PHÁT TRIỂN NHÓM NGHIÊN CỨU MẠNH CẤP ĐẠI HỌC QUỐC GIA HÀ NỘI
"KHAI PHÁ DỮ LIỆU VÀ CÔNG NGHỆ TRI THỨC"

Giới thiệu chung

Tên nhóm nghiên cứu: Khai phá dữ liệu và công nghệ tri thức
Trực thuộc đơn vị: Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

Trưởng nhóm nghiên cứu và các thành viên

2.1. Trưởng nhóm: PGS. TS. Hà Quang Thụy, Trưởng phòng Thí nghiệm Công nghệ Tri thức (KTLab), Trường ĐHCN

Thông tin cơ bản về hoạt động KH-CN của PGS. Hà Quang Thụy (2009-nay):

(1) Chủ trì đề tài QG.10.38 vượt mức mục tiêu, chủ trì đề tài nhánh KC.01.02/06-10; (2) Hoàn thành 3 giáo trình (2 chủ biên, 1 đồng tác giả); (3) Công bố 02 bài tạp chí ISI-covered journals (01 bài SCI), 06 bài báo ISI-covered conferences, 19 bài Scopus indexed, 8 bài báo quốc tế và quốc gia khác; (4) Hướng dẫn NCS Cù Thu Thủy nhận bằng Tiến sỹ năm 2013 và đang hướng dẫn 4 NCS khác; (5) Trình bày báo cáo tại hội nghị quốc tế ICCCI 2012; (6) Thành viên Ban biên tập của The Springer Vietnam Journal of Computer Science, tham gia Ban chương trình các hội nghị khoa học quốc tế ICCCI 2011-2014, ICCSAMMA 2013-2014, phản biện cho các tạp chí IEEE Transactions on Knowledge and Data Engineering-TKDE (SCI), Data & Knowledge Engineering-DKE (SCIE); (7) Trưởng nhóm nghiên cứu mạnh cấp ĐHQGHN

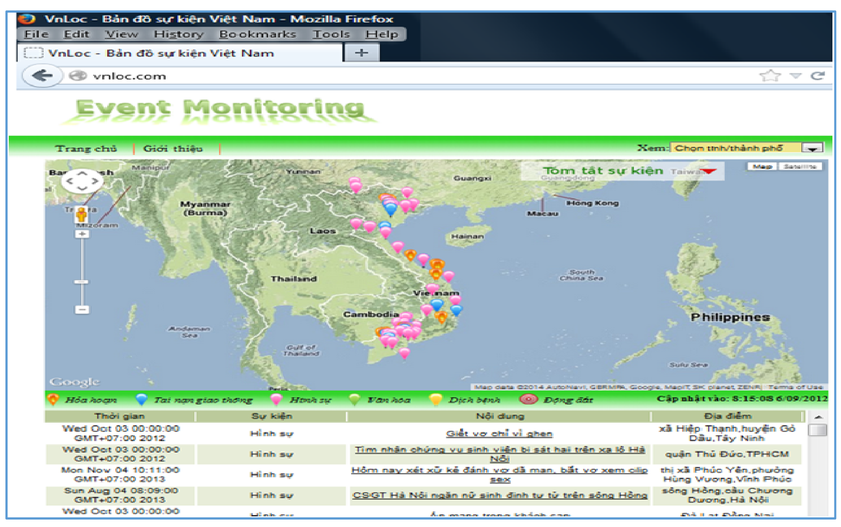
2.2. Phó trưởng nhóm nghiên cứu: TS. Phan Xuân Hiếu, Phòng Thí nghiệm KTLab, Trường ĐHCN

Thông tin cơ bản về hoạt động KH-CN của TS. Phan Xuân Hiếu:

(i) Công bố 21 bài báo khoa học quốc tế (6 bài báo tạp chí SCI & SCIE, 02 bài báo tạp chí quốc tế khác, 13 bài báo kỷ yếu thuộc Scopus – indexed; (ii) Đồng tác giả giáo trình Khai phá dữ liệu web (NXB Giáo dục, 2009); (iii) Là tác giả (cùng TS. Nguyễn Cẩm Tú) bộ công cụ GibbsLDA++: C/C++ implementation of LDA with Gibbs sampling thuộc danh sách bài báo và phần mềm mô hình chủ đề điển hình http://www.cs.princeton.edu/~mimno/topics.html; thành viên Ban chương trình của nhiều hội nghị khoa học quốc tế: Hanoi-ICT 2007, Hanoi-ICT 2008, PRICAI 2008, PRICAI EMALP 2008, WISM 2009, WISM 2010, WISM 2011, KSE 2011, KSE 2012; (iv) Giải thưởng "Quả cầu vàng Việt Nam" TW Đoàn TNCS Hồ Chí Minh năm 2013.

2.3. Danh sách thành viên:

3. TS. Nguyễn Cẩm Tú, Trường ĐHCN, cán bộ nghiên cứu (CBNC) chủ chốt,
4. TS. Nguyễn Trí Thành, Trường ĐHCN (Phó Chủ nhiệm Bộ môn HTTT), CBNC chủ chốt
5. PGS. TS. Nguyễn Hà Nam, Trường ĐHCN (Giám đốc Trung tâm TSK), CBNC kiêm nhiệm
6. PGS. TSKH. Nguyễn Hùng Sơn, Trường ĐHCN & Univesity of Warsaw, CBNC chủ chốt kiêm nhiệm
7. PGS. TSKH. Nguyễn Anh Linh, Trường ĐHCN & Univesity of Warsaw, CBNC chủ chốt kiêm nhiệm,
8. TS. Đặng Thanh Hải, Trường ĐHCN, CBNC kiêm nhiệm,
9. TS. Trần Thị Oanh, Trường ĐHCN (Phòng Thí nghiệm KTLab), CBNC kiêm nhiệm
10. TS. Đoàn Sơn, Trường ĐHCN & UC San Diego, CBNC chủ chốt kiêm nhiệm
11. TS. Nguyễn Việt Cường, Trường ĐHCN & HPC Systems, Inc., CBNC kiêm nhiệm
12. NCS. Trần Mai Vũ, Trường ĐHCN (Phòng Thí nghiệm KTLab), CBNC chủ chốt
13. ThS. Lê Hoàng Quỳnh, Trường ĐHCN (Phòng Thí nghiệm KTLab), CBNC,
14. CN. Lê Đức Trọng, Trường ĐHCN (Phòng Thí nghiệm KTLab), CBNC
15. CN. Vũ Trọng Hóa, Trường ĐHCN (Phòng Thí nghiệm KTLab), CBNC
16. CN. Phí Văn Thủy, Trường ĐHCN (Phòng Thí nghiệm KTLab), CBNC
17. TS. Trần Trọng Hiếu, Trường ĐHKHTN, CBNC kiêm nhiệm
18. ThS. Nguyễn Thị Thùy Linh, Trường ĐHCN (Phòng Thí nghiệm KTLab), CBNC kiêm nhiệm
19. ThS. NCS. Lê Diệu Thu, Trường ĐHCN (Phòng Thí nghiệm KTLab), đang làm NCS tại University of Triete (Italia), CBNC kiêm nhiệm
20. ThS. NCS. Vũ Tiến Thành, Trường ĐHCN (Phòng Thí nghiệm KTLab), đang làm NCS tại Open University (England), CBNC kiêm nhiệm
21. CN. NCS. Nguyễn Thanh Sơn, Trường ĐHCN (Phòng Thí nghiệm KTLab), đang làm NCS tại Singapore Management University (Singapore), CBNC kiêm nhiệm



Hình 1: Hệ thống giám sát sự kiện VNLoc

Hướng nghiên cứu

- Phát triển các mô hình và thuật toán tiên tiến về (i) Khai phá dữ liệu (Text, Web, phương tiện xã hội, quy trình kinh doanh); (ii) Xử lý ngôn ngữ tự nhiên và trích xuất thông tin dựa trên mô hình chủ đề (mô hình xác suất) và các mô hình tích hợp; (iii) Học máy dựa trên lôgic mô tả, học máy đa nhãn đa thể hiện; (iv) Tích hợp dữ liệu, tích hợp tri thức;
- Áp dụng các mô hình và thuật toán đã được phát triển vào thực tiễn Việt Nam.

Kết quả và sản phẩm khoa học tiêu biểu (2009-nay)

Trong giai đoạn từ năm 2009 tới nay, nhóm tập trung nghiên cứu phát triển mô hình và thuật toán trong lĩnh vực khai phá dữ liệu văn bản, khai phá dữ liệu quá trình, xử lý ngôn ngữ tự nhiên, học máy theo lôgic mô tả.

4.1. Sản phẩm khoa học công nghệ

- Hoàn thành 3 giáo trình: Khai phá dữ liệu web, Hệ điều hành UNIX-Linux, Khai phá dữ liệu
- Công bố 01 bài báo tạp chí SCI, 03 bài báo tạp chí SCIE, 06 bài báo kỷ yếu ISI-converted conferences, 20 bài báo Scopus indexed, và 03 bài báo công bố quốc gia. Trong 3 năm gần đây (2010-2013), nhóm có 10 công bố khoa học [3, 10, 11, 12, 14, 16, 17, 19, 20, 37] chung với 5 nhà khoa học người nước ngoài (cố GS. Susumu Horiguchi: JAIST & Tohoku University, GS. Andrzej Szalas: Linköping University, GS. Nigel Collier: NII & European Bioinformatics Institute, GS. Takeshi Tokuyama và TS. Jinhee Chun: Tohoku University).
- Hướng dẫn 07 Nghiên cứu sinh, trong đó Nghiên cứu sinh Cù Thu Thủy đã bảo vệ và nhận bằng Tiến sỹ năm 2013. Hướng dẫn 26 học viên đã bảo vệ luận văn Thạc sỹ (có 3 người làm luận án Tiến sỹ, trong đó 01 người đã nhận bằng Tiến sỹ năm 2014), đang hướng dẫn 35 học viên Cao học làm luận văn Thạc sỹ.

4.2. Kết quả và các hoạt động khác

- Đạt giải NHÌ giải thưởng khoa học –công nghệ của Trường ĐHCN năm 2010.
- Hướng dẫn 25 nhóm sinh viên NCKH và có 11 nhóm đạt giải cấp Trường ĐHCN (2009: 01 giải NHÌ và 03 giải BA, 2010: 02 giải NHÌ và 01 giải BA, 2011: 1 giải NHẤT và 1 giải NHÌ, 2013: 02 giải BA)
- Tham gia Ban tổ chức, Ban chương trình nhiều hội nghị khoa học quốc tế như KSE 2009-2013, RIFV 2012-2013, ICCCI 2011-2014, ICCSAMMA 2013-2014 ... Đang triển khai chủ trì một hội thảo tại hội nghị ACML'14 tại Nha Trang (Tháng 11/2014).

4.3. Đối sánh với tiêu chí nhóm nghiên cứu mạnh hướng chuẩn quốc tế của ĐHQGHN

Số liệu thống kê cho thấy hoạt động của nhóm đáp ứng về cơ bản các tiêu chí chính về nhóm nghiên cứu mạnh hướng chuẩn quốc tế của ĐHQGHN theo Hướng dẫn HD1206 (hướng dẫn số 1206/HD-ĐBCLGD ngày 23/4/2013) của ĐHQGHN.

STT	Tiêu chí	HD 1206		Nhóm NC 2009-14
		2013	2015	
1	Số bài báo, báo cáo trong nước và quốc tế trung bình trên cán bộ khoa học hàng năm	0.5	1	0.8
2	Số lượng trích dẫn/bài báo trong Scopus trong 5 năm gần đây (37/35)	1.2	2.0	1.1
3	Sách chuyên khảo mỗi năm (02 giáo trình: "Khai phá dữ liệu web" và "Khai phá dữ liệu")	×	×	0.4
4	Sản phẩm KH-CN tiêu biểu (Giải NHÌ giải thưởng KH-CN Trường ĐHCN năm 2010)	×	×	0.2
5	Số lượng nhà khoa học được mời báo cáo mời tại hội nghị khoa học quốc gia mỗi năm	×	×	×
6	Số lượng nhà khoa học được mời báo cáo mời tại hội nghị khoa học quốc tế mỗi năm	×	×	×
7	Hợp tác nghiên cứu quốc tế có công bố chung trong vòng 3 năm gần đây (5 nhà khoa học, 10 bài báo)	×	×	5 người 10 bài báo

Bảng 1: Bảng đối sánh với tiêu chí nhóm nghiên cứu mạnh hướng chuẩn quốc tế của ĐHQGHN

4.4. Các thành tích nổi bật của nhóm nghiên cứu

- Tăng nhanh công bố khoa học quốc tế có uy tín (1 tạp chí SCI, 3 tạp chí SCIE, 6 kỷ yếu ISI conferences, 20 Scopus),
- Tích hợp hiệu quả đào tạo chất lượng cao với nghiên cứu khoa học – triển khai công nghệ, góp phần đào tạo tài năng: từ 2009-nay, có 6 cán bộ trẻ (Lê Diệu Thu, Trần Thị Oanh, Vũ Tiến Thành, Nguyễn Thanh Sơn, Lê Đức Trọng, Phí Văn Thủy) được nhận học bổng đào tạo sau đại học (4 TS, 2 ThS) tại các cơ sở đào tạo tiên tiến ở nước ngoài.
- Nâng cấp trình độ hợp tác quốc tế về khoa học – công nghệ trong việc phối hợp đề xuất đề tài-dự án quốc tế, cộng tác nghiên cứu và công bố các kết quả nghiên cứu có uy tín cao: 5 nhà khoa học quốc tế với 10 công bố khoa học có uy tín. Các thành viên trong nhóm nghiên cứu tích cực tham gia các hoạt động tại các tạp chí khoa học, các hội nghị khoa học quốc tế.
- Nâng cấp trình độ hợp tác trong nước về khoa học – công nghệ với các nhóm nghiên cứu của Tổng cục Tình báo (Bộ Công an), Trường ĐHBKHN, Trung tâm CNTT thuộc Viện Dầu khí, Công ty Hài Hòa trong việc đề xuất và thực hiện đề tài, dự án để bổ sung kinh phí cho hoạt động KH-CN trình độ cao (từ năm 2009-nay, nhóm nghiên cứu nhận được khoảng hai trăm (200) triệu đồng từ ĐHQGHN).
- Phát triển được một số phần mềm ứng dụng thử nghiệm (khai phá quan điểm trên phương tiện xã hội, tóm tắt văn bản, ...) trong đó có hệ thống giám sát sự kiện VNLoc (http://vnloc.com/ - Hình 1)

Sản phẩm khoa học dự kiến trong 5 năm tới

Mục tiêu phát triển nhóm nghiên cứu của nhóm là phát triển bền vững nhóm nghiên cứu "Khai phá dữ liệu và công nghệ tri thức" của ĐHQGHN theo hai nội dung:

- Đạt chuẩn quốc tế theo các tiêu chí về công bố khoa học quốc tế uy tín cao, sản phẩm công nghệ tiêu biểu và các tiêu chí khác theo Hướng dẫn HD1206 của ĐHQGHN,
- Phát triển cộng tác KH-CN quốc tế và liên kết hàn lâm – doanh nghiệp (trước mắt với Công ty Tìn học Hài Hòa và Trung tâm CNTT thuộc Viện dầu khi) nhằm tăng cường nguồn lực (bao gồm nguồn lực tài chính) đảm bảo điều kiện vật chất và tinh thần ngày càng được cải thiện.

STT	Sản phẩm	Mục tiêu nghiên cứu và chế tạo	Nội dung triển khai (dự kiến)	Các hoạt động để tạo sản phẩm	Thời gian thực hiện (dự kiến)
1.	Công trình công bố khoa học quốc tế	08 bài báo ISI-converted journals và 25 bài báo Scopus - indexed	Phát triển mô hình, thuật toán tiên tiến về (i) mô hình chủ đề; (ii) khai quá quá trình và áp dụng thử nghiệm tại Việt Nam; (iii) Học khái niệm, xử lý truy vấn dựa trên luật và tích hợp tri thức dựa trên lôgic; (iv) Học máy (đa nhãn) và Khai phá tri thức từ dữ liệu Text/Web hướng đến phát triển các dịch vụ trực tuyến thông minh	- Tích hợp đào tạo và nghiên cứu với NCS và tập thể hướng dẫn là lực lượng lao động nòng cốt; - Tích hợp nghiên cứu nội tại với cộng tác KH-CN quốc tế, - Thực hiện các đề tài Nafosted và ĐHQGHN	Trung bình mỗi năm công bố 1,6 bài báo ISI-converted journals và 05 bài báo Scopus - indexed
2.	Đào tạo	5 luận án Tiến sỹ, 20 luận văn Thạc sỹ, 02 sách chuyên khảo	Như trên	Như trên	Mỗi năm 1 luận án và 4 luận văn; hai năm 01 sách chuyên khảo
3.	Hệ thống khai phá dữ liệu hỗ trợ doanh nghiệp tự động thu thập, tổng hợp quan điểm, quản lý danh tiếng trên Internet	Hoàn thành hệ thống chạy trên Internet như VnLoc.com với chất lượng tương đương với hệ thống iSearch (Spock.com) hoặc hệ thống Zoominfor (Zoominfo.com)	(1) Thi hành (i) mô hình chủ đề; (ii) Học máy (đa nhãn) và Khai phá tri thức từ dữ liệu Text/Web hướng đến phát triển các dịch vụ trực tuyến thông minh và (2) lập trình và cài đặt hệ thống	Như dòng (1) và cộng tác hàn lâm-công nghiệp (chẳng hạn, công ty Tin học Hài Hòa)	2014 -2017
4.	Hệ thống khai phá dữ liệu quá trình hỗ trợ tăng cường tài nguyên tri thức doanh nghiệp	Đáp ứng các yêu cầu về phát hiện quá trình, kiểm tra phù hợp và tăng cường quá trình theo yêu cầu của doanh nghiệp.	(1) Thi hành (i) mô hình chủ đề; (ii) khai quá quá trình và áp dụng thử nghiệm tại Việt Nam; (iii) Học máy (đa nhãn) và Khai phá tri thức từ dữ liệu Text/Web hướng đến phát triển các dịch vụ trực tuyến thông minh; (2) lập trình và cài đặt hệ thống	Như dòng (1) và cộng tác hàn lâm-công nghiệp (chẳng hạn, công ty Tin học Hài Hòa)	2015 -2017
5.	Hệ thống khai phá dữ liệu hỗ trợ tự động thu thập, chia sẻ và hình thành cơ sở dữ liệu về di sản, bản sắc vùng Tây Bắc	Đáp ứng các yêu cầu về phát hiện quá trình, kiểm tra phù hợp và tăng cường quá trình theo yêu cầu của cơ quan quản lý	(1) Thi hành (i) mô hình chủ đề; (ii) Học máy (đa nhãn) và Khai phá tri thức từ dữ liệu Text/Web hướng đến phát triển các dịch vụ trực tuyến thông minh và (2) lập trình và cài đặt hệ thống	Như dòng (1) bao gồm tham gia Chương trình Tây Bắc; cộng tác hàn lâm-công nghiệp (chẳng hạn, công ty Tin học Hài Hòa)	2016 -2018

Bảng 2: Danh mục sản phẩm khoa học dự kiến 2014-2018