi một cách đặc thù khác nhau. Ví dụ như các vùng biển Cà Mau, Kiên Giang, Vũng Tàu, thì đặc thù về hình thái thời tiết ở các vùng này như thế nào? Các vùng Cần Thơ – Hậu Giang, vùng Tây Nam Bộ nói chung, thì có các đặc điểm hình thái thời tiết nào thường hay biến đổi,…với mục tiêu là kết hợp với tri thức của chuyên gia, dựa trên dữ liệu thực tế của các vùng miền này nhằm hỗ trợ ra quyết định trong các lĩnh vực về kinh tế, xã hội,...

Từ những nguyên nhân và thách thức trên, chúng tôi đã tiến hành tìm hiểu, thu thập dữ liệu về các cấp hình thái thời tiết của vùng. (tính từ đầu năm 2007 cho đến tháng 12 năm 2009), sau đó chúng tôi tiến hành số hóa dữ liệu đã thu thập được, rồi từng bước đưa dữ liệu vào thực nghiệm trên hệ thống đã xây dựng (khai thác và đánh giá chất lượng mẫu tuần tự tương tác dựa trên hướng tiếp cận mạng Bayes). Bảng 1 mô tả dữ liệu biến đổi về thời tiết của vùng Kiên Giang.

Bảng 1: CÁC SỰ KIỆN BIẾN ĐỔI TRÊN VÙNG KIÊN GIANG

TT	Sự kiện	Giá trị	Ký hiệu	Diễn giải
1	Mưa	1	A	Mưa rào nhẹ
2		2	B	Mưa rào
3		3	C	Mưa vừa
4		4	D	Mưa to và rất to
5	Mây	2	E	Mây thay đổi
6		3	F	Nhiều mây
7		4	G	Đầy mây
8	Lốc	1	H	Có lốc xoáy
9	Dông	1	I	Có dông
10	Gió	3	J	Gió nhẹ tốc độ từ 12-19 km/h
11		4	K	Tốc độ gió từ 20-28 km/h, cây nhỏ có lá bắt đầu lay động
12		5	L	Tốc độ gió từ 29-38 km/h, biển hơi động. Thuyền đánh cá bị chao nghiêng
13		6	M	Tốc độ gió từ 39-49 km/h, cây cối rung chuyển. Khó đi ngược gió
14		7	N	Tốc độ gió từ 50-61 km/h, biển động. Nguy hiểm đối với tàu, thuyền

5.2 Thực nghiệm khai thác mẫu tuần tự

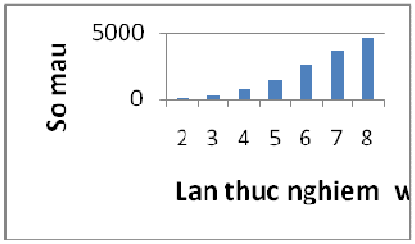
Tiến hành chạy thực nghiệm trên tập dữ liệu của vùng Kiên Giang, ứng với các giá trị thuộc tính được mô tả ở bảng 1, phương pháp lấy mẫu được thực hiện theo mô hình ở phần II. Tùy theo mục đích của các chuyên gia mà ta có thể điều chỉnh tham số w thích hợp. Đối với tham số w càng lớn thì sẽ cho kết quả càng nhiều mẫu, nhưng vấn đề cần quan tâm là các mẫu đạt được có chất lượng hay không? Kết quả các lần thực nghiệm được thống kê theo bảng 2 và hình 9

Bảng 2: THỐNG KÊ CÁC LẦN KHAI THÁC MẪU

Lần thực nghiệm	Số mẫu thu được
2	56
3	234
4	701
5	1396
…	…

Bảng 3: MINH HỌA KẾT QUẢ KHAI THÁC MẪU

TT	Các mẫu tuần tự
1	F I B (F: nhiều mây, B: mưa rào)
2	F I M
3	F I I (F: nhiều mây, I: có dông)
4	F I B I M



Hình 16: Minh họa các lần thực nghiệm

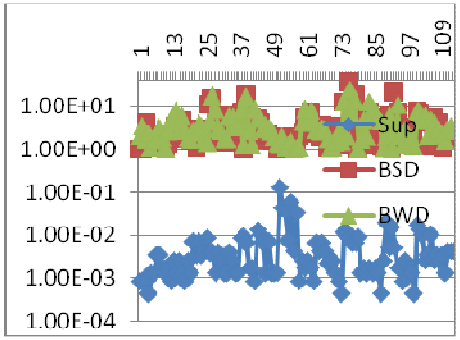
Bài toán khai thác mẫu tuần tự tùy vào tham số w sẽ đạt được số mẫu khác nhau, tham số w càng lớn thì sẽ khai thác được càng nhiều mẫu. Tùy theo tri thức của các chuyên gia mà sử dụng tham số w thích hợp, nhằm thu được các mẫu đạt chất lượng để ứng dụng trong thực tiễn. Ví dụ khai thác được mẫu F|B tương ứng với trời nhiều mây (F) | dẫn đến mưa rào (B), có được quan tâm trong lĩnh vực khí tượng thủy văn của vùng Kiên Giang hay không ? Đây là sự cần thiết để đánh giá chất lượng mẫu.

5.3 Thực nghiệm đánh giá chất lượng mẫu

Đánh giá chất lượng mẫu là việc không thể thiếu trong quá trình tìm kiếm và trích lọc tri thức. Do đó, sau khi thực nghiệm khai thác mẫu, tiến hành thực nghiệm đánh giá mẫu nhằm lọc lại những mẫu thực sự có chất lượng, từ đó giúp

cho các chuyên gia nắm bắt các qui luật chung nhất trong các lĩnh vực thực tế của xã hội hiện nay.

Qua kết quả khai thác mẫu, tiến hành thực nghiệm đánh giá chất lượng mẫu và đạt được kết quả như Hình 17.



Hình 17: So sánh độ đo phụ thuộc theo mạng Bayes và độ đo Support

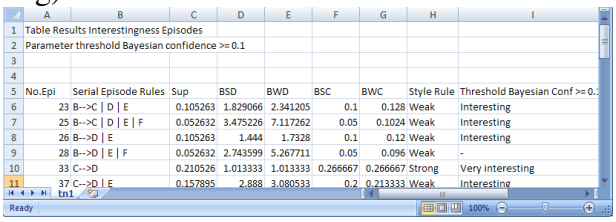
Qua Hình 17 nhận thấy, giá trị độ đo Support của mẫu luôn thấp hơn giá trị độ đo theo Bayesian dependency. Đối với mẫu tuần tự có giá trị độ đo Support quá thấp và tương đương nhau, thì rất khó để đánh giá mẫu, trong khi sử dụng độ đo phụ thuộc thì dễ dàng đánh giá mẫu hơn.

Bảng 4: MINH HỌA ĐỘ ĐO CỦA MẪU TRÊN DỮ LIỆU VÙNG KIÊN GIANG

Mẫu	Sup	BSD	BWD
F | C | N (Nhiều mây, mưa vừa, gió cấp 7)	0.001734	2.091906	1.778121
H | I | L (Lốc xoáy, dông, gió cấp 5)	0.004335	2.275922	2.844902

5.4 Thực nghiệm phân loại và đánh giá mẫu tiềm năng

Đối với một mẫu tuần tự đã đạt chất lượng, ta tiến hành phân loại mẫu. Việc phân loại mẫu nhằm mục đích nhận dạng xem mẫu mạnh hay yếu. Việc đánh giá chất lượng mẫu dựa vào độ đo phụ thuộc theo mạng Bayes (BSD, BWD), nên để phân loại mẫu ta xét tỉ lệ độ đo phụ thuộc BSD/BWD. Nếu mẫu α đạt chất lượng và có tỉ lệ BSD/BWD ≥ 1 thì được phân loại là **mẫu mạnh** (strong), ngược lại α đạt chất lượng và có tỉ lệ BSD/BWD <1 thì được phân loại là **mẫu yếu** (weak). Việc đánh giá mẫu tiềm năng được thực hiện: nếu một mẫu được xem là mẫu mạnh và có độ tin cậy theo Bayesian (BSC) vượt qua ngưỡng η thì mẫu được đánh giá là rất tiềm năng (very interesting); nếu một mẫu được xem là yếu và có độ tin cậy theo Bayesian (BWC) vượt qua ngưỡng η thì mẫu được đánh giá là tiềm năng (interesting)



Hình 18: Minh họa giao diện kết quả phân loại và đánh giá sự tiềm năng

Bảng 5: MINH HỌA KẾT QUẢ PHÂN LOẠI VÀ ĐÁNH GIÁ

Serial Episode	Conf	BSC	BWC	Style	Eval
M | I | G | C	0.0142	0.01428	0.00578	Strong	Very interesting

E | A | J	0.0129	0.01299	0.01385	Weak	Interesting
G | C | M | I	0.2777	0.27777	0.19728	Strong	Very interesting
G | M | I	0.3888	0.38888	0.39176	Weak	Interesting

Bảng 5 đã minh họa các mẫu đạt chất lượng (tri thức có ích), có giá trị ứng dụng trong thực tiễn nhằm giúp hỗ trợ ra quyết định.

Ví dụ mẫu chất lượng G | C | M | I : đầy mây (G), mưa vừa (C), gió cấp 6 (M), có dông (I). Đây là tri thức có ích mang tính ứng dụng thực tiễn cao nhằm hỗ trợ cho các chuyên gia nên quan tâm đến các sự kiện thời tiết thay đổi này ở vùng Kiên Giang.

5.5 Thực nghiệm đánh giá luật

Bên cạnh khai thác mẫu tuần tự, thì luật tuần tự cũng rất đáng được quan tâm khai thác. Bài toán với mục đích tìm mẫu có tiềm năng sinh ra luật tốt nhất, nhằm ứng dụng cho các hệ thống dự báo theo nhiều lĩnh vực trong tương lai. Khi một mẫu tuần tự $\alpha =< g1.g2 >$ đạt chất lượng, thì luật chất lượng có dạng là $g1 \to g2$; với g1= $< a_1,a_2,\ldots,a_j >$, g2= $< a_{j+1},\ldots, a_n >$. Do đó nếu mẫu α đạt chất lượng và có $Lift(\alpha) \geq \lambda$ thì luật $g1 \to g2$ được xem là chất lượng tốt (Best rule). Một luật được xem là chất lượng tốt và có độ tin cậy $\geq \eta$ thì luật được đánh giá là tốt và có tiềm năng (Interesting).

Nếu quan tâm khai thác mẫu, thì việc đánh giá mẫu sẽ được dựa vào độ đo theo Bayesian dependency là lý tưởng. Trường hợp chú trọng đến khai thác luật, thì việc đánh giá luật sẽ được dựa vào độ đo Lift. Những luật có mẫu giống nhau, giá trị độ đo theo Bayesian dependency cũng giống nhau, nhưng nếu vế trái của luật thay đổi sẽ ảnh hưởng đến độ đo Lift của luật. Do đó tiêu chí để đánh giá luật từ các mẫu đạt chất lượng giống nhau sẽ được dựa vào độ đo Lift, nhằm trích lọc luật tốt từ mẫu chất lượng.

Bảng 6: MINH HỌA HAI LUẬT CÙNG MẪU CHẤT LƯỢNG

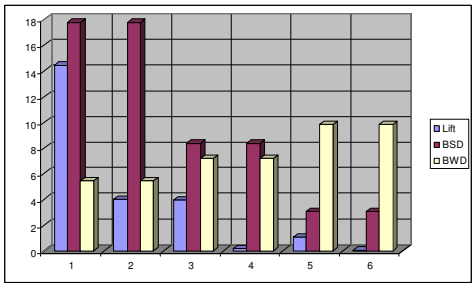
Serial Episode	Episode Rules	Sup	Lift	BSD	BWD
L | H | I | F	L | H --> I | F	0.0021	0.660744	2.596321	3.8175
L | H | I | F	L --> H | I | F	0.0021	0.276897	2.596321	3.8175
F | C | N	F | C --> N	0.0013	4.27037	2.708761	1.7336
F | C | N	F --> C | N	0.0013	0.991117	2.708761	1.7336

Bảng 6 cho thấy hai luật (cùng mẫu tuần tự) có cùng độ đo phụ thuộc và độ đo support, chỉ khác nhau độ đo Lift, nên luật có độ đo Lift cao hơn sẽ được chọn.

Ví dụ xét mẫu F | C | N có cùng độ đo phụ thuộc theo mạng Bayes là BSD = 2.708 và BWD = 1.7336, do đó độ đo Lift sẽ được sử dụng để khai thác luật. Đối với các luật có mẫu giống nhau như bảng VI, thì luật nào có độ đo Lift cao hơn sẽ được quan tâm hơn. Cụ thể giữa hai luật F→C | N và F | C → N, thì luật F | C → N sẽ được quan tâm hơn vì có độ đo Lift cao hơn.

Lift (F | C → N) = 4.27037 > Lift (F → C | N)= 0.991

Xét ý nghĩa thực tiễn cần quan tâm đến các sự kiện khí tượng thủy văn của vùng Kiên Giang như sau: trời nhiều mây (F), sau đó sẽ có mưa vừa (C) và kế đến là có gió (N). Rất có ích, hỗ trợ cho các chuyên gia ra quyết định trong thực tiễn tại vùng miền này.



Hình 19: Minh họa đánh giá luật qua độ đo Lift

5.6 Đánh giá kết quả thực nghiệm

Kết quả thực nghiệm đáp ứng và phù hợp với mục tiêu ban đầu đặt ra về đánh giá chất lượng mẫu tuần tự. Chức năng khai thác mềm dẻo, tùy vào mục đích của các chuyên gia mà có thể hiệu chỉnh tham số w thích hợp. *Giá trị thực nghiệm độ đo các mẫu tuần tự theo Bayesian dependency luôn đạt kết quả cao* hơn giá trị độ đo support, confidence. Do đó việc đánh giá và phân loại mẫu dựa vào độ đo theo hướng tiếp cận của mạng Bayes là hoàn toàn phù hợp và dễ dang thực hiện.

Kết quả không chỉ đạt ở mức đánh giá chất lượng mẫu, mà còn trích lọc được các luật tốt từ những mẫu đạt chất lượng. Việc này rất có ích và mang tính ứng dụng rất cao trong các hệ thống dự báo, hỗ trợ ra quyết định,…trong nhiều lĩnh vực khác nhau như: thị trường chứng khoán, giao dịch dữ liệu, mạng viễn thông, khí tượng thủy văn…

6 KẾT LUẬN

Sau quá trình nghiên cứu và thực hiện, chúng tôi đã xây dựng được các mẫu tuần tự tương tác với tham số w tùy chọn và đã đánh giá được chất lượng các mẫu này trên cơ sở tiếp cận mạng Bayes. Công cụ cài đặt đảm bảo thực hiện được các chức năng chính như khai thác mẫu tuần tự, tính toán các độ đo của mẫu tuần tự. Song hơn thế nữa là chức năng tính toán các độ đo chất lượng của mẫu theo hướng tiếp cận mạng Bayes, để nhằm hỗ trợ cho việc đánh giá chất lượng mẫu tuần tự qua các độ đo.

Kết quả thực nghiệm trên các tập dữ liệu thu thập thực tiễn đã thể hiện được mục tiêu ban đầu. Kết quả các độ đo của mẫu theo sự phụ thuộc luôn đạt cao hơn so với các giá trị độ đo Support, Confidence. Nên hướng tiếp cận mạng Bayes để đánh giá chất lượng mẫu là phù hợp và đạt được mục tiêu đề ra. Kết quả đạt được rất có ích và mang tính ứng dụng rất cao trong các hệ thống dự báo, hỗ trợ ra quyết định,…trong nhiều lĩnh vực khác nhau như: thị trường chứng khoán, giao dịch dữ liệu, mạng viễn thông, khí tượng thủy văn…

Bên cạnh những kết quả đã đạt được, theo chúng tôi nội dung nghiên cứu cần được phát triển với mô hình rộng hơn. Trong tương lai, có thể nghiên cứu hướng tiếp cận khác như phép toán tích hợp, mạng nơ ron,… để tính các độ đo của mẫu tuần tự. Sau đó so sánh với hướng tiếp cận đang áp dụng, từ đó lựa chọn giải pháp (hướng tiếp cận) thích hợp nhằm đánh giá chất lượng mẫu tuần tự một cách lý tưởng hơn nữa. Hướng nghiên cứu của chúng tôi không chỉ sử dụng mạng Bayes

để đánh giá chất lượng mẫu tuần tự, mà còn mở rộng hướng tiếp cận này để đánh giá chất lượng luật kết hợp. Ngoài ra, không chỉ dừng lại ở phạm vi nghiên cứu khai phá tri thức từ mẫu tuần tự, mà có thể mở rộng mô hình nghiên cứu các mẫu (episodes) phức tạp hơn như mẫu *tuần tự và kết hợp*. Hơn thế nữa là số hóa các tập dữ liệu thực nghiệm sao cho có nhiều sự kiện với độ dài chuỗi cao hơn, thiết thực hơn và mang lại nhiều ứng dụng cao trong các lĩnh vực kinh tế, xã hội, …. ở vùng Đồng Bằng Sông Cửu Long.

TÀI LIỆU THAM KHẢO

[1] R. Agrawal and R. Srikant, "Mining Sequential Patterns", In Proceedings of the 7th International Conference on Data Engineering, pp.3-14, 1995.

[2] R. Agrawal and R. Srikant, "Mining Sequential Patterns: Generalizations and Performance Improvements", In Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology, pp. 3-17, 1996.

[3] H. Mannila, H. Toivonen and A. I. Verkamo, "Discovery of Frequent Episodes in Event Sequences", Data Mining and Knowledge Discovery, Vol.1, no.3, pp. 259-289, 1997.

[4] Ohm Sornil and Sunatashee Poonvutthikul, "Constructing Bayesian Networks from Association Analysis", Pacific Rim International Conference on Artificial Intelligence (PRICAI'06), LNCS 4099, pp. 231-240, 2006.

[5] Kuralmani Vellaisamy and Jinyan Li, "Bayesian Approaches to Ranking Sequential Patterns Interestingness", Pacific Rim International Conference on Artificial Intelligence (PRICAI'06), LNCS 4099, pp. 241-250, 2006.

PHÁT HIỆN MẪU TUẦN TỰ VỚI KÍCH THƯỚC THAY ĐỔI BẰNG GIẢI THUẬT DYNEPI

Nguyễn Bá Diệp[1], *Huỳnh Xuân Hiệp*[2], *Julien Blanchard*[3]

TÓM TẮT

Khai phá mẫu tuần tự phổ biến trong một tập hợp dữ liệu có thứ tự về thời gian là một vấn đề quan trọng trong khai khoáng dữ liệu. Việc khai phá mẫu tuần tự bao gồm khám phá tất cả các chuỗi sự kiện mà sự xuất hiện của mỗi kiện có mối liên hệ về thời gian. Những giải thuật trước đây xác định tiềm năng của mẫu episode dựa vào một kích thước cửa sổ cố định do người dùng định nghĩa trước. Trong bài báo này chúng tôi trình bày một giải thuật linh động để khai phá mẫu tuần tự episode hiệu quả hơn. Kết quả thực nghiệm cho thấy cách tiếp cận này hữu dùng và có nhiều thuận lợi.

Từ khóa: Chuỗi tuần tự, Mẫu tuần tự, phân tích mẫu tuần tự, sinh mẫu tuần tự kích thước thay đổi, DYNEPI.

Title: Discovery of sequential patterns by dynamic size DYNEPI algorithms

ABSTRACT

The discovery of frequent sequential patterns in a time ordered collection data is an important data mining issue. Sequential pattern mining consists in discovering all sequences of events, where each event has an associated time of occurrence .Most previous algorithms decide the interestingness of an episode from a fixed user-speciefied window width. In this paper, we present a new flexible algorithms to efficiently discover episodes. Experimental results confirm that our approach results useful and advantageous.

Keywords: Event sequence, Serial Episode, Analysis Serial Episode, Generate Serial Episode, DYNEPI.

1. GIỚI THIỆU

Khai thác mẫu tuần tự (Sequential Patterns) [2, 3, 4, 5] là một phương pháp mới để khai khoáng dữ liệu có liên hệ về thời gian. Mục tiêu của phương pháp này là tìm ra chuỗi dữ liệu con phổ biến (frequent subsequences) hay được gọi là mẫu trong cơ sở dữ liệu mang đặc trưng về thời gian. Những mẫu phổ biến này là một chuỗi các tập sự kiện (itemset) có trong cơ sở dữ liệu với một quan hệ thứ tự về mặt thời gian. Một hướng tiếp cận khác với phương pháp trên là khai phá mẫu Episode (episode mining) [1,6] với mục tiêu tìm ra các mẫu episode phổ biến (frequent episode). Các mẫu episode phổ biến bao gồm tập hợp các sự kiện (event) xảy ra thường xuyên trong một chuỗi dữ liệu sự kiện tuần tự (event sequences) và được tìm ra bằng giải thuật WINEPI hoặc MINEPI [Mannila, et al., 1997]. Cả hai giải thuật đều sử dụng cửa sổ thời gian với độ rộng do người dùng định trước để tìm mẫu, do đó kích thước của mẫu tìm được sẽ bị giới hạn tương ứng với độ rộng của cửa sổ. Những mẫu có kích thước lớn hơn kích thước cửa sổ sẽ không xuất hiện trong cửa sổ do đó tập mẫu tìm được sẽ không chứa những mẫu đó do đó chỉ có thể tìm được một phần của mẫu. Mặt khác những mẫu có kích thước nhỏ hơn cửa sổ sẽ xuất hiện trong nhiều cửa sổ do đó sẽ làm giảm tính chính xác của độ đo ủng hộ mẫu (support) và độ tin cậy (confidence) của luật tuần tự (sequential rule) tương ứng với

[1] Khoa CNTT, Trường ĐH Đồng Tháp, Số 783 Phạm Hữu Lầu, P.6, Tp. Cao Lãnh, dqbao@dthu.edu.vn
[2] Phòng Thanh Tra Đào Tạo, Trường ĐH Đồng Tháp, thle@staff.dthu.edu.vn
[3] Bộ môn Khoa Học Máy Tính, khoa CNTT&TT, Trường ĐH Cần Thơ, dtnghi@cit.ctu.edu.vn

mẫu đó. Giải thuật WINEPI và MINEPI [1] được chia làm hai bước, với bước đầu tiên là tạo các mẫu ứng viên (candidate episode) có kích thước n bằng cách kết hợp từng mẫu kích thước n-1 với tất cả các sự kiện của các mẫu còn lại do đó sẽ sinh ra số lượng mẫu không phổ biến lớn. Với phương pháp MINEPI sử dụng độ đo ủng hộ bằng sự xuất hiện trực tiếp mẫu trong cửa sổ, ngưỡng phân loại mẫu phổ biến dựa vào chính xác số lần xuất hiện của mẫu gây khó khăn trong việc lựa chọn mức phân ngưỡi phổ biến. Ngoài ra cả hai phương pháp trên sẽ làm xuất hiện những mẫu người dùng không quan tâm đến gây nhiễu cho tập luật tuần tự.

Chúng tôi đề xuất giải thuật DYNEPI nhằm giải quyết các vấn đề đã nêu trên. DYNEPI sử dụng một cửa sổ trượt với kích thước w và biên độ thời gian phụ thuộc vào kích thước mẫu người dùng cho trước để tìm ra những mẫu tuần tự phổ biến thoả tiêu chí của người dùng. Với khả năng mềm dẻo và biên độ thời gian phụ thuộc vào kích thước mẫu DYNEPI sẽ không bỏ sót mẫu có kích thước lớn hơn kích thước cửa sổ w đồng thời chỉ đếm chính xác mẫu có trong cửa sổ chứ không đếm mẫu có trong các cửa sổ chồng lắp như WINEPI [1]. DYNEPI có thể sử dụng được trên chuỗi dữ liệu tuần tự đa sự kiện (multi event sequences) với nhiều sự kiện khác nhau cùng xảy ra trong một đơn vị thời gian trong khi WINEPI chỉ có thể sử dụng trên chuỗi dữ liệu đơn sự kiện (event sequences) với mỗi thời điểm chỉ có một sự kiện diễn ra. DYNEPI còn hỗ trợ tính năng tìm những mẫu phổ biến con của mẫu mà người dùng đã cho ban đầu.

Phần tiếp theo của bài viết này được trình bày như sau: phần 2 trình bày về mô hình giải thuật DYNEPI. Phần 3 là kết quả thực nghiệm và đánh giá. Phần 4 là kết luận và hướng phát triển.

2. GIẢI THUẬT DYNEPI

Cho một tập E là tập hợp các sự kiện. Một sự kiện (e,t) là một sự kiện khi và chỉ khi e thuộc tập E và t là một số nguyên dương chỉ thời gian xảy ra sụ kiện này. Sự kiện e có thể chứa một hoặc nhiều thuộc tính nhưng để đơn giản cho việc trình bày chúng ta giả định mỗi sự kiện chỉ chứa một thuộc tính. Theo đó ta có tập sự kiện trên chuỗi sự kiện :{ (e, t) | e ∈ E} với t > 0.

Một chuỗi tuần tự S trên tập E là một bộ 3(S, T_s ,T_e) trong đó T_s và T_e là số nguyên dương chỉ thời gian trên S. T_s là thời gian bắt đầu, T_e là thời gian kết thúc và $T_s \leq t_i \leq T_e$ với mọi i = 1,…,n và S={ (A_1,t_1) , (A_2,t_2) , (A_3,t_3) , …. , (A_n,t_n) }là một chuỗi tuần tự các sự kiện mà $A_i \in E \ \forall$ i = 1,…,n và $t_i \leq t_{i+1} \ \forall$i=1,….,n-1. Định nghĩa này có một chút khác biệt với định nghĩa chuỗi tuần tự S' trong [1] với S'={ (A'_1,t'_1) , (A'_2,t'_2) , (A'_3,t'_3) , …. , (A'_n,t'_n) } $A'_i \in E \ \forall$ i = 1,…,n và $t'_i < t'_{i+1} \ \forall$i=1,….,n-1. Như vậy S có thể chứa nhiều sự kiện xảy ra tại một thời điểm t so với S' chỉ có thể chứa 1 sự kiện tại một thời điểm t'.

Một mẫu (Episode) được xét là một bộ 3 (V , ≤ ,g) trong đó V là tập hợp của các nút, ≤ là một thứ tự bán phần trên V và ánh xạ g V $\rightarrow$ E là một ánh xạ kết hợp giữa các nút với các loại sự kiện. kích thước của α kí hiệu |α| và kích thước V kí hiệu(|V|). Episode α là tuần tự khi mối quan hệ ≤ là có thứ tự (ví dụ: x ≤ y hoặc y ≤ x, ∀ x,y∈ V).

Tần suất của mẫu X trên chuỗi S với kích thước cửa sổ là k có công thức như sau:

$$freq_{\mathcal{S},k}(X) = |W(X,\mathcal{S},k)|,\ supp_{\mathcal{S},k}(X) = \frac{freq_{\mathcal{S},k}(X)}{|W(\mathcal{S},k)|}$$

Trong đó $freq_{S,k}(X)$ là tần suất xuất hiện của mẫu X trên chuỗi S với kích thước cửa sổ k, $supp_{S,k}(X)$ là độ đo ủng hộ của mẫu X trên S với kích thước cửa sổ k và số cửa sổ có kích thước k trên S |W (S ,k)| .

Gọi σ là ngưỡng độ hỗ trợ nhỏ nhất với 0 <σ< 1. Khi đó một mẫu X là mẫu phổ biến khi và chỉ khi supp (X) ≥ σ.

Sau đây là phần trình bày giải thuật DYNEPI:

Đầu vào:

chuỗi S có chiều dài n, kích thước cửa sổ w, tập dữ liệu chứa các sự kiện và thời điểm xảy ra tương ứng với từng sự kiện E =∅, tập mẫu cần quan tâm C, tập mẫu tuần tự cần tìm F = ∅

Giải thuật:

```
1.  while i< n do {
2.      if S[i] ∉ E then{
3.          E := E U S[i]
4.      }End if
5.      E[S[i]] := E[S[i]] U t[i]
6.  }End while
7.  While i < |C| do{
8.      h := (|Ci| – 1) x w
9.      While j < |E[Ci[0]] | do {
10.         t1 := E[Ci[0]]j.t
11.         t2 := t1+h
12.         t3 := t1 - h
13.         If Ci ∈ S[t1,t2] then
14.             F := F U (Ci |*)
15.         Endif
16.         If Ci ∈ S[t3,t1] then
17.             F := F U (*|Ci )
18.         Endif
19.     EndWhile
20. EndWhile
```

Đầu ra:

Tập mẫu tuần tự F chứa các mẫu mới có kích thước lớn hơn các mẫu trong C và có dạng (C | *) hoặc (* | C).

Ta có S là chuỗi dữ liệu tuần tự có chiều dài n. Một mảng E chứa danh sách các sự kiện và từng thời điểm tương ứng của sự kiện. Mảng C chứa mẫu cần quan tâm do người dùng cho trước, mảng F chứa tập mẫu được sinh ra từ những mẫu được cho trước C. Kích thước cửa sổ w cho trước để xác định biên độ thời gian cho quá trình phát hiện mẫu. Giải thuật được chia làm 2 bước, với bước đầu tiên là sử dụng mảng E lập chỉ mục các sự kiện trong chuỗi S. Việc lập chỉ mục các sự kiện giúp phát hiện mẫu trên chuỗi dữ liệu nhanh hơn không phải qua duyệt nhiều lần trên chuỗi dữ liệu làm giảm đi độ phức tạp của giải thuật.

Đối với từng mẫu quan tâm trong C giá trị h được xác định bằng kích thước mẫu -1 và nhân với w. Xét duyệt mảng E chứa sự kiện bắt đầu của mẫu tìm các biên độ thời gian t1, t2, t3 với giá trị t1 là từng thời điểm sự kiện đầu tiên của mẫu đang xét, t2 = t1+h, t3=t1-h. Nếu mẫu đang xét thuộc khoảng [t1,t2] trên chuỗi thì tìm những mẫu có dạng (C | *) trên chuỗi dữ liệu và thêm vào mảng F, nếu mẫu đang xét thuộc khoảng [t3,t1] thì tìm những mẫu có dạng (* | C) và thêm vào mảng F.

Độ phức tạp của DYNEPI là O (|C| log n)

Với vòng lặp đầu tiên của bước 1 bắt đầu dòng 1 đến dòng 6 có độ phức tạp O(|S|) = O(n) (i). Bước 2 bao gồm 2 vòng lặp với vòng lặp ngoài có độ phức tạp O (|C|)và độ phức tạp của vòng lặp trong là O (|E[C[0]]|) = O (|C|.n). Nhưng do các phần tử trong E[C[0]] được lập chỉ mục và sắp xếp tăng dần nên độ phức tạp của vòng lặp trong là O (log n). Vậy độ phức tạp của cả 2 vòng lặp ở bước 2 là O (|C| log n) (ii). Từ (i) và (ii) ta có độ phức tạp của DYNEPI là O (|C| log n)

3. KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ

Để có thể đánh giá hiệu quả của giải thuật, chúng tôi cài đặt giải thuật DYNEPI bằng ngôn ngữ lập trình Java. Để có thể so sánh với giải thuật WINEPI và MINEPI chúng tôi đồng thời sử dụng giải thuật trên tập dữ liệu chạy thử nghiệm và quan sát kết quả của WINEPI, MINEPI và DYNEPI. Tất cả các kết quả đều được thực hiện trên một máy tính cá nhân chạy hệ điều hành Window. Dữ liệu thử nghiệm là kết quả quan sát thời tiết quận Ninh Kiều thành phố Cần Thơ trong 10 ngày từ ngày 10/10/2009 – 20/10/2009 từ 7h sáng đến 5h chiều. Mỗi đơn vị thời gian tương ứng 3 giờ thực tế

time	event		time	event
4	Nắng gắt		60	Nắng
6	Độ ẩm cao		61	Gió mạnh
7	Gió mạnh		62	Mây
10	Mây		64	Độ ẩm cao
11	Mưa		65	Mưa
15	Mây		66	Gió hiu hiu
16	Độ ẩm thấp		67	Nắng
18	Nắng		72	Gió mạnh
20	Gió hiu hiu		77	Mưa
22	Mây		78	Bão
23	Sương mù		80	Mây
25	Mây		81	Gió mạnh
28	Gió mạnh		83	Mưa
29	Mưa		85	Mây
34	Gió mạnh		86	Gió mạnh
37	Bão		88	Mưa
38	Mây		89	Mây
40	Độ ẩm thấp		90	Gió mạnh
41	Nắng		91	Mưa
43	Gió hiu hiu		92	Mây
45	Gió mạnh		93	Gió mạnh
47	Mây		94	Mưa
48	Mưa		95	Nắng
49	Mưa		97	Gió hiu hiu
53	Nắng		98	Bão
56	Gió hiu hiu		99	Mây
59	Độ ẩm cao			

Bảng 2: Kết quả mẫu thu được từ giải thuật WINEPI, MINEPI và DYNEPI

Frequent threshold	**Kích thước cửa sổ**	**WIN EPI**	**WINEPI Frequent episode**	**MIN EPI**	**MINEPI Frequent episode**	**DYN EPI**	**DYNEPI Frequent episode**
0.2	4	7	4	12	7	13	10
0.2	5	10	5	20	11	24	16

Bảng 1 : dữ liệu thử nghiệm

Qua bảng kết quả cho thấy giải thuật DYNEPI sinh được nhiều mẫu tuần tự hơn đồng thời các mẫu này có tần suất xuất hiện vượt ngưỡng phân loại tần suất nhiều hơn.

Frequent threshold	**Kích thước cửa sổ**	**%frequent episode / WINEPI episode**	**%frequent episode / MINEPI episode**	**%frequent episode / DYNEPI episode**
0.2	4	51.14%	58.33%	76.92%
0.2	5	50%	55%	66.66%

Bảng 3 : Minh họa tỷ lệ mẫu phổ biến sinh ra trên tổng số mẫu được sinh ra.

Đánh giá kết quả thực nghiệm

Kết quả thực nghiệm cho thấy giải thuật DYNEPI sinh ra nhiều mẫu tuần tự hơn và các mẫu này có tần suất xuất hiện vượt ngưỡng nhiều hơn giải thuật WINEPI và MINEPI [1]. Qua đó cho thấy giải thuật DYNEPI cho kết quả tin cậy hơn. Mặt khác những mẫu do giải thuật DYNEPI sinh ra là những mẫu có liên hệ với mẫu người dùng quan tâm do đó đem lại nhiều thông tin có giá trị hơn.

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi vừa trình bày giải thuật phát hiện mẫu với kích thước động DYNEPI với độ tin cậy và hiệu quả hơn [1]. Ý tưởng phát hiện mẫu trực tiếp trên chuỗi dữ liệu giúp cải thiện thời gian chạy giải thuật, đồng thời giúp sinh ra những

mẫu có tần suất xuất hiện cao nhiều hơn qua đó thu được số lượng mẫu phổ biến nhiều hơn.

Qua quá trình nghiên cứu chúng tôi đề xuất cải tiến thêm hiệu năng và tính năng của giải thuật bằng cách xây dựng mô hình mới thích ứng với nhiều dạng mẫu hơn.

TÀI LIỆU THAM KHẢO

1 H. Mannila, H .Toivonen, A. I. Verkamo: Discovery of frequent episodes in event sequences, DataMining and Knowledge Discovery 1, 259–289, 1997

2 R. Agrawal and R.Srikant, "Mining Sequential Patterns". *In Proceedings of the 7th International Conference on Data Engineering* (ICDE'95), pp. 3-14, 1995.

3 R. Agrawal and R.Srikant, "Mining Sequential Patterns: Generalizations and Performance Improvements", *In Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology* (EDBT'96), pp. 3-17, 1996.

4 J. Pei, J. Han , B. Mortazavi- Asi, J. Wan g, H. Pinto, Q. Che , U. Dayal, M.-C. Hsu: "Mining sequential patterns by pattern-growth: The PrefixSpan approach", IEEE Trans. Knowledge and Data Engineering 16, 1–17, 2004.

5 J. Wang , J. Han : BIDE: "Efficient mining of frequent closed sequences", Proc. 20th ICDE , 2004.

6 C. Bettini , S. Wang, S. Jajodia, J.- L. Lin : "*Discovering frequent event patterns with multiple granularities in time sequences*", IEEE Tran s. Knowledge and Data Engineering 10, 222–237, 1998.

7 G. Casas – Garriga: "*Discovering Unbounded Episodes in Sequential Data*", KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2003 Lecture Notes in Computer Science, 2003, Volume 2838/2003, 83-94, DOI: 10.1007/978-3-540-39804-2_10

ỨNG DỤNG ONTOLOGY TRONG HỆ THỐNG ĐA TÁC TỬ

Phan Khoa Anh

Tóm tắt

Bài báo này trình bày tổng quan về Ontology được sử dụng trong truyền thông giữa các tác tử trong hệ thống đa tác tử. Tôi đã bước đầu tiếp cận cách thành phần Ontology cần thiết cho việc trao đổi thông tin giữa các tác tử, Ontology được xem như một thư viện để các tác tử tham chiếu nhằm mục đích hiểu nội dung tin nhắn được truyền gửi, từ đó suy luận và đưa ra các hành động tiếp theo của tác tử.

Abstract

This article generally presents Ontology used in communication between agents in multiagent system. I initially approached the elements (components) of Ontology which are necessary for exchanging information between the agents, Ontology is considered as a library for the agents' references in order to understand the message's content sent, thence bring out the next actions of agents.

1. ĐẶT VẤN ĐỀ

Khoảng giữa thập niên 90, tác tử và hệ thống đa tác tử được biết đến rộng rãi [1]và thu hút nhiều quan tâm của giới nghiên cứu và được ứng dụng vào các hệ thống tự động hóa. Trong các hệ thống tự động, các chức năng riêng lẻ được xem như một tác tử, có khả năng suy luận và hoạt động một cách độc lập với con người.

Tại Việt Nam, nghiên cứu về hệ thống đa tác tử vẫn còn khá mới mẻ. Các nghiên cứu tập trung vào các ứng dụng mô phỏng thế giới thực, từ đó đưa ra các tiên đoán và giải pháp giải quyết vấn đề, ví dụ: mô tả hệ thống giao thông, theo dõi dịch bệnh trên tôm, lan truyền dịch bệnh rầy nâu [5]… Việc xây dựng các hệ thống có khả năng tự động hoạt động độc lập còn dừng lại ở mức chưa cao. Giao tiếp giữa các tác tử được xây dựng thông qua các quy luật cố định cho trước.

Mục tiêu của bài báo là cung cấp một cách nhìn tổng quát về các thành phần Ontology phục vụ trong quá trình trao đổi thông tin cho các tác tử trong hệ thống. Xây dựng phần mềm tạo Ontology để thiết lập các quy tắc trao đổi thông tin trong hệ thống một cách trực quan. Bài báo giúp các bạn bước đầu hiểu rõ hơn về lĩnh vực hệ thống đa tác tử.

Nội dung bài báo được chia làm bốn phần. Phần thứ nhất giới thiệu vấn đề. Phần hai cung cấp khái niệm cơ bản về ontology. Phần ba trình bày ứng dụng ontology trong truyền thông giữa các tác tử trong hệ đa tác tử và giới thiệu phần mềm tạo ontology ứng dụng trong hệ thống mua bán. Kết luận được trình bày trong phần cuối.

2. KHÁI NIỆM VỀ ONTOLOGY

Ontology là một thuật ngữ của triết học đã được sử dụng một cách rộng rãi trong lĩnh vực trí tuệ nhân tạo và đã có nhiều định nghĩa khác nhau, trong đó định nghĩa của T. Gruber là được chấp nhận rộng rãi nhất. Theo T. Gruber, Ontology là một đặc tả hình thức về khái niệm. Các yêu cầu cho biểu diễn ontology [6] là:

- Các khái niệm được dùng trong ontology và các ràng buộc giữa các khái niệm đó phải được định nghĩa một cách rõ ràng.
- Ontology phải là dạng thông tin sao cho máy có thể hiểu được.
- Thông tin biểu diễn trong ontology phải có tính phổ quát, nghĩa là thông tin đó không chỉ cho một thành phần mà cần được chấp nhận bởi một nhóm các thành phần khác nhau.

Có thể nói, ontology xác định tập các thuật ngữ dùng để mô tả và biểu diễn các khái niệm dựa trên mối quan hệ qua lại (hoặc ràng buộc) giữa các khái niệm đó. Các khái niệm trong ontology giúp cho việc dùng chung và chia sẻ tri thức giữa hai miền tri thức khác nhau.

Các thành phần của ontology bao gồm:

- Các cá thể: thành phần cơ bản, nền tảng của Ontology. Các cá thể trong một ontology có thể bao gồm các đối tượng cụ thể như: con người, động vật, cái bàn…
- Các lớp: là các nhóm, tập hợp các đối tượng trừu tượng. Chúng có thể chứa các cá thể, các lớp khác, hay sự phối hợp cả hai.
- Các thuộc tính: đối tượng trong ontology có thể được mô tả thông qua việc khai báo các thuộc tính của chúng. Mỗi 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.
- Các mối quan hệ: thể hiện sự liên quan giữa các đối tượng. Đặc trưng là mối quan hệ loại riêng biệt mà quy định cụ thể chiều hướng các đối tượng này có liên quan đến các đối tượng khác trong ontology.

3. ỨNG DỤNG CỦA ONTOLOGY TRONG HỆ THỐNG ĐA TÁC TỬ

Trong hệ thống đa tác tử, cách thức trao đổi tin nhắn là các thức trao đổi thông tin phổ biến nhất. Ngoài ra còn cách thức sử dụng bảng đen, tuy nhiên cách thức này kém hiệu quả hơn. Các tin nhắn được truyền gửi trong hệ thống đa tác tử chủ yếu được sử dụng bằng hai ngôn ngữ đó là: KQML và FIPA. Hai ngôn ngữ này đều được xây dựng trên lý thuyết lời nói (Speech Acts).

Lý thuyết lời nói quy định các hành vi lời nói và theo Austin, mỗi hành vi lời nói bao gồm ba phương diện[6]:

- Locution (kiểu nói, cách nói): định nghĩa một cấu trúc lời nói.
- Illocution: hành vi ngôn ngữ thực hiện bởi người nói, mục đích.
- Perlocution: sự thuyết phục, lôi cuốn của lời nói, kết quả của lời phát biểu.

Tuy nhiên, một hành vi lời nói đầy đủ không chỉ định nghĩa cấu trúc lời nói mà còn xác định hành động liên quan đến lời nói đó. Chúng ta sẽ tập trung vào phần Illocution, nó đóng vai trò quan trọng trong việc xác định ngữ nghĩa mục đích và hành vi của một thông điệp được truyền đi. Theo Searle Illocution được chia thành năm kiểu performative như sau [6]:

- Representatives (Mô tả/diễn tả/ thông báo)
- Directives (ra lệnh): yêu cầu người nghe làm gì đó
- Commisive (hứa hẹn)
- Expressive (lời nói mag tính chất biểu cảm)
- Declarations (tuyên bố)

Ví dụ:

Ngôn ngữ thường	**Ngôn ngữ FIPA**
Người mua yêu cầu người bán: “Tôi muốn mua CD có tên là Alias”	(Request :sender Buyer :receiver Seller :ontology MusicShop :language FIPA :content (buy CD:name “Alias”)

Thông điệp trên được gửi đi từ một agent có chức năng là Buy (mua) đến agent Sell (bán). Thông điệp thuộc dạng Representatives (thông báo), và Buy một hành động, CD là khái niệm có thuộc tính là name tất cả đều được định nghĩa trong Ontology MusicShop nhằm mục đích bên nhận có thể suy luận và hiểu nội dung thông điệp.
Qua ví dụ trên ta nhận thấy rằng Ontology được sử dụng trong phần mềm hướng đa tác tử bao gồm những Ontology sau:
Domain Ontology: Ontology về các khái niệm sử dụng trong hệ thống.
Performative Ontology: trình bày các mục đích lời nói trong Speech Acts.
Protocol Ontology: định nghĩa các thể hiện của Performative.
Agent Role Ontology:

4. CÁC ONTOLOGY THÀNH PHẦN

4.1. Domain ontology

Ontology này định nghĩa các từ vựng được sử dụng trong quá trình trao đổi của các agent trong hệ thống. Ví dụ: giá, sản phẩm, câu hỏi, trả lời, người, địa chỉ,…. Bên cạnh đó, Ontology còn định nghĩa thuộc tính cũng như các mối quan hệ của các khái niệm trong miền được quan tâm, ví dụ: một sản phẩm có một giá, một câu hỏi không có trả lời hay có nhiều hơn một câu trả lời, một người có một tên và địa chỉ, sản phẩm này được một người nào đó bán,….

4.2. Performative ontology

Trong Ontology này, định nghĩa các Performative đã được trình bày trong phần Speech Act Theory. Trong trao đổi thông tin giữa các tác tử, để tránh sự nhọc nhằn trong ngữ nghĩa, các thông điệp được đính kèm theo các loại động từ Performative. Vd: một tin nhắn có nội dung như sao (CD:name "Alias") được gửi tới servic musicshop, kèm theo Performative Buy thì sẽ được hiểu một agent muốn mua một CD có tên là Alias. Nếu không có các Performative kèm theo thì sẽ rất khó hiểu nội dung của thông điệp.

Performative	**Mô tả**	**Ví dụ**
Representatives (Mô tả/Diễn giải/Thông báo)	Thông báo bên nhận một số trạng thái của vấn đề	Giá của một CD là 23$
Commisive (hứa hẹn)	Một lời hứa hẹn với bên nhận	Nếu mua 10 CD, thì giá sẽ là 200$
Declarations (tuyên bố)	Một tuyên bố chính xác cho bên nhận	Tôi đồng ý bán CD này
Directives (ra lệnh)	Ra lệnh cho bên nhận	Hãy đưa ra giá CD
Expressive (lời nói mag tính chất biểu cảm)	Trình bày một cảm xúc cho bên nhận	Tôi cám ơn đã bán CD đó

4.3. Protocol ontology

Trong khi Performative Ontology định nghĩa các Performative được sử dụng trong trao đổi thông điệp, thì Protocol định nghĩa các thể hiện của Performative chính xác trong miền được quan tâm của hệ đa tác tử.

Các ví dụ được mô tả trong bảng sau:

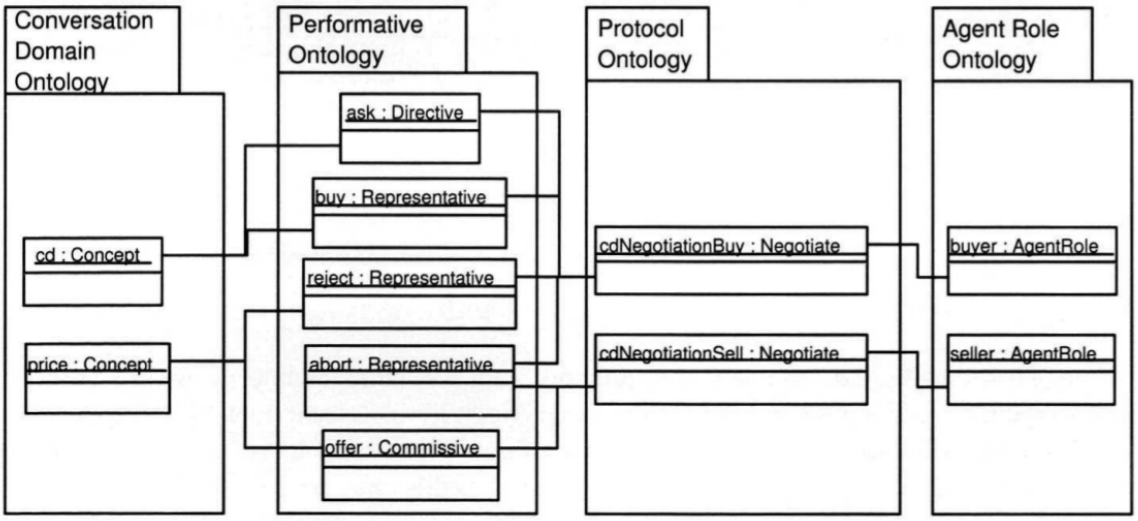
Protocol	**Mô tả**	**Dạng Performative**	**Ví dụ**
Query	Hỏi về thông tin	A to B (directive) B to A (commissive) B to A (representative)	"CD này giá bao nhiêu?" "Đồng ý trả lời." "Giá của CD là 23$."
Request	Yêu cầu hành động	A to B (directive) B to A (commissive) B to A (commissive)	"Có thể đặt hàng CD này không?" "Sẽ đồng ý đặt hàng." "CD sẽ tới vào tuần tới."
Auction	Đấu giá	A to B, C (directive) B to A (commissive) C to A (commissive) A to B, C (representative)	"Ai muốn mua CD này?" "Tôi trả giá 23$." "Tôi trả giá 10$." "Đã bán cho B với giá 23$."
Negotiate	Thương lượng	A to B (directive) B to A (commissive) A to B (representative) B to A (commissive) A to B (representative)	"Giá CD này bao nhiêu?" "30$." "Tôi từ chối, cao quá." "23$." "Đồng ý."
Supervise	Phối hợp với một agent hoặc tiến trình	M to A (declarative) A to M (commissive) M to A (directive) A to M (representative) M to A (directive) A to M (representative) M to A (directive) M to A (expressive)	"Theo hướng dẫn của tôi." "Đồng ý." "Tìm một CD Shop." "Tôi đã tìm thấy rồi." “Cố gắng mua 1 CD thấp hơn 25$.” "Tôi đã mua với giá 23$." "Gửi CD đó tới địa chỉ X cho tôi." "Bạn làm tốt lắm."

4.4. Agent role ontology

Ontology này định nghĩa các luật của các tác tử trong hệ thống. Mỗi tác tử sẽ được quy định rõ ràng từng hành động riêng của mình. Ví dụ: agent mua mới có quyền hỏi mua sản phẩm, nhưng không có quyền bán sản phẩm.

4.5. Tổng hợp các ontology

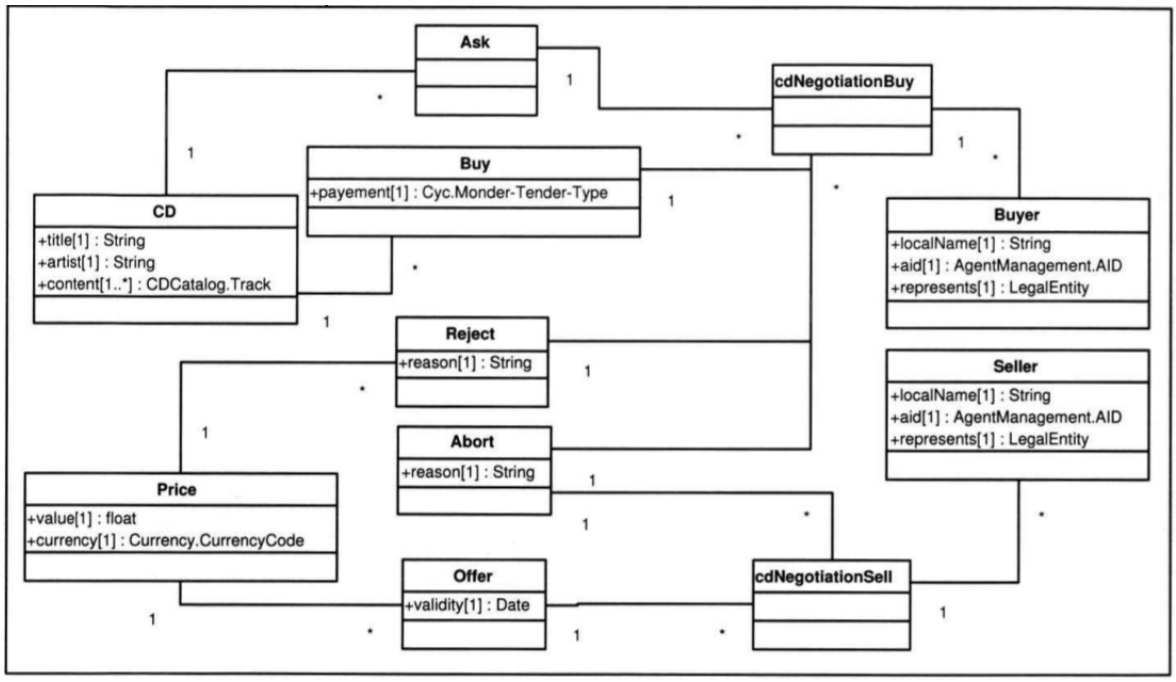
Trong phần này sẽ trình bày một Ontology về một cửa hàng bán CD trên mạng được sử dụng để tham chiếu trong quá trình gửi nhận thông tin từ người bán và người mua.
Các khái niệm trong Ontology bao gồm các từ vựng được sử dụng trong miền quan tâm, các Performative hệ thống sẽ sử dụng, các giao thức tương tác giữa các tác tử, và cuối cùng các luật chung giữa các tác tử trong hệ thống.
Domain: CD, giá của CD, nhãn của CD, nghệ sỹ trình bày trong CD…
Perforamtive: ask, answer, offer…
Protocol: NegotiateBuy, NegotiateSell,…
Role: buyer role, seller role
Và các Ontology có mối quan hệ như hình sau:



Hình 1: Mối quan hệ giữa các Ontology

Hình 1 mô tả các mối quan hệ giữa các Ontology. Ví dụ khái niệm CD trong Domain Ontology được sử dụng trong các hành động ask và buy được định nghĩa trong Performative, và hai hành động này được xác định vào dạng thương lượng trong quá trình bán ở Protocol Ontology, và cuối cùng quá trình thương lượng bán được gán quyền cho agent buyer trong Agent Role Ontology.

Dựa trên các mối quan hệ đã xác định như trên, ta xác định lược đồ quan hệ cho cả Ontology trong hệ thống như hình 2.



Hình 2: Thiết kế chi tiết của các Ontology trong hệ thống

Hình 2 là bảng thiết kế chi tiết của toàn bộ Ontology trong hệ thống. Mô tả các thuộc tính của từng khái niệm và các mối quan hệ giữa các khái niệm với nhau. Ngoài ra, để phục vụ mô tả chi tiết được các vấn đề trong hệ thống ta cần sử dụng thêm các Ontology khác. Ví dụ trong các thuộc tính của khái niệm CD, có thuộc tính nội dung, tuy nhiên một CD sẽ bao gồm nhiều bài hát, mỗi bài hát sẽ có thông tin chi tiết,… nên ta cần kết hợp với một Ontology khác nữa để mô tả rõ vấn đề này. Việc xây dựng các Ontology kết hợp này hoàn toàn giống với Domain Ontology.

5. KẾT LUẬN

Trong bài báo này, tôi đã nghiên cứu cơ bản các thành phần cần có của Ontology trong hệ thống đa tác tử. Khi sử dụng vào thực tế, Ontology này có nhiều cách khác nhau để thể hiện. Bằng ngôn ngữ OWL, trong CSDL quan hệ, và phổ biến nhất là bằng ngôn ngữ Java kết hợp với Zade Framework (framework hỗ trợ xây dựng hệ thống đa tác tử). Tuy còn nhiều hạn chế nhưng đây là tiền đề rất quan trọng giúp các bạn có những nghiên cứu rộng hơn trong lĩnh vực này

Chân thành cảm ơn Ths. Trương Thị Thanh Tuyền, Giảng viên Khoa Công Nghệ Thông Tin và Truyền Thông đã góp ý và giúp hoàn chỉnh bài báo.

TÀI LIỆU THAM KHẢO

[1] Ngô Quốc Hưng. Quy trình xây dựng Ontology. Trường ĐHCNTT – ĐHQG – HCM.

[2] Lê Tuấn Hùng, Từ Minh Phương, Huỳnh Quyết Thắng. Tác tử công nghệ phần mềm hướng tác tử. Trường Đại Học Bách Khoa Hà Nội.

[3] Message Content Ontology

http://en.wikipedia.org/wiki/Massive_%28software%29#See_also

[4] Nguyễn Thanh Tuấn. Mô phỏng giao thông sử dụng hệ thống đa tác tử. Tạp chí Khoa Học và Công Nghệ - Đại học Đà Nẵng, Số 5(40).2010

[5] Trương Thị Thanh Tuyền. Bài giảng hệ đa tác tử. Khoa CNTT&TT-ĐHCT. 2011

[6] Wooldridge. An introduction to Multiagent System (chapter 6,8). 2002