

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

HỘI NGHỊ
SINH VIÊN NGHIÊN CỨU KHOA HỌC
LẦN THỨ XXVIII
NĂM HỌC 2010 – 2011

THÁNG 5 NĂM 2011

BAN TỔ CHỨC HỘI NGHỊ

BAN TỔ CHỨC

1. Trưởng ban: TS. Nguyễn Kim Khánh
2. Phó trưởng ban: ThS. Trần Tuấn Vinh
3. Ủy viên: TS. Trần Đức Khanh
4. Ủy viên: KS. Đỗ Bá Lâm
5. Ủy viên: KS. Nguyễn Tuấn Hải
6. Đại diện các Bộ môn, trung tâm

BAN CHƯƠNG TRÌNH

GS. Nguyễn Thanh Thủy	TS. Tạ Tuấn Anh
PGS.TS. Đặng Văn Chuyết	TS. Trần Đức Khanh
PGS.TS. Nguyễn Thị Hoàng Lan	TS. Trương Thị Diệu Linh
TS. Vũ Tuyết Trinh	ThS. Đỗ Văn Uy
TS. Lê Thanh Hương	ThS. Bùi Trọng Tùng
TS. Nguyễn Hồng Quang	ThS. Huỳnh Thị Thanh Bình
TS. Vũ Thị Hương Giang	ThS. Lương Ánh Hoàng
TS. Cao Tuấn Dũng	ThS. Ngô Tuấn Phong
TS. Đỗ Phan Thuận	ThS. Nguyễn Đức Tiến
TS. Hà Quốc Trung	ThS. Nguyễn Duy Hiệp
TS. Ngô Hồng Sơn	ThS. Nguyễn Mạnh Tuấn
TS. Ngô Quỳnh Thu	ThS. Nguyễn Thị Thu Trang
TS. Nguyễn Hữu Đức	ThS. Nguyễn Tiến Thành
TS. Nguyễn Khánh Văn	ThS. Phạm Ngọc Hưng
TS. Nguyễn Linh Giang	ThS. Phạm Văn Thuận
TS. Nguyễn Nhật Quang	ThS. Trần Nguyên Ngọc
TS. Nguyễn Thị Oanh	ThS. Trần Tuấn Vinh
TS. Phạm Đăng Hải	KS. Hoàng Văn Hiệp
TS. Phạm Huy Hoàng	KS. Phạm Hồng Phong
TS. Tạ Hải Tùng	KS. Đỗ Bá Lâm

LỜI GIỚI THIỆU

Hội nghị Sinh viên nghiên cứu khoa học (SVNCKH) là sự kiện thường niên được tổ chức tại Đại học Bách Khoa Hà Nội, trong đó có sự tham gia của Viện Công nghệ thông tin và Truyền thông (CNTT&TT). Chính từ những hội nghị này, nhiều công trình nghiên cứu xuất sắc có tính khoa học cao đã được phát hiện và bồi dưỡng để tham gia và đạt giải ở các cuộc thi cao hơn như Giải thưởng SVNCKH của Bộ Giáo dục và Đào tạo, Giải thưởng sáng tạo WIPO dành cho sinh viên... Hội nghị SVNCKH hàng năm là một sân chơi kích thích niềm sáng tạo, giúp sinh viên làm quen với thử thách của sự nghiệp nghiên cứu, tìm tòi tri thức mới.

Phát huy các kết quả đã đạt được, trong năm học 2010 – 2011, Viện CNTT&TT tiếp tục tổ chức sự kiện này nhằm tìm ra các công trình xuất sắc để trao giải và đề cử tham gia cuộc thi SVNCKH cấp Bộ. Có 43 công trình đã gửi báo cáo để đăng trong kỉ yếu chung của Hội nghị. Các báo cáo được phân công phản biện bởi các giảng viên trong Viện CNTT&TT. Kết quả phản biện được sử dụng làm cơ sở để chọn ra những công trình có chất lượng tốt nhất, trình bày chính thức trước hội đồng chấm giải của Viện.

Quyển kỉ yếu Hội nghị SVNCKH – Viện CNTT&TT thể hiện một kết quả làm việc nghiêm túc, đầy nỗ lực của cả sinh viên và giảng viên hướng dẫn trong nghiên cứu và giảng dạy năm học 2010 – 2011. Cuốn kỉ yếu này là một kỉ niệm đẹp, đánh dấu một mốc son bắt đầu sự nghiệp nghiên cứu khoa học đối với các em sinh viên có công trình nghiên cứu được đăng tải.

Chúc các em luôn luôn sáng tạo, biết phát huy tri thức trong học tập và làm việc!

Thay mặt Ban Tổ Chức

TS. Nguyễn Kim Khánh

Phó Viện Trưởng Viện CNTT&TT

Đại học Bách Khoa Hà Nội

Mục lục

STT	CÔNG TRÌNH - TÁC GIẢ	TRANG
1	ExpertRank: Thuật toán lặp đánh giá chuyên môn người dùng và chất lượng câu trả lời trong các hệ thống hỏi đáp cỡ lớn <i>Nguyễn Văn Đông Anh, Phạm Tuấn Long, Nguyễn Thị Thanh Vi</i>	1
2	Chống trùng lặp địa danh trong hệ thống khai thác thông tin bất động sản <i>Nguyễn Trung Kiên, Đinh Anh Tuấn</i>	7
3	Xây dựng JOO framework chuẩn hóa mô hình lập trình ứng dụng web trên hệ thống phân tán cỡ lớn <i>Bùi Kim Dung, Bùi Anh Dũng, Bùi Trung Hiếu</i>	14
4	Truyện tranh trên di động <i>Nguyễn Thị Thuyên</i>	20
5	Hệ thống tổng hợp tiếng nói tiếng Việt chất lượng cao <i>Nguyễn Trọng Hiếu, Lê Quang Thắng, Lê Anh Tú, Đỗ Văn Thảo, Nguyễn Hữu Thuận</i>	24
6	Giải pháp ngữ nghĩa – Tích hợp dữ liệu, gợi ý và tìm kiếm thông tin cho hệ thống hướng dẫn du lịch thông minh <i>Phan Thanh Hiền, Nguyễn Anh Đức</i>	32
7	Tích hợp nội dung web phổ dụng <i>Phan Văn Hùng, Vũ Mạnh Hùng, Trần Đắc Long</i>	41
8	Các vấn đề an toàn bảo mật cho điện thoại di động, phần mềm bảo mật trên nền Android <i>Trần Ngọc Khải</i>	46
9	Botnet Tracking Framework – Framework hỗ trợ theo dõi và giám sát mạng botnet <i>Triệu Minh Tuân</i>	51
10	RSED: Môi Trường Giả Lập Mạng Giống Thực Tế Phục Vụ Cho Nghiên Cứu Tấn Công Từ Chối Dịch Vụ (DDoS) <i>Trương Thảo Nguyên</i>	56
11	Ứng dụng công nghệ GPS, GIS xây dựng hệ thống theo dõi và quản lý xe buýt Hà Nội	61

Vũ Ngọc Thành

12	Xây dựng hệ mờ nhận dạng biển số xe <i>Đoàn Hồng Quân</i>	66
13	Các gợi ý cá nhân hóa được gửi tự động cho người dùng di động <i>Hoàng Minh Thuần, Tạ Thị Quỳnh Lan</i>	74
14	Hệ thống lưu trữ và chia sẻ dữ liệu Lindax <i>Nguyễn Đức Huy, Nguyễn Thị Khen, Phạm Việt Linh</i>	79
15	Chương trình tạo video 3D từ mô hình 3D sử dụng công nghệ GPGPU <i>Trịnh Quốc Việt, Nguyễn Hữu Dũng</i>	84
16	Mô hình dịch vụ điện toán đám mây Bkloud <i>Lê Quang Hiếu, Hoàng Quốc Nam, Lưu Thị Thùy Nhung</i>	88
17	Hệ điều hành hiệu năng cao HPOS <i>Cao Minh Quỳnh, Nguyễn Đắc Minh, Ngô Văn Vī</i>	94
18	Giải thuật di truyền lai giải bài toán phủ đỉnh <i>Nguyễn Hữu Phước</i>	100
19	Hệ thống nhận diện Virus máy tính theo hành vi <i>Trần Minh Quang</i>	105
20	Phân cụm tài liệu sử dụng độ tương đồng dựa trên cơ sở các cụm từ <i>Nguyễn Kim Thuật, Cao Mạnh Đạt</i>	109
21	Hệ thống trích rút thông tin cho việc xây dựng cơ sở tri thức từ văn bản tiếng Việt <i>Nguyễn Hữu Thiện, Nguyễn Quang Vinh, Nguyễn Thị Minh Ngọc</i>	114
22	Phát triển nền tảng NS2 nhằm phục vụ mô phỏng các giao thức định tuyến trên mạng cảm biến không dây <i>Bùi Tiến Quân, Nguyễn Trung Hiếu</i>	119
23	Xây dựng thư viện khung song song dữ liệu cho hệ thống nhiều bộ xử lý đồ họa <i>Nguyễn Minh Tháp, Ngô Huy Hoàng</i>	124
24	Hệ thống giám sát năng lượng tòa nhà sử dụng công-tơ điện tử và hệ thống truyền tin trên đường điện lưới	129

Nguyễn Trọng Nhật Quang

25	Nâng cao chất lượng tín hiệu tiếng nói <i>Nguyễn Đức Hải</i>	134
26	Xây dựng hệ thống bán hàng tương tác dựa trên nền tảng mạng cảm biến không dây <i>Phạm Đức Anh, Trương Quốc Tú</i>	139
27	Nghiên cứu và xây dựng mô hình mạng của Network-on-Chip <i>Tạ Thị Hà Thư</i>	144
28	Nghiên cứu mạng cảm biến không dây và ứng dụng trong hệ thống xếp chỗ tự động <i>Trần Duy Phương</i>	150
29	Giải pháp camera giám sát giao thông trên nền tảng mạng 3G <i>Trịnh Thị Mây</i>	156
30	Hệ thống định vị - hỗ trợ quản lý học sinh tiểu học trên nền tảng GPS-GSM/GPRS <i>Dinh Thanh Tùng, Đặng Thanh Huyền</i>	161
31	Hệ thống định vị qua bước chân người trong môi trường không có GPS với chi phí thấp <i>Nguyễn Đình Thuận</i>	166
32	Phát triển hệ thống dẫn đường bằng giọng nói và giám sát từ xa bằng camera sử dụng công nghệ 3G trên nền tảng kit friendlyarm <i>Nguyễn Thành Luân</i>	171
33	Xây dựng nền tảng phát triển ứng dụng quảng cáo dựa trên công nghệ Led 3d <i>Nguyễn Thị Phương Ly, Mai Xuân Chiến</i>	176
34	Ứng dụng xác thực khuôn mặt trong kiểm tra hộ chiếu <i>Nguyễn Viết Thành Trung</i>	181
35	Hệ thống thu thập tài liệu theo chủ đề cho tiếng Việt <i>Nguyễn Xuân Hòa</i>	186
36	Xây dựng thiết bị tích hợp dịch vụ phục vụ cho hệ thống mạng doanh nghiệp vừa và nhỏ <i>Bạch Hà Duy, Hoàng Xuân Nam</i>	192

37	Phát hiện và theo vết đối tượng chuyển động <i>Phạm Đức Long, Trương Thị Tâm</i>	197
38	Hệ thống xác thực khuôn mặt hỗ trợ quản lý thẻ thư viện <i>Bùi Thị Minh Yến</i>	202
39	Xây dựng ứng dụng tổng đài nội bộ thoại và hội nghị VOIP trên nền Asterisk <i>Nguyễn Văn Nhẫn, Nguyễn Trung Hiếu</i>	207
40	Bộ thu thập trang Web ẩn theo chủ đề <i>Vũ Thành Đô, Bùi Anh Đức</i>	212
41	Kỹ thuật định vị dựa trên wifi và ứng dụng <i>Chu Bảo Trung, Phạm Hữu Hoàng</i>	218
42	Nghiên cứu, đánh giá và cải tiến hiệu quả sử dụng năng lượng và hiệu suất truyền gói tin của các giao thức định tuyến trong mạng cảm biến không dây <i>Nguyễn Sơn Thủy, Nguyễn Đình Minh</i>	223
43	Nghiên cứu lý thuyết và xây dựng hệ thống phát hiện xâm nhập <i>Nguyễn Xuân Quang</i>	228

ExpertRank: Thuật toán lặp đánh giá chuyên môn người dùng và chất lượng câu trả lời trong các hệ thống hỏi đáp cỡ lớn

Nguyễn Văn Đông Anh, Phạm Tuấn Long, Nguyễn Thị Thanh Vi

Tóm tắt - Nghiên cứu này trình bày thuật toán lặp đánh giá chất lượng câu trả lời và trình độ chuyên môn của người dùng về một lĩnh vực nào đó trong các hệ thống hỏi và trả lời cỡ lớn, mà cụ thể là mạng cộng đồng chia sẻ tri thức Việt Nam BkProfile. Việc đánh giá chất lượng câu trả lời sẽ giúp người dùng chọn ra được câu trả lời đáng tin cậy nhất cho một câu hỏi, trong khi việc đánh giá chuyên môn người dùng sẽ giúp họ có thể chứng minh được kiến thức chuyên môn của mình trong hồ sơ nghề nghiệp của họ trong hệ thống. Hai việc đánh giá này là động lực quan trọng thúc đẩy hoạt động của các hệ thống Hỏi & Đáp và có liên quan chặt chẽ với nhau: câu trả lời chất lượng cao sẽ đóng góp nhiều hơn cho hồ sơ nghề nghiệp (profile) người trả lời và ngược lại, câu trả lời từ người dùng có hồ sơ nghề nghiệp tốt sẽ đáng tin cậy hơn. Chúng tôi đã dựa trên thuật toán phân loại trang web của máy tìm kiếm Google có tên là PageRank, và mô hình chuỗi Markov để chuyển các bài toán trên thành các mô hình xác suất, từ đó xây dựng thuật toán lặp đánh giá cùng một lúc hai đại lượng trên. Thuật toán của chúng tôi được thiết kế trên mô hình Map-Reduce nên có thể được áp dụng cho các hệ thống phân tán cỡ lớn. Chúng tôi đã thử nghiệm nó trên hệ thống mã nguồn mở có tên là Hadoop Map Reduce và triển khai nó chạy ổn định trên ứng dụng web BkProfile tại địa chỉ <http://www.bkprofile.com>. Các kết quả của thuật toán cũng có thể được đóng gói như là một tham số tin cậy sử dụng cho các hệ thống đánh giá trong các ứng dụng chia sẻ tri thức cỡ lớn khác.

Từ khóa - Iterative method, Markov chain, MapReduce, PageRank.

1 GIỚI THIỆU

1.1 Việc xếp hạng chuyên gia và đánh giá chất lượng câu trả lời trong các hệ thống hỏi đáp

Trong những năm gần đây, các hệ thống hỏi đáp, viết tắt là

Công trình này được thực hiện dưới sự bảo trợ của nhóm BKProfile, <http://www.bkprofile.com>, và được hướng dẫn bởi PGS. TS. Huỳnh Quyết Thắng, Ths. Lê Quốc.

Nguyễn Văn Đông Anh, sinh viên lớp Công nghệ phần mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 84-982-933-787, e-mail: anhvny@bkprofile.com).

Phạm Tuấn Long, sinh viên lớp Công nghệ phần mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 84-972-889-760, e-mail: longpham@bkprofile.com).

Nguyễn Thị Thanh Vi, sinh viên lớp Công nghệ phần mềm, khóa 52, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 84-1688-329-541, e-mail: vinguyen@bkprofile.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

Q&A, đã ra đời và phát triển rất mạnh với mục tiêu là hỗ trợ tạo ra các câu trả lời mới vừa cập nhật và vừa sát yêu cầu của câu hỏi; đồng thời lưu trữ được tri thức dưới định dạng mà có thể dễ tìm kiếm lại được. Một số hệ thống hỏi đáp nổi tiếng trên thế giới phải kể đến như Yahoo answers, Google Confucius, Google answers, Google AardVark, Quora, Baidu Zhidao,... Thậm chí một nền tảng web mã nguồn mở có tên là Question2Answer được tạo ra để hỗ trợ việc xây dựng các trang web Hỏi và đáp; hiện nay đã giúp tạo ra hơn 1000 trang web Q&A về các lĩnh vực khác nhau.

Thực tế đã chứng minh là các trang web Q&A chỉ thành công khi các câu hỏi phải nhận được câu trả lời vừa nhanh và vừa có chất lượng cao. Để làm được điều đó, chúng cần tạo ra được động lực cho người dùng trả lời các câu hỏi không những nhanh chóng mà còn cẩn thận. Google Confucius đã thống kê các loại động lực của các trang Q&A phổ biến[3], trong đó có 2 động lực chính là giá trị ảo như việc kêt bạn trực tuyến, thể hiện bản thân,... và giá trị vật chất mà cụ thể là tiền bạc. Cả hai loại động lực này đều cần một hệ thống tự động đánh giá chuyên môn người dùng và chất lượng các câu trả lời một cách công bằng. Với động lực là tiền bạc, việc đánh giá chuyên môn người dùng và chất lượng câu trả lời sẽ là căn cứ để tính ra số tiền cần phải trả cho một câu trả lời nào đó. Với động lực là giá trị ảo thì việc đánh giá chuyên môn khách quan đi kèm với bản hồ sơ cá nhân công bố rộng rãi trong cộng đồng sẽ kích thích người dùng đóng góp nhiều hơn cho hệ thống.

1.2 Đánh giá chuyên môn chuyên gia bằng tiền cù và sử dụng nó để đánh giá chất lượng câu trả lời

Việc đánh giá chuyên môn của một người nào đó có thể được thực hiện trực tiếp bằng cách đánh giá chất lượng công việc của người ấy hay gián tiếp bằng sự tiền cù của các chuyên gia khác (rất phổ biến ở các nước phương Tây như Mỹ, Canada,...). Trong hệ thống Q&A thi việc đánh giá chất lượng câu trả lời một cách trực tiếp có thể thực hiện bằng các phương pháp xử lý ngôn ngữ tự nhiên. Tuy nhiên, cách đó đòi hỏi các biện pháp rất phức tạp để đạt được độ chính xác cao. Với việc tiền cù giữa các chuyên gia thì thông thường nó ẩn chứa dưới hoạt động của các chuyên gia trong hệ thống hỏi đáp. Ví dụ như khi chuyên gia Alice bình chọn cho một câu trả lời của chuyên gia Bob thì việc này có thể được hiểu ngầm là nếu cần tiền cù một người về lĩnh vực của câu trả lời ấy thì có một xác suất nào đó, Alice sẽ tiền cù Bob.

Sau khi đã có chất lượng chuyên môn của người dùng, chúng ta có thể dùng nó để đánh giá chất lượng các câu trả lời. Một câu trả lời được coi là tốt nếu nó được viết và bình chọn bởi các chuyên gia có thứ hạng cao trên hệ thống.

1.3 ExpertRank

Chúng tôi đưa ra một khái niệm có tên là ExpertRank để đánh giá chuyên môn của người dùng. ExpertRank được đánh giá dựa trên sự tiến cử của những người dùng khác trên hệ thống. Người có ExpertRank càng cao thì giá trị của sự tiến cử của họ càng lớn. Đặc biệt, giả sử chỉ có một loại tiến cử thì nếu hai người có cùng ExpertRank, một người tiến cử 5 người, một người tiến cử 20 người thì chất lượng tiến cử của người thứ nhất sẽ cao hơn.

1.4 PageRank và mối liên hệ với ExpertRank

PageRank [1] là tên một thuật toán nổi tiếng được ứng dụng trong máy tìm kiếm Google để sắp xếp được các kết quả tìm kiếm không những theo mật độ từ khóa như các phương pháp tìm kiếm văn bản thông thường mà còn dựa trên các độ tin cậy của trang web. Độ tin cậy này được tính gần đúng bằng việc tiến cử giữa các trang web với nhau và thông qua các đường liên kết giữa chúng. Ví dụ như trang web BKProfile.com có chứa liên kết tới website của Viện Công nghệ thông tin & Truyền thông thì cũng có nghĩa là BKProfile tin tưởng trang web này và trang web này có thêm điểm cho độ tin cậy.

Nếu ta nhìn một trang web như một chuyên gia, và liên kết giữa các trang web giống như việc tiến cử giữa các chuyên gia thì hai hệ thống này tương tự nhau. Hơn nữa cả hệ thống hỏi đáp và hệ thống tìm kiếm đều hướng tới những bài toán với dữ liệu cỡ lớn. Điều đó gợi ý rằng ta có thể áp dụng phương pháp tính toán tầm quan của các trang web mà PageRank đã sử dụng để tính toán trình độ chuyên môn của các chuyên gia. Tuy nhiên, có một số vấn đề tiềm năng mà ExpertRank cần phải quan tâm khi áp dụng tu相似 thuật toán của PageRank. Đó là cấu trúc mạng đầu vào của ExpertRank có thể khác UserRank; độ thưa của mạng cũng như quy mô của mạng cũng là một vấn đề tiềm tàng, đặc biệt khi PageRank sử dụng các quy luật xác suất, vốn chỉ áp dụng được với các số lớn. Để khắc phục điều này, bên cạnh so sánh với PageRank, chúng tôi đã mô hình hóa ExpertRank theo chuỗi Markov để phân tích tính đúng đắn của nó.

Tóm lại, trong phần tiếp theo của bài báo, chúng tôi sẽ trình bày ExpertRank như là một phiên bản mở rộng của PageRank cho việc đánh giá chuyên môn người dùng. Bên cạnh việc bám sát các chi tiết của PageRank để làm căn cứ cho tính đúng đắn, chúng tôi cũng hiệu chỉnh các bước cho phù hợp với điều kiện mới, kết hợp việc phân tích tính hợp lý của ExpertRank và chỉ rõ tính đúng đắn của nó dựa trên mô hình chuỗi Markov.

Chúng tôi đã áp dụng ExpertRank để thiết kế một giải pháp đánh giá chất lượng chuyên gia trên dịch vụ web chia sẻ tri thức Việt Nam BKProfile, sử dụng mô hình lập trình MapReduce, cài đặt trên nền tảng mã nguồn mở Hadoop cho phép phân tán hệ thống tính toán để có thể xử lý dữ liệu lớn. Thành công bước đầu của BKProfile chính là minh chứng cho tính hiệu quả của ExpertRank.

2 CÁC NGHIÊN CỨU LIÊN QUAN

Trong bài báo [4], AardVark thực hiện việc xếp hạng kết quả tìm kiếm người sau khi nhận được truy vấn của người dùng. Công việc tính toán này thông thường phải mất một khoảng thời gian nhất định, với trường hợp của AardVark là một vài phút để tìm ra được những người phù hợp nhất và gửi câu hỏi đi.

Google Confucius [3] thì xếp hạng người dùng trong bước đánh chỉ mục và sử dụng thuật toán HITS [2] với đầu vào là quan hệ người hỏi & người trả lời. Lập luận của Google là số lượng bình chọn của người dùng cho câu trả lời là không đủ để có thể tin cậy được. Hơn nữa, việc tính toán được đẩy cho quá trình đánh chỉ mục làm thời gian truy vấn giảm xuống chỉ còn chưa đến 1 giây.

Tuy nhiên, trong một hệ thống khác là Quora thì số lượng bình chọn của người dùng là rất lớn. Lý do là Quora xây dựng hệ thống Q&A của mình theo mô hình một mạng cộng đồng mà ở đó, người dùng có thể trả lời hoặc bày tỏ quan điểm của mình bằng việc bình chọn cho câu trả lời mà họ thấy đúng.

Một hệ thống khác cũng liên quan đến việc đánh giá chất lượng chuyên gia thông qua tiến cử là hệ thống đánh giá chất lượng các bài viết khoa học thông qua danh mục tài liệu tham khảo [7] nhưng các phương pháp này thường không nhắm tới các hệ thống xử lý dữ liệu rất lớn.

Bài báo này sẽ chỉ tập trung xây dựng thuật toán đánh giá chất lượng người dùng cho hệ thống hỏi đáp dạng cộng đồng cỡ lớn, tức là hoạt động tương tác của người dùng với hệ thống là đủ nhiều để mang ý nghĩa. Việc đánh chỉ mục cũng cần được thực hiện vì hệ thống sẽ không tự động điều hướng câu hỏi mà chỉ gợi ý cho người dùng tự chọn chuyên gia, trong khi thời gian gợi ý không thể quá lâu được.

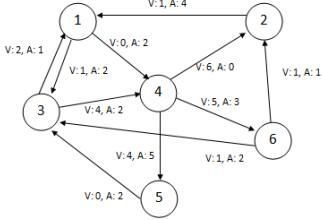
3 XẾP HẠNG CHUYÊN MÔN NGƯỜI DÙNG

3.1 Đồ thị tiến cử giữa các chuyên gia

Như trên đã mô tả, trong các hệ thống Q&A, các chuyên gia tiến cử nhau ngầm trong hoạt động của họ trên hệ thống, tiêu biểu như các lý do sau để người dùng A tiến cử người dùng B:

- B từng trả lời câu hỏi của A
- A đã từng bình chọn cho câu trả lời của B
- A tiến cử B trực tiếp như là một chuyên gia trong một lĩnh vực nào đó.
- A mời B vào hệ thống và chuyên môn của B có liên quan đến lĩnh vực đang xét.
- A và B có chung nhau một số thuộc tính có liên quan đến chuyên môn chung của hai người như cùng lớp, cùng trường, cùng nhóm dự án,...

Bằng các quan hệ ở trên ta có thể xây dựng được một mạng lưới quan hệ giữa các người dùng với nhau, một người sẽ tiến cử người kia với một trọng số nào đó, tùy thuộc vào lý do tiến cử. Nếu coi mỗi chuyên gia là một nốt mạng, việc tiến cử giữa người này đến người kia là một cung có hướng thì toàn bộ hệ thống giống như một đồ thị có hướng và có trọng số.



Hình 1. Đồ thị tiên cử của nhóm gồm 6 chuyên gia với hai loại tiên cử: từng được trả lời (A) và từng bình chọn (V)

3.2 Công thức tính ExpertRank dạng đơn giản

Giả sử u là một chuyên gia trong hệ thống Q&A. Gọi F_u là tập hợp chúa những chuyên gia mà u tiên cử, B_u là tập hợp chúa những chuyên gia tiên cử u .

Gọi $f(u,v)$ là đại lượng đo mức độ tiên cử của chuyên gia u với chuyên gia v mà ở đó $\sum_{v \in F_u} f(u,v) = 1$. $f(u,v)$ đóng vai trò hàm phân phối chuyên môn của chuyên gia u tới chuyên gia v .

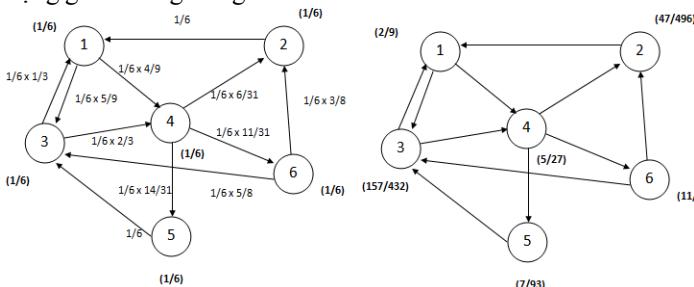
Đặt $R(u)$ là ExpertRank của u , ở đó $\sum_u R(u) = 1$, thì

$$R(u) = \sum_{k \in B_u} R(k) * f(k,u) \quad (1)$$

3.3 Sự lan truyền của tiên cử thông qua mạng lưới tiên cử

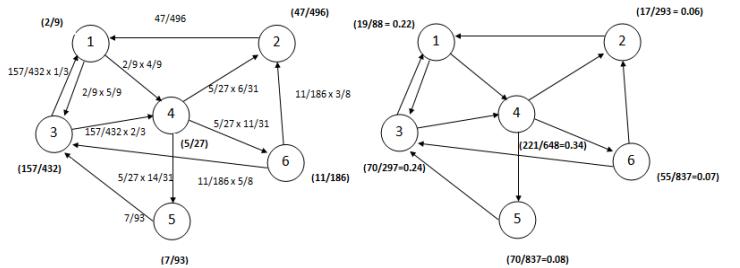
Một người có chuyên môn cao sẽ làm cho chuyên môn của những người mà người này tiên cử tăng lên, và quá trình này tiếp tục cho những người khác mà được tiên cử bởi những người được người này tiên cử. Hiện tượng này chúng tôi gọi là sự lan truyền của tiên cử thông qua mạng lưới tiên cử. Sau một quá trình lặp nào đó, thuật toán có thể hội tụ ở một trạng thái ổn định mà giá trị tại mỗi nốt mạng sẽ là ExpertRank của nó.

Các hình vẽ dưới đây biểu diễn hai quá trình lặp đầu tiên của đồ thị tiên cử chuyên gia được mô tả trong hình 1 mà ở đó: số ở trong dấu ngoặc đơn là ExpertRank của nốt tương ứng trong lần lặp hiện thời; trên mũi tên có biểu thức mô tả việc truyền ExpertRank trong mạng, mỗi biểu thức có hai thửa số: thửa số thứ nhất chính là ExpertRank của nốt tương ứng, thửa số thứ hai đo mức độ tin cậy của nốt được tiên cử với nốt tiên cử; mức độ tin cậy này phụ thuộc vào lý do tiên cử: là do đã từng trả lời hay đã được bình chọn; tỷ lệ trọng số tầm quan trọng giữa chúng trong hình vẽ là 2:1.



Hình 2. Trong lần lặp đầu, các chuyên gia có chuyên môn như nhau và bằng nghịch đảo của tổng số chuyên gia.

Hình bên phải là kết quả của lần lặp đầu.



Hình 3. Dữ liệu đầu vào của lần lặp thứ hai là kết quả của lần lặp thứ nhất

Trong ví dụ trên, sau lần lặp thứ hai, ta có thể thấy chuyên gia số 4 có điểm số cao hơn các chuyên gia còn lại bởi vì chuyên gia số 4 trả lời nhiều câu hỏi và được bình chọn bởi những chuyên gia có chất lượng nhất.

3.4 Mô hình khách hàng xin tư vấn ngẫu nhiên

Thuật toán mô tả một cách định tính ở trên có thể mô hình hóa một cách toán học dưới dạng xác suất mà một khách hàng xin tư vấn ngẫu nhiên viếng thăm một chuyên gia trong hệ thống hỏi đáp để hỏi về một lĩnh vực nào đó.

Giả sử một khách hàng xin tư vấn ngẫu nhiên cần được tư vấn về một lĩnh vực nào đó, như Java chẳng hạn, người đó có được một danh sách các chuyên gia về lĩnh vực Java trong hệ thống. Người đó nhặt ngẫu nhiên một cái tên và bắt đầu xin tư vấn với chuyên gia tương ứng, ví dụ như Alice. Sau khi tham khảo ý kiến của Alice, người này vẫn muốn xin thêm tư vấn nên đã nhờ Alice giới thiệu cho mình các chuyên gia khác mà Alice tin tưởng. Alice căn cứ vào lịch sử làm việc của mình trên hệ thống để giới thiệu một số chuyên gia. Căn cứ vào mức độ giới thiệu này mà người xin tư vấn sẽ có những xác suất viếng thăm những chuyên gia này khác nhau. Giả sử khách hàng đó chọn thăm Bob sau Alice. Khi đó, quá trình làm việc với Bob sẽ hoàn toàn tương tự khi làm việc với Alice: kết thúc phiên làm việc với Bob, khách hàng lại yêu cầu Bob giới thiệu thêm chuyên gia để xin tư vấn tiếp. Lưu ý là khi tới Bob thì giả sử khách hàng sẽ chỉ quan tâm tới lời khuyên của Bob mà không quan tâm tới lời khuyên mà trước đó Alice đã khuyên, điều này là quan trọng cho việc kết nối mô hình này với chuỗi Markov được trình bày trong phần sau.

Sau một số lần đú lớn di chuyển thì người dùng sẽ không còn nhớ được là mình đã bắt đầu ở đâu và xác suất tới thăm một chuyên gia bất kỳ sẽ ổn định. Lúc đó, xác suất khách hàng xin tư vấn tới thăm một chuyên gia bất kỳ sẽ đại diện cho mức độ được tiên cử của chuyên gia đó; và theo đó là đại diện cho chuyên môn của chuyên gia.

3.5 Sự tương ứng với chuỗi Markov và khả năng hồi tụ

Trong mô hình trên, nếu coi mỗi chuyên gia là một trạng thái, việc tiên cử từ chuyên gia này đến chuyên gia kia là quá

trình chuyển trạng thái, và hàm chuyển trạng thái chỉ phụ thuộc vào trạng thái hiện tại mà không phụ thuộc vào các trạng thái trước đó thì quá trình trên chính là một chuỗi Markov [6].

Chuỗi Markov sẽ hội tụ tại một trạng thái duy nhất nếu như từ một trạng thái có thể đến tất cả các trạng thái còn lại bao gồm cả trạng thái đầu [6]. Quy đổi sang ExpertRank thì điều kiện trên trở thành: tiền cù của một chuyên gia có thể tác động tới việc đánh giá chuyên môn toàn bộ những chuyên gia còn lại của hệ thống bao gồm cả chính chuyên gia ban đầu.

Trong quá trình nghiên cứu tính chất mạng của ExpertRank cũng như đối chiếu với PageRank thì việc không hội tụ chủ yếu do những nốt hoặc nhóm nốt chỉ nhận mà không phân phối ExpertRank. Chúng đóng vai trò như những cái bẫy ExpertRank làm ExpertRank bị tắc hoặc biến mất ở trong đó. Hình ở dưới đây minh họa hai trường hợp cho việc ExpertRank bị mất và bị tắc trong các bẫy ExpertRank:



Hình 4. Bẫy ExpertRank là các nốt không có đầu ra hay là một chu trình mà chỉ lấy ExpertRank từ bên ngoài mà không phân phối lại ra ngoài

Nếu loại bỏ hoặc giảm thiểu được hiệu ứng của điều này thì nói chung hệ thống sẽ đảm bảo được hai tính chất trên theo nguyên lý Small World Phenomenon [5]. Tư tưởng của nguyên lý này là tồn tại một con đường bạn bè có độ dài nhỏ hơn hoặc bằng sáu nối giữa hai người bất kỳ trên thế giới. Khi hệ thống hỏi đáp đủ lớn thì cộng đồng người dùng sẽ mô tả được nhiều đặc điểm của một cộng đồng xã hội nên cũng sẽ tuân theo nguyên lý trên, tức là sẽ đảm bảo được hai nguyên lý hội tụ của chuỗi Markov. Với hệ thống nhỏ thì có thể các điều kiện hội tụ trên không được đảm bảo nhưng chúng ta có thể cố định số lần lặp đủ lớn để tìm được các kết quả chính xác ở mức chấp nhận được.

Để giảm thiểu hiệu ứng của các bẫy trong PageRank, PageRank đã sử dụng một nhân tố gọi là $E(u)$ đại diện cho xác suất một trang web có khả năng được nhảy tới một cách ngẫu nhiên, chứ không phải do tiền cù bởi các trang web khác. Với ExpertRank cũng tương tự, người dùng xin tư vấn không phải bao giờ cũng hoàn toàn nghe lời gợi ý của các chuyên gia, họ có thể thăm một chuyên gia khác với một tỉ lệ ngẫu nhiên nào đó. Yếu tố ngẫu nhiên này sẽ làm cho ExpertRank không hoàn toàn bị giữ lại ở trong bẫy mà có thể thoát ra ngoài để quay trở về cộng đồng với một xác suất nào đó.

3.6 Công thức tính ExperRank dạng đầy đủ

Công thức ExpertRank dạng đầy đủ chính là công thức ExpertRank dạng cơ bản cộng thêm nhân tố $E(u)$:

$$R(u) = c \sum_{k \in B_u} (R(k) * f(k,u)) + cE(u) \quad (2)$$

Ở đó:

c: Tham số chuẩn hóa để đảm bảo $\sum_u R(u) = 1$

$E(u)$: đại lượng đặc trưng cho yếu tố ngẫu nhiên khi người xin tư vấn ở một hoàn cảnh bất kỳ viếng thăm chuyên gia u. Cách đơn giản nhất là chọn

$$E(u) = const \quad \forall u$$

Tuy nhiên, việc $E(u)$ có thể tùy biến được là một lợi thế cho việc tùy biến thứ tự kết quả trả về trong các trường hợp khác nhau. Việc này sẽ được đề cập trong mục 7, cá nhân hóa kết quả tìm kiếm.

4 CÀI ĐẶT EXPERT-RANK

4.1 Hệ thống chia sẻ tri thức Việt Nam BKProfile

BKProfile là hệ thống chia sẻ tri thức Việt Nam, được xây dựng ban đầu hướng tới đối tượng là sinh viên và cựu sinh viên Bách Khoa. Hệ thống phục vụ nhu cầu tìm kiếm tri thức có tính tập trung, chất lượng và độ tin cậy cao từ người dùng.

Hiện tại hiện hệ thống mới chỉ được công bố rộng rãi trong sinh viên của Viện CNTT&TT và có khoảng 250/2000 sinh viên sử dụng. Trong tương lai, chúng tôi muốn mở rộng BKProfile tới toàn bộ cộng đồng tri thức Việt nên việc chuẩn bị một hệ thống chạy được với dữ liệu lớn là rất cần thiết.

Hiện tại, chúng tôi đã thực hiện cài đặt thuật toán với số lượng câu hỏi thực tế trên hệ thống tính đến thời điểm hiện tại là 128, số câu trả lời là 204, số bình chọn là 434. Tỉ lệ bình chọn/tỉ lệ câu trả lời là khoảng 2/1, ở mức chấp nhận được.

4.2 Cài đặt ExpertRank bằng Hadoop MapReduce trên BKProfile

Để mở rộng được thì ExpertRank cần phải có khả năng phân tán. Trong số những cách lập trình phân tán thì MapReduce là mô hình lập trình được sử dụng rộng rãi, đặc biệt là với sự trợ giúp của hệ thống mã nguồn mở Hadoop, vốn đang được sử dụng cho nhiều hệ thống lớn như Facebook, Yahoo,...

Mô hình key-value mà MapReduce sử dụng cho định dạng đầu vào và ra phù hợp với việc cài đặt hệ thống tiền cù chuyên gia cài đặt thuật toán ExpertRank, trong đó, mỗi key là một id của chuyên gia và value là một đối tượng mang các thông số đặc trưng người dùng. Mỗi phân phối chuyên môn của các cá nhân có liên quan cho người dùng được lưu trong một value và mô hình MapReduce sẽ thực hiện gom nhóm các value dựa vào key và thực hiện tính toán ExpertRank thông qua việc tổng hợp các phân phối chuyên môn từ các cá nhân có liên quan. Các đoạn mã giả dưới đây mô tả phiên bản đã được đơn giản hóa của ExpertRank.

Calculate ExpertRank distribution

```
function map(expertid,expert){
    foreach(Expert E in related people){
        calculate E's partial rank
        emit(E.id,E)
    }
}
```

Calculate ExpertRank

```
function reduce(expertid, list<expert> expertPartials){  
    foreach(expert in experts){  
        calculate expert's sum rank  
    }  
    emit(expertid,expert)  
}
```

Quá trình thực hiện tính toán ExpertRank cho các người dùng trên hệ thống được tính toán thông qua việc lặp nhiều lần đến khi thuật toán hội tụ.

Initiate Graph

```
loop
```

Calculate ExpertRank

$$\delta = \sum_{E \in Experts} E(NewRank) - \sum_{E \in Experts} E(OldRank)$$

```
while(  $\delta > \epsilon$  )
```

5 SỰ HỘI TỤ

Biểu đồ dưới đây mô tả quá trình hội tụ khi hệ thống chạy với dữ liệu gồm có 250 người dùng và 434 bình chọn. Hệ thống đã hội tụ sau 16 lần lặp. Hơn nữa, theo như thuật toán PageRank thì tốc độ hội tụ thuật toán lặp tỉ lệ với logarit của số nốt trong mạng. Điều đó có nghĩa là PageRank và theo đó là ExpertRank cho phép mở rộng với dữ liệu lớn.



Hình 5. Biểu đồ sự hội tụ của ExpertRank

6 ÚNG DỤNG EXPERT-RANK TRÊN BKPROFILE

ExpertRank trên BkProfile ngoài việc được hiển thị trong hồ sơ cá nhân của người dùng (dạng đã được chuẩn hóa) còn được tích hợp với máy tìm kiếm văn bản mã nguồn mở có tên là SOLR để cho ra các kết quả tìm kiếm được sắp xếp không những theo độ phù hợp của từ khóa mà còn độ tốt của đối tượng tìm kiếm. Có ba ứng dụng tiêu biểu của việc kết hợp ExpertRank và máy tìm kiếm:

-Tiến cử người trả lời tiêm năng: Người trả lời tiêm năng trong một lĩnh vực trước tiên phải là người có chuyên môn đủ tốt nên ExpertRank được sử dụng trong bước lọc đầu tiên.

-Tuyển dụng: ExpertRank là một thông số quan trọng để đánh giá sự phù hợp của ứng viên cho một công việc nào đó.

-Câu trả lời tốt nhất: Nhờ ExpertRank, hệ thống có thể tính được chất lượng câu trả lời. Điều này cung cấp căn cứ cho người dùng chọn ra câu trả lời tốt nhất hoặc để người dùng chọn đọc các câu hỏi mà có các câu trả lời chất lượng cao.

7 THẢO LUẬN VÀ CÔNG VIỆC TIẾP THEO

Công thức mở rộng của ExpertRank có sự xuất hiện của đại lượng $E(u)$, có thể trở thành một công cụ để tùy biến hóa các kết quả trả về. Điều này trong một số trường hợp rất hữu ích. Ví dụ như khi nhận được một truy vấn từ một sinh viên Bách Khoa về việc tìm ra một chuyên gia về lĩnh vực Java thì hệ thống sẽ ưu tiên hơn các chuyên gia mà cũng thuộc trường Đại học Bách Khoa Hà Nội bằng cách sử dụng bộ chỉ mục với $E(u)$ lớn khi u là sinh viên Bách Khoa. Rõ ràng điều này có thể làm tăng khả năng câu hỏi được trả lời bởi vì mối quan hệ của hai người có thể kích thích người trả lời trả lời hơn.

Hướng tiếp cận của ExpertRank dựa trên việc gián tiếp liên kết giữa trọng số của tương tác của người dùng và chuyên môn của họ, ví dụ như nếu người dùng muốn bình chọn của mình có trọng lượng lớn hơn người khác thì cần phải trả lời nhiều câu hỏi với chất lượng cao để làm tăng chuyên môn của mình. Điều này phù hợp cho hệ thống dựa trên tương tác người dùng.

ExpertRank đánh giá chất lượng chuyên gia dựa trên quan hệ giữa các chuyên gia với nhau. Đây là một hướng mới mẻ và cho kết quả ổn định, công bằng. Tuy nhiên, việc áp dụng thêm các phương pháp xử lý ngôn ngữ tự nhiên như những gì Google Confucius và Google AardVark đã làm cũng là rất cần thiết để đánh giá chất lượng câu trả lời, từ đó đánh giá chuyên môn người trả lời một cách tốt hơn.

Một số chi tiết trong ExpertRank chưa được nêu cụ thể như cách thức lựa chọn tham số c trong công thức ExpertRank dạng đầy đủ, phương pháp hiệu quả kiểm tra tính hội tụ,... Trong BkProfile, chúng tôi sử dụng phương pháp hỏi ý kiến chuyên gia, là các bạn tình nguyện viên, để tìm ra những tham số gần đúng. Tuy nhiên, chúng ta có thể làm những công việc đó tốt hơn với những nghiên cứu tiếp theo.

Tóm lại, trong bài báo này, chúng tôi đã trình bày thuật toán lặp ExpertRank để đánh giá chuyên môn người dùng và sau đó sử dụng nó để đánh giá chất lượng câu trả lời. Giải pháp mà ExpertRank cung cấp là khách quan, không tôn kem và có thể mở rộng được. ExpertRank có thể được đóng gói thành một tham số tin cậy để kết hợp với các phương pháp khác để tạo

nên một giải pháp toàn diện hơn. Bên cạnh đó, chúng tôi đã nêu ra một số ứng dụng của ExpertRank trong các hệ thống Q&A, cũng như khả năng cá nhân hóa kết quả tìm kiếm rất độc đáo của ExpertRank. Cuối cùng, ExpertRank được thiết kế để triển khai trên các hệ thống phân tán như Hadoop MapReduce nên có thể ứng dụng cho các hệ thống hỏi đáp cỡ lớn.

8 LỜI TRI ÂN

Chúng tôi xin được gửi lời cảm ơn chân thành tới nhóm BKProfile, PGS. TS. Huỳnh Quyết Thắng và ThS. Lê Quốc đã tận tình hỗ trợ chúng tôi trong quá trình viết bài báo này.

9 TÀI LIỆU THAM KHẢO

- [1] NBrin, S. and Page, L. (1998) *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia, <http://infolab.stanford.edu/pub/papers/google.pdf>
- [2] Jon Kleinberg, Authoritative Sources in a Hyperlinked Environment, *Journal of the ACM (JACM, Volume 46 Issue 5, Sept. 1999,* <http://www.cs.cornell.edu/home/kleinber/auth.pdf>
- [3] Xianc Si, Edward Y. Chang, Zoltán Gyongyi, Maosong Sun, *Confucius and Its Intelligent Disciples: Integrating Social with Search, Proceedings of the VLDB Endowment, Volume 3 Issue 1-2, September 2010,* <http://infolab.stanford.edu/~echang/Confucius-VLDB10.pdf>
- [4] Damon Horowitz, Sepandar D. Kamvar, *The Anatomy of a Large Scale Social Search Engine, WWW '10 Proceedings of the 19th international conference on World wide web,* <http://vark.com/aardvarkFinalWWW2010.pdf>
- [5] Travers, Jeffrey & Stanley Milgram. 1969. "An Experimental Study of the Small World Problem." *Sociometry*, Vol. 32, No. 4, pp. 425-443.
- [6] A.A. Markov. "Extension of the limit theorems of probability theory to a sum of variables connected in a chain". reprinted in Appendix B of: R. Howard. *Dynamic Probabilistic Systems, volume 1: Markov Chains*. John Wiley and Sons, 1971
- [7] Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science* 178, 1972, www.garfield.library.upenn.edu/essays/V1p527y1962-73.pdf
- [8] Amit Goyal, Francesco Bonchi, Laks V. S. Lakshmanan, *Learning Influence Probabilities In Social Networks, WSDM '10 Proceedings of the third ACM international conference on Web search and data mining* <http://research.yahoo.com/files/wsdm339-goyal.pdf>

Chống trùng lặp địa danh trong hệ thống khai thác thông tin bất động sản

Nguyễn Trung Kiên, Đinh Anh Tuấn

Tóm tắt - Nhu cầu bán hoặc cho thuê bất động sản luôn nóng. Hiện nay, sự phát triển mạnh mẽ của mạng internet đã tạo điều kiện rất tốt cho sự phát triển của những giao dịch liên quan tới nhà đất. Người có nhu cầu bán hoặc cho thuê, có thể dễ dàng đăng tin rộng rãi trên mạng internet, còn người mua cũng có thể tra cứu được thông tin. Người đưa thông tin mong muốn chỉ dẫn cho người xem tới một địa chỉ duy nhất. Nhưng bởi tính đa nghĩa của tiếng Việt, cùng với thói quen đặt tên địa danh có sự trùng lặp, người đọc hoàn toàn có thể xác định sai vị trí mà người đưa thông tin hướng tới. Mặt khác, với thông tin dạng ngôn ngữ tự nhiên chúng ta không thể tận dụng được sức mạnh tính toán của máy tính để giúp xử lý những truy vấn phức tạp từ người dùng đến hệ thống khai thác thông tin bất động sản.

Với mong muốn áp dụng bài toán phát hiện địa danh trong văn bản và xử lý chống trùng lặp địa danh trong tiếng Việt vào thực tế để khắc phục lượng thông tin bất động sản to lớn, không ngừng gia tăng trên mạng internet. Đề tài nghiên cứu này đi sâu vào việc tìm hiểu các phương pháp để nhận diện địa danh nhắc tới trong văn bản, chống trùng lặp và đưa ra vị trí chính xác mà địa danh nhắc tới, các thuật toán đã, đang và có thể được áp dụng để giải quyết vấn đề trên nhằm giúp người sử dụng dễ dàng trong việc tìm kiếm thông tin, nghiên cứu cũng hướng tới việc tận dụng sức mạnh mới to lớn cho các hệ thống truy vấn thông tin địa lí được xây dựng trong tương lai

Từ khóa—Named Entity Recognition (NER), Support vector machine (SVM), Toponym Disambiguation, Map based.

1. GIỚI THIỆU

Chúng ta đã quen thuộc với khái niệm về địa điểm trong cuộc sống hàng ngày. Một địa điểm là không chỉ là một không gian mà còn mang trong nó những ý nghĩa riêng, ý nghĩa này phụ thuộc vào văn hóa, tục lệ và những quan điểm về địa điểm đó. Một thành phố là một địa điểm được giới hạn bởi ranh giới hành chính, hình thành một cộng đồng dân cư cư ngụ bên trong nó,

Công trình được thực hiện dưới sự hướng dẫn của thầy Hoàng Anh Việt, Bộ môn Công nghệ phần mềm, Viện CNTT&TT, ĐH Bách Khoa Hà Nội (email: hoanganhviet@gmail.com)

Nguyễn Trung Kiên, sinh viên lớp Công nghệ phần mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (e-mail: kiennghien.hut@gmail.com).

Đinh Anh Tuấn, sinh viên lớp Công nghệ phần mềm, khóa 52, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (e-mail: tuanad121@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

cũng là một không gian bởi bên trong nó chứa những tòa nhà và những địa điểm khác như đường xá, công viên. Mỗi ngày, chúng ta đi từ nơi này đến nơi kia để làm việc, học tập, gặp gỡ bạn bè, đi du lịch những khi rảnh rỗi và để hoàn thành rất nhiều những công việc khác. Thông tin chúng ta tiếp nhận hàng ngày thường về 1 sự kiện đã xảy ra ở nơi. Ta sẽ không thể đưa ra những hành động thích hợp khi không biết tên hay vị trí của địa điểm đó.

Ngày nay, các địa điểm được mô tả bằng các địa danh. Những địa danh có mặt khắp nơi trên web: hầu hết tin tức đều đề cập đến một nơi nào đó trên trái đất. Sự mơ hồ trong ngôn ngữ của chúng ta chính là thách thức lớn nhất trong lĩnh vực xử lý ngôn ngữ tự nhiên. Chỉ xét riêng các địa danh, sự mơ hồ có thể có rất nhiều kiểu: Một cái tên nhưng được sử dụng cho nhiều lớp thực thể khác nhau. Có thể lấy ví dụ: "London" có thể chỉ tên nhà văn "Jack London" hoặc thủ đô nước Anh. Hoặc nhiều thực thể thuộc cùng lớp có thể trùng tên (ví dụ: "London" cũng là một thành phố của Canada).

Áp dụng nói riêng cho bài toán thông tin mua bán bất động sản. Nhu cầu thuê hoặc mua bán bất động sản luôn gắn liền với một số yêu cầu cá nhân của riêng khách hàng. Người thuê mướn bằng bản hàng muốn đến gần với khách hàng, những vùng dân cư có thu nhập cao. Người mua nhà để ở muốn mua nhà ở những nơi có cơ sở hạ tầng phát triển có trường học, siêu thị, bệnh viện... Một công ty lớn sẽ chọn những vị trí đẹp trong thành phố để làm văn phòng, quảng bá hình ảnh... Tất cả đều dẫn đến một nhu cầu chung của khách hàng. Họ muốn biết vị trí chính xác của những bất động sản này, vị trí mà sẽ biểu diễn duy nhất 1 điểm trên bản đồ, việc quan sát trực tiếp trên bản đồ sẽ tạo điều kiện tốt nhất để ta nhận xét và đối sánh vị trí của các bất động sản với nhau và với nhu cầu của mình. Đoạn tin tức có thể như sau :

"Tôi muốn bán nhà tại đường Minh Khai, cách cầu Mai Động 200m"

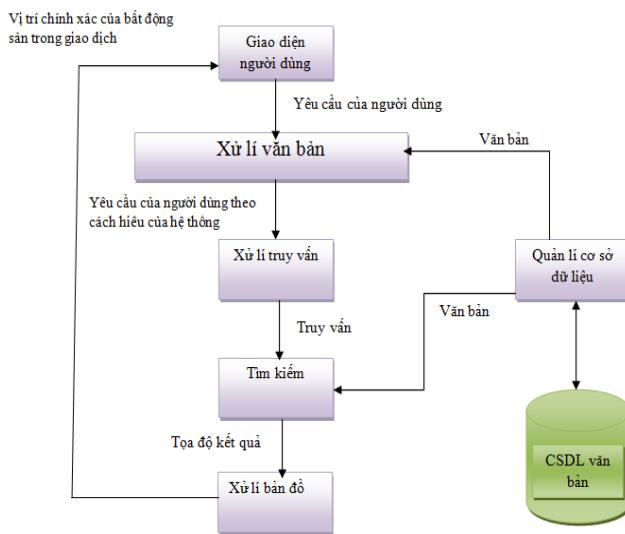
Công việc của bài báo là phát hiện ra các địa danh được nhắc tới trong đoạn tin mà người dùng đưa lên là "đường Minh Khai" và "cầu Mai Động". Tuy nhiên người đọc không phải ai cũng có thể phân biệt được "đường Minh Khai" là đường tại thành phố Hà Nội hay tại TP. Hồ Chí Minh, thậm chí là Đà Nẵng? vị trí trên bản đồ cụ thể như thế nào, có gần trung tâm hay không?

Trong khuôn khổ báo cáo này chúng tôi chỉ quan tâm tới phương pháp trắc địa chỉ sử dụng thông tin về tọa độ và khoảng cách giữa các điểm, phục vụ bài toán **chống sự trùng lặp địa danh trong văn bản tiếng Việt**. Ta sẽ cố gắng khám phá đặc điểm của những địa danh chứa trong tập thông tin mẫu, qua đó chỉ ra những khó khăn khi giải quyết những địa danh trùng lặp và những đặc điểm thực sự giá trị giúp chúng ta xử lý được sự mơ hồ trong việc phân biệt các địa danh trong văn bản Tiếng Việt.

2. MÔ HÌNH ĐỀ XUẤT

2.1. Mô hình chung

Mô hình đề xuất chung giả quyết bài toán nhập nhằng trong tìm kiếm bất động sản được mô tả như trong hình 1. Trong đó yêu cầu từ người dùng được đưa vào thông qua giao diện người dùng, qua bộ xử lý văn bản được ứng dụng các thuật toán nhằm tách lọc địa danh và yêu cầu của khách hàng, từ những yêu cầu chính được tách lọc sẽ thông qua bộ xử lý truy vấn – tìm kiếm đưa ra thông tin chính xác cho người tìm kiếm kết quả chi tiết hoặc gợi ý trên bản đồ.



Hình 1. Hệ thống tìm kiếm bất động sản và nhà đất để xuất

Hiện tại theo chúng tôi được biết, chưa có nghiên cứu nào tại Việt Nam để xuất xây dựng mô hình và dữ liệu cho bộ từ điển địa danh hoàn chỉnh, cùng với tập luật nhận biết địa danh.

Do đó, Trong khuôn khổ của bài báo, ta chỉ tập trung vào khái xử lý văn bản liên quan đến việc tách lọc các địa danh trong văn bản và các thuật toán chống nhập nhằng các địa danh này.

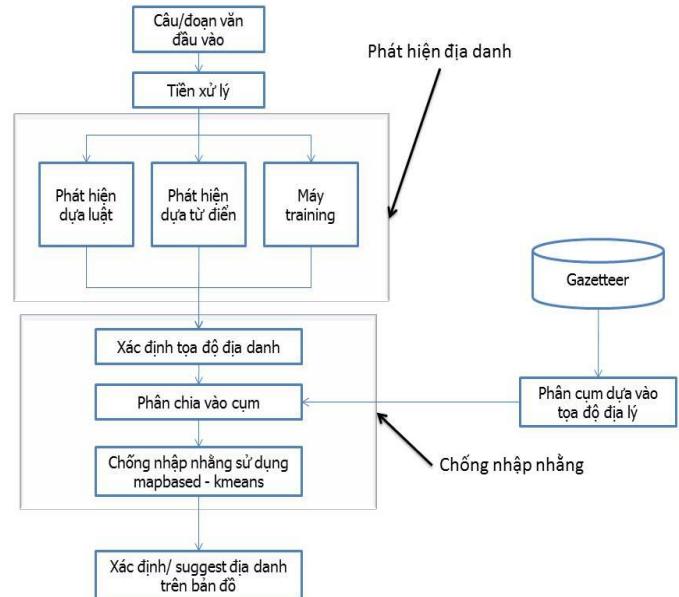
Với bài toán tách lọc địa danh trong văn bản tiếng Việt, chúng tôi đưa ra mô hình xử lý như trong hình 2, trong đó áp dụng các phương pháp:

- Phân tách địa danh từ văn bản :
- Xây dựng từ điển địa danh : đưa ra phương án xây dựng từ điển địa danh cho lãnh thổ Việt Nam và các địa danh trên thế giới. Bộ dữ liệu bao gồm hơn 12.000 địa danh trên lãnh thổ Việt Nam và khoảng 12 triệu dữ liệu địa danh trên thế giới
- Phương pháp phân tách dựa luật : Xây dựng và áp dụng hơn 40 luật dành riêng cho trích rút thông tin dành cho địa danh.
- Phương pháp học máy: Sử dụng phương pháp học máy máy hỗ trợ Vector (Support vector machine) nhằm làm chính xác hóa quá trình phát hiện và phân tách địa danh trong văn bản tiếng Việt.
- Chống nhập nhằng địa danh :
- Phương pháp map-based chống nhập nhằng địa danh: Map-based là tập những phương pháp sử dụng những mô

tả tường minh của địa điểm trên bản đồ. Một trong những phương pháp này được Smith và Crane đưa ra vào năm 2001, dựa vào tọa độ địa lý của địa điểm để xóa bỏ mờ hồ cho địa điểm đó: Các vị trí xuất hiện cùng với địa danh mờ hồ trong văn bản (được gọi là ngữ cảnh) được định vị trên bản đồ, được đánh trọng số theo số lần chúng xuất hiện. Sau đó ta sẽ tính toán trọng tâm cho cụm những địa danh này rồi so sánh với những vị trí địa danh mờ hồ có thể nhận. Vị trí nằm gần tâm này nhất sẽ được lựa chọn là nơi mà địa danh nhắc đến. Bài báo này sẽ đề xuất một phương pháp map-based dựa vào thuật toán k-means.

- Thuật toán K-means trong phân cụm địa danh : K-means là thuật toán được sử dụng rộng rãi trong phân cụm dữ liệu. Bài báo này sẽ sử dụng K-means để chia các địa danh trong từ điển dữ liệu thành K cụm địa danh Việt Nam theo tọa độ của chúng. Mỗi cụm này sẽ chứa các địa danh nằm sát nhau trên bản đồ.

- Phương pháp map-based đề xuất: Ta dựa trên cơ sở trong các bản tin bất động sản, do người viết muốn chỉ dẫn cho người đọc tới 1 địa điểm xác định.. Các địa danh được họ sử dụng trong bản tin sẽ tập trung xung quanh vị trí này. Nên với các địa danh trong bản tin nhà đất, ta dựa vào từ điển địa danh liệt kê mọi khả năng (tọa độ địa điểm) mà chúng có thể nhận. Sau đó ta phân các khả năng này vào K cụm địa danh Việt Nam bằng cách xét vị trí của chúng gần tâm của cụm nào nhất. Ta tìm cụm địa danh chứa nhiều khả năng nhất, những khả năng thuộc cụm đó chính là những kết quả ta cần tìm.



Hình 2. Vị trí của Chống trùng lặp trong hệ thống

2.2. Phân tách địa danh

2.2.1. Từ điển địa danh

Qua tìm hiểu, hiện nay tại Việt Nam chưa có đề xuất nào liên quan đến từ điển địa danh dành cho văn bản tiếng Việt, việc xây dựng đưa ra một chuẩn mới, có thể áp dụng cho các bài toán nghiên cứu sau này. Áp dụng trong bài toán chống trùng lặp địa danh cụ thể, từ điển được xây dựng với hơn **12.000** địa danh

trên lãnh thổ Việt Nam và hơn **12 triệu** địa danh trên thế giới có cấu trúc như sau :

Tên địa danh – mã địa danh – tọa độ

Có thể lấy ví dụ một địa danh như :

Hoàn kiếm VN01805 ADM2 21.0375135 105.8491500

Trong đó :

- Tên địa danh là tên gọi dưới dạng tiếng Việt của địa danh
- Mã địa danh bao gồm :
 - Mã quốc gia : 2 kí tự theo chuẩn quốc tế **ISO 3166-1** 2-letter country code. Ví dụ : Việt nam có mã là VN. Bảng mã có thể tham khảo tóm tắt dưới hình 4.
 - Mã đơn vị hành chính cấp 1 : tương đương với cấp Tỉnh tại Việt Nam
 - Mã đơn vị hành chính cấp 2 : tương đương với đơn vị quận/huyện tại Việt Nam
 - Mã đơn vị hành chính cấp 3: tương đương với xã/phường tại Việt Nam
 - Mã đơn vị hành chính cấp 4 : tương đương với làng/xóm/bản tại Việt Nam (Có thể có hoặc không)
 - Loại hình địa danh : Mã mô tả loại hình của địa danh, có bảng hướng dẫn đi kèm theo khi đọc : Ví vụ PPL : vùng có người sinh sống...Bảng tóm tắt như dưới bảng 5.
- Tọa độ : Bao gồm kinh/vĩ độ của địa danh, được thu thập thông qua môi trường internet, với sự hỗ trợ của Google APIs

VIET NAM	VN
VANUATU	VU
VENEZUELA	VE
AUSTRALIA	AU
UNITED STATES	US
JAPAN	JP

Bảng 4. Bảng mã quốc gia 2 kí tự theo chuẩn ISO 3166-1

ADM1	Đơn vị hành chính cấp 1 – Tỉnh Việt Nam
ADM2	Đơn vị hành chính cấp 2 – Huyện Việt Nam
SEA	Biển, đại dương
PPL	Khu vực sinh sống phô biển
CHN	Con kênh

Bảng 5. Bảng tóm tắt loại hình địa danh

Phương pháp xây dựng từ điển địa danh này có những ưu điểm như sau:

- Chỉ ra được quan hệ cha con trong quan hệ giữa các địa danh. Như : quận Hoàn Kiếm thuộc TP.Hà Nội, Tp.Hà Nội thuộc nước Việt Nam
- Chỉ ra được loại hình địa danh đang nói tới : Minh Khai là tên đường, Cửu Long là một con sông, Tân Viên là tên của một ngọn núi...
- Đưa ra được tọa độ chính xác cho từng địa danh phân biệt
- Đưa ra được phương pháp bổ sung địa danh, phương

pháp xác định tọa độ dựa vào Google API, các API thao tác trên bộ từ điển

2.2.2. Phương pháp phân tách dựa luật

Đưa ra hơn 40 luật trích rút địa danh trong văn bản tiếng Việt, trong đó chủ yếu sử dụng phương pháp phát hiện sử dụng tiền tố và hậu tố dành cho địa danh. Trong tiếng Anh, người ta sử dụng các hậu tố như “city”, “river”, “town”... để phát hiện các địa danh. Cụ thể các tiền tố trong tiếng Việt có thể mô tả được như

Tôi đang đi trên đường Minh Khai, phía bên gần sông Hồng. Thị tiền tố “đường” và “sông” chính là 2 tiền tố giúp phát hiện ra các địa danh đi sau nó.

Phương pháp dựa luật sử dụng thêm 2 bộ phân tích từ là : JvnSegmenter (1) và VietTagger (2) nhằm nâng cao độ chính xác của việc phát hiện. Cụ thể JvnSegmenter sẽ giúp tách đoạn văn thành các từ tiếng Việt, và VietTagger sẽ gán nhãn loại từ sau khi phân tách.

Chúng tôi đưa ra khái niệm gọi là trọng số của từ so với tiền tố hay hậu tố trong quá trình nhận biết. Trọng số này là khoảng cách số từ tính từ tiền tố đến danh từ riêng (Np) đứng sau nó. Cụ thể

Tôi	Có	Chung cư	Tại	Khu vực	Minh Khai
			Tiền tố	*	Np

Thì số lượng từ trong ô * sẽ được tính là trọng số của tiền tố đến danh từ riêng có khả năng là địa danh đứng đằng sau nó.

Trọng số này sẽ hỗ trợ luật đưa ra được khả năng có được địa danh một cách chính xác hơn. Một trọng số bao gồm khả năng phát hiện từ là địa danh chính xác và khoảng có khả năng chính xác lớn. Việc đưa ra trọng số trong bộ luật thực hiện có thể tùy chỉnh được trong quá trình cho chương trình học từ tập dữ liệu đã được gán sẵn nhằm cải thiện độ chính xác của luật. Một luật sẽ được mô tả trong file luật rule.xml như sau

```
<rule name="prefix">
  <expression>
    <name> thành phố </name>
    <type>ADM1</type>
    <weight> 0.3 </weight>
  </expression>
</rule>
```

Như trong ví dụ dưới đây, khả năng chính xác đối với trọng số có tiền tố là 0, có độ chính xác cao hơn nếu trọng số nằm trong khoảng từ 0 đến 3, còn nếu trọng số lớn hơn 3, khả năng từ là địa danh sẽ được xếp vào loại có độ chính xác không cao

Mã <type> là loại hình của địa danh đứng sau tiền tố, có thể tra thông tin như trong bảng 5.

2.2.3. Phương pháp sử dụng học máy

Có nhiều phương pháp học máy cung cấp khả năng học cho phép phát hiện các Named entity nói chung và địa danh nói riêng như : Conditional Random fields (CRF), Maximum Entropy (ME), Support vector machine (SVM). Tuy nhiên SVM là phương pháp cung cấp khả năng phân lớp với số chiều lớn, nhanh và có hiệu suất tổng hợp, hiệu suất tính toán cao, một ưu điểm khác là SVM giải quyết được vấn đề overfitting rất tốt.

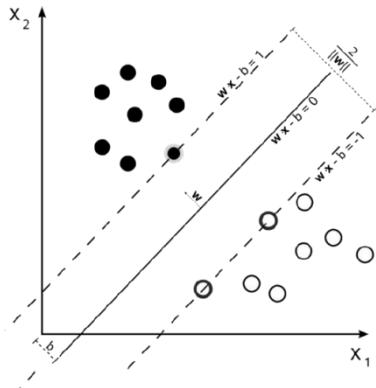
Trong phương pháp sử dụng học máy cho việc phân tách địa danh. Chúng tôi sử dụng phương pháp phân lớp sử dụng máy Vector hỗ trợ (SVM) sử dụng vector siêu phẳng N - chiều. Để mô tả đơn giản hơn, trong phần dưới đây, chúng tôi sẽ chỉ mô tả SVM đối với phân lớp nhị phân

Giả định rằng chúng ta đã có một tập học được gán nhãn là $D = \{(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)\}$

Với $y_i \in \{-1, 1\}$ là một số nguyên xác định lớp của x_i , trong đó x_i là một mô tả dưới dạng vector đặc trưng của một đoạn văn bản đầu vào. Bộ phân lớp tuyến tính sẽ được duyệt thông qua một siêu phẳng có dạng

$$f(x) = w \cdot x - b = 0$$

Trong đó w là vector pháp tuyến của siêu phẳng còn b là tham số mô hình



Mục đích là tìm ra hàm quyết định rằng với một vector đầu vào (chứa các đặc trưng tình huống của đoạn văn tiếng Việt cần xác định địa danh). Hàm quyết định này đưa ra được rằng vector đầu vào ấy thuộc lớp nào $f(x) = 1$ và ngược lại.

Trong quá trình sử dụng SVM, chúng tôi sử dụng bộ thư viện libSVM (3). libSVM là một bộ thư viện do nhiều tác giả thực hiện, trong đó phần java do các thành viên đại học quốc gia Đài Loan (National Taiwan University) viết, cung cấp thư viện gắn vào các chương trình khác, cho phép việc training và phân lớp thông qua các hàm viết sẵn, thư viện java được cung cấp sẵn với các đặc trưng đầu vào dành cho quá trình học và phân lớp bao gồm:

- Có tiền tố hay không tương đương với giá trị số 1 và 0
- Có danh từ riêng đứng sau tiền tố hay không
- Trọng số của danh từ riêng đối với tiền tố

2.3. Chống trùng lắp địa danh

Đóng góp lớn nhất là việc sử dụng linh hoạt thuật toán k-mean để loại bỏ những cõi lặp ra những vị trí đúng đắn nhất mà các địa danh mờ hồ có thể nhận. Ban đầu ta sẽ phân thành các cụm lớn để loại dần các khả năng ở xa vị trí cần tìm, do cụm này có kích thước lớn vẫn chứa những khả năng trùng lắp nhau, ta sẽ tiếp tục sử dụng thuật toán kmeans để thu được cụm địa danh nhỏ hơn.

Kết quả ta thu được là 1 bộ địa danh duy nhất, loại bỏ mọi khả năng trùng lắp không chính xác.

3. DỮ LIỆU THỬ NGHIỆM

Dữ liệu thử nghiệm đầu vào cho quá trình nhận biết và chống trùng lắp địa danh của chúng tôi bao gồm

- 1) Bộ dữ liệu địa danh đi kèm tọa độ lấy từ google maps

với khoảng 1000 dữ liệu làm đầu vào cho quá trình chống trùng lắp địa danh.

- 2) Dữ liệu các câu đầu vào chứa địa danh đang trong quá trình xây dựng. Hiện tại có khoảng 500 câu đầu vào dành cho quá trình thử nghiệm
- 3) Từ điển địa danh bao gồm 12.000 địa danh trên lãnh thổ Việt Nam và hơn 12 triệu địa danh tại các quốc gia trên thế giới.
- 4) Từ điển tiếng Việt phục vụ cho quá trình phân tách địa danh được bổ sung nhiều hơn của JvnSegmenter nâng cao độ chính xác, bao gồm hơn 74.000 từ.
- 5) Bộ dữ liệu đầu vào cho việc training phát hiện địa danh, được thu thập và gán nhãn bằng tay. đang trong quá trình xây dựng với cấu trúc xây dựng như hình 3 dưới đây.

Từ	Phong cách viết	Loại từ	Từ điển địa danh	Trọng số	Nhãn
Tôi	Hoa đầu từ	P	NON	-1	
có	Thường	T	NON	-1	
chung_cư	Thường	N	NON	-1	
trên	Thường	A	NON	-1	
đường	Thường	N	NON	-1	prefix
Minh_Khai	Hoa toàn bộ	NP	LOCATION	0	LOCATION

- Đối với quá trình chống trùng lắp địa danh, dữ liệu đầu vào sử dụng hơn 12.000 địa danh trong từ điển làm đầu vào xác định trùng lắp. Các dữ liệu địa danh này được tiến hành phân cụm thông qua bước xử lý đầu vào bằng thuật toán Kmeans, chia toàn bộ lãnh thổ Việt Nam thành các cụm có chứa địa danh, cụm ít địa danh sẽ được phân thành một cụm lớn, các cụm chứa nhiều địa danh sẽ tùy vào mức độ sử dụng cũng như yêu cầu xác định đầu vào sẽ tiếp tục được phân cụm tới mức nhỏ hơn.

- Thủ nghiệm với địa chỉ sau: số 1 Đại Cồ Việt, Hai Bà Trưng, Hà Nội.

Bước 1: Tra cứu trong từ điển địa danh ta thu được các khả năng sau:

STT	Địa danh	Ghi chú
1	Đại Cồ Việt	Thuộc thủ đô Hà Nội
2	Hà Nội	Hà Tây cũ
3	Hà Nội	Thủ đô Hà Nội
4	Hai Bà Trưng	Thuộc Hà Tây cũ
5	Hai Bà Trưng	Thuộc thủ đô Hà Nội
6	Hai Bà Trưng	Thuộc thành phố Hồ Chí Minh

Bước 2: Dùng thuật toán k-means phân các địa danh của

Việt Nam thành 8 cụm. Tính toán ra tâm của các cụm đó.

Tam_i(x_i, y_i): tâm cụm i, có tọa độ x_i, y_i

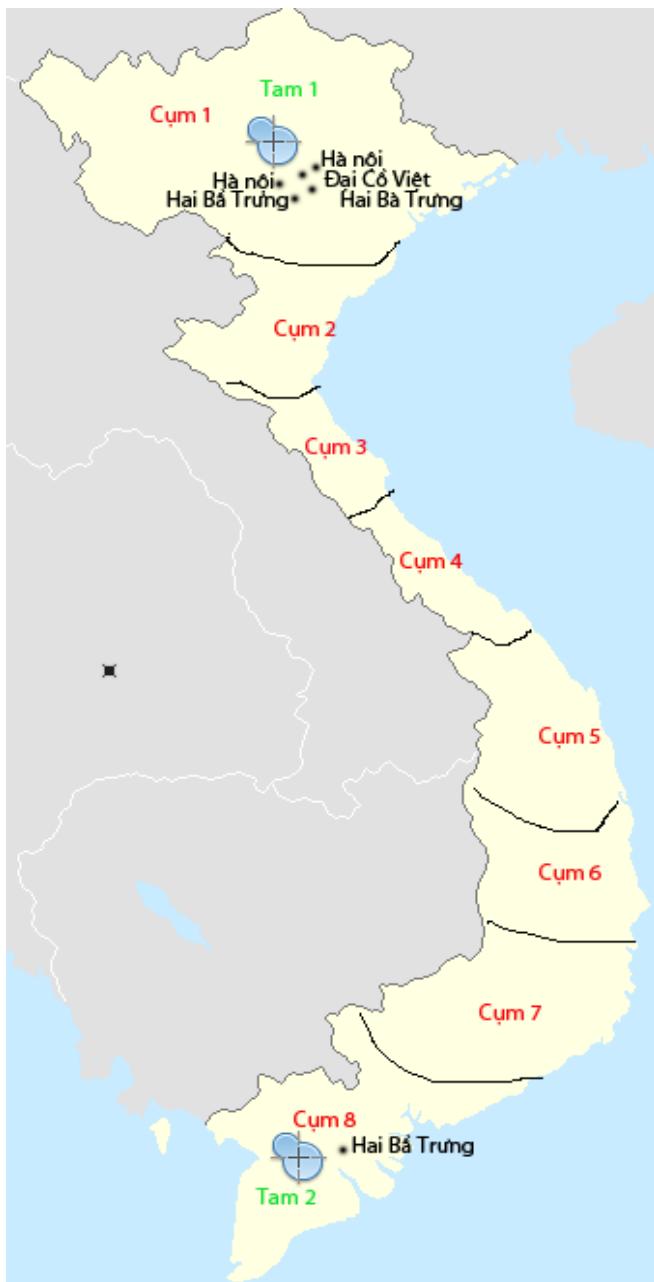
$$x_i = \frac{\sum_1^n x}{n} \quad y_i = \frac{\sum_1^n y}{n}$$

$\sum_1^n x$ là tổng hoành độ của mọi điểm thuộc cluster i

$\sum_1^n y$ là tổng tung độ của mọi điểm thuộc cluster i

n: số điểm thuộc cluster i

Bước 3: Phân các khả năng vào 1 trong 8 cụm trên dựa vào việc tính toán khoảng cách giữa chúng và tâm cụm.



Ta chọn cụm 1 vì nó chứa nhiều khả năng nhất.

Bước 4: Do cụm 1 còn tồn tại các địa danh trùng nhau như Hà Nội và Hải Bà Trưng nên ta tiếp tục phân nhỏ nó thành 2 cụm



Ta lựa chọn cụm 1 vì nó chứa nhiều khả năng hơn.

4. KẾT QUẢ THỬ NGHIỆM

4.1. Quá trình chống nhập nhằng địa danh

Với bộ dữ liệu cho thuật toán kmeans là từ điển dữ liệu, cùng với các địa danh nhập nhằng được đưa vào từ danh sách địa danh được lấy từ google maps, các địa danh nhập nhằng ta đã loại bỏ thành công được các địa danh trùng lặp cuối cùng chỉ thu về 1 tập các địa danh duy nhất đúng với kết quả dự kiến.

Ta gọi tập các địa danh cùng xuất hiện với địa danh mơ hồ trong văn bản là ngữ cảnh của địa danh đó. Thử nghiệm cho thấy phương pháp này càng chính xác khi kích thước ngữ cảnh càng lớn và trong tập ngữ cảnh có địa danh không mơ hồ. Nếu địa danh mơ hồ đứng riêng 1 mình thì khả năng gỡ bỏ nhập nhằng là không thể.

Kích thước ngữ cảnh	Số thử nghiệm	Số nhận biết thành công
0	50	2
1	50	20
≥ 2	50	42

- Khi ngữ cảnh bằng 0, ta chỉ nhận biết được chính xác vị trí cho địa danh khi bắn thân địa danh đó không mơ hồ.

- Khi ngữ cảnh bằng 1, địa danh mơ hồ được nhận biết thành công khi ngữ cảnh của nó là 1 địa danh rõ ràng. Ví dụ trong trường hợp: địa danh mơ hồ Hoàng Hoa Thám có ngữ cảnh {Hà Nội} thì Hà Nội là địa danh quá rõ ràng khi được dùng để giải thích cho Hoàng Hoa Thám ở đâu.

- Khi ngữ cảnh lớn hơn 2, ví dụ: địa danh mơ hồ Đại Cồ Việt có ngữ cảnh {Hai Bà Trưng, Hà Nội}. Những trường hợp sai ở đây là do trong ngữ cảnh chỉ chứa những địa danh mơ hồ.

4.2. Quá trình phát hiện địa danh

Kết quả được chúng tôi đánh giá thông qua các độ đo thông dụng như sau

- Precision (P): Số lượng kết quả chính xác chia cho tổng số nhãn gán được
- Recall(R): Số lượng kết quả chính xác chia cho tổng số nhãn đúng mong muốn
- Độ đo cân bằng F-measure = $2RP/(R+P)$

Qua quá trình thử nghiệm với 50 câu đầu vào chứa 50 địa danh được lấy từ trang thông tin điện tử Tuoitre có chứa địa danh trong từ điển. Khả năng phát hiện đối với các địa danh đạt 41 địa danh nhận biết chính xác và nhận biết nhầm 8 địa danh. Các địa danh được thử nghiệm với 50 câu đầu vào, trong đó nhận biết chính xác được 38 địa danh.

Các sai số không nhận biết được địa danh, nhận biết nhầm các từ không phải địa danh nhưng vẫn gán nhãn là địa danh có nguyên nhân một phần khi sử dụng bộ phân tích từ JvnSegmenter hoạt động không chính xác dẫn đến sai số trong

quá trình nhận biết, hầu hết sai số nằm tại các từ phụ thuộc vào lỗi văn nói, thói quen của người viết và không có trong từ điển tiếng Việt. Một số ví dụ nhận biết chính xác và nhầm lẫn tiêu biểu được mô tả trong bảng 7

Số lượng câu thử nghiệm	Địa danh có trong từ điển	Số nhận biết đúng	Không nhận biết được	Số nhận biết nhầm
50	✓	41	9	8
50	X	38	12	11

Bảng 6. Thống kê số lượng nhận biết

Câu đầu vào	Kết quả	Không nhận biết được	Nhận biết nhầm
Ông Phạm Bá Điểm, phó chủ tịch UBND huyện vùng cao biên giới Mường Lát - một địa phương trọng điểm của thực trạng thiêu đói lương thực ở Thanh Hóa	Mường Lát, Thanh Hóa		
Hàng ngàn người dân ở các tỉnh, thành phố lân cận đã đổ về TP.HCM bắt đầu làm việc, chủ yếu là xe máy, khiến cho tình hình giao thông thành phố phức tạp	TP, HCM		TP
Tại khu vực ngã tư Bình Triệu , ngay sau khi tình hình hìnìn ùn tắc xe xảy ra, lực lượng cảnh sát giao thông đã có mặt để phân luồng, điều tiết giao thông.	Bình Triệu		
Các nước phát triển trên thế giới rất coi trọng việc tuyên dụng vào khu vực này, điển hình như Hàn Quốc và Singapore	Hàn Quốc, Singapore, trên thế		Trên thế

Bảng 7. Chi tiết đầu vào nhận biết địa danh

Với bộ dữ liệu thử nghiệm, chúng tôi tính toán được kết quả thông qua các độ đo Precision (P), Recall(R), F-measure (F) như sau

	P (%)	R (%)	F
Nhận biết địa danh	79.79	79.00	79.39

Bảng 8. Kết quả theo độ đo

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo nhằm giới thiệu một hướng tiếp cận mới trong việc hỗ trợ các hệ thống thu thập và trích rút thông tin đối với bất động sản nói riêng và các hệ thống trích rút thông tin nói chung, một trong những bài toán nhỏ của hệ thống nhận dạng thực thể dành cho tiếng Việt. Nghiên cứu của chúng tôi có thể được áp

dụng hiệu quả vào hệ thống thu thập thông tin bất động sản mà mục tiêu là đưa đến cho người dùng khả năng truy xuất thông tin từ internet một cách chính xác và nhanh chóng nhất với khối lượng thông tin lớn.

Với một số kết quả đạt được bao gồm:

- Dựa ra được phương pháp cấu trúc bộ từ điển địa danh, phương pháp xây dựng và 12.000 dữ liệu trên lãnh thổ Việt Nam và hơn 12 triệu dữ liệu cho toàn thế giới.
- Dựa ra được bộ luật gồm hơn 40 luật phục vụ cho việc trích rút địa danh từ văn bản tiếng Việt dựa vào luật
- Dựa ra được mô hình trong đó tận dụng được khả năng tính toán của máy tính vào việc chuẩn hóa các thông tin về địa danh trên internet.
- Đề ra được mô hình xác định duy nhất sự tồn tại của địa danh hoặc gợi ý địa danh cho người dùng thông qua sự áp dụng linh hoạt thuật toán Kmeans và phương pháp dựa bản đồ
- Xây dựng được bộ thư viện cho phép áp dụng và tách lọc thông tin áp dụng trong các hệ thống thu thập và trích rút thông tin bất động sản.

Mặc dù đã đạt được những kết quả nhỏ, nhưng trong tương lai chúng tôi mong muốn đạt được những kết quả có tính chất quan và có giá trị từ những nghiên cứu này, tập trung vào việc cải thiện độ chính xác cho quá trình phân tách địa danh và việc chống trùng lặp địa danh

- Từ điển địa danh: Từ điển địa danh là một phần quan trọng trong nghiên cứu, nhằm mục tiêu đánh giá chỉ đến từng địa danh nhỏ nhất trên toàn lãnh thổ Việt Nam và xa hơn là quốc tế. Trong tương lai, chúng tôi mong muốn mở rộng từ điển địa danh thông qua sự bổ sung bằng do cộng đồng đóng góp, các Webcrawler giúp xây dựng một từ điển đầy đủ phục vụ cho các nghiên cứu sau này
- Cải thiện tốc độ phân tách và chống trùng lặp thông qua cải thiện tốc độ thực hiện các quá trình phân tách từ, gán nhãn từ và phân cụm địa danh. Cải thiện độ chính xác của quá trình nhận biết và chống trùng địa danh
- Kết hợp và so sánh độ chính xác của các phương pháp áp dụng trong nghiên cứu, như so sánh phương pháp SVM với các phương pháp khác tương đương như CRF, ME..., Phương pháp phân cụm Kmeans và các phương pháp tương đương khác

ACKNOWLEDGEMENT

Chúng tôi xin gửi lời cảm ơn đến thầy Hoàng Anh Việt, các thầy cô trong Bộ môn Công Nghệ Phần Mềm, Viện Công Nghệ Thông Tin và Truyền Thông, đại học Bách Khoa Hà Nội vì những lời khuyên và chỉ bảo về mặt kiến thức cũng như tinh thần trong suốt quá trình nhóm thực hiện đề tài.

Xin gửi lời cảm ơn đến nhóm phát triển JvnSegmenter, VietTagger vì những công cụ đóng góp trong mô hình của chúng tôi.

6. TÀI LIỆU THAM KHẢO

- [1] Tri Tran Q, Thao Pham TX, Hung Ngo Q, Dien Dinh, Nigel Collier - “Named entity recognition in Vietnamese”, 2007.
- [2] Thanh Nghi Do, Jean Daniel Fekete - “Large Scale Classification with Support Vector Machine Algorithms”, INRIA Futurs, Univ, Paris-Sud

- [3] Andrei Mikheev, Marc Moens, Claire Grover -“Named entity recognition without Gazetteer”, *HCRC Language group, University of Edinburgh, UK*, 1999
- [4] Chh Wei Hsu, Chj Chung Chang, Chih Jen Lin - “A Practical Guide to Support Vector Classification”, *National Taiwan University, Taipei 106, Taiwan*, 2003
- [5] Andrew W. Moore- “Support VectorMachines” , *School of Computer ScienceCarnegie Mellon University*, 2003
- [6] David A. Smith and Gregory Crane. Disambiguation geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries, volume 2136 of Lecture Notes in Computer Science*, page 127-137, Springer, Berlin,2001
- [7] VDC Goldenkey - <http://danhba.vdc.com.vn>
- [8] Geonames - <http://www.geonames.org/>
- [9] LibSVM - <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

Xây dựng JOO framework

Chuẩn hóa mô hình lập trình ứng dụng web trên hệ thống phân tán cỡ lớn

Bùi Kim Dũng, Bùi Anh Dũng, Bùi Trung Hiếu

Tóm tắt - Mục tiêu của chúng tôi là xây dựng JOO - framework nhằm chuẩn hóa lại mô hình lập trình các ứng dụng web, quản lý và tối ưu hóa hiệu năng xử lý dữ liệu ở máy chủ, truyền dữ liệu phân tán qua mạng... JOO ra đời giải quyết những trở ngại còn tồn tại trong quá trình xây dựng các ứng dụng này: có khả năng tương đương với những ứng dụng thông thường trên máy tính cá nhân, tương thích với nhiều loại nền tảng và phần cứng, hiệu năng truyền tải dữ liệu qua mạng, hiệu năng xử lý dữ liệu ở máy chủ, khả năng phân tán... Chúng tôi đã nghiên cứu trên các mô hình lập trình web truyền thống và một số framework tiêu biểu, sau đó thiết kế các giải pháp nhằm khắc phục những điểm yếu của chúng. Kiến trúc hệ thống của JOO là kết quả của quá trình nghiên cứu và tích hợp các giải pháp thu được – đáp ứng hầu hết các thuộc tính về mặt chất lượng của ứng dụng web. JOO đồng thời chuẩn hóa mô hình lập trình và mô hình xử lý dữ liệu ở các ứng dụng loại này. JOO đã được triển khai thực tế trên hệ thống BKProfile, hệ thống kết nối chia sẻ tri thức. Với những lí do trên, JOO có khả năng tạo ra một hướng tiếp cận hiệu quả cho việc lập trình ứng dụng web, đáp ứng được những hệ thống phân tán cỡ lớn, đồng thời giải quyết những trở ngại trong lĩnh vực này.

Từ khóa - Ajax-based Single-page Web Applications, Javascript object oriented, JOOframework, Web application standard.

1. GIỚI THIỆU

Ngày nay, cùng với sự phổ biến của Internet, sự phát triển của các loại thiết bị điện tử cá nhân có sức mạnh tính toán cao và xu hướng công nghệ điện toán đám mây, các ứng dụng web đa nền ngày càng trở nên phổ biến: dữ liệu được đặt ở các máy chủ phân

Công trình này được thực hiện dưới sự bảo trợ của nhóm BkProfile, <http://www.bkprofile.com>.

Bùi Kim Dũng, sinh viên lớp Công nghệ phần mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 84-904-924-914, e-mail: kimdung@bkprofile.com).

Bùi Anh Dũng, sinh viên lớp Công nghệ phần mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 84-1275-065-837, e-mail: dungba@bkprofile.com).

Bùi Trung Hiếu, sinh viên lớp Công nghệ phần mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 84-915-585-266, e-mail: hieubui@bkprofile.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

tán, ứng dụng được truy cập và tất cả các thao tác được thực thi trên các trình duyệt, thông tin được truyền qua mạng giữa máy khách với nhiều máy chủ. Tuy nhiên việc xây dựng các ứng dụng này còn gặp rất nhiều trở ngại: khả năng tương tác với người dùng rất phức tạp do phải đáp ứng nhiều loại nền tảng phần cứng khác nhau, hiệu năng xử lý dữ liệu máy chủ cũng như hiệu năng truyền tải dữ liệu còn rất thấp (tốc độ tải trang trung bình xấp xỉ 4s) do mô hình xử lý dữ liệu truyền thống còn nhiều điểm bất hợp lý.

Từ khi mô hình truyền tải dữ liệu bắt đầu lần đầu tiên được triển khai thực tế trong bộ Java Applet của Java vào năm 1995, đã có rất nhiều công nghệ dựa trên mô hình này được phát triển như mô hình Spar[7] để xây dựng ứng dụng web dạng một trang (Single-page application)[1], framework GWT của Google[2], các nền tảng RIA như Flash, Java Applet ... Những nền tảng này đã giải quyết được một phần lớn những trở ngại kể trên.

Dựa trên cơ sở đó, chúng tôi đã nghiên cứu những công nghệ có sẵn đã nói ở trên, phân tích các ưu nhược điểm của chúng để đề xuất ra một giải pháp có tính thuyết phục thông qua việc xây dựng framework JOO với kiến trúc tốt, hỗ trợ việc xây dựng ứng dụng web đa nền một cách hiệu quả, đã được ứng dụng và kiểm nghiệm thực tế trên hệ thống BKProfile – Hệ thống chia sẻ tri thức (www.bkprofile.com), và một số hệ thống khác.

Trong khi những mô hình trước thường cần nhiều điều kiện đặc biệt kèm theo, như Java Applet và Flash cần phải cài thêm plugin cho trình duyệt và không hoàn toàn chạy tốt trên mọi nền tảng (Flash chạy trên Linux rất nặng và kém chất lượng), hay như GWT yêu cầu phải có một máy chủ Java Servlet trong quá trình phát triển và chỉ hỗ trợ tốt cho Java... thì JOO framework đã khắc phục được hầu hết những nhược điểm đó. JOO framework được xây dựng hoàn toàn bằng Javascript, HTML5 và CSS3, dựa vào đó cơ chế lập trình hướng đối tượng đã được chuẩn hóa và thân thiện hơn, mô hình xây dựng ứng dụng web dễ dàng cấu hình và mở rộng, một thư viện các thành phần hiển thị đã được tích hợp sẵn và nhiều ưu điểm được kế thừa từ những nền tảng trước.

Tóm lại, trong bài nghiên cứu này chúng tôi đề xuất một framework để lập trình các ứng dụng web một cách nhanh chóng đồng thời chuẩn hóa quá trình hoạt động của ứng dụng nhằm cải thiện tốc độ, khả năng tương tác người dùng, nâng cao hiệu suất truyền tải và hiệu năng phía máy chủ.

2. PHÂN TÍCH ƯU NHƯỢC ĐIỂM CỦA CÁC CÔNG NGHỆ TRƯỚC

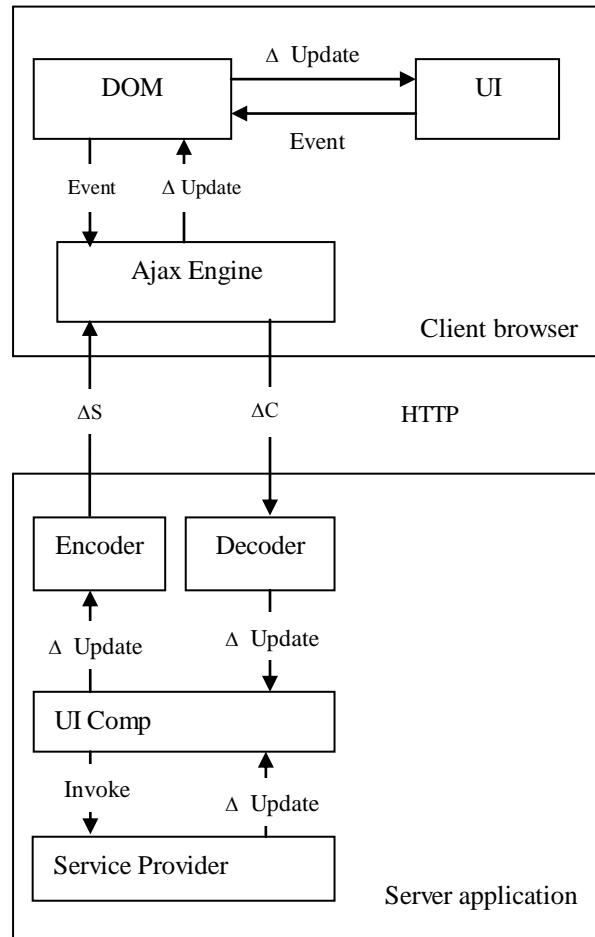
Trước khi phân tích ưu nhược điểm của các framework trước, chúng ta sẽ xem xét lại mô hình lập trình web truyền thống, hay còn gọi là mô hình lập trình web nhiều trang (multi-page website)[2].

Một website thông thường được cấu tạo từ nhiều trang, mỗi trang phục vụ cho một mục đích khác nhau (ví dụ: một website bán hàng trực tuyến thường có các trang để tìm kiếm hàng hóa, trang để đặt mua qua mạng và trang để thanh toán). Mô hình này hoạt động theo cơ chế thay đổi trạng thái giữa các trang. Người dùng từ trang này sẽ truyền đến trang khác bằng các đường link. Mỗi khi truyền trạng thái, phía máy chủ sẽ xử lý lại toàn bộ trang web. Mô hình này tuy có ưu điểm là việc lập trình đơn giản, tuy nhiên lại có nhiều nhược điểm về hiệu năng của hệ thống:

- Mỗi khi chuyển trang, phía máy chủ và máy khách đều phải xử lý lại toàn bộ trang web: Phía máy chủ phải xử lý lại trang web và gửi lại máy khách nội dung, phía máy khách thì phải tải lại HTML và xử lý lại javascript. Trong khi đó, thông thường các trang web đều có những thành phần không thay đổi qua các trang.
- Dữ liệu truyền giữa máy chủ và máy khách rất lớn (bao gồm cả dữ liệu HTML, javascript, CSS và các thành phần liên quan). Việc lập trình các trang web “động” không tận dụng được cơ chế cache của browser.
- Khả năng tương tác người dùng không cao, do người dùng phải đợi một thời gian lâu khi chuyển trang.

Cơ chế tải nội dung trang web không đồng bộ (async loading of content) lần đầu tiên được ứng dụng thực tế trong bộ Java Applet của Java năm 1995. Kể từ đó đã có nhiều công nghệ mới được phát triển dựa trên cơ chế bất đồng bộ, như IFrame được giới thiệu trong IE năm 1996, XMLHttpRequest năm 1999 và sau là XMLHttpRequest. Thuật ngữ Ajax ra đời năm 2006 đánh dấu một bước nhảy vọt trong công nghệ lập trình web[4][3], tuy rằng Ajax không phải một thứ mới mẻ tại thời điểm đó.

Tận dụng mô hình truyền tải bất đồng bộ, một hướng tiếp cận mới được ra đời năm 2005, khởi xướng bởi Steve Yen là Single-page application (SPA)[1]. Theo hướng tiếp cận này, toàn bộ nội dung của trang web được tải về trong một lần duy nhất, bao gồm HTML, Javascript, CSS. Khi cần thay đổi nội dung, phía máy khách chỉ cần tải thêm những nội dung mà thay đổi. Một mô hình thực tế của hướng tiếp cận này là Spiar [2], được Fielding giới thiệu năm 2000. Xem hình 1.



Hình 1. Mô hình xử lý dữ liệu của Spiar

Trong mô hình này, cả server và client đều lưu trữ thành phần tương tác người dùng, nhưng server chỉ lưu trữ dữ liệu và các hoạt động của thành phần này, còn client chỉ lưu trữ phần hiển thị của chúng (HTML). Mỗi khi có một sự kiện được phát sinh ở client, client thông báo cho server, server dựa vào sự kiện, cùng với trạng thái hiện thời của các thành phần tương tác người dùng để quyết định trạng thái tiếp theo. Sau đó server thông báo lại cho client sự thay đổi. Điểm đặc biệt của mô hình này là ở bộ Mã hóa/Giải mã Thay đổi (Delta Encoder/Decoder). Nhờ có bộ này, dữ liệu truyền gửi giữa client và server chỉ là sự thay đổi giữa 2 trạng thái kế tiếp chứ không phải toàn bộ trang web. Mô hình này rất giống với các mô hình truy cập từ xa [7].

Tuy có những ưu điểm như trên, nhưng mô hình này vẫn còn nhược điểm: Server phải xử lý khá nhiều công việc, bao gồm cả việc quản lý trạng thái của các thành phần tương tác người dùng trên client. Mặc dù nó thích hợp cho việc xây dựng các ứng dụng truy cập từ xa, khi mà số lượng người dùng nhỏ, nhưng lại không thích hợp với các ứng dụng có số lượng người dùng lớn (như một trang web thông thường).

Để khắc phục nhược điểm đó, Google đã đưa ra một mô hình khác dựa trên mô hình Spiar trong framework GWT [1][2] của mình

Mô hình này gần như giống hoàn toàn mô hình Spiar, chỉ có một điểm khác biệt là việc quản lý hoạt động của các thành phần tương tác được chuyển về phía client, làm giảm bớt công việc của server.

GWT có một điểm mạnh khác là nó phân tách được thời gian biên dịch (bằng Java) và thời gian chạy (bằng Javascript). Việc phát triển bằng GWT dễ dàng hơn vì cấu trúc của một chương trình được phân tách thành các thành phần sử dụng lại được. Cơ chế phân tách này cũng là một điểm nổi bật mà chúng tôi đã áp dụng trong JOO Framework, dưới hình thức portlet.

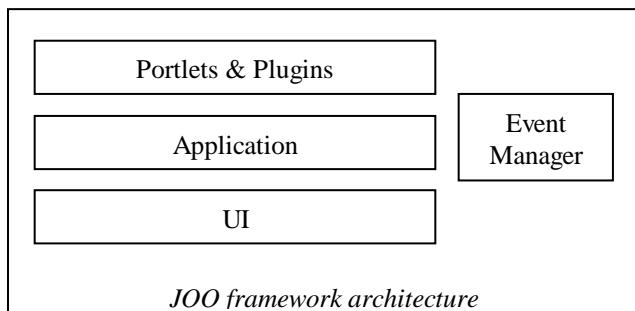
Tuy nhiên GWT được xây dựng trên nền tảng Java, việc lập trình và biên dịch đều bằng Java, đôi lúc không tận dụng được thế mạnh của Javascript và mặt khác, nó lại không hỗ trợ khá nhiều bộ thư viện hay dùng của Java. Ngoài ra, mặc dù cũng là một Ajax framework, nhưng GWT tập trung vào việc xây dựng các ứng dụng trên client cho Java Servlet. Quá trình phát triển ứng dụng cũng hơi phức tạp do yêu cầu phải có server bằng Java.

Từ việc nghiên cứu những ưu nhược điểm của các mô hình và framework hiện tại, JOO Framework đã tận dụng được một số ưu điểm của các framework kể trên, như cơ chế Ajax[10], thiết kế trang web dạng SPA[7], tổ chức cấu trúc chương trình thành các thành phần độc lập, liên kết lỏng lẻo với nhau và dễ dàng cấu hình. Phần sau sẽ phân tích rõ hơn kiến trúc của framework cũng như cấu trúc của một ứng dụng được xây dựng trên nền framework.

3. KIẾN TRÚC JOO FRAMEWORK

3.1. Kiến trúc hệ thống

Dựa trên những phân tích ưu nhược điểm của các framework và mô hình xây dựng ứng dụng web ở phần 2, JOO được thiết kế nhằm khắc phục nhược điểm và kế thừa những ưu điểm của các giải pháp trước đây. Sơ đồ dưới đây thể hiện mô hình kiến trúc tổng quan của JOO:



Hình 3. Sơ đồ tổng quan hệ thống

Kiến trúc hệ thống bao gồm các 4 thành phần chính:

* **Portlet & Plugins:** tiếp nhận các yêu cầu từ người dùng dưới dạng các sự kiện (event), sau đó gửi các Ajax request lên máy chủ và nhận về dữ liệu dưới dạng JSON. Các plugin & portlet sẽ xử lý dữ liệu dạng JSON gửi cho tầng Application, thông qua UI hiển thị dữ liệu. Điểm khác biệt giữa portlet và plugin là: portlet là các thành phần hiển thị và xử lý dữ liệu độc lập và được gọi khi trang web được tải, còn plugin chỉ được gọi với những sự kiện nhất định.

* **Event Manager:** bộ quản lý sự kiện: sử dụng mẫu thiết kế Observer, giúp các portlet, plugin có thể tương tác với nhau thông qua sự kiện, làm cho cấu trúc chương trình trở nên lỏng lẻo và các portlet, plugin có thể hoạt động độc lập với nhau. Các sự kiện có thể được sinh ra do tương tác của người dùng hoặc do chính portlet/plugin sinh ra.

* **Application:** bao gồm 3 thành phần chính:

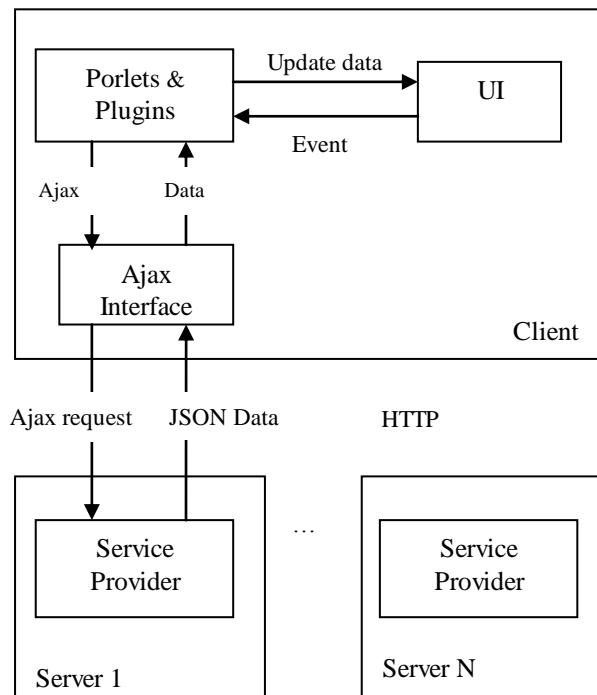
- Bộ quản lý tài nguyên (Resource Manager)
 - Bộ xử lý yêu cầu (Request Handler)
 - Bộ quản lý porlet & plugin (Porlet&Plugin Manager)

Ở tầng này xử lý các template và layout, thực thi các câu lệnh và nạp portlet và plugin vào bộ nhớ. Thông qua lớp Bootstrap tiếp nhận các yêu cầu, lớp xử lý yêu cầu gửi các yêu cầu này đến plugin & porlet tương ứng. Các lớp Plugin Manager, Porlet Container kiểm tra và nạp (load) các plugin, porlet tương ứng.

* **UI:** bao gồm các lớp và các thành phần (widget) hỗ trợ xây dựng giao diện giao tiếp với người dùng. UI trực tiếp nhận các yêu cầu của người dùng và thông qua bộ quản lý sự kiện (Event Manager) gửi sự kiện đến tầng ứng dụng (Application) và tầng Porlets & Plugins để xử lý.

3.2. Kiến trúc một ứng dụng xây dựng bằng JOO

Sau đây, chúng ta sẽ tìm hiểu kiến trúc một ứng dụng được xây dựng bằng JOO framework. Hình 4 mô tả mô hình xử lý dữ liệu của một ứng dụng như vậy:



Hình 4. Mô hình xử lý dữ liệu của một ứng dụng xây dựng trên JOO framework

Quá trình xử lý dữ liệu có thể được mô tả như sau: ở lần request (yêu cầu) đầu tiên máy chủ gửi về toàn bộ trang ứng dụng và sẽ được cache lại tại trình duyệt. Mỗi khi có một sự kiện phát sinh ở phía client, sự kiện này được gửi đến bộ Porlet & Plugin của JOO. Bộ Porlet & Plugin sẽ tiếp nhận sự kiện này và sinh ra một yêu cầu Ajax (Ajax request), thông qua bộ Ajax Interface gửi đến server. Ajax request này có thể được gửi đến một hay nhiều server. Server tiếp nhận yêu cầu và xử lý dữ liệu, sau đó gửi về client bằng giao thức REST. Client sau đó sẽ nhận được phần dữ liệu kết quả ở định dạng JSON (hoặc JSONP nếu như có nhiều máy chủ phục vụ) và xử lý nó thành dữ liệu HTML tương ứng và hiển thị nó lên client.

Thay vì yêu cầu dữ liệu trả về dưới dạng các chuẩn như HTTP, HTML, Cascading Style Sheets, Javascript, Document Object Model, phía client chỉ yêu cầu phía server trả về dữ liệu dưới dạng JSON (hoặc JSONP) thông qua giao thức REST. Sau đó, client xử lý nó thành dữ liệu theo các chuẩn như trên, và hiển thị theo mô hình giao diện tương tác người dùng một trang (Single-page user interface interaction). Tương tự như các ứng dụng trên desktop, mô hình tương tác này cũng được chia thành các thành phần riêng biệt. Sự thay đổi dữ liệu trên từng thành phần này diễn ra độc lập mà không cần tải lại toàn bộ trang.

3.3. Phân tích tính hợp lý của JOO framework

Để đánh giá chất lượng về chất lượng của một mô hình hay framework lập trình web, thông thường người ta dựa vào 6 yếu tố sau[7]:

- Khả năng tương tác người dùng
- Độ trễ trong cảm nhận người dùng: Khoảng thời gian kể từ lúc người dùng tiến hành 1 thao tác đến thời điểm **đầu tiên** họ nhận được phản hồi của hệ thống
- Hiệu năng của việc truyền gửi dữ liệu qua mạng: Được tính theo dung lượng được truyền gửi giữa máy khách và máy chủ
- Hiệu năng của máy chủ: Được tính theo khối lượng công việc mà máy chủ phải xử lý mỗi khi nhận được yêu cầu từ phía máy khách hoặc các tác vụ chạy nền liên quan.
- Chi phí công sức để xây dựng ứng dụng: Thời gian để hoàn thiện ứng dụng trên nền framework.
- Tính khả chuyển: Khả năng tương thích giữa nhiều trình duyệt và nhiều loại nền tảng.

Theo những tiêu chí kể trên, chúng tôi sẽ phân tích các đặc tính nổi bật của JOO Framework mà đáp ứng được các tiêu chí chất lượng này.

Các đặc tính nổi bật của JOO gồm có

- Tương tác bắt đồng bộ với server: Mọi cơ chế gửi nhận đều dựa trên AJAX.
- Dữ liệu truyền gửi giữa client và server nhỏ: Không bao gồm các thành phần hiển thị.
- Single page application: Mỗi khi thay đổi trang chỉ tải thêm

các thành phần mới mà không cần tải lại toàn bộ trang. Đặc tính này tận dụng triệt để cơ chế cache của trình duyệt: Dữ liệu được tải về một lần và cache tại trình duyệt máy khách, do đó những lần tiếp theo vào trang web, máy khách hầu như không phải tải lại dữ liệu trang web lần nữa.

- Việc sinh ra và quản lý các thành phần hiển thị đều ở phía client, giảm bớt công việc cho server
- Phân tách trang thành các porlet và plugin, dễ quản lý và cấu hình. Mỗi portlet là một thành phần độc lập, liên kết lỏng lẻo với các thành phần khác. Đây chính là điểm mạnh của GWT đã được áp dụng trong JOO Framework.
- Scalable: Việc trao đổi dữ liệu giữa client và server được thông qua Ajax Interface, đảm bảo tính độc lập giữa client và server. Khi đó một client có thể được phục vụ bởi nhiều server mà không ảnh hưởng đến kết quả[8].
- Sử dụng bộ UI widget để xây dựng các thành phần giao diện (UI component), giảm bớt thời gian xây dựng các phần này. Hình 5 mô tả ảnh hưởng của các đặc tính kể trên đến các tiêu chí chất lượng của ứng dụng. [7] Trong đó ngoại trừ tính khả chuyển thì JOO đều đáp ứng được.

	User Interactivity	User-perceived Latency	Network Performance	Server Performance	Development Effort	Portability
Single Page Interface	✓					
Asynchronous Interaction	✓	✓				
Delta Communication	✓	✓	✓	✓		
Client-side processing	✓	✓		✓		
UI Component-based	✓				✓	
Web standard-based					✓	✓

Hình 5. Mối quan hệ giữa các thuộc tính chất lượng của ứng dụng và các đặc tính của framework

4. KẾT QUẢ THU ĐƯỢC

Ở phần này, chúng tôi sẽ trình bày những kết quả đạt được của hệ thống khi ứng dụng để xây dựng một hệ thống hoàn chỉnh, đó là hệ thống BKProfile.

Hệ thống BKProfile (www.bkprofile.com) là một hệ thống chia sẻ kiến thức dành cho sinh viên, cựu sinh viên và giảng viên Bách Khoa. Hệ thống được phát triển từ năm 2009 dưới sự hướng dẫn của thầy Huỳnh Quyết Thắng và thầy Lê Quốc thuộc Viện Công nghệ thông tin và truyền thông.

Ban đầu hệ thống BKProfile được xây dựng trên nền Zend Framework, với tất cả thao tác xử lý nằm ở phía máy chủ.

Qua một thời gian phát triển, Zend Framework không còn đáp ứng được nhu cầu phát sinh khi hệ thống bắt đầu được ra mắt cho các bạn sinh viên Bách Khoa sử dụng.

Do đó, JOO đã được triển khai trên hệ thống BKProfile, và thực nghiệm đã chứng minh ưu điểm nổi trội của JOO so với hệ thống trước, cụ thể thông qua các số liệu sau:

- Dữ liệu truyền gửi giữa máy chủ và máy khách nhỏ: trung bình khoảng 65KB cho trang chủ với khá nhiều nội dung trong điều kiện trình duyệt đã cache lại ứng dụng ở lần truy cập đầu tiên (khoảng 1.6MB). Trong đó phần dung lượng dành cho câu hỏi (phần trọng tâm của hệ thống) vào khoảng 50KB. Trong tương lai, chúng tôi sẽ cố gắng rút gọn dung lượng tải của phần này xuống nhiều nhất có thể.
- Thời gian tải trang nhanh hơn. Hình 6 so sánh tốc độ tải trang giữa hệ thống khi dùng Zend Framework so với tốc độ tải trang khi dùng JOO. Việc thử nghiệm được tiến hành trên hệ điều hành Windows 7 ở 2 máy tính cá nhân có cấu hình tương đương và sử dụng trình duyệt Chrome phiên bản 11, được chạy 10 lần ngẫu nhiên trong cùng một thời điểm.

Zend Framework	JOO
3.3 s	495 ms
3.6 s	447 ms
3.2 s	374 ms
4.0 s	410 ms
3.3 s	394 ms
3.9 s	505 ms
4.1 s	441 ms
3.7 s	423 ms
5.8 s	412 ms
4.3 s	420 ms
Trung bình: 3.92s	Trung bình: 432 ms

- Thời gian phát triển ứng dụng khá nhỏ (từ lúc bắt đầu đến lúc ra mắt phiên bản Alpha tại Hội nghị học tốt 2011 là khoảng 2 tuần).
- Hệ thống hiện đang chạy ổn định trên trang web www.bkprofile.com sau gần 2 tháng ra mắt.

Tuy nhiên, khi hệ thống ra mắt và có nhiều người dùng, JOO đã nảy sinh một số lỗi trong thiết kế:

- Framework chưa tính đến SEO (Tối ưu hóa bộ máy tìm kiếm). Do đó các liên kết nội bộ trong hệ thống hầu như không được bộ thu thập dữ liệu của các hệ thống tìm kiếm đánh chỉ mục.
- Hệ thống BKProfile không chạy ổn định trên trình duyệt IE8 trở xuống (chạy tốt trên IE9).

Chúng tôi đã tính đến các phương pháp khắc phục mà dự định sẽ đưa vào trong phiên bản tiếp theo của framework. Phương pháp khắc phục những nhược điểm trên đã được đề cập như sau:

- Google có đề cập đến việc sử dụng các bản HTML snapshot dành riêng cho bộ thu thập dữ liệu của Google, nhằm tăng tính thân thiện cho việc tối ưu hóa bộ máy tìm kiếm.[11]
- Bộ thu viện hỗ trợ việc tương thích trên IE [12], khắc phục nhiều nhược điểm của IE trong việc xử lý HTML và CSS.

5. KẾT LUẬN

JOO framework thực hiện một hướng tiếp cận – không hoàn toàn mới so với những mô hình lập trình trước, mà kế thừa những ưu điểm, đồng thời đưa ra và thực thi những cải tiến hiệu quả,

khắc phục nhược điểm của những mô hình này – mang lại hiệu năng tốt nhất về mặt hiệu năng cũng như tính bảo mật cho các ứng dụng trên nền web. Những cải tiến này đồng thời chuẩn hóa lại mô hình lập trình các ứng dụng trên nền web, quản lý các luồng xử lý và truyền dữ liệu giữa máy khách với một máy chủ hoặc nhiều máy chủ (đối với hệ thống phân tán) nhằm cải thiện tốc độ xử lý và truyền dữ liệu. JOO cũng góp phần chuẩn hóa mô hình lập trình hướng đối tượng với ngôn ngữ Javascript.

Theo các số liệu thống kê ở phần 4, so với các mô hình lập trình web truyền thống (mô hình truyền – nhận dữ liệu đồng bộ, mô hình lập trình web trên nhiều trang), JOO đã đạt hiệu quả hơn hẳn về hiệu năng và khả năng, cũng như tính đơn giản và khả năng tái sử dụng. So với một số hướng tiếp cận khá hiệu quả gần đây (mô hình lập trình ứng dụng Spiar, GWT), JOO đã khắc phục được nhược điểm khi áp dụng như mô hình này cho việc lập trình ứng dụng web.

Mô hình kiến trúc của JOO đạt yêu cầu hầu hết đối với các thuộc tính chất lượng của ứng dụng. Những kết quả và số liệu thực tế đạt được khi thử nghiệm với hệ thống Bkprofile đã chứng minh tính hiệu quả của JOO. Việc chuẩn hóa mô hình lập trình và mô hình hoạt động của ứng dụng web như phân tích ở phần 3 đồng thời cải thiện thời gian phát triển (bao gồm từ giai đoạn thiết kế, cài đặt, đén kiểm thử) và thời gian nâng cấp một ứng dụng được xây dựng trên nền tảng JOO.

Bên cạnh những ưu điểm nổi bật kể trên, hệ thống vẫn còn tồn tại một số nhược điểm như: yêu cầu trình duyệt phải hỗ trợ Javascript và không tương thích với nhiều loại trình duyệt như mô hình web truyền thống (đặc biệt là Internet Explorer), các liên kết không thân thiện với máy chủ tìm kiếm. Các phương pháp khắc phục những nhược điểm này đã được tính đến và sẽ được hoàn thiện trong phiên bản tiếp theo của hệ thống.

Trong tương lai gần, với những ưu thế hiện thời của JOO, chúng tôi đã sẽ mở rộng JOO framework theo mô hình mã nguồn mở để thu nhận nhiều phản hồi của người sử dụng cũng như mời những lập trình viên trong cộng đồng nguồn mở cùng tham gia phát triển, để framework ngày càng hoàn thiện hơn. Chúng tôi đã tính đến việc sử dụng Google code hoặc Git Hub để quản lý mã nguồn, và quảng bá framework tới những lập trình viên mã nguồn mở trong cộng đồng Việt Nam nói riêng và trên thế giới nói chung để thu hút các đóng góp từ bên ngoài.

6. LỜI TRI ÂN

Chúng tôi xin được gửi lời chân thành cảm ơn tới nhóm BKProfile, PGS. TS. Huỳnh Quyết Thắng, ThS. Lê Quốc đã tạo điều kiện để giúp chúng tôi hoàn thành nghiên cứu này. Chúng tôi cũng xin gửi lời cảm ơn đến những cá nhân đã góp ý trong quá trình phát triển JOO để chúng tôi có thể tạo ra được một framework ngày càng hoàn thiện với một kiến trúc tốt.

7. TÀI LIỆU THAM KHẢO

- [1] Ali Mesbah, Analysis and Testing of Ajax-based Single-page Web Applications, <http://homepages.cwi.nl/~arie/phds/Mesbah.pdf>
- [2] Ali Mesbah, Arie van Deursen, A Component- and Push-based Architectural Style for Ajax Applications, <http://homepages.cwi.nl/~arie/papers/spci/spiar-jss.pdf>
- [3] Ali Mesbah, Arie van Deursen, An Architectural Style for Ajax, June 2006, <http://arxiv.org/ftp/cs/papers/0608/0608111.pdf>
- [4] OpenAjax <http://openajax.org> (15/03/2011)
- [5] Joomla. <http://www.Joomla.org> (17/03/2011)
- [6] Louenas Hamdi, Huaigu Wu, Serhan Dagtas. Ajax for Mobility: MobileWeaver Ajax Framework, April 2008, <http://www2008.org/papers/pdf/p1077-hamdi.pdf>
- [7] A. Mesbah, K. Broenink, A. van Deursen, Spiar: An architectural style for single page internet applications, April 2006
- [8] Ken Birman, Krzysztof Ostrowski, Storing and accessing Live Mashup Content in the Cloud, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.160.326&rep=rep1&type=pdf>
- [9] Google Closure <http://code.google.com/intl/vi-VN/closure/> (17/03/2011)
- [10] Jun Zhang, Optimising Ajax Web Applications with Communication Restructuring <http://www.cs.ubc.ca/~ericazhj/papers/ma.pdf>
- [11] Making AJAX Applications Crawlable, <http://code.google.com/intl/vi-VN/web/ajaxcrawling/docs/getting-started.html> (03/05/2011)
- [12] Ie7-js Project, <http://code.google.com/p/ie7-js/> (03/05/2011)

Truyện tranh trên di động

Nguyễn Thị Thuyên

Tóm tắt: Đọc truyện tranh là một trong những loại hình giải trí rất phổ biến đặc biệt là ở nhiều nước châu Á trong đó có Việt Nam. Hàng ngày có hàng nghìn bản in truyện tranh được xuất bản. Hầu hết trong số đó được số hóa đưa lên các website cho phép người dùng đọc từ các thiết bị có khả năng kết nối Internet như máy tính, PDA hay điện thoại di động. Xu hướng gần đây cho thấy nhu cầu đọc truyện tranh trên điện thoại di động ngày càng trở nên phổ biến. Tuy nhiên, người dùng gặp nhiều khó khăn khi muốn đọc toàn bộ các trang truyện số hóa trên màn hình di động, do hạn chế về kích thước màn hình và tốc độ kết nối của điện thoại di động. Bài báo này giới thiệu các giải pháp khắc phục các vấn đề trên, cụ thể như sau:

Giải pháp phân tích nội dung web tự động nhằm hỗ trợ người dùng sưu tập tự động truyện tranh từ các website cung cấp truyện miễn phí như <http://manga24h.com/>, <http://truyentranh.com/>

Giải pháp tách biên dựa trên công thức tính gradient để phân tách trang truyện thành các frame có kích thước nhỏ hơn phù hợp với kích thước màn hình di động

Giải pháp thêm và tách biên hạn chế mất thông tin nhằm tiếp tục phân tách các trang truyện, các frame (kích thước vẫn khá lớn sau khi tách lần thứ nhất) có khung truyện chồng lênh nhau không có các đường biên rõ ràng.

Bài báo này cũng giới thiệu kết quả cài đặt và triển khai thử nghiệm của trình đọc truyện tranh trên thiết bị di động có sử dụng các giải pháp nói trên, đồng thời so sánh kết quả đạt được với các trình đọc truyện tranh trên thiết bị di động hiện có.

Từ khóa: Phân tách, Truyện tranh, Truyện tranh trên di động,

1. GIỚI THIỆU

Đọc truyện tranh là hình thức giải trí không chỉ dành cho trẻ em mà còn dành cho người lớn, nhưng thay vì đọc truyện trên các bản in mọi người ngày càng có xu hướng thích đọc truyện tranh trên các thiết bị di động đặc biệt là điện thoại. Mặc dù cấu hình ngày càng được cải thiện nhưng những hạn chế về kích thước màn hình, bộ nhớ và tốc độ truyền tải dữ liệu vẫn là những khó khăn trong việc đọc truyện tranh trên điện thoại di động.

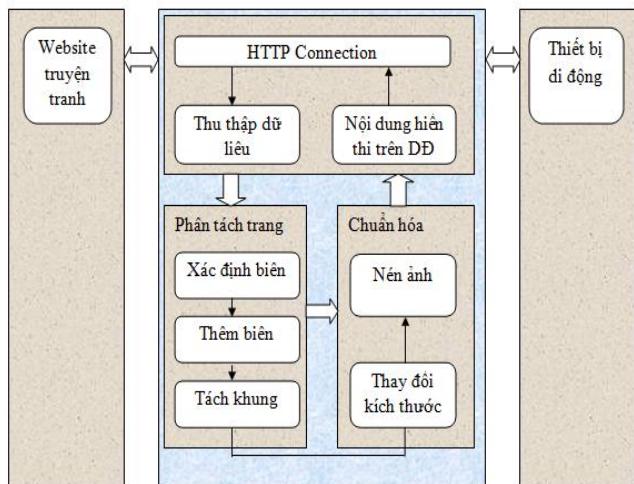
Nguyễn Thị Thuyên, sinh viên lớp Công nghệ phần mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0122-824-7721, e-mail: thuyen183@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

Hiện tại có khá nhiều giải pháp được đặt ra nhằm giải quyết vấn đề trên, như trong bài báo của Kohei Arai và Herman Tolle [2] đã dựa trên việc tính toán số lượng điểm ảnh màu đen để tiến hành phân tách truyện, phương pháp này thực hiện tốt khi các trang truyện không bị mờ, đường biên rõ ràng. Trong báo cáo “Comic Viewer for iphone” của Cheung Kam Shun và Sung Siu Hang Aaron [3] đã sử dụng công thức tính histogram để phân tách trang truyện, phương pháp này tách nhanh nhưng không sử dụng được trong trường hợp các khung lồng nhau. Trong phạm vi bài báo này chúng ta sẽ giới thiệu một phương pháp cho phép tách trang truyện dựa trên sự biến thiên gradient của các điểm ảnh. Giải pháp này đã được đề cập trong bài báo của Yusuke In, Takashi Oie, Masakazu Higuchi, Shuji Kawasaki, Atushi Koike và Hitomi Murakami [1], phân tách chính xác tuy nhiên quá trình tính toán và xử lý khá phức tạp và mất thời gian. Ở đây ta sẽ đơn giản hóa quá trình tính toán và xử lý mà không làm mất đi hiệu quả của giải thuật. Giải thuật tách tốt đối với các trang truyện có các khung rõ ràng và cả các trang truyện khi scan bị mờ, đối với các trang truyện có khung chồng lênh nhau, không có đường biên chúng ta sẽ sử dụng kết hợp thuật toán trên cùng với giải thuật OCR (tách kí tự trong file ảnh) để thêm biên sau đó sử dụng phép phân tách tương tự

Nội dung tiếp theo được tổ chức như sau: Phần 2: mô tả kiến trúc dịch vụ truyện tranh trên di động, Phần 3: mô tả giải thuật tách biên, Phần 4: dịch vụ truyện tranh trên di động, Phần 5: đánh giá, so sánh với một số giải thuật khác, Phần 6: kết luận

4. MÔ TẢ KIẾN TRÚC DỊCH VỤ TRUYỆN TRANH TRÊN DI ĐỘNG “MCOMIC”



Hình 1. Kiến trúc dịch vụ truyện tranh trên di động

Hình vẽ trên mô tả kiến trúc của dịch vụ đọc truyện tranh trên di động. Quy trình xử lí bắt đầu khi người sử dụng dùng di động gửi yêu cầu tới hệ thống

Thu thập dữ liệu: các trang truyện tranh sẽ được lấy về từ các website cung cấp truyện miễn phí thông qua kết nối HTTP và quy trình xử lí phân tách nội dung trang web. Sử dụng SQL server 2005 để lưu trữ dữ liệu sau khi thu thập về

Phân tách truyện: dựa trên các đặc tính đặc biệt của các trang truyện (thường các trang truyện được phân thành các miền nhỏ hơn thông qua việc sử dụng các khung) để tách trang truyện thành các frame có kích thước phù hợp khi hiển thị trên thiết bị di động. Đối với các trang truyện có các khung lồng nhau (đường biên phân tách không rõ ràng) sẽ sử dụng thêm quá trình phân tích để thêm đường biên một cách phù hợp nhất giảm thiểu hạn chế khả năng mất mát thông tin đến mức có thể. Chuẩn hóa: sử dụng một số giải thuật nén ảnh (như chuyển file ảnh từ .png về dạng .jpg), thay đổi kích thước trong phạm vi cho phép (giảm kích thước ảnh nhưng không làm mất thông tin)

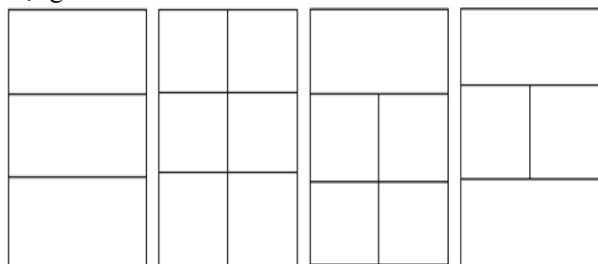
3. GIẢI THUẬT PHÂN TÁCH BIÊN

Có sự khác biệt rất lớn về kích thước màn hình trên di động và màn hình trên desktop, do đó rất khó để hiển thị toàn bộ trang truyện trên màn hình di động. Ở đây ta sẽ đưa ra giải pháp cho việc tách một trang truyện thành các miền nhỏ để hiển thị trên di động.

Dựa vào một số đặc tính của các trang truyện chúng ta sẽ thu hẹp khoảng giá trị cần phải tính toán (các trang truyện được tách theo các biên ngang hoặc biên đứng, trang truyện được đọc từ trái sang phải)

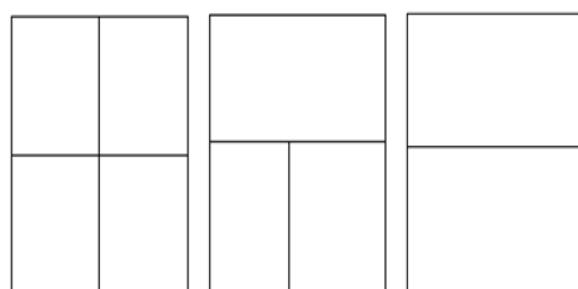
Trang truyện sẽ được tách theo một số dạng sau đây

Dạng 1



Hình 2. Dạng phân tách 1

Dạng 2.



Hình 3. Dạng phân tách 2

Công thức

Đường biên là tập hợp các điểm tại đó hàm ảnh biến thiên và bao gồm những điểm với biên độ biên cao

Đường biên và các phần của nó luôn trực giao với hướng của gradient

Dựa vào công thức tính gradient để xác định đường biên tách các phần trên trang truyện

Độ lớn và hướng của gradient tại một điểm của hàm

$$G_x = \frac{\partial s(x,y)}{\partial x}; G_y = \frac{\partial s(x,y)}{\partial y} \quad (1)$$

$$G = \sqrt{G_x^2 + G_y^2}$$

Vì ảnh là tập các điểm rời rạc nên chúng ta sẽ dùng một số công thức tính gần đúng sau để xác định gradient

$$G_x = \frac{s(m,n) - s(m-k,n)}{k} \quad G_y(m, n) = \frac{s(m,n) - s(m,n-k)}{k} \quad (2)$$

hoặc

$$G_x(m, n) = \frac{s(m+k,n) - s(m,n)}{k} \quad G_y(m, n) = \frac{s(m,n+k) - s(m,n)}{k} \quad (3)$$

Thường k=1

Gradient cho đường thẳng P(x1,y1,x2,y2): tổng giá trị gradient của tất cả các điểm thuộc đường thẳng đó. Đường biên sẽ là đường có giá trị gradient lớn nhất.

$$P(x1, y1, x2, y2) = \sum_{x=x1}^{x2} \sum_{y=y1}^{y2} G(x, y) \quad (4)$$

Ở đây chúng ta sẽ sử dụng công thức tính gradient gần đúng (2)

Với s là giá trị xám tại điểm ảnh có tọa độ m,n

Để tính giá trị xám chúng ta sử dụng công thức

$$s(m, n) = 0.3 * red + 0.59 * green + 0.11 * blue \quad (5)$$

Tách biên

Bước 1:

Xác định đường biên ngang thứ nhất(từ trên xuống) trước. Xét giá trị cho đường biên này trong khoảng từ(h/4;h/2). Nếu tọa độ y của đường biên này nằm trong khoảng(h/3+h/6;h/2) chúng ta sẽ tách tiếp theo dạng số 2 ngược lại tiếp tục tách theo dạng 1.

Với dạng 1 tiếp tục xác định đường biên ngang thứ hai. Giá trị y cho đường biên này nằm trong khoảng từ (h/2;3h/4)

Với dạng 2 xác định đường biên dọc. Giá trị x cho đường biên này nằm trong khoảng (w/4;3w/4)

(w : độ rộng của ảnh gốc, h: chiều cao của ảnh gốc)

Bước 2:

Đối với từng dạng(1,2) tùy theo việc xác định biên để xác định chia theo số phần khác nhau(với dạng 1 chia làm 4,5 hoặc 6 phần ; với dạng 2 chúng ta chia làm 2,3 hoặc 4 phần Trong các bước tính toán để xác định các đường biên này công thức sử dụng chủ yếu là (4).

Việc xác định đường biên dọc ngoài tính toán dựa trên gradient ta còn sử dụng thêm một giá trị khác để xác định. Giá trị này tính toán dựa trên mức xám của các điểm lân cận. Giới hạn để

xác nhận là đường biên khi khoảng 80% các điểm lân cận có mức xám >240

Thêm biên

Áp dụng với trường hợp ảnh không có các biên để tách trang thành các frame có kích thước nhỏ hơn, hoặc sau khi tách biên lần thứ nhất kích thước của frame vẫn còn lớn nhưng không có biên để tách tiếp

Bước 1:

Xác định được đường biên ứng viên thông qua xác định các đường có sự biến thiên gradient lớn nhất trong vùng từ $w/4$ đến $3w/4$ (w : độ rộng của ảnh gốc, h : chiều cao của ảnh gốc)

Bước 2:

Tính toán số điểm đen về 2 phía của đường biên ứng viên trên, nếu số điểm đen đủ lớn chúng ta sẽ tiếp tục bước theo, ngược lại thuật toán dừng (không thể thêm được đường biên)

Bước 3:

Dùng thuật toán OCR(Optical character recognition) để xác định có chữ hay không giữa vùng không chứa đường biên ứng viên. Nếu không có đổi các điểm đen trên đường biên ứng viên thành trắng và tiếp tục bước 4, ngược lại dừng thuật toán(không thể thêm đường biên)

Bước 4:

Tiến hành cắt ảnh bình thường theo phương pháp tách biên phía trên.

Trong các bước tính toán trên ngoài sử dụng công thức (4) để so sánh giá trị gradient xác định đường biên ứng viên ta còn sử dụng công thức (5) để xác định lại các đường biên ứng viên này có phù hợp hay không.

4. DỊCH VỤ TRUYỆN TRANH TRÊN DI ĐỘNG “MCOMIC”

Dịch vụ truyện tranh trên di động “EComic” xây dựng dựa trên giải thuật thu thập truyện trên mạng và giải thuật phân tách trang truyện trên. Dịch vụ này cho phép đọc truyện tranh, bình luận, chia sẻ, đọc các tin tức, sự kiện liên quan đến truyện tranh trên các thiết bị di động hỗ trợ java, 3G

Cài đặt

Phía server: sử dụng framework JDK1.6, IDE Eclipse, server Glassfish

Phía di động: J2me :Cài đặt giao diện: cài đặt bộ thư viện j2mepolish

Dịch vụ cung cấp cho các thiết bị di động có hỗ trợ java và 3G
Hình ảnh ví dụ về dịch vụ “truyện tranh trên di động” được chạy trên localhost với bộ giả lập của wtk



Hình 4. Kết quả sử dụng giải thuật tách biên truyện



Hình 5. Dịch vụ truyện tranh trên di động

(Trích từ truyện bạo quyền hung tinh – chương 1)

Tách tốt nhất để hiển thị trên di động là khi trang truyện được tách thành 6 phần, khi đó mỗi khung nhỏ này có thể hiển thị được gần như toàn bộ trên màn hình di động. Tuy nhiên không phải trang truyện nào khi được vẽ cũng tách rõ ràng như vậy. Trong một số trường hợp có thể không có biên giữa các phần, hay số đường biên không nhiều, khi đó nếu cố tình tách thành 6 phần việc mất mát thông tin là không tránh khỏi. Do đó trong những trường hợp này chúng ta sẽ tách với số phần nhỏ hơn (từ 2->5 phần) và sử dụng thêm các thanh cuộn phía di động để đảm bảo đọc được toàn bộ nội dung của truyện

5. SO SÁNH, ĐÁNH GIÁ

Phương pháp tách biên truyện dựa trên biến đổi gradient trên tuy có chậm hơn so với một số phương pháp khác nhưng kết quả tách tốt cho cả trường hợp các trang truyện scan bị mờ. Kết quả thu được sau khi thực hiện phân tách trên một số truyện tranh

Tên truyện	Số trang	Tách đúng	Tỉ lệ thành công
Chú bé rồng chương 130	45	38	84%
Jindo đường dẫn đến khung thành tập 5	145	133	91,7%
Onepice chương 1	39	35	89,78%

Bảng 1: Một số kết quả khi tách trang truyện theo giải thuật sử dụng

So sánh với phương pháp tách của Kohei Arai và Herman Tolle [2]

Kết quả thu được của cả 2 phương pháp sau khi cùng thực hiện phân tách trên truyện "One Piece chương 1"

Phương pháp của	Số trang đúng	Tỉ lệ phân tách đúng
Ta sử dụng	35/39	89,74%
Kohei Arai & Herman Tolle	35/39	89,74%

Bảng 2: Kết quả so sánh giải thuật ta sử dụng với giải thuật của Kohei Arai và Herman Tolle

So sánh với phương pháp tách của Yusuke In, Takashi Oie, Masakazu Higuchi, Shuji Kawasaki, Atushi Koike and Hitomi Murakami [1]: cùng dựa trên ý tưởng sử dụng biến thiên gradient phương pháp của chúng ta tuy đạt độ chính xác kém hơn nhưng tính toán, xử lý nhanh hơn và vẫn đảm bảo lượng thông tin bị mất trong giới hạn cho phép (truyện tranh vẫn đọc hiểu tốt sau khi phân tách)

So sánh với phương pháp sử dụng histogram của Cheung Kam Shun và Sung Siu Hang Aaron [3]: phương pháp sử dụng histogram thời gian xử lý nhanh hơn lượng tính toán ít hơn, nhưng ngoài sử dụng công thức tính histogram còn phải kết hợp thêm một số công thức tính toán khác để tăng thêm độ chính xác, sử dụng không tốt cho trường hợp các khung truyện bị lỏng (không có biên rõ ràng)

KẾT LUẬN

Dịch vụ truyện tranh trên di động "MComic" đã thực hiện thu thập được một số lượng tương đối truyện tranh trên một số trang web như manga24h.com, truyentranh.com (hiện việc thu thập tiến hành trên PC cá nhân do đó số lượng truyện thu thập

về khoảng 50 truyện), đồng thời thực hiện việc tách nội dung trang truyện theo kích thước phù hợp để hiển thị trên di động. Phương pháp tách nội dung này thực hiện khá tốt trên hầu hết các trang truyện tranh hiện nay, truyện được hiển thị tốt trên di động với khoảng 90% không bị mất thông tin. Dịch vụ vẫn đang tiếp tục được mở rộng bộ thư viện truyện (thêm các loại truyện gốc tiếng Anh hoặc tiếng Nhật) thông qua phân tách các trang cung cấp truyện của nước ngoài như <http://www.onemanga.com/>, <http://www.mangareader.net/>. Cải tiến thuật toán phân tách trang truyện, giảm sự mất mát thông tin (giảm trường hợp các lời thoại trong truyện bị cắt đứt quãng)

Kết hợp thuật toán đang sử dụng, giải thuật OCR và thuật toán được đề cập trong bài báo của Kohei Arai, Herman Tolle để tách các khung chữ trong trang truyện sau đó phân tách kí tự khỏi trang truyện, nhằm hỗ trợ việc hiển thị khi kí tự trên ảnh quá nhỏ hoặc quá mờ khi hiển thị trên di động. Phát triển thêm phần đọc truyện tiếng nước ngoài bằng việc kết hợp sử dụng với google translate

TÀI LIỆU THAM KHẢO

- [1] Yusuke IN, Takashi Oie, Masakazu Higuchi, Shuji Kawasaki, Atushi Koike and Hitomi Murakami : "Using Fast Frame Decomposition and Sorting by Contour Tracing Mobile Phone Comic Imaging System", Internation journal of systems application , engineering and development, Issue 2 Volume 5, 2011
- [2] Kohei Arai & Herman Tolle, Automatic E-Comic Content Adaptation, International Journal of Ubiquitous Computing (IJUC) Volume (1), Issue (1) 11.
- [3] Final year report project "Comic Viewer for iphone" Cheung Kam Shun , Sung Siu Hang AaronDepartment of Computer Science and Engineering The Chinese University of Hong Kong, 2010-2011
- [4] Bài giảng "Xử lý ảnh" thầy Nguyễn Linh Giang.

Hệ thống tổng hợp tiếng nói tiếng Việt chất lượng cao

Nguyễn Trọng Hiếu, Lê Quang Thắng, Lê Anh Tú, Đỗ Văn Thảo, Nguyễn Hữu Thuận

Tóm tắt – Tổng hợp tiếng nói là một lĩnh vực có ứng dụng rộng rãi và được rất nhiều quan tâm nghiên cứu trên thế giới cũng như ở Việt Nam. Hiện nay tại Việt Nam đã phát triển nhiều bộ tổng hợp tiếng nói dành riêng cho tiếng Việt. Tuy nhiên, chất lượng tiếng nói tổng hợp sao cho dễ nghe và tự nhiên vẫn là điều mà các nhà nghiên cứu đang hướng tới. Nghiên cứu này tập trung vào toàn bộ thành phần của một bộ tổng hợp tiếng nói, để xuất và cài đặt các thuật toán và mô hình cho việc cải thiện chất lượng tiếng nói tổng hợp thông qua cải tiến từng thành phần trong hệ thống tổng hợp tiếng nói. Mục tiêu cuối cùng là có thể xây dựng một hệ thống tổng hợp tiếng nói hoàn chỉnh các thành phần với chất lượng tốt.

Từ khóa – Tổng hợp tiếng nói, chuẩn hóa, phân tích cú pháp, trường độ, cao độ, lựa chọn đơn vị.

1. GIỚI THIỆU

Tổng hợp tiếng nói là quá trình tạo ra tiếng nói nhân tạo của người trên máy tính từ văn bản. Đây là một đề tài có tính ứng dụng thực tiễn cao nên được nghiên cứu nhiều trên thế giới và Việt Nam từ rất sớm. Ứng dụng của tổng hợp tiếng nói có thể dễ dàng thấy trong nhiều hệ thống, như hệ thống hỗ trợ đọc văn bản cho người khuyết tật, hệ thống trả lời tự động tại các tổng đài hay robot, hệ thống chỉ đường trong các phương tiện vận tải...

Bộ tổng hợp tiếng nói được chia làm hai phần chính: tổng hợp mức cao và tổng hợp mức thấp. Nhiệm vụ phần tổng hợp mức cao là chuẩn hóa văn bản, phát sinh thông tin về ngữ âm, ngữ điệu.

Thông tin về nhóm tác giả:

Nguyễn Trọng Hiếu, Lê Quang Thắng, Lê Anh Tú, Đỗ Văn Thảo, Nguyễn Hữu Thuận. Nhóm sinh viên lớp Công nghệ phần mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 01677225100, e-mail: tronghieubk@gmail.com).

Giáo viên hướng dẫn:

TS. Trần Đỗ Đạt, Trung tâm nghiên cứu Mica.
ThS. Nguyễn Thị Thu Trang, Viện CNTT-TT.

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

Phần tổng hợp mức thấp dựa vào các thông tin phía trên sẽ tiến hành tìm kiếm và lựa chọn đơn vị âm, thực hiện ghép nối và làm trọn tín hiệu, cho ra tiếng nói cần tổng hợp.



Hình 1 Hệ thống tổng hợp tiếng nói

2. CHUẨN HÓA VĂN BẢN

Trong hệ thống tổng hợp tiếng nói, việc chuẩn hóa văn bản là công đoạn đầu tiên có ảnh hưởng quan trọng trong việc đảm bảo văn bản được đọc một cách đúng đắn. Hiện tại đã có một số nghiên cứu về chuẩn hóa văn bản trong tiếng Việt, nhưng kết quả chủ yếu mới chỉ dừng lại ở những tập luật cơ bản áp dụng cho những trường hợp đặc biệt, chưa giải quyết được bài toán một cách hệ thống.

Trong nghiên cứu này, chúng tôi xem xét bài toán một cách tổng quát để đưa ra giải pháp tổng thể cho việc chuẩn hóa văn bản tiếng Việt. Các vấn đề về các dạng chưa chuẩn và các tình huống nhập nhằng được giải quyết.

2.1 Non standard words

Các trường hợp cần phải chuẩn hóa được quan sát và phân loại vào các dạng khác nhau gọi là “các loại từ chưa chuẩn hóa” hay Non-standard Word (NSW). Mỗi loại từ chưa chuẩn hóa có cách xử lý riêng. Việc phân loại các từ chưa chuẩn hóa được thể hiện trong bảng sau:

Nhóm	Loại	Mô tả	Ví dụ
Số	NTIM	Thời gian/giờ	1:30
	NDAT	Ngày/tháng/năm	17/3/87
	NDA Y	Ngày và tháng	17/3, 03-05/3
	NMON	Tháng và năm	3/88, 5/2011
	NNUM	Số/số học	2009, 70.000
	NTEL	Số điện thoại	0915.334.577
	NDIG	Số hiệu, mã số	VN534
	NSCR	Tỉ số	Tỉ số là 3-5
	NRNG	Miền giá trị	Từ 3-5 ngày
	NPER	Số phần trăm	93%, 30-40%,
	NFRC	Phân số	34/6, 6/145
	NCOM	Hỗn hợp	2x2x3, 18+, 2*3
	NADD	Địa chỉ	Ngách 128/27/2A
	NSIG	Kí hiệu	m2, m3
Chữ	LW RD	Từ mượn	London, NATO
	LSEQ	Dãy các ký tự	ODA, GDP
	GREK	Số Hi Lạp	I, II
	LABB	Viết tắt	TS (tiến sĩ)
Khác	PUNC	Dấu câu đọc được	... () [] ‘ ’ “ ” - /
	SENT	Dấu phân tách câu	. ? ! ...
	URLE	Địa chỉ url, email	http://soict.hut.vn
	MONY	Tiền tệ	2\$, \$2, 100¥,
	DURA	Trường độ (nghỉ)	“-” in scores (2-3)
	NONE	Bỏ qua	ascii art...

Bảng 1 Bảng phân loại NSW

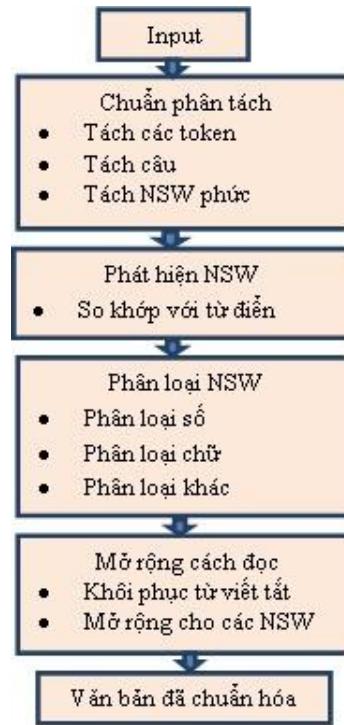
Văn bản đầu vào là văn bản lấy trong thực tế, ban đầu rất hỗn độn vì nó chứa nhiều dạng từ chưa chuẩn hóa khác nhau. Ván đề là nhận ra và phân loại đúng những từ này. Vì mỗi loại có cách đọc khác nhau nên khi phân loại sai có thể sẽ khiến cách đọc sai và người nghe hiểu sai nội dung văn bản. Ví dụ “phản XI” đọc lên là “phản mười một”, nếu không nhận đúng số la mã “XI” sẽ đọc là “phản xi”!

Để đảm bảo tính nguyên vẹn của văn bản đầu vào mà vẫn bỏ xung đột thông tin sau quá trình chuẩn hóa, chúng tôi tổ chức lại văn bản và thông tin bỏ xung đột dạng cấu trúc XML. Các thông tin là thuộc tính của các thẻ XML, khi bỏ qua các thẻ này ta nhận được văn bản gốc.

2.2 Các bước của quy trình chuẩn hóa.

Chuẩn phân tách: Văn bản đầu vào trước hết được xử lý bằng regex để nhận ra và đánh dấu các tổ hợp thuộc nhóm số, URLE bởi các nhóm này sẽ được xử lý riêng. Tiếp đó các dấu trắng thừa trong văn bản được loại bỏ, thêm dấu trắng vào trước và sau các dấu câu, các khoảng trắng trong một tổ hợp số được thay bởi dấu chấm “.” để tiện cho việc xử lý về sau. Các câu trong văn bản được phân tách và đánh dấu, phục vụ cho việc khai thác

ngữ cảnh và đưa ra nhịp điệu đọc phù hợp cho tiếng nói tổng hợp. Cuối cùng các NSW phức được tách ra chuẩn bị cho bước sau.



Hình 2 Quy trình chuẩn hóa văn bản

Phát hiện NSW: Các token được tìm kiếm trong một từ điển các âm tiết tiếng Việt có thể đọc được, nếu không thấy thì đánh dấu là NSW.

Phân loại NSW: Các NSW nhóm số, nhóm chữ, nhóm khác được phân nhóm đơn giản bởi các luật nhận ra định dạng của chúng trong regex. Việc phân loại các NSW nhóm số, nhóm chữ phức tạp hơn vì tiềm ẩn nhiều trường hợp nhập nhằng nên được giải quyết bằng các kỹ thuật data mining. Các dạng trong nhóm số được phân loại bằng cây quyết định, với các thuộc tính định dạng và ngữ cảnh. Định dạng gồm số các chữ, số các số, miền giá trị, ngữ cảnh gồm 2 token liền trước và liền sau của NSW. Các dạng trong nhóm chữ không có định dạng đặc trưng để nhận ra, ngữ cảnh cũng kém rõ ràng hơn dạng số. Vì thế dạng chữ được phân loại bằng việc tính xác suất của một NSW thuộc về LWRD, LSEQ hay LABB dựa trên một tập huấn luyện. NSW được phân loại theo xác suất lớn nhất. Các trường hợp thuộc nhóm khác được phân loại dựa vào định dạng, SENT được phân loại bởi việc phân tách câu bước chuẩn phân tách.

Mở rộng cách đọc: Ván đề khôi phục từ viết tắt cũng tiềm ẩn rất nhiều nhập nhằng. Cùng một NSW có thể tìm được nhiều từ đầy đủ thỏa mãn nó. Một danh sách các từ viết tắt và các từ đầy đủ được sử dụng. Với mỗi LABB sẽ duyệt tìm các từ đầy đủ có thể của nó, sau đó dùng mô hình ngôn ngữ để khai thác ngữ cảnh và tính xác suất từ đầy đủ và entropy cho LABB đó. Tổng hợp lại

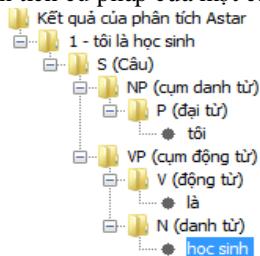
ta sẽ lựa chọn lấy từ đây đủ là trường hợp có tích xác suất và entropy lớn nhất.

Cách đọc các nhóm số đưa ra bởi các luật khá đơn giản. Từ mượn được đọc dựa vào từ điển từ mượn, từ viết tắt được đọc theo từ đầy đủ của nó, dãy chữ và dãy số được đọc từng kí tự, các dấu đọc theo cách phát âm thông thường của nó, dấu phân tách câu không được đọc.

3. PHÂN TÍCH CÚ PHÁP.

Trong tổng hợp tiếng nói, phân tích cú pháp đóng một vai trò rất quan trọng trong công đoạn xử lý văn bản của hệ thống. Phân tích cú pháp chuẩn xác sẽ đưa ra cho hệ thống một cái nhìn toàn cảnh về cấu trúc của văn bản, các cụm từ trong văn bản từ phức tạp cho đến đơn giản nhất, đồng thời các vị trí âm tiết trong cụm từ cũng được đưa ra luôn.

Mục đích của bộ phân tích cú pháp là đưa ra được cây phân tích cú pháp của văn bản đầu vào. Dưới đây là một ví dụ về cây phân tích cú pháp của một câu:



Trong phần này, chúng tôi đã nghiên cứu cách thức để có thể cài tiến đầu ra cho bộ phân tích cú pháp cả về mặt tốc độ cũng như chất lượng.

3.1 Mô hình xác suất PCFG.

Mô hình PCFG là một mô hình văn phạm phi ngữ cảnh dùng để biểu diễn và quản lý tập luật cú pháp tiếng Việt. Mô hình PCFG là một tập bao gồm 5 tham số $G=(N, \Sigma, P, S, D)$ ^[5], trong đó :

- N : tập các nhãn từ loại, $\{N^i\}$, $i=1, \dots, n$
- Σ : tập các từ được tách từ văn bản, $\{W^k\}$, $k=1, \dots, V$
- P : tập các luật có dạng $\{N^i \rightarrow \zeta^j\}$, $\zeta^j \in (\Sigma \cup N)^*$
- S : ký hiệu khởi đầu, tượng trưng cho một câu.
- D : tính xác suất cho mỗi luật tương ứng P .

Với PCFG, ta có xác suất của một cây phân tích cú pháp

$$P(T) = \sum_{i=1}^m \lg(r_i),$$

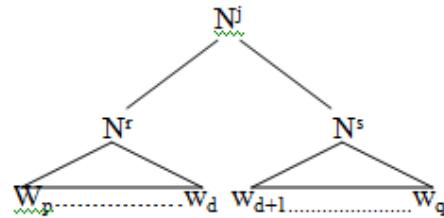
trong đó r_i là các luật sử dụng trong cây

Với một câu đầu vào, sẽ có nhiều cây phân tích cú pháp đầu ra, cây nào có xác suất cao nhất sẽ là cây đầu ra thích hợp nhất. Vậy vấn đề chúng ta đặt ra ở đây là phải tìm được giải thuật giúp được cây phân tích cú pháp có xác suất lớn nhất với thời gian ngắn nhất.

3.1.1 Xác suất inside.

Nếu có một nút N^j được tạo ra bởi luật $N^j \rightarrow N^r N^s$ thì inside của nó sẽ được tính bằng :

$$\text{inside}(N^j) = \lg(P(N^j \rightarrow N^r N^s)) + \text{inside}(N^r) + \text{inside}(N^s)$$



Hình 3 Xác suất inside

inside của một nút ở đây mang ý nghĩa là xác suất của nút đó, trong trường hợp nút này là nút gốc S , thì inside chính là xác suất của cây.^[1]

3.1.2 Xác suất outside.

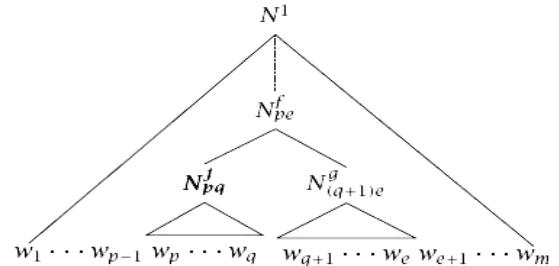
Giả sử ta có nút D và có hai luật $A \rightarrow B D$ và $C \rightarrow D E$ thì outside của D sẽ được tính bằng:

$$P1 = \lg P(A \rightarrow B D) + \text{inside}(B) + \text{outside}(A)$$

$$P2 = \lg P(C \rightarrow D E) + \text{inside}(E) + \text{outside}(C)$$

$$\text{Outside}(D) = \max(P1, P2);$$

Outside của một nút ở đây mang ý nghĩa tương trung cho độ liên kết của nút với nút gốc, hay nói một cách khác nó là xác suất của một nút về việc từ nút đó có bao nhiêu khả năng tìm được nút gốc.^[2]



Hình 4 Xác suất outside

3.2 Giải thuật A-star.

Bộ phân tích sẽ tạo ra hai tập : agenda và chart. Trong đó agenda là tập nút đang xếp hàng chờ được xem xét, còn chart là tập các nút đã xét qua. Bộ phân tích cú pháp sẽ lấy ra nút có độ ưu tiên cao nhất, kết hợp với các phần tử trong chart, tạo ra một tập các nút mới, các nút này sẽ được thêm vào agenda để chờ xử lý tiếp. Bộ phân tích sẽ dừng lại khi tìm được $S[1, n+1]$ là đáp án cuối cùng hoặc không còn nút để xét. Nếu kết thúc giải thuật, tìm được đáp án cuối cùng thì quá trình phân tích thành công, ngược lại quá trình phân tích thất bại.^[3]

Vấn đề lớn nhất ở đây chính là hàm mục tiêu để chọn ra phần tử có độ ưu tiên cao nhất. Hàm mục tiêu trong A-star bao giờ cũng có hai thành phần là chi phí và ước lượng. Chi phí là quãng đường đã xét duyệt qua, ở đây ta có thể dùng inside như một hàm chi phí. Còn về ước lượng quãng đường đi đến đích, bản thân outside đặc trưng cho khả năng tìm được đường đến nút gốc nên outside là một sự lựa chọn tuyệt vời. Vậy công thức hàm mục tiêu của chúng ta sẽ như dưới đây:

$$f(\text{nút}) = \text{inside}(\text{nút}) + \text{outside}(\text{nút}).$$

3.3 Kết quả và đánh giá.

Xác suất của các luật trong bộ phân tích cú pháp được tính toán dựa vào việc thống kê từ tập huấn luyện viettreebank gồm 2000 câu đã được phân tích đúng 100%.

Tập dữ liệu test	Phần trăm phân tích được	Thời gian
630 câu bất kì	92%	20mins

Tập dữ liệu test	Phần trăm phân tích chính xác	Thời gian
500 câu bất kì	60-70%	15mins

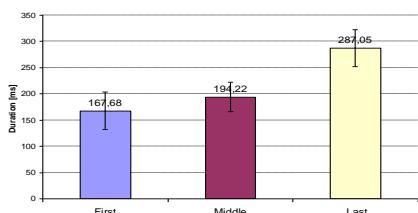
4. MÔ HÌNH HÓA TRƯỜNG ĐỘ ÂM TIẾT TIẾNG VIỆT.

Trong tất cả các hệ thống tổng hợp tiếng nói, để có thể đạt được độ tự nhiên cao cho tiếng nói tổng hợp, một vấn đề cần phải xử lý đó là dự đoán được trường độ cho các âm tiết. Một trong những phương pháp truyền thống trong việc sử mô hình hóa trường độ đó là sử dụng các tập luật, nhưng việc xây dựng và chọn lựa được các luật là một công việc rất khó và đạt độ chính xác không cao.

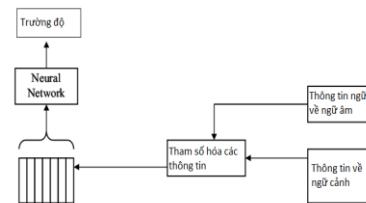
Cùng với sự phát triển của trí tuệ nhân tạo, và học máy đã có nhiều phương pháp huấn luyện đạt được độ chính xác cao trong việc dự đoán trường độ như sử dụng cây quyết định hoặc mạng Neuron. Trong đó mạng Neuron là phương pháp đạt được độ chính xác cao nhất đối với việc đưa ra trường độ của âm tiết.

4.1 Các yếu tố ảnh hưởng đến trường độ âm tiết

Có nhiều yếu tố ảnh hưởng đến trường độ của âm tiết trong tiếng Việt, các yếu tố này có thể phân thành các nhóm chính. Ngữ âm, ngữ cảnh, và vị trí của âm tiết. Ví dụ đối với cùng một âm tiết thì nếu như âm tiết đó đứng ở vị trí cuối câu hoặc cuối từ thì trường độ của âm tiết đó dài hơn hẳn so với các thẻ hiện của âm tiết đó ở vị trí khác.[8]



Hình 5 Ảnh hưởng của vị trí đến trường độ của âm tiết

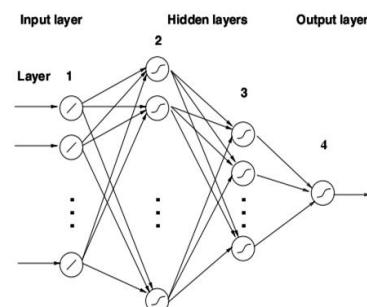


Hình 6 Mô hình hóa trường độ

Bộ phân tích cú pháp sẽ xử lý đối với các văn bản đầu vào để đưa ra thông tin về ngữ cảnh của âm tiết như là các âm tiết liền trước, liền sau, vị trí của âm tiết trong câu... Về mặt ngữ âm, thì một âm tiết tiếng Việt thuộc một trong tám loại V, VC, CV, CVC, VV, VVC, CVV, CVVC. Trong đó V là nguyên âm và C là phụ âm (Vowel and Consonant)[26]. Dựa vào đó, các âm tiết sẽ được phân tích để đưa ra các thông tin về mặt ngữ âm như các thành phần cấu tạo nên âm tiết, số lượng thành phần của âm tiết, thanh điệu. Các thông tin này được tham số hóa về trong khoảng (0,1) để làm đầu vào cho mạng neuron. Giá trị đầu ra của mạng Neuron chính là giá trị về trường độ của âm tiết.

4.1 Mạng Neuron

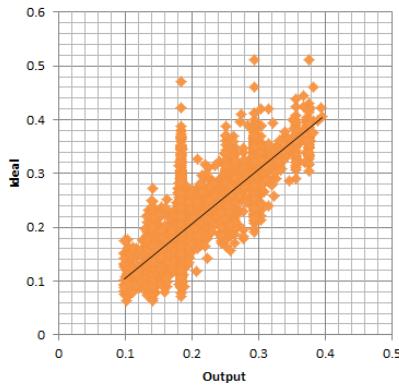
Mạng Neuron sử dụng trong quá trình huấn luyện là mô hình mạng dẫn tiến nhiều lớp. Sử dụng giải thuật học lan truyền ngược (Backpropagation)[27]. Với một tầng đầu vào, hai tầng ẩn và một đầu ra. Kiến trúc của mạng Neuron được lựa chọn theo phương pháp thử sai.



Hình 7. Mạng Neuron

4.2 Kết quả

Trường độ của âm tiết được dự đoán đạt độ chính xác 85% với dữ liệu đã được huấn luyện. Và độ chính xác 82% với các âm tiết nằm ngoài tập huấn luyện. Kết quả này đạt độ chính xác tương đối cao so với các hệ thống hệ thống sử dụng CART 77% [8].



5 MÔ HÌNH HÓA CAO ĐỘ

Trong các hệ thống tổng hợp tiếng nói, việc sinh ra đường F0 là một vấn đề thiết yếu để có thể thu được những âm thanh tổng hợp tự nhiên. Hiện nay đã có một số mô hình sinh ra đường tần số cơ bản (F0) của tiếng Việt. Các mô hình này đã đạt được những kết quả đáng chú ý, nhưng điều mới là mô hình ngữ điệu cho câu khẳng định.

Trong nghiên cứu này, chúng tôi xem xét đến ảnh hưởng của hai yếu tố lên chất lượng ngữ điệu của câu hỏi là ngữ điệu của toàn câu và ngữ điệu của phần cuối câu. Đầu tiên một bộ dữ liệu được xây dựng, sau đó chúng tôi thực hiện biến đổi F0 của các câu này theo hai yếu tố và thực hiện một bài kiểm tra cảm thụ để đánh giá được vai trò của chúng.

5.1 Khác nhau giữa ngữ điệu câu hỏi và câu khẳng định

Theo như trong [8], ngữ điệu của câu hỏi có xu hướng đi lên cao ở cuối câu mà không bị ảnh hưởng bởi âm tiết cuối là mang thanh điệu nào. Đồng thời trong [8], [24] và [25], cũng đưa ra kết luận là câu hỏi được nói với một âm vực cao hơn câu hỏi. Do đó, trong nghiên cứu này, chúng tôi sẽ xem xét ảnh hưởng của hai yếu tố này trong ngữ điệu của câu hỏi, nhằm để xuất ra được một mô hình ngữ điệu cho câu hỏi.

5.2 Bộ dữ liệu

Bộ dữ liệu bao gồm 25 câu khẳng định, với nội dung được trích ra từ các đoạn hội thoại trong cuộc sống hàng ngày. Mục đích của việc lựa chọn các câu đối thoại là có thể dễ dàng biến đổi sang các câu hỏi nghi vấn với nội dung giống hệt nhưng với ý định hỏi để xác nhận.

5.3 Thí nghiệm thay đổi F0 để đạt ngữ điệu câu hỏi

5.3.1 Ảnh hưởng của ngữ điệu toàn câu

Phương pháp

Bài test thứ nhất nhằm xác định ảnh hưởng của ngữ điệu toàn câu với trường hợp của câu hỏi. Chúng tôi thực hiện việc nâng cao ngữ điệu của 25 câu đã có lên lần lượt 2 mức là 10% F0 trung bình của cả câu và 20% F0 trung bình của cả câu. Việc nâng cao ngữ điệu câu được thực hiện tự động bằng phần mềm Praat. Sau đó 50 câu này được trộn với 25 câu gốc tạo ra một bộ dữ liệu test gồm 75 câu. Bộ dữ liệu này sau đó được 9 nam và 9 nữ nghe và đánh giá xem những câu này có giống câu hỏi không, và mức độ tin tưởng của họ khi đưa ra những đánh giá đó.

Kết quả

Mức độ tăng	Tỉ lệ chọn câu hỏi	Độ lệch chuẩn của tỉ lệ chọn	Mức độ tin tưởng
10%	50.22%	23.93%	75.82%
20%	76.00%	17.66%	82.28%

Bảng 2 Thống kê kết quả đối với 9 nam

Mức độ tăng	Tỉ lệ chọn câu hỏi	Độ lệch chuẩn của tỉ lệ chọn	Mức độ tin tưởng
10%	35.56%	20.14%	77.59%
20%	73.33%	12.81%	77.99%

Bảng 3 Thống kê kết quả đối với 9 nữ

Mức độ tăng	Tỉ lệ chọn câu hỏi	Độ lệch chuẩn của tỉ lệ chọn	Mức độ tin tưởng
10%	42.89%	22.74%	76.70%
20%	74.66%	15.02%	80.13%

Bảng 4 Thống kê kết quả chung

5.3.2 Ảnh hưởng của ngữ điệu cuối câu

Bài test thứ hai cũng được thực hiện tương tự bài test thứ nhất. Điểm khác biệt duy nhất là 25 câu đã có được nâng cao ngữ điệu của âm tiết cuối lên lần lượt 2 mức là 10% F0 trung bình của cả câu và 20% F0 trung bình của cả câu. Công việc này cũng được thực hiện hoàn toàn tự động bằng phần mềm Praat.

Kết quả

Mức độ tăng	Tỉ lệ chọn câu hỏi	Độ lệch chuẩn của tỉ lệ chọn	Mức độ tin tưởng
10%	55.11%	22.34%	78.61%
20%	80.44%	16.78%	85.55%

Bảng 5 Thống kê kết quả đối với 9 nam

Mức độ tăng	Tỉ lệ chọn câu hỏi	Độ lệch chuẩn của tỉ lệ chọn	Mức độ tin tưởng
10%	43.11%	15.59%	75.32%
20%	72.89%	13.08%	78.44%

Bảng 6 Thống kê kết quả đối với 9 nữ

Mức độ tăng	Tỉ lệ chọn	Độ lệch chuẩn	Mức độ tin tưởng
10%			
20%			

	câu hỏi	của tỉ lệ chọn	
10%	49.11%	19.68%	76.97%
20%	76.67%	15.10%	82.00%

Bảng 7 Thống kê kết quả chung

5.4 Kết luận

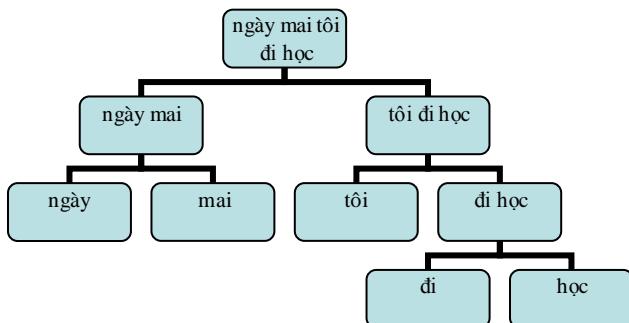
Dựa vào kết quả của hai bài test ta có thể thấy vai trò của hai yếu tố trong ngữ điệu của câu hỏi là gần tương đương nhau. Ngữ điệu của cuối câu khi được nâng lên cho kết quả cao hơn một chút so với khi ngữ điệu của cả câu được nâng lên. Khi mức nâng là 20% với cả hai trường hợp thì tỉ lệ chọn là câu hỏi đã lên đến xấp xỉ 75%, đây là kết quả khá tốt trong trường hợp câu hỏi không có từ để hỏi. Do vậy, khi áp dụng vào hệ tổng hợp tiếng nói, ta cần kết hợp giữa cả hai yếu tố này sao cho hợp lý để đạt kết quả cao nhất.

6. TỔNG HỢP MỨC THẤP

Tổng hợp mức thấp là quá trình lựa chọn, tìm kiếm đơn vị âm trong cơ sở dữ liệu đơn vị âm và ghép nối chúng để tạo nên tiếng nói cần tổng hợp. Các loại đơn vị âm có thể dùng để ghép nối với chiều dài tăng dần là âm vị, bán âm tiết, âm đầu/vần, âm tiết, cụm từ. Theo [8], trong tiếng Việt, loại đơn vị âm có thể dùng để tổng hợp cho chất lượng tốt là bán âm tiết, âm tiết, cụm từ. Quan điểm được thừa nhận rộng rãi hiện nay là đơn vị âm càng lớn thì chất lượng tiếng nói tổng hợp càng cao do giảm thiểu được số điểm ghép nối tín hiệu, tuy nhiên đôi lại là sự tăng lên về kích thước cơ sở dữ liệu. Với mục tiêu là nâng cao chất lượng tiếng nói tổng hợp, giải pháp là sử dụng thuật toán lựa chọn đơn vị không đồng nhất kết hợp với phương pháp TD-PSOLA để điều khiển các tham số ngữ điệu tiếng nói.

6.1. Lựa chọn đơn vị

Các loại đơn vị được sử dụng với mức ưu tiên giảm dần là cụm từ, âm tiết, bán âm tiết. Câu cần tổng hợp sẽ được chia ra thành các cụm từ theo các mức khác nhau nhờ quá trình phân tích cú pháp. Ví dụ như hình minh họa dưới đây cho câu “Ngày mai tôi đi học”.



Hình 8 Cây phân tích cú pháp

Quá trình tìm kiếm sẽ được bắt đầu từ gốc, sau đó đi xuống các nhánh. Việc tìm kiếm sẽ dừng lại ở mức cao nhất có thể ngay khi tìm thấy cụm từ hoặc đi tới mức lá là các âm tiết. Cách thức phân chia để tìm kiếm này làm tăng xác suất tìm thấy của những cụm từ có độ dài lớn hơn một âm tiết hơn là việc chọn ngẫu nhiên cụm từ theo một độ dài xác định nào đó để tìm kiếm. Đây là ý tưởng chủ đạo trong thuật toán lựa chọn đơn vị không đồng nhất.

Trong trường hợp không tìm thấy ứng viên nào ở mức lá, âm tiết còn lại sẽ được tổng hợp ở mức bán âm tiết. Theo [8], việc tổng hợp ở mức bán âm tiết có thể tổng hợp được hầu hết các âm tiết trong tiếng Việt.

6.2. Tổ chức dữ liệu

Dữ liệu được dùng trong tổng hợp là các đoạn văn, hội thoại tiếng Việt được thu âm bởi một giọng đọc duy nhất. Dữ liệu văn bản bao gồm 250 đoạn của 630 câu với 10852 mẫu của khoảng 1600 âm tiết phân biệt, được tổ chức theo cấu trúc XML. Các thông tin của âm tiết như về thành phần cấu tạo, năng lượng, thanh điệu, trường độ được phân tích offline và lưu trữ để phục vụ cho quá trình lựa chọn và ghép nối đơn vị. Dữ liệu âm thanh là các file wav tương ứng với các đoạn văn bản, kích thước 68M, độ dài 37 phút. Kích thước cơ sở dữ liệu như vậy được đánh giá là có thể chấp nhận được đối với bộ tổng hợp tiếng nói.

6.3. Hàm chi phí

Kết quả của quá trình tìm kiếm đơn vị sẽ là tập các mẫu của các đơn vị âm tìm thấy. Một đơn vị âm có thể có nhiều mẫu. Nhiệm vụ là cần chọn ra mẫu tốt nhất trong đó để ghép nối. Theo [8], mẫu được chọn sẽ dựa theo tiêu chí có sự khác biệt nhỏ nhất về ngữ điệu với đơn vị âm đích mong muốn. Sự sai khác này được lượng hóa thành hàm chi phí. Hàm chi phí là tổng của tất cả các độ méo bao gồm hai loại độ méo chính:

- Độ méo của đơn vị âm thể hiện bằng sự khác nhau giữa đơn vị âm được lựa chọn với đơn vị âm cần tổng hợp. Đây gọi là chi phí đích.

- Độ méo về sự liên tục được thể hiện bằng khoảng cách giữa đơn vị âm được chọn so với đơn vị âm trước đó. Đây gọi là chi phí ghép nối.

Thuật toán Viterbi được sử dụng để chọn ra các đơn vị âm có hàm chi phí nhỏ nhất.

6.4 Thuật toán TD-PSOLA

Mục tiêu của việc tổng hợp mức cụm từ là giảm thiểu số điểm ghép nối, nâng cao chất lượng tiếng nói tổng hợp. Tuy nhiên, một điều dễ nhận thấy là ta không thể xây dựng tập dữ liệu có đầy đủ các âm tiết tiếng Việt. Vì vậy, ta cần tổng hợp sử dụng bán âm tiết. Việc này sẽ dẫn tới hiện tượng không liên tục tại điểm ghép nối giữa các đơn vị âm (về cao độ, về phô, về pha). Sự không liên tục này xảy ra do sự khác nhau về ngữ cảnh của các đơn vị âm hoặc do quá trình phân đoạn tiếng nói. Chúng ta cần một kĩ thuật cho phép điều khiển các tham số ngữ điệu của đơn vị âm cần tổng hợp để khi ghép nối giảm được tối thiểu sự không liên tục giữa chúng. Cụ thể mục tiêu là thay đổi biên độ, trường độ và cao độ

của đoạn tiếng nói. Việc sửa đổi biên độ có thể dễ dàng được thực hiện bởi bộ nhân trực tiếp, tuy nhiên trường độ và cao độ không đơn giản như vậy. Kỹ thuật được đề xuất là PSOLA (Pitch Synchronous Overlap and Add). Đây là một kỹ thuật dùng rất phổ biến trong các chương trình tổng hợp tiếng nói tiếng Việt và các tiếng khác.

Phương pháp TD-PSOLA là phiên bản trên miền thời gian của PSOLA trong đó những thay đổi trong tín hiệu tiếng nói được thực hiện trực tiếp trên miền thời gian, tín hiệu được tổng hợp bằng cách đơn giản là sao chép tín hiệu phân tích tương ứng.

6.5 Kết quả tổng hợp

Chương trình đã tổng hợp một đoạn văn bản với số lượng 11 câu. Kết quả tiếng nói tổng hợp ban đầu tương đối dễ nghe, tuy nhiên về mức độ tự nhiên và ngữ điệu còn hạn chế. Việc này cần được cải tiến bằng cách thay đổi các tham số trong hàm chi phí, sử dụng thuật toán làm mượt các điểm ghép nối.

7. KẾT LUẬN

Một mô hình hệ thống tổng hợp tiếng nói với năm module được đề xuất là mô hình tiên tiến nhất trong tổng hợp tiếng nói chất lượng cao dựa trên phương pháp ghép nối. Việc chuẩn hóa văn bản đầu vào giúp tổng hợp được nhiều trường hợp viết tắt và nhập nhằng trong tiếng Việt. Phân tích cú pháp và lựa chọn đơn vị ghép nối đảm bảo độ dễ nghe của âm thanh nhờ giảm thiểu được chi phí ghép nối. Và nhờ vào phân tích ngữ điệu (trường độ và cao độ) âm thanh tổng hợp có được độ tự nhiên cao.

Trong thời gian tới, nhóm chúng tôi sẽ tập trung vào cải tiến các module, ráp nối và hoàn thiện để hệ thống hoạt động ổn định, đạt chất lượng cao.

8. LỜI TRI ÂN

Chúng tôi xin gửi lời cảm ơn chân thành TS. Trần Đỗ Đạt tại Trung tâm nghiên cứu Mica và ThS. Nguyễn Thị Thu Trang – Bộ môn Công nghệ phần mềm – Viện Công nghệ thông tin và truyền thông đã hết lòng hướng dẫn và chỉ bảo chúng tôi trong suốt quá trình nghiên cứu.

Chúng tôi cũng bày tỏ lòng biết ơn trung tâm nghiên cứu Mica đã tạo điều kiện về cơ sở vật chất cho chúng tôi trong quá trình học tập và nghiên cứu.

Cuối cùng, chúng tôi xin cảm ơn các thầy cô giáo trong Viện Công nghệ thông tin và truyền thông đã giảng dạy cho chúng tôi các kiến thức hữu ích trong suốt những năm vừa qua.

9. TÀI LIỆU THAM KHẢO

- [1] Fei Xia, “Inside-Outside algorithm”, LING572.
- [2] Michael Collins, “Head-Driven Statistical Models for Natural Language Parsing”, MIT Computer Science and Artificial Intelligence Laboratory.

[3] Dan Klein and Christopher D. Manning. 2003. “A* parsing: Fast exact Viterbi parse selection. In Proceedings of the Human Language Technology Conference and the North American Association for Computational Linguistics”(HLT-NAACL).

[4] Adam Pauls and Dan Klein, “K-Best A* Parsing”, Computer Science Division University of California, Berkeley.

[5] Hoàng Anh Việt, “Phân tích cú pháp tiếng Việt sử dụng mô hình xác suất PCFG”, đồ án tốt nghiệp đại học năm 2006.

[6] Phạm Thị Nhung, “Phân tích cú pháp tiếng Việt sử dụng beam search”, đồ án tốt nghiệp đại học năm 2009.

[7] Đỗ Bá Lâm, Lê Thành Hương, “Implementing a Vietnamese syntactic parser using HPSG”, Khoa Công nghệ thông tin, trường Đại học Bách khoa Hà Nội.

[8] Trần Đỗ Đạt, “Synthèse de la parole a partir du texte en langue Vietnamienne”, Thèse en cotutelle international MICA, Hanoi, 2007.

[9] Minghui Dong, Kim-Teng Lua, Haizhou Li, “A Unit Selection-based Speech Synthesis Approach for Mandarin Chinese”, Institute for Infocomm Research.

[10] Vũ Hải Quân, Cao Xuân Nam, “Tổng hợp tiếng nói tiếng Việt, theo phương pháp ghép nối cụm từ”.

[11] Mạc Đăng Khoa, “Modeling the prosody of Vietnamese language for speech synthesis”, Thesis for the degree of MASTER OF SCIENCE, 2007.

[12] Trần Đỗ Đạt, Eric Castelli, “Generation of F0 contours for Vietnamese speech synthesis”.

[13] Nguyễn Thị Thu Trang, Phạm Thị Thanh, Trần Đỗ Đạt, “A method for Vietnamese Text Normalization to improve the quality of speech synthesis”.

[14] Taylor Paul. “Text-To-Speech Synthesis. s.l : Cambridge University Press, 2009”

[15] Language Technologies Institute Carnegie Mellon University, “Non – Standard Word and Homograph Resolution for Asian Language Text Analysis”

[16] Richard Sproat, Alan Black, Stanley Chen, Shankar Kumar, Mari Osten-dorf, Christopher Richards, “Normalization of Non-Standard Words. Computer Speech and Language, Volume 15, Issue 3. July 2001”.

[17] Stanley F. Chen, Joshua Goodman, “An Empirical Study of Smoothing Techniques for Language Modeling”

[18] K. Sreenivasa Rao, B. Yegnanarayana, “Modeling durations of syllables using neural network”, ScienceDirect, 2006.

[19] Martti Vainio, “Artificial Neural Network Based Prosody Models for Finnish”, University of Helsinki, Department of Phonetics.

[20] Marcello Balestri, Alberto Pacchiotti, Silvia Quazza, Pier Luigi Salza, Stefano Sandri, “Choose the best to modify the least: a new generation concatenative synthesis system”, CSELT - Centro Studi e Laboratori Telecomunicazioni Sp.A, Torino, Italy.

[21] Min Chu, Hu Peng, Hong-yun Yang, Eric Chang, “Selecting non-uniform units from a very large corpus for concatenative speech synthesizer”, Microsoft Research China, Beijing.

[22] Paul Taylor, “Text-to-SpeechSynthesis”, University of Cambridge, Cambridge University Press.

[23] Mark Tatham, Katherine Morton, “Development in Speech Synthesis”, Wiley, 2005.

[24] Đỗ T.D., Trần T.H., et Boulaiki G, “*Intonation in vietnamese*”, Intonation systems: A survey of 22 languages, Hirst & Di Cristo (ed.), Cambridge U.P, 1998.

[25] Nguyễn T.T.H, “*Contribution à l'étude de la prosodie du vietnamien. Variations de l'intonation dans les modalités: assertive, interrogative et impérative*”, Thèse 2004, Doctorat de Linguistique Théorique, Formelle et Automatique.

[26] Trần Ngọc Dũng, “*Cẩm nang văn phạm tiếng Việt*”, 2010.

[27] Christopher M. Bishop, “*Neural network for pattern recognition* ”.

Giải pháp ngữ nghĩa – Tích hợp dữ liệu, gợi ý và tìm kiếm thông tin cho hệ thống hướng dẫn du lịch thông minh

Phan Thanh Hiền, Nguyễn Anh Đức

Tóm tắt: Ngành du lịch phát triển mạnh mẽ, kéo theo đó là những yêu cầu ngày càng cao của khách du lịch về thông tin du lịch nơi mà họ muốn tới. Tuy nhiên tại thời điểm hiện tại, khách du lịch chủ yếu tiếp cận thông tin một cách hạn chế thông qua các tourism guidebook. Cũng có một số website triển khai cung cấp hỗ trợ thông tin du lịch cho khách hàng, khách hàng đã có thể xem được hình ảnh cũng như video. Tuy nhiên dữ liệu được trích xuất rời rạc theo từng kiểu dữ liệu cũng như giữa chúng không có mối liên hệ nào cả. Bên cạnh đó thông tin các tour du lịch thường là cố định và dùng chung cho tất cả các đối tượng mà không để ý đến sở thích hay mối quan tâm của khách hàng. Trước tình hình thực tế đó nhóm nghiên cứu đã xây dựng một hệ thống thông minh được bao bọc bởi một tầng dữ liệu ngữ nghĩa có khả năng cung cấp được cho người dùng những chức năng gợi ý, tìm kiếm thông minh đồng thời đưa ra kết quả trực quan, sinh động và có mối quan hệ với nhau. Các chức năng được triển khai ở trên nhiều môi trường khác nhau (Smart phone, WEB). Hệ thống được thiết kế theo hướng có thể mở rộng dần, dữ liệu có thể tích hợp, bổ sung dễ dàng từ nhiều nguồn phân tán, chức năng có thể thêm bớt, nâng cấp mà không ảnh hưởng tới các thành phần khác của hệ thống. Những phần được nêu trong bài báo này chỉ là những phần cốt lõi nhất, đặt nền tảng cho hệ thống hoàn chỉnh sau này chứ chưa phải là toàn bộ hệ thống cuối cùng mà nhóm nghiên cứu muốn đạt tới.

Từ Khóa: Du lịch, semantic web, tìm kiếm, thông minh, tích hợp dữ liệu

1. GIỚI THIỆU

Ngày nay, khi chất lượng cuộc sống được nâng cao, con người có nhiều điều kiện hơn trong việc khám phá những điều mới lạ, các nền văn hóa hay các kỳ quan thiên nhiên. Theo số liệu thống kê của tổ chức du lịch thế giới (UNWTO) thì số lượng khách du lịch trên thế giới tăng đều đặn trong 10 năm qua, dự kiến du lịch thế giới sẽ tăng trưởng 200% trong vòng 20 năm tới.

Song song với sự phát triển của ngành du lịch thì nhu cầu tiếp cận thông tin du lịch cũng ngày càng tăng lên. Du khách không những muốn biết những thông tin sơ lược được ghi trong các guidebook mà họ còn muốn được xem trước hình ảnh, video của nơi mà họ muốn đến ở nhiều góc cạnh. Hiện tại cũng đã có nhiều website cung cấp dữ liệu đa phương tiện

về địa điểm du lịch cho khách hàng, tuy nhiên hầu hết các hệ thống này đều lưu trữ dữ liệu theo phương pháp truyền thống. Việc tìm kiếm dữ liệu trở nên bị hạn chế, phân tách theo từng loại dữ liệu, không có mối quan hệ với nhau và không chọn lọc. Bên cạnh đó thông tin về các tour du lịch cũng hầu hết cố định và không quan tâm đến sở thích của từng du khách hoặc từng nhóm du khách. Việc xây dựng các kho dữ liệu phục vụ cho các hệ thống trên không đơn giản, với mỗi một hệ thống khác nhau lại phải dùng một cơ sở dữ liệu khác nhau và gần như không có khả năng tích hợp với nhau.

Cùng với sự phát triển của semantic web (semantic web là cụm từ dùng để nói về công nghệ mô tả dữ liệu sao cho máy tính có thể hiểu được) trong những năm gần đây, trên thế giới đã có một số mô hình thông minh cung cấp cho người dùng những thông tin giàu ý nghĩa hơn và dễ dàng tích hợp, mở rộng dữ liệu. Cụ thể như **hệ thống hướng dẫn tham quan bảo tàng hướng ngữ cảnh**[4] của nhóm nghiên cứu gồm Fabien và các cộng sự của ông. Hệ thống dựa trên những thông tin khách hàng nhập vào qua PDA được phát khi vào bảo tàng, từ đó đưa ra những gợi ý tham quan phù hợp với sở thích, mối quan tâm của khách. Hệ thống được phát triển dựa trên dữ liệu của bảo tàng khoa học tự nhiên quốc gia – một trong những bảo tàng lớn nhất Đài Loan. **Hệ thống tìm kiếm ngữ nghĩa thông tin di sản văn hóa trên di động dựa vị trí địa lý**[5] của nhóm nghiên cứu thuộc đại học Amsterdam – Hà Lan. Hệ thống dựa trên định vị GPS sẽ có thể biết được người sử dụng đang ở đâu qua đó cho biết nơi đó ở cạnh những di sản văn hóa nào, ở đây đã từng xảy ra sự kiện lịch sử gì và những nhân vật lịch sử nào được sinh ra ở đây, những tác phẩm nghệ thuật nào đã được lấy cảm hứng ở khu vực này ...v.v. Thông tin được truy vấn ngữ nghĩa từ nhiều nguồn.

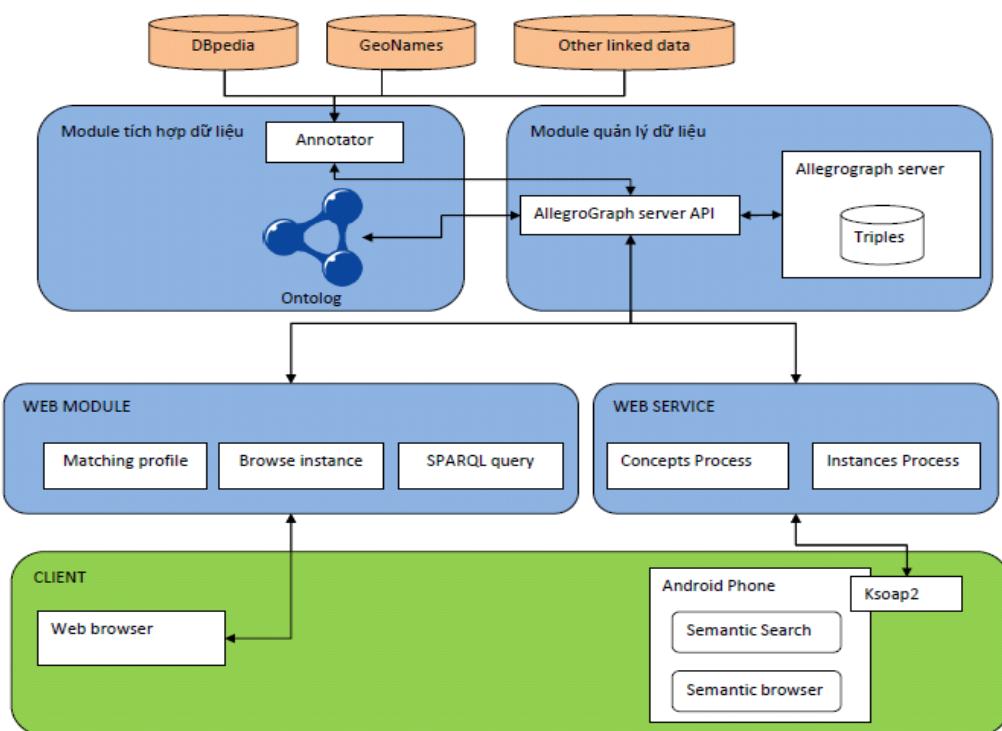
Từ thực tế phân tích ở trên và qua việc khảo sát các hệ thống đã có trên thế giới, nhóm nghiên cứu đã xây dựng hệ thống VTIO (Vietnam Travel Information Online) và đã đạt được những thành công bước đầu. Hệ thống sử dụng công nghệ semantic web mô tả tầng dữ liệu ngữ nghĩa nhằm phục vụ cho việc tích hợp dễ dàng dữ liệu từ nhiều nguồn ngữ nghĩa phân tán (dbpedia, geonames...v.v.), đồng thời cung cấp được cho người dùng các tính năng thông minh mà một hệ thống thông thường không thể làm được như tìm kiếm thông tin chính xác nhờ các ràng buộc, gợi ý đường đi có dựa vào trọng số sở thích, duyệt dữ liệu theo cú pháp ngôn ngữ tự nhiên đơn giản SVO (Subject/Verb/Object – Chủ ngữ/Động từ/Tân ngữ) ...v.v. Cụ thể giải pháp sẽ được mô tả ở phần tiếp theo.

2. GIẢI PHÁP VÀ KIẾN TRÚC HỆ THỐNG

Mô hình dữ liệu: Để có thể dễ dàng mở rộng, tích hợp dữ liệu cũng như hệ thống có khả năng suy diễn, nhóm đã sử dụng một tầng dữ liệu ngữ nghĩa để mô hình hóa và lưu trữ dữ liệu, mô hình này sẽ mô tả tất cả các khái niệm trong lĩnh vực du lịch cũng như các khái niệm liên quan (gọi là ontology). Hệ thống sử dụng ngôn ngữ RDF (Resource Description Framework – Khung đặc tả tài nguyên) để tạo nên mô hình này. Dựa trên ontology này toàn bộ dữ liệu được nhập vào hệ thống như là các thể hiện (Instance) của các khái niệm (Concept). Giả sử chúng ta có thông tin: “Nhà hát lớn(Hanoi Opera House) là một **nhà hát nổi tiếng**, ở **cạnh hồ Hoàn Kiếm** và **liên quan đến chủ đề kiến trúc Pháp**” thì sẽ được lưu trữ dưới dạng:

```
<owl:Thing rdf:type="resource="#Theater"/>
<rdfs:label>
  Hanoi opera house (Nha Hat Lon Ha Noi)
</rdfs:label>
<isWellKnown>true</isWellKnown>
<nearBy rdf:resource="#hoan-kiem-lake"/>
<relatedToTopic rdf:resource="#france-architecture-topic"/>
</owl:Thing>
```

Thiết kế hệ thống:



Hình 1 : Kiến trúc hệ thống VTIO - mức cao

Module tích hợp dữ liệu sẽ có nhiệm vụ lọc tách dữ liệu từ các nguồn dữ liệu ngữ nghĩa phân tán và đưa vào Allegrograph Server để quản lý thông qua allegrograph server API.

Module quản lý dữ liệu: Hệ thống sử dụng allegroGraph server để quản lý dữ liệu dạng triples (Một phát biểu được tách thành 3 phần Subject – Predicate – Object gọi là một triple). Mọi thao tác với dữ liệu được thông qua AllegroGraph server API.

Web module: Ở đây các thao tác chức năng của hệ thống như: khớp nối profile của người dùng để gợi ý hành trình du lịch, truy vấn SPARQL phục vụ cho môi trường WEB được thực hiện.

Web service: Nhằm phục vụ cho môi trường phân tán, đa môi trường – hệ thống cung cấp các dịch vụ để thao tác với hệ thống. Tuy nhiên các dịch vụ này đều ở mức thấp, thao tác với ontology (Các khái niệm, các thể hiện).

Client: Hệ thống triển khai trên cả môi trường WEB lẫn Android smart phone. Tận dụng tối đa ưu nhược điểm của mỗi môi trường. Client kết nối với server thông qua mạng không dây đối với di động.

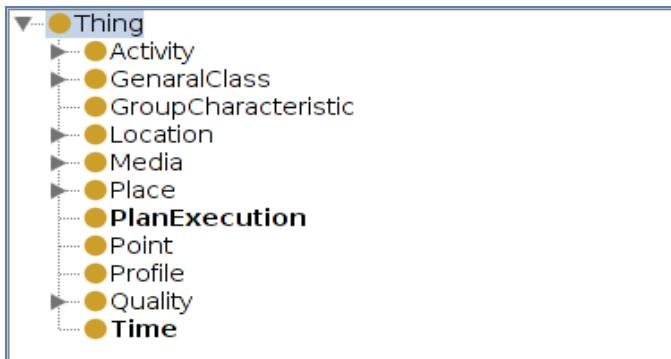
Nhân của hệ thống là một ontology lõi – hình thức hóa bằng ngôn ngữ RDF++ định nghĩa các khái niệm và mối quan hệ giữa chúng (RDF++ là ngôn ngữ mở rộng từ RDF hỗ trợ các xử lý nâng cao, có cho phép sử dụng một số thuộc tính quan trọng của OWL nhưng lại nhẹ hơn OWL).

3. XÂY DỰNG ONTOLOGY LÕI

Việc xây dựng ontology được thiết kế theo quy trình top-down (làm mịn dần từ trên xuống, đồ dần từ những khái niệm chung nhất đến những khái niệm chi tiết). Trong quá trình xây dựng có tham khảo một số ontology cùng lĩnh vực du lịch. Cụ thể ở đây có ontology của hệ thống CRUZAR (tây ban nha) và Travel ontology v1.0 của Holger Knublauch. Ontology được xây dựng ở dạng mở và luôn luôn được bổ sung, cập nhật các khái niệm cũng như

các thuộc tính mới nhằm phục vụ cho quá trình phát triển hệ thống. Ontology của hệ thống được xây dựng dựa trên ngôn ngữ **RDF++** với sự hỗ trợ của thuộc tính **owl:sameAs** (Thuộc tính cho phép tham chiếu tới các tài nguyên ở các nguồn dữ liệu ngữ nghĩa khác nhau) giúp cho việc liên kết, tích hợp với các nguồn dữ liệu ngữ nghĩa phong phú hiện nay như dbpedia,

geonames trở nên đơn giản, dễ dàng. Đây là điểm đặc biệt của hệ thống so với các nghiên cứu khác trong cùng lĩnh vực. Hiện tại ontology bao gồm một số lớp và thuộc tính sau đây:

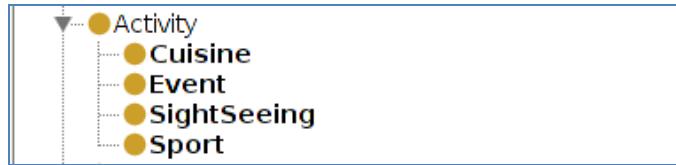


Hình 2 : Các khái niệm mức 1 của hệ thống

- **Activity:** Là khái niệm mô tả các hoạt động liên quan đến du lịch.
- **Place:** Là khái niệm mô tả các địa điểm liên quan đến du lịch
- **Location:** Là khái niệm mô tả các vị trí địa lý
- **Point:** Là khái niệm mô tả một tọa độ địa lý (lat, long)
- **Profile:** Là khái niệm mô tả hồ sơ du lịch của khách, từ đây để có thể có căn cứ lên tour phù hợp cho khách du lịch sau này.
- **Quality:** là khái niệm đánh giá chất lượng của nhà hàng, khách sạn. Đồng thời cho biết phong cách của một địa điểm.
- **Media:** Là khái niệm mô tả các tài nguyên đa phương tiện liên quan đến một địa điểm du lịch nào đó.
- **GeneralClass:** Là khái niệm được xây dựng nhằm mục đích gộp các lớp phụ phục vụ làm ObjectProperty. Hỗ trợ làm rõ các khái niệm chính. Tránh việc làm loãng ontology khi tách riêng chúng ra.
- **PlanExecution:** khái niệm mô tả loại hình du lịch
- **Time:** khái niệm mô tả về thời gian

Các khái niệm ở mức chi tiết: đây là các khái niệm được phân rã nhỏ hơn của các khái niệm chính nêu trên.

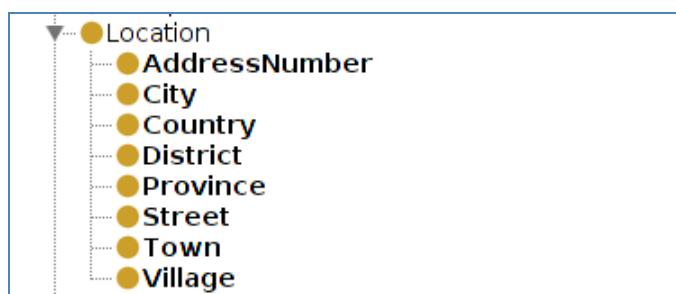
Các lớp con của Activity:



Hình 3 Các lớp con của Activity

- **Cuisine:** Mô tả các hoạt động ẩm thực.
- **Event:** Các sự kiện đặc biệt, chỉ xảy ra trong 1 khoảng thời gian nhất định.
- **SightSeeing:** Các hoạt động tham quan hàng ngày.
- **Sport:** Các hoạt động thể thao

Các lớp con của Location:



Hình 4 Các lớp con của Location

- **AddressNumber:** Dùng mô tả số nhà, số đường ...v.v.
- **City:** Mô tả thành phố
- **Country:** Mô tả một quốc gia
- **District:** Mô tả một quận, huyện
- **Province:** Mô tả một tỉnh
- **Town:** Mô tả một thị trấn
- **Village:** Mô tả xã, phường

Các lớp con của Place



Hình 5 Các lớp con của Place

- **TravelAgent:** Mô tả các công ty cung cấp dịch vụ du lịch.
- **Accommodation:** Mô tả những địa điểm là chỗ nghỉ cho khách du lịch (Khách sạn, nhà nghỉ ...v.v.).
- **CommercialResource:** Mô tả các địa điểm thương mại, được chia nhỏ thành chợ, shop ...v.v.
- **DiningService:** Mô tả các địa điểm cung cấp dịch vụ ăn uống. Được chia nhỏ làm 2 lớp con là nhà hàng sang trọng và nhà hàng bình dân.
- **HealthService:** Những địa điểm chăm sóc sức khỏe, phục vụ cho khách du lịch khi có vấn đề về sức khỏe.
- **TouristResource:** Dùng để mô tả các tài nguyên du lịch, đây chính là những địa điểm chính trong kế hoạch du lịch của khách. Được chia nhỏ ra làm nhiều lớp tùy theo loại hình của địa điểm đó (Bãi biển, Hồ, Viện bảo tang ...v.v.).

Các lớp con của Media:



Hình 6 Các lớp con của Media

- **Image:** Hình ảnh về địa điểm, hoạt động nào đó
- **Video:** Video về địa điểm, hoạt động nào đó

Các lớp con của Quality:



Hình 7 Các lớp con của Quality

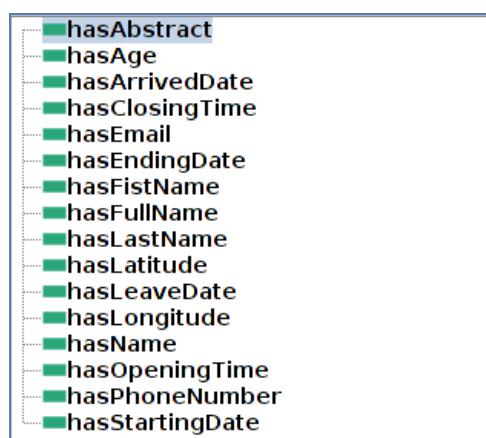
- **Rating:** Mô tả đánh giá chất lượng 1 nhà hàng, khách sạn nào đó (1 sao, 2 sao, 3 sao, 4 sao ...v.v.).
- **Style:** Mô tả phong cách của một địa điểm, hoạt động nào đó.

Các khái niệm được mô tả ở trên được đặt trong mối quan hệ ràng buộc bởi các thuộc tính:

Các thuộc tính nguyên thủy:

- **hasAbstract:** là thuộc tính chung của tất cả các lớp.
- **hasAge, hasEmail, hasFirstName, hasLastName, hasPhoneNumber, hasArrivedDate, hasLeaveDate:** là các thuộc tính của lớp **Profile**
- **hasClosingTime, hasOpeningTime:** là thuộc tính của lớp **Activity**
- **hasStartingDate, hasEndingDate:** là thuộc tính của một **Event**
- **hasLatitude, hasLongitude:** là thuộc tính của một **Point**

Các thuộc tính đối tượng:



Hình 8 Các thuộc tính đối tượng

Với hệ thống ontology lõi với khoảng 90 khái niệm (concept), 40 thuộc tính và gần 1000 triples (bộ ba SVO) hệ thống đã bước đầu xây dựng được một nền tảng dữ liệu ngữ nghĩa mô tả được hầu hết các tài nguyên liên quan đến lĩnh vực du lịch. Ontology sử dụng ngôn ngữ RDF++ có hỗ trợ

cho việc liên kết, tích hợp các nguồn dữ liệu ngữ nghĩa phân tán (owl:sameAs) nên có thể dễ dàng mở rộng và bổ sung sau này.

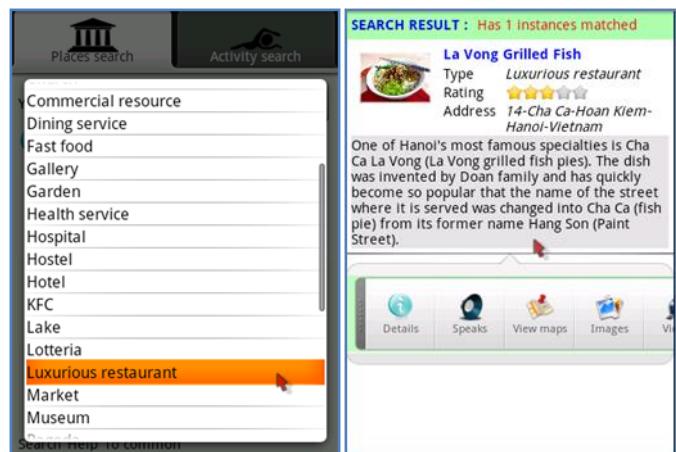
4. PHÁT TRIỂN ÚNG DỤNG HƯỚNG DẪN DU LỊCH ĐA NỀN TẢNG

Hệ thống đã xây dựng các chương trình phục vụ người dùng cuối cá ở smartphone Android và trên nền web. Trên nền Android, hệ thống tận dụng các giao diện kiểu mới của điện thoại như động, cảm ứng... để xây dựng ứng dụng cho phép người dùng có thể tìm kiếm nhanh về địa điểm, hoạt động mong muốn. Đồng thời hiển thị các thông tin ở nhiều hình thức khác nhau như văn bản, hình ảnh, âm thanh và đặc biệt là trên bản đồ sử dụng Google maps, hỗ trợ hướng dẫn bằng giọng nói sử dụng công nghệ text to speech của Android. Bên cạnh đó, trên nền web, hệ thống cho phép người dùng khai báo các sở thích và khuyến nghị một hành trình đi qua các tài nguyên du lịch của thành phố sao cho phù hợp nhất với các sở thích đó, đồng thời kết hợp với việc tối ưu hóa quãng đường đi.

4.1. Úng dụng tìm kiếm ngữ nghĩa thông tin du lịch trên điện thoại Android

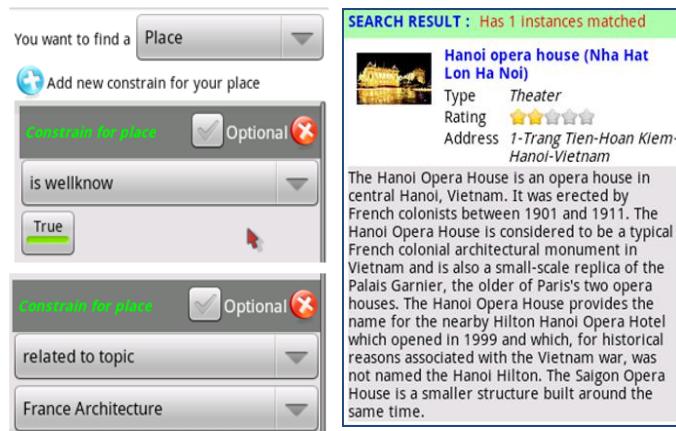
Trên di động android, nhóm nghiên cứu tập trung vào khả năng truy cứu dữ liệu mọi lúc mọi nơi và phát huy tối đa khả năng trình diễn dữ liệu của smart phone (Text, image, video, maps ...v.v.). Ngoài việc hiển thị dữ liệu thông thường, hệ thống còn sử dụng công nghệ **text to speech** (Phát âm văn bản) nhằm phục vụ du khách tiếp cận thông tin một cách dễ dàng chỉ với một vài cái chạm tay. Có thể thấy được sự vượt trội của việc tìm kiếm trên hệ thống của chúng tôi qua lần lượt các ví dụ sau:

Ví dụ 1: “Tìm một nhà hàng sang trọng” - đây là kiểu tìm kiếm đơn giản nhất không có thêm bất cứ ràng buộc nào ngoại trừ khái niệm định nghĩa chính nó (“nhà hàng sang trọng”). Người dùng chỉ đơn giản là lựa chọn kiểu cho địa điểm cần tìm, hệ thống sẽ tự động tìm kiếm tất cả những “nhà hàng sang trọng” trong hệ thống (kể cả những địa điểm thuộc các lớp con của “nhà hàng sang trọng” cũng sẽ được tìm thấy)



Hình 9 Tìm nhà hàng sang trọng

Ví dụ 2: “Tìm một địa điểm nổi tiếng và liên quan đến kiến trúc” – Đây là kiểu tìm kiếm có nhiều ràng buộc, tuy nhiên các ràng buộc này chỉ được áp dụng trên một đối tượng (địa điểm cần tìm).



Hình 10 “Tìm kiếm có ràng buộc trên một đối tượng”

Với ví dụ trên, khi người dùng chọn thêm ràng buộc, hệ thống sẽ tự động tạo truy vấn tìm những thuộc tính có thể của khái niệm được chọn (property) đồng thời tạo giao diện động để có thể tạo ràng buộc một cách thuận tiện nhất. Khi property được chọn, nếu là một thuộc tính đối tượng (có thể nhân giá trị là một đối tượng) thì sẽ tự động lây ra các lớp, các thể hiện có thể là giá trị của thuộc tính đó. Nếu người dùng không chọn một đối tượng cụ thể trong hệ thống làm giá trị mà chọn một khái niệm, hệ thống sẽ lập tức tạo biến mới và sinh ràng buộc phù hợp. chẳng hạn ở ví dụ trên, các ràng buộc thực tế là: *x là place cần tìm, x thì nổi tiếng, x liên quan đến chủ đề y và y là một thể hiện của France Architecture (hoặc là thể hiện của một lớp con)*.

Ví dụ 3: “Tìm một địa điểm nằm tham quan nằm trên đường nào đó của quận Hoàn Kiếm” – Với ví dụ này các ràng buộc không còn chỉ đặt lên riêng địa điểm cần tìm kiếm

mà còn ràng buộc tới các đối tượng liên quan, ở đây là “*con đường nào đó*”.

The screenshot shows the Dorigo search interface. On the left, there's a search bar labeled "You want to find a Place". Below it, a button says "Add new constrain for your place". A constraint panel is open, showing "has location" set to "Street" and "is part of" set to "Ba Dinh district". On the right, the search results are displayed under "SEARCH RESULT : Has 1 instances matched". The result is "One pillar pagoda (Chua Mot Cot)", with details: Type: Pagoda, Rating: ★★★★☆, Address: chua mot cot street-Ba Dinh-Hanoi-Vietnam. A description follows: "The One Pillar Pagoda is a historic Buddhist temple in Hanoi, the capital of Vietnam. It is regarded alongside the Perfume Temple, as one of Vietnam's two most iconic temples." Below the result are buttons for "Details", "Speaks", "View maps", "Images", and "V"

Hình 11 Tìm kiếm có ràng buộc trên nhiều đối tượng

Ở ví dụ trên, hệ thống tự động tạo các biến bổ sung để thỏa mãn ràng buộc: “*x là place cần tìm, x có location là Street y, y là một phần của Ba Dinh district*”

4.2. Giới thiệu điểm du lịch và khuyến nghị hành trình trên nền Web

Trên nền web, hệ thống có khả năng đưa ra hành trình khuyến nghị cho người dùng dựa trên sở thích, mối quan tâm của họ. Ngoài ra còn cho phép người dùng tìm kiếm nhanh thông tin qua giao diện tìm kiếm đơn giản.

Chức năng khuyến nghị hành trình: Đề xuất hành trình khuyến nghị dựa vào lựa chọn của người dùng trên một tập các sở thích được lấy từ các concept và instance trong ontology. Hệ thống sẽ tìm kiếm các địa điểm, sự kiện trong thành phố thỏa mãn các sở thích của người dùng bằng cách so khớp mang tính ngữ nghĩa chứ không phải các so khớp từ vựng truyền thống. Ngoài ra, người dùng được tùy chọn mức độ thích thú của bản thân đối với từng loại sở thích, tỷ lệ giữa việc tối ưu về khoảng cách với tối ưu về sở thích.

Dorigo đã đề xuất thuật toán giải quyết bài toán Người du lịch (Travelling Saleman Problem) về việc tìm đường đi ngắn nhất xuất phát từ một điểm, đi qua một tập điểm và trở lại điểm ban đầu. Ông dựa trên ý tưởng mô phỏng việc tìm đường đi từ tổ đến nơi có mồi của bầy kiến, trong đó các con kiến khi di chuyển thì để lại trên đường đi một chất bay hơi gọi là vết mùi, và chúng thường lựa chọn đường đi theo con đường có độ đậm đặc mùi cao hơn. Thuật toán này có ưu điểm là có khả năng tìm được lời giải tốt trong những trường hợp dữ liệu cực lớn.

Ở đây ta sẽ áp dụng thuật toán của Dorigo đồng thời có bổ sung yếu tố về độ phù hợp của địa điểm với sở thích của người sử dụng.

Ta có tập $I = \{I_j\}$ biểu diễn cho các concept là con trực tiếp của concept “Topic” trong ontology.

Một profile sở thích của người dùng được biểu diễn bởi tập $W = \{W_j\}$ trong đó W_j là trọng số ứng với concept I_j trong ontology. W_j được xác định bằng số sao mà người dùng dành cho sở thích ứng với concept I_j .

Tập $M = \{m_j\}$ biểu diễn sự tương hợp ngữ nghĩa của một điểm du lịch P với tập sở thích W của người dùng. Trong đó:

- $m_j = 1$ nếu P được mô tả ngữ nghĩa là liên quan đến concept I_j hoặc các con của I_j . Trong thuật toán ban đầu này chúng tôi không phân biệt về mức độ tương hợp ngữ nghĩa giữa các concept cha và con với chủ thể. Ví dụ : P liên quan đến chủ đề âm thực dân tộc Việt Nam thì suy ra P liên quan đến chủ đề âm thực, do đó độ tương hợp của P với chủ đề âm thực $m_j = 1$.

- $m_j = 0$ nếu P không liên quan đến concept I_j hoặc các con của I_j .

Độ tương hợp ngữ nghĩa L của điểm P với sở thích W của người dùng được đề xuất như sau:

$$L = \sum m_j W_j . \quad (1)$$

Độ tương hợp ngữ nghĩa được bổ sung vào công thức tính xác suất lựa chọn đỉnh kế tiếp trên hành trình trong thuật toán của Dorigo như sau:

$$P_{u,v} = k \frac{(T_{u,v})^\alpha (\eta_{u,v})^\beta}{\sum_{w \in UV(u)} [(T_{u,w})^\alpha (\eta_{u,w})^\beta]} + (1-k) \frac{L_v}{\sum_{w \in UV(u)} L_w} \quad (2)$$

trong đó:

$T_{u,v}$: nồng độ mùi trên cạnh (u,v)

L_v : độ phù hợp của đỉnh v với sở thích của khách du lịch.

k : tỷ lệ do người dùng chọn giữa tối ưu theo khoảng cách và tối ưu theo sở thích.

$UV(u)$: là tập các đỉnh láng giềng của u chưa được con kiến hiện tại đi qua.

$$\eta_{u,v} = \frac{1}{d_{u,v}} \quad (3)$$

$$\text{Thành phần } \frac{(T_{u,v})^\alpha (\eta_{u,v})^\beta}{\sum_{w \in UV(u)} [(T_{u,w})^\alpha (\eta_{u,w})^\beta]} \text{ được lấy từ công thức}$$

của Dorigo biểu diễn cho xác suất lựa chọn của con kiến liên quan đến khoảng cách và độ đậm đặc của vết mùi, còn thành phần $\frac{L_v}{\sum_{w \in UV(u)} L_w}$ được bổ sung để biểu diễn cho độ phù hợp của

định v với sở thích của khách du lịch.

Định tiếp trong hành trình được lựa chọn ngẫu nhiên theo xác suất (tức là đỉnh có xác suất cao hơn sẽ có khả năng được chọn cao hơn, tuy nhiên không có nghĩa là đỉnh với xác suất thấp không bao giờ được chọn). Điều này được thực hiện bằng qua kỹ thuật Bánh xe xổ số (Lottery Wheel) như sau: sinh ra một số ngẫu nhiên $k \in (0,1]$ rồi chọn i nhỏ nhất sao

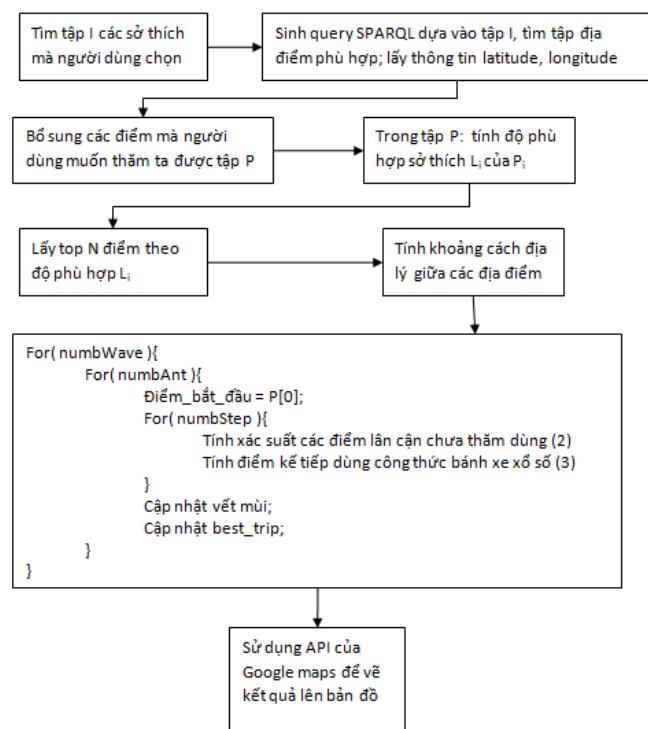
cho $\sum_{j=1}^i P_j \geq k$ (3). Như vậy, các con kiến từ một đỉnh xuất

phát, lần lượt tới thăm các đỉnh tiếp theo quy tắc trên (thăm xong đánh dấu chúng lại) cho đến thăm tới đỉnh cuối cùng và quay về đỉnh ban đầu, kết thúc một hành trình. Quá trình này được lặp đi lặp lại, hành trình tốt hơn (có chiều dài ngắn hơn) sẽ được cập nhật cho đến một khoảng thời gian đủ tốt. Sau mỗi vòng lặp (các con kiến đều tìm được hành trình riêng của mình), vết mùi trên mỗi cạnh được cập nhật lại theo công thức sau :

$$T_{i,j} \leftarrow \rho T_{i,j} + \Delta_{i,j} \quad (4); \quad \Delta_{i,j} \leftarrow \Delta_{i,j} + Q/S_k \quad (5)$$

trong đó: $\rho \in (0,1)$: tham số bay hơi, Q là một hằng số, S_k là độ dài hành trình của con kiến thứ k .

Nhờ việc cập nhật nồng độ mùi mà cùng một con kiến có thể sẽ không chọn một cạnh mà ở bước lặp trước nó đã chọn. Nhờ vậy thuật toán có khả năng tìm được lời giải tốt với tập dữ liệu lớn.



Hình 12 Giải thuật sinh hành trình khuyến nghị

Ví dụ 4: “Giả sử người dùng có sở thích về kiến trúc, ẩm thực, văn hóa nghệ thuật & giải trí. Đưa ra hành trình khuyến nghị cho người dùng trên”.

Planning your visit now

Preferences

Rate between distance or preference ?

Distance	<input type="text" value="0"/>	<input type="text" value="100"/>	<input type="text" value="20"/>	Preference
<input checked="" type="checkbox"/> Architecture ★★★★★				
<input checked="" type="checkbox"/> Cuisine ★★★★★				
<input checked="" type="checkbox"/> Culture-Art ★★★★★				
<input checked="" type="checkbox"/> Entertainment ★★★★★				
<input type="checkbox"/> History ★★★★★				
<input type="checkbox"/> Religion ★★★★★				

Well-known Places

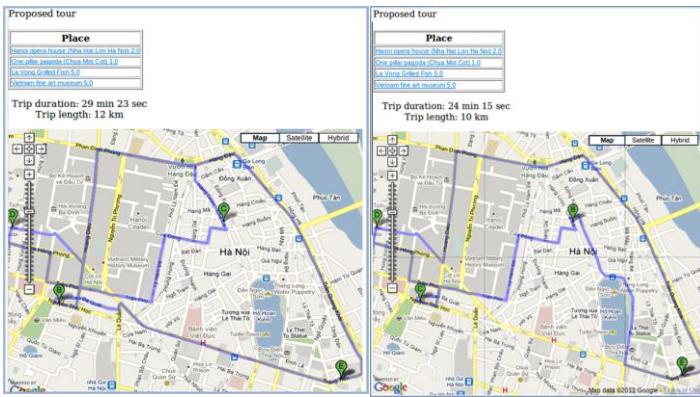
<input type="checkbox"/> Hoan-Kiem-lake	<input type="checkbox"/> Lon-church
<input type="checkbox"/> Trang-Tien-plaza	<input type="checkbox"/> hanoi-opera-house
<input type="checkbox"/> one-pillar-pagoda	<input type="checkbox"/> turtle-tower
<input checked="" type="checkbox"/> lavong-grilled-fish-restaurant	<input type="checkbox"/> vietnam-fine-art-museum

[Create route](#) [Clean form](#)

Hình 13 Giao diện khai báo sở thích và ràng buộc

Giao diện trên cho phép người dùng khai báo sở thích và mức độ quan tâm đối với từng loại sở thích. Chọn tỉ lệ tối ưu giữa khoảng cách và sở thích (Nếu chỉ quan tâm đến khoảng cách thì tỉ lệ là 100, ngược lại nếu chỉ quan tâm đến sở thích

thì để 0). Ngoài ra người dùng có thể đánh dấu các địa điểm nổi tiếng muốn có trong hành trình của mình. Hình dưới đây thể hiện sự khác biệt khi người dùng thay đổi trọng số giữa khoảng cách và sở thích.



Hình 14 Khuyến nghị đường đi khi thay đổi trọng số

(14a. Đường đi tối ưu; 14b. Đường đi tối ưu có tính trọng số sở thích)

Chức năng tìm kiếm đơn giản: Bên cạnh giao diện tìm kiếm ngữ nghĩa đầy đủ được giới thiệu trong mô đun ứng dụng điện thoại – nơi mà người dùng có thể diễn đạt những câu hỏi về những nội dung họ quan tâm qua một giao diện đồ họa động - tự nhiên, hợp về logic sử dụng ontology, chúng tôi muốn cung cấp chức năng tìm kiếm ngữ nghĩa đơn giản cho những người dùng vốn quen tìm kiếm thông qua việc gõ bàn phím. Mục đích của hệ thống này là cố gắng ánh xạ các câu hỏi đơn giản của người sử dụng về 1 câu hỏi ở dạng bộ ba Subject – Property - Object. Và chuyên nó sang truy vấn ngữ nghĩa.

Ontology sử dụng các phát biểu dưới dạng mô hình bộ ba bao gồm: Chủ ngữ - Vị ngữ - Tân ngữ tương ứng với Subject – Property – Object/Literal. Trong đó, Subject là các thể hiện của các định nghĩa khái niệm (Concept), Property có thể nhận các giá trị là các thể hiện của các Concept khác nhau hoặc là các giá trị nguyên thủy như: chuỗi, số hoặc thời gian... Trong việc tìm kiếm ta sử dụng so các khớp bộ ba mang tính ngữ nghĩa trên chứ không phải là các từ vựng truyền thống.

Hình 14 Chọn subject

Hình 15 Chọn property

Hình 16 Chọn object

property	value
has abstract	Turtle Tower, also called Tortoise Tower is a small tower in the middle of Sword Lake, Hanoi, Vietnam.
label	Turtle tower (Tháp rùa)
is wellknow	true

property	object
related to topic	vietnam history
has location	le-thai-to-street
has activity	general sightseeing

Image

Hình 17 Kết quả truy vấn

5. KẾT QUẢ VÀ ĐÁNH GIÁ

Hệ thống được xây dựng với mục đích triển khai trở thành một ứng dụng thiết thực, thay thế và bổ sung những khiếm khuyết trong các hệ thống sử dụng phương pháp lưu trữ truyền thống hiện nay. Trong những kết quả ban đầu mà nhóm nghiên cứu đạt được đã cho thấy tính khả thi cao của hệ thống.

Log			
Time	pid	tag	Message
03-23	D 441	GET ALL CLASS	TIME START : CURRENT TIME = 3:20:38:16
03-23	D 441	GET ALL CLASS	TIME FINISH: CURRENT TIME = 3:20:38:18
03-23	D 441	GET ALL PROPERTY	START CURRENT TIME = 3:20:38:28
03-23	D 441	GET ALL PROPERTY	FINISH CURRENT TIME = 3:20:38:31
03-23	D 441	QUERY	START TIME <CURRENT TIME = 3:20:39:14>
03-23	D 441	QUERY	FINISH TIME <CURRENT TIME = 3:20:39:14>
03-23	D 441	GET INSTANCE DATA	START TIME <CURRENT TIME = 3:20:39:14>
03-23	D 441	GET INSTANCE DATA	FINISH TIME <CURRENT TIME = 3:20:39:14>
03-23	D 441	GET INSTANCE DATA	START TIME <CURRENT TIME = 3:20:39:14>
03-23	D 441	GET INSTANCE DATA	FINISH TIME <CURRENT TIME = 3:20:39:15>
03-23	D 441	GET INSTANCE DATA	START TIME <CURRENT TIME = 3:20:39:15>
03-23	D 441	GET INSTANCE DATA	FINISH TIME <CURRENT TIME = 3:20:39:15>

Hình 12 Time log về thời gian truy vấn của hệ thống

Với việc sử dụng AllegroGraph server 4.2 để quản lý và xử lý dữ liệu ngữ nghĩa, tốc độ truy vấn được cải thiện nhiều so với các mô hình suy diễn thông thường. Với chức năng về đường đi phụ thuộc vào các trọng số về sở thích của khách hàng, quãng đường đi đã làm cho hệ thống trở nên hữu ích hơn đối với người dùng đồng thời thể hiện tính vượt trội của hệ thống so với các hệ thống hiện có. Đối với phần ứng dụng client triển khai trên di động Android thông qua mạng không dây để kết nối web service, tốc độ truy vấn, truyền tải dữ liệu cũng đã được nhóm quan tâm và không ngừng cải tiến để thu được kết quả ngày càng tốt hơn. Theo số liệu ban đầu, trung bình một câu truy vấn, suy diễn từ lúc gửi truy vấn cho tới lúc nhận dữ liệu trả về thông qua mạng không dây là 1/10s – đây là một con số rất khả quan so với tốc độ truy vấn và suy diễn hiện tại của các mô hình suy diễn.

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Những kết quả ban đầu mà nhóm nghiên cứu đã đạt được, một lần nữa khẳng định sức mạnh của công nghệ semantic web trong việc đưa ra các giải pháp thông minh cho các bài toán thực tế. Hệ thống đã tạo nên sự khác biệt khi có thể giải quyết được những vấn đề mà các hệ thống hiện có không thể làm được, cụ thể như khả năng tích hợp dữ liệu từ nhiều nguồn phân tán. Khả năng cung cấp thông tin ngữ nghĩa một cách thông minh, chính xác và “ngữ nghĩa” hơn. Sự kết hợp giữa dịch vụ web, công nghệ web ngữ nghĩa và việc khai thác các kho dữ liệu ngữ nghĩa mở trên Internet hứa hẹn đem lại những kết quả tích cực so với các hệ thống truyền thống.

So với các hệ thống cùng sử dụng giải pháp ngữ nghĩa đã giới thiệu ở phần đầu bài viết, hệ thống đã kế thừa được những điểm mạnh khi sử dụng công nghệ semantic web, bên cạnh đó còn đưa ra được những kết quả mới giúp hệ thống trở nên thông minh hơn như sử dụng linked-data (kết nối dữ liệu), tích hợp được công nghệ ngữ nghĩa với các dịch vụ ưu việt khác hiện nay của google, youtube trên nền tảng di động. Đề xuất và đưa ra được thuật toán lên hành trình du lịch có tham số phù hợp ngữ nghĩa.

Trong thời gian sắp tới, hệ thống sẽ được mở rộng về mặt chức năng, cải thiện năng và làm giàu dữ liệu. Về mặt chức năng, phía client di động nhóm sẽ phát triển khả năng suy

diễn, gợi ý và sàng lọc kết quả dựa vào location của du khách thông qua hệ thống định vị GPS, trên nền web nhóm sẽ phát triển chức năng biên tập dữ liệu cho phép quản lý việc thêm bớt dữ liệu hệ thống một cách hiệu quả, an toàn. Ngoài ra các truy vấn sẽ được xử lý thông minh hơn sao cho có thể đưa ra được các kết quả xấp xỉ thỏa mãn ràng buộc của yêu cầu người dùng. Về việc tăng hiệu năng, nhóm sẽ tập trung tối ưu hóa các câu truy vấn tương tác với ontology để mang lại tốc độ nhanh nhất với kết quả tốt nhất.

7. LỜI TRI ÂN

Để hoàn thành bài báo này, nhóm chúng em xin chân thành cảm ơn sự hướng dẫn tận tình của Ts. Cao Tuấn Dũng – bộ môn công nghệ phần mềm – Viện công nghệ thông tin - trường ĐH Bách Khoa Hà Nội, cảm ơn những đóng góp cũng như giúp đỡ quý báu từ Ths. Trịnh Tuấn Đạt về các vấn đề chuyên môn cũng cũng như kĩ thuật. Cảm ơn ông Gary King – Franz Inc, đã hỗ trợ giải quyết các vấn đề kĩ thuật liên quan đến AllegroGraph server. Xin cảm ơn nhóm các em khóa dưới đang nghiên cứu về semantic web đã hỗ trợ nhóm trong suốt quá trình nghiên cứu, xây dựng hệ thống.

8. TÀI LIỆU THAM KHẢO

- [1] Bergur Päsl Gylfason, “The future of the web – the semantic web”, REYKJAVIK UNIVERSIT, 25-March 2010.
- [2] Sean B. Palmer, Semantic web: introduction <http://infomesh.net/2001/swintro/>
- [3] W3schools, Semantic web tutorial, <http://www.w3schools.com/semweb/default.asp>
- [4] Shih-Chun Chou, Wen-Tai Hsieh, Fabien L. Gandon and Norman M. Sadeh – “Semantic Web Technologies for Context-Aware Museum Tour Guide Application”
- [5] Chris van Aart, Bob Wielinga and Willem Robert van Hage – “Mobile cultural heritage guide: location-aware semantic search”
- [6] M. Dorigo and T. Stützle. Ant Colony Optimization, MIT Press., Cambridge, MA, 2004
- [7] Franz inc, Allegrograph - <http://www.franz.com/agraph/allegrograph/>
- [8] Heiko Haller, “QuiKey – a Demo”, SemSearch 2008, CEUR workshop proceedings, ISSN 1613-0073, June 2008

Tích hợp nội dung web phổ dụng

Phan Văn Hùng, Vũ Mạnh Hùng, Trần Đắc Long

Tóm tắt-- Ngày nay, các website liên quan đến mọi khía cạnh của cuộc sống đã trở thành các kênh thông tin không thể thiếu đối với chúng ta. Tuy nhiên, không phải ai cũng có khả năng tiếp cận được với thông tin đăng trên các website một cách dễ dàng cũng như sử dụng các dịch vụ tích hợp trong các website này một cách hiệu quả. Có thể chỉ ra một số nguyên nhân chính như: các website không được chuẩn hóa về mặt nội dung, chưa hỗ trợ người dùng tương tác phù hợp với khả năng của họ hay chưa tương thích được với các thiết bị đầu cuối khác nhau.

Để giải quyết các vấn đề trên, chúng tôi đề xuất “Mô hình tích hợp nội dung web phổ dụng” (universal web content mashup) cho phép xây dựng nội dung trang web có khả năng đáp ứng tính truy cập cho nhiều đối tượng người dùng khác nhau mà không phụ thuộc vào năng lực hành vi, bối cảnh sử dụng hay phương pháp tiếp cận thông tin của người dùng. Chúng tôi đã áp dụng mô hình này để xây dựng một ứng dụng web trên cơ sở tích hợp các dịch vụ tư vấn chọn trường, định hướng nghề nghiệp, đào tạo nghề và giới thiệu việc làm cho các đối tượng người dùng bình thường và khuyết tật. Công nghệ được sử dụng để xây dựng ứng dụng này là công nghệ Web service, cơ sở hạ tầng điện toán đám mây Google App Engine và hệ quản trị nội dung mã nguồn mở DotNetNuke. Ứng dụng này có thể truy cập dễ dàng từ máy tính và thiết bị di động. Ngoài ra, giao diện di động còn cho phép người dùng dễ dàng định vị trên bản đồ các cơ sở đào tạo nghề và các địa điểm tuyển dụng lân cận với vị trí hiện tại của người dùng.

Từ khóa—Cloud computing, WCAG, SOA, universal usability.

1. GIỚI THIỆU

Tính dễ truy cập và dễ sử dụng là hai yếu tố quan trọng nhất của một website nhằm thu hút người dùng. Sự phát triển của

Công trình này được thực hiện với sự hướng dẫn của tiến sĩ Vũ Thị Hương Giang.

Phan Văn Hùng, sinh viên lớp Công nghệ phần mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0988188836, e-mail: hungpv1988@gmail.com).

Vũ Mạnh Hùng, sinh viên lớp Công nghệ phần mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 01692640187 e-mail: hungvubk06@gmail.com).

Trần Đắc Long, sinh viên lớp Công nghệ phần mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 01664253266 e-mail: longtdbk@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

công nghệ đã cho phép các trang web hiển thị được trên các thiết bị đầu cuối khác nhau như máy tính, điện thoại di động hay PDA. Tuy nhiên, các trang web hiện nay mặc dù có giao diện bắt mắt nhưng chưa thực sự chú trọng đến tính dễ truy cập và dễ sử dụng, dẫn đến việc một số lượng lớn người dùng chán hạn như người lớn tuổi, người khuyết tật... gặp rất nhiều khó khăn khi truy cập Web. Hiện nay đã có nhiều xu hướng công nghệ cho phép giải quyết bài toán nâng cao tính truy cập và sử dụng của các ứng dụng kể cả về mặt cơ sở lý thuyết và ứng dụng.

Xu hướng đầu tiên quan tâm đến việc chuẩn hóa các trang web theo chuẩn truy cập nội dung Web (Web content Accessibility Guidelines - WCAG 2.0 <http://www.w3.org/WAI/>). Chuẩn này bao gồm tập các khuyến cáo về cách áp dụng các kỹ thuật liên quan đến âm thanh, hình ảnh, video, văn bản, bố cục... để hỗ trợ người dùng truy cập các trang web. Ở Việt Nam, đã xuất hiện một số sản phẩm được xây dựng theo xu hướng này, điển hình như các trang Web <http://www.nguoikhuyettat.org/> hay <http://tamhonvietnam.net/>.

Xu hướng thứ hai tập trung vào việc tích hợp nội dung Web theo mô hình kiến trúc hướng dịch vụ (Service Oriented Architecture-SOA <http://www.service-architecture.com>). Kiến trúc này cho phép kết nối các dịch vụ thực hiện các qui trình nghiệp vụ khác nhau một cách độc lập với nền tảng hệ thống. Hướng tiếp cận này phù hợp với việc thiết kế và xây dựng các phần mềm mới trên cơ sở các phần mềm hiện có đúng theo quan hệ giữa khách hàng và nhà cung cấp

Xu hướng tiếp theo – điện toán đám mây (cloud computing) mặc dù mới xuất hiện những năm gần đây nhưng rất được ưa chuộng. Xu hướng này cho phép khách hàng bên ngoài có thể sử dụng theo nhu cầu- các sức mạnh tính toán, lưu trữ, các nền tảng và các dịch vụ ảo hóa. Nhờ đó, khả năng truy cập nội dung Website trở nên phong phú hơn, giảm thiểu tình trạng quá tải hệ thống.

Những xu hướng trên mặc dù đã cải thiện đáng kể tính truy cập và sử dụng trang web, nhưng nếu chỉ đơn thuần đi theo 1 xu hướng, sẽ tồn tại nhiều vấn đề chưa được giải quyết.

Chuẩn WCAG đã tạo ra nhiều kỹ thuật hỗ trợ người dùng truy cập web, đặc biệt với lớp người khuyết tật và lớp người cao tuổi. WCAG với phiên bản 2.0 đã giúp cho các trang web trở nên dễ truy cập và dễ thao tác hơn rất nhiều. Mặc dù vậy, việc áp dụng WCAG để xây dựng các website vẫn chưa đáp ứng được yêu cầu về tính dễ truy cập và dễ sử dụng trên nền mobile, ví dụ nội dung trang Web thường bị vỡ khi hiển thị trên các thiết bị di động khiến người dùng không thể truy cập web bình thường được.

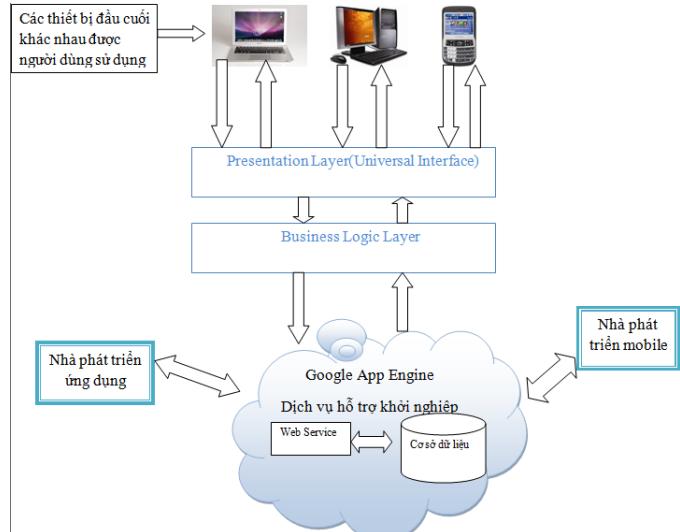
Các ứng dụng theo mô hình SOA có thể tích hợp các luồng thông tin liên quan đến nhau lại thành 1 khối tổng thể. Ứng dụng có thể sử dụng các dịch vụ của bên cung cấp để tạo ra 1 luồng nghiệp vụ gắn kết chặt chẽ, giúp người dùng nắm bắt nội dung thông tin mong muốn dễ dàng và tiện lợi hơn mà không phải tiêu tốn thời gian truy cập các web khác nhau để tập hợp thông tin. Tuy nhiên, các ứng dụng này vẫn chưa thoát được tình cảnh quá tải hệ thống. Nhiều thời điểm, lượng bandwidth hết, nhu cầu sử dụng của người dùng đang cao nhưng họ không thể truy cập được trang web trong 1 khoảng thời gian dài, điều này ảnh hưởng lớn đến quá trình sử dụng web của người dùng.

Mô hình cloud đã giúp bên phát triển giảm bớt chi phí triển khai do không phải chi trả phí hosting và lưu trữ dữ liệu. Quan trọng hơn, tình trạng quá tải do bùng nổ lượng truy cập nhiều đã được xóa bỏ. Truy cập web của người dùng không bị đứt quãng kể cả khi nhu cầu sử dụng trang web quá lớn. Tuy nhiên, mô hình cloud thuận tiện lại không thể tích hợp các nội dung web có liên quan đến nhau. Người dùng phải tự tổng hợp các thông tin gắn kết với nhau từ nhiều nguồn để ra được 1 luồng thông tin thống nhất.

Như vậy, mỗi mô hình chỉ giải quyết được 1 phần của bài toán nâng cao tính dễ truy cập và dễ sử dụng và đều có những nhược điểm nhất định ảnh hưởng đáng kể đến tính truy cập và sử dụng của trang web. Do đó, chúng tôi đề xuất mô hình “**Tích hợp nội dung web phổ dụng**”, kết hợp các ưu điểm của tất cả các cơ sở lý thuyết ở trên để giải quyết triệt để bài toán nâng cao tính truy cập và sử dụng của Website. Để chứng minh những tiện ích mà mô hình đem lại, chúng tôi sẽ áp dụng mô hình để giải quyết bài toán về tính truy cập và sử dụng cho ứng dụng hỗ trợ người dùng khởi nghiệp.

Nội dung tiếp theo được tổ chức như sau: Phần 2 mô tả kiến trúc mô hình đề xuất, theo sau là phần 3: trình bày cách áp dụng mô hình để xây dựng ứng dụng hỗ trợ người dùng khởi nghiệp, tiếp theo đến phần 4 tổng hợp lại các kết quả đạt được, và cuối cùng là phần 5 hướng phát triển trong tương lai.

2. MÔ HÌNH TÍCH HỢP WEB PHỔ DỤNG



Hình 1: Tích hợp nội dung Web phổ dụng

Hình vẽ trên miêu tả kiến trúc của mô hình.

Mô hình tích hợp nội dung Web phổ dụng kết hợp ưu điểm của các xu hướng công nghệ về SOA, WCAG 2.0, cloud computing và các kỹ thuật đảm bảo tương thích web với di động để thiết kế những ứng dụng có khả năng truy cập cao, tính sử dụng tốt, không những cho phép kết nối những luồng thông tin nghiệp vụ khác nhau lại thành một khối mà còn đáp ứng được nhu cầu truy cập cao của người dùng.

Mô hình được chia thành 3 tầng chính:

Tầng presentation: Người dùng sẽ tương tác với ứng dụng thông qua tầng này. Cách bố trí và hiển thị nội dung sẽ tuân theo các khuyến cáo WCAG 2.0 nhằm đảm bảo tính truy cập ứng dụng Web. Để giao diện hiển thị rõ ràng trên các thiết bị di động, lập trình viên sẽ áp dụng thêm các kỹ thuật tương thích Web, tác động lên thành phần CSS và JavaScript của ứng dụng, nhờ vậy, nội dung web sẽ hiển thị theo màn hình và thiết bị phần cứng của di động hợp lý.

Tầng business logic: Các kỹ thuật chịu lỗi của chuẩn WCAG 2.0 sẽ được áp dụng trong tầng này. Thông tin thu thập được từ người dùng sẽ được tập hợp lại tại tầng này. Sau đó, business logic sẽ gọi thực thi các dịch vụ được triển khai trên nền cloud computing để thực hiện các chức năng nghiệp vụ mà người dùng yêu cầu.

Tầng Service: Được triển khai trên nền cloud computing. Service sẽ được viết theo ngôn ngữ Python và triển khai dưới nền tảng PaaS Google App Engine do Google cung cấp. Do service được triển khai trên nền cloud, ứng dụng không còn bận tâm về vấn đề bandwidth khi truy cập mà vẫn có thể tích hợp nội dung liên quan lại với nhau thành 1 khối thống nhất. Ngoài ra, các ứng dụng mobile hay web từ các hãng phát triển khác có thể sử dụng

các dịch vụ của ứng dụng. Điều này tiết kiệm chi phí xây dựng và tăng độ gắn kết của dịch vụ với các ứng dụng khác.

Nhận xét:

Ứng dụng có giao diện nhất quán giữa các nền, do đó, người dùng có thể tương tác với ứng dụng từ các thiết bị đầu cuối khác nhau(desktop, laptop, mobile). Ngoài ra, mỗi đối tượng người dùng sẽ được hỗ trợ các kỹ thuật trợ giúp phù hợp với bối cảnh sử dụng, hiện trạng sức khỏe để tương tác với Web, đảm bảo tính dễ truy cập của ứng dụng.

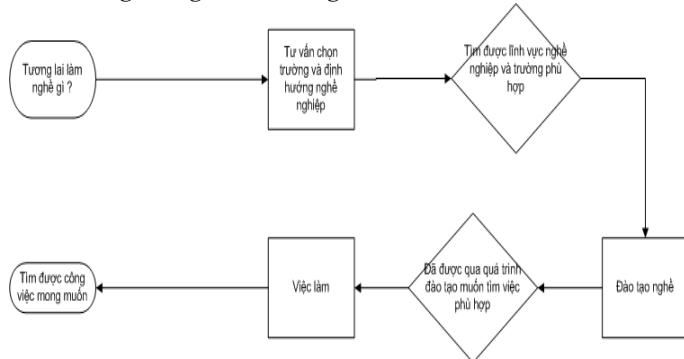
Các tầng được tách biệt rõ ràng, thuận lợi cho quá trình xây dựng và triển khai ứng dụng. Hơn nữa, ứng dụng có thể tích hợp những nội dung liên quan lại thành 1 luồng đi thống nhất qua các dịch vụ, giúp người dùng tiếp nhận thông tin hiệu quả hơn. Áp dụng mô hình, người dùng sẽ không bị gián đoạn quá trình sử dụng Web do vấn đề bandwidth gây ra.

3. ÁP DỤNG MÔ HÌNH XÂY DỰNG ỦNG DỤNG HỖ TRỢ NGƯỜI DÙNG KHỎI NGHIỆP.

Nhằm chứng minh những lợi ích to lớn mà mô hình mang lại, chúng tôi sẽ áp dụng mô hình xây dựng website trợ giúp người dùng khởi nghiệp trên cơ sở tích hợp các dịch vụ tư vấn chọn trường, định hướng nghề nghiệp, đào tạo nghề và giới thiệu việc làm. Hình 2 mô tả logic nghiệp vụ của ứng dụng: Chức năng tư vấn chọn trường và định hướng sẽ giúp người dùng tìm được lĩnh vực nghề nghiệp phù hợp và chọn được trường (đại học, cao đẳng hay học viện) để theo học ngành nghề yêu thích. Sau đó, người dùng có thể học kiến thức về các ngành nghề tại Web thông qua chức năng đào tạo nghề. Sau khi đã trải qua quá trình đào tạo và đã nắm vững kiến thức về ngành nghề, người dùng có thể bắt đầu tìm việc qua chức năng tìm kiếm việc làm.

Ứng dụng cung cấp 3 dịch vụ chính

3.1. Luồng thông tin hệ thống



Hình 2: Luồng thông tin hệ thống

3.2. Định Hướng Nghề Nghiệp Và Tư Vấn Chọn Trường

Phần này được xây dựng nhằm hỗ trợ người dùng quyết định chọn cho mình được lĩnh vực nghề nghiệp phù hợp và chọn được trường (đại học, cao đẳng hay học viện) để theo học ngành nghề yêu thích.

3.3. Hỗ Trợ Đào Tạo Nghề

Phần này được xây dựng với mục đích hỗ trợ người dùng học nghề. Cơ sở đào tạo có thể đưa nội dung các giáo trình, câu hỏi và bài kiểm tra lên Web và sau đó, người dùng có thể học kiến thức về các nghề trong cuộc sống qua các giáo trình giảng dạy, bài kiểm tra đã được đưa lên.

3.4. Giới Thiệu Việc Làm

Phần này có các service đã được triển khai trực tiếp trên nền cloud computing. Người dùng có thể tìm kiếm được những thông tin về việc làm phù hợp với họ. Đồng thời dịch vụ cũng cung cấp một cách trực quan nhất có thể các vị trí của các nhà tuyển dụng xung quanh vị trí hiện tại của người dùng thông qua GPS để giúp người dùng lựa chọn được các công việc tốt hơn theo điều kiện địa lý của họ.

3.5. Công nghệ sử dụng:

Chúng tôi có áp dụng control được phát triển bởi hãng phát triển Telerik và công nghệ Ajax cùng với các kỹ thuật thêm động control trong ASP.NET để xây dựng các chức năng ứng dụng, giúp người dùng thao tác với dữ liệu và sử dụng chức năng dễ dàng, hiệu quả hơn. Ngoài ra, sự tiện dụng của các công nghệ này đã giúp chúng tôi đơn giản hóa được nhiều xử lý nghiệp vụ phức tạp.

4. KẾT QUẢ ĐẠT ĐƯỢC

Áp dụng mô hình, chúng tôi đã xây dựng thành công ứng dụng có các luồng thông tin và chức năng nghiệp vụ gắn kết chặt chẽ đến nhau, giúp người dùng liên kết các thông tin lại và tìm ra cho mình 1 hướng khởi nghiệp phù hợp với bản thân và mong muốn. Ứng dụng được triển khai tại địa chỉ <http://cungkhoinghiep.net/>.

Mô hình đã giúp chúng tôi tạo ra một giao diện dễ sử dụng và có tính truy cập cao (chúng tôi đã áp dụng 153 kỹ thuật trong tổng số 342 kỹ thuật trong chuẩn WCAG2.0 và các kỹ thuật tương thích di động để xây dựng giao diện). Điều này cho phép người dùng tương tác với web thuận tiện và dễ dàng hơn, không bị ảnh hưởng bởi các yếu tố như bối cảnh sử dụng, thiết bị đầu cuối, hiện trạng sức khỏe. Cách thức tổ chức hợp lý các tầng của mô hình giúp chúng tôi giảm thiểu thời gian xây dựng do các module được tái sử dụng nhiều và công việc các phần được tách biệt, không chồng chéo. Chi triển khai ứng dụng được giảm thiểu đáng kể. Đã không còn hiện tượng lãng phí bandwidth do lượng truy cập ít hay quá tải do lượng truy cập nhiều (khiến người dùng không tiếp cận thông tin trong thời gian dài, dẫn đến việc bỏ website). Nội dung web hiển thị tốt trên các thiết bị di động đã cho phép chúng tôi đã tận dụng cơ chế định vị GPS, hiển thị vị trí các nhà tuyển dụng, cơ sở đào tạo xung quanh người dùng giúp họ chọn lựa được nhà tuyển mộ hay cơ sở đào tạo dễ dàng hơn. Sau đây là một số kết quả so sánh sản phẩm của chúng tôi với các sản phẩm hiện có

4.1. So sánh về mặt nội dung và luồng thông tin, nghiệp vụ cho người dùng

Tên trang	Định hướng người dùng	Đào tạo nghề	Việc Làm
http://vietnamwork.com/	không	không	Có
http://cungkhoinghiep.net	Có	Có	Có
http://timviecnhanh.com	Không	Không	Có
http://www.cione.com.vn	Không	Có	Không

Bảng 1: So sánh nghiệp vụ

Qua bảng trên ta có thể thấy tính vượt trội về luồng nghiệp vụ của ứng dụng so với các trang web ở trên. Mô hình nội dung web phổ cập đã cho phép chúng tôi tích hợp các luồng thông tin lại thành 1 luồng thống nhất. Hiện nay, ở Việt Nam chưa có 1 website nào có thể hỗ trợ người dùng tìm việc làm trong từ những bước đầu khởi nghiệp đến khi tìm được việc làm ứng ý như ứng dụng.

4.2. So sánh tính truy cập của ứng dụng với các trang web hỗ trợ người khuyết tật khác:

Chúng tôi sẽ so sánh dựa trên Website <http://achecker.ca/checker/index.php> được xếp hạng cao nhất về khả năng kiểm tra tính truy cập web của W3C. Bảng dưới đây sẽ liệt kê số lỗi trung bình được mà mỗi ứng dụng mắc phải.

Tên trang	Known Problems	Potential Problems	Likely Problem
http://www.nghilucsong.net	112	501	35
http://cungkhoinghiep.net	30	285	0
http://tamhonvietnam.net	45	363	0

Bảng 2: So sánh tính truy cập

Loại lỗi nghiêm trọng ảnh hưởng lớn đến tính truy cập của ứng dụng là 30. Loại lỗi có thể xảy ra của ứng dụng là 0 và với loại lỗi phát sinh, ứng dụng có 490 lỗi. Các con số này ít hơn rất nhiều so với các loại lỗi mà 2 ứng dụng còn lại mắc phải, điều này chứng minh cho tính dễ sử dụng của ứng dụng.

4.3. So sánh kỹ thuật trợ giúp người khuyết tật :

Tên trang	Người khiếm thị	Người khiếm thính	Người khó khăn vận động
http://www.nghilucsong.net	không	không	không
http://cungkhoinghiep.net	JAW (trình đọc màn hình)	phụ đề cho video	bàn phím áo, phím tắt điều hướng

http://tamhonvietnam.net	JAW	không	phím tắt điều hướng.
--------------------------	-----	-------	----------------------

Bảng 3: So sánh kỹ thuật trợ giúp

Từ bảng so sánh, ta thấy trang web <http://cungkhoinghiep.net> đã cung cấp được một số công cụ quan trọng nhất hỗ trợ các đối tượng người khuyết tật.

4.4. So sánh khả năng đáp ứng truy cập người dùng

Tên trang	Triển khai trên nền cloud
http://vietnamwork.com/	Không
http://cungkhoinghiep.net	Có
http://timviecnhanh.com	Không
http://www.cione.com.vn	Không

Bảng 4: So sánh khả năng đáp ứng truy cập

Do service được triển khai trên nền cloud, do đó không bị hạn chế về mặt bandwidth. Nền tảng cloud có thể đáp ứng được nhu cầu sử dụng cao của người dùng trong mọi thời điểm. Điều này giúp người dùng không bị gián đoạn quá trình sử dụng do hiện tượng hết bandwidth gây ra.

4.5. So sánh khả năng truy cập trang web bằng điện thoại

Tên trang	Hiển thị theo độ phân giải
http://www.nghilucsong.net	không
/	
http://cungkhoinghiep.net	Có
http://tamhonvietnam.net	không

Bảng 5: So sánh khả năng truy cập Web bằng điện thoại

Nội dung Web sẽ được hiển thị rõ ràng và chuẩn theo chế độ phân giải của điện thoại

5. KẾT LUẬN

Bài báo đã đề xuất về mô hình universal web content mashup cho việc xây dựng trang web có thể đáp ứng được yêu cầu về tính dễ truy cập, tính dễ sử dụng của người dùng trên những thiết bị đầu cuối khác nhau. Áp dụng mô hình, chúng tôi đã xây dựng thành công ứng dụng hỗ trợ người dùng khởi nghiệp đáp ứng được yêu cầu về tính dễ sử dụng và dễ truy cập trên các thiết bị đầu cuối khác nhau. Ứng dụng được triển khai thử nghiệm tại <http://cungkhoinghiep.net/>, các dịch vụ được triển khai trên nền tảng PAAS google app engine. Trong thời gian sắp tới, chúng tôi sẽ tiếp tục triển khai toàn bộ service các phần định hướng và học nghề lên nền cloud computing. Ngoài ra, nhóm sẽ nghiên cứu thêm về chuẩn WBMP dành riêng cho di động để nâng cao hơn nữa tính truy cập trên nền di động. Tính chính xác về các câu hỏi định hướng sẽ được nâng cao qua việc tìm hiểu và áp dụng các số liệu thống kê của các tổ chức xã hội về nghề nghiệp với các cá nhân trong xã hội

6. LỜI TRI ÂN

Emxin gửi lời cảm ơn sâu sắc đến tiến sĩ Vũ Thị Hương Giang, giảng viên Viện Công Nghệ Thông Tin và Truyền Thông, Đại học Bách Khoa Hà Nội đã tận tình hướng dẫn em thực hiện công trình này.

7. TÀI LIỆU THAM KHẢO

- [1] Ian Sommerville. Software Engineering. 2006. Addison Wesley (8th edition)
- [2] Munindar P. Singh and Michael N. Huhns. Service-Oriented Computing: Semantics, Processes, Agents. 2005. Prentice-Hall
- [3] Patterns: Service-Oriented Architecture and Web Services. 2004. IBM Redbook
- [4] Peter Van Roy. Programming Paradigms for Dummies: What Every Programmer Should Know, New Computational Paradigms for Computer Music. 2009. IRCAM/Delatour France
- [5] Wrox Professional DotNetNuke Module Programming
- [6] Building WebSite with DotNetNuke 5 – Micheal Washington , Ian Lackey
- [7] <http://universalusability.com/>
- [8] Access by Design: A Guide to Universal Usability for Web Designers by Sarah Horton
- [9] O'reilly - Programming Google Apps Engine (Nov 2009)
- [10] Building.Android.Apps.with.HTML.CSS.and.JavaScript
- [11] Sullivan, Terry and Matson, Rebecca (2000): Barriers to Use: Usability and Content Accessibility on the Web's Most Popular Sites

Các vấn đề an toàn bảo mật cho điện thoại di động, phần mềm bảo mật trên nền Android

Trần Ngọc Hải

Tóm tắt: Đề tài nghiên cứu này tập trung đi sâu vào nghiên cứu những vấn đề an toàn, bảo mật cho điện thoại di động như:

- Vấn đề tin nhắn rác.
- Vấn đề chặn cuộc gọi quấy rối trên điện thoại di động.
- Vấn đề lưu trữ thông tin cá nhân, bảo mật, sao lưu thông tin cá nhân nhạy cảm.
- Vấn đề virus trên điện thoại di động.
- Giải pháp bảo mật cho các vấn đề trên: phần mềm bảo mật trên điện thoại di động (trong đề tài này là phần mềm bảo mật được viết cho nền tảng Android)

Từ khóa: Android, an toàn, bảo mật, mobile security

1. Giới thiệu

Điện thoại thông minh từ lâu đã trở thành mục tiêu của giới tội phạm mạng với mục đích chủ yếu là ăn cắp các thông tin quan trọng của người sử dụng. Ngày nay smartphone có thể truy cập internet nhanh và tiện lợi như trải nghiệm trên máy tính PC thông thường với sự hỗ trợ của mạng viễn thông thế hệ mới 3G hay LTE. Vì vậy việc lây lan virus trên nền tảng di động trở nên dễ dàng hơn vì kênh Internet phổ biến hơn nhiều so với Bluetooth, qua MMS, phần mềm chat trên di động... Với việc smartphone đang dần đạt tới tốc độ tính toán của PC và kiểu dáng, kích thước nhỏ gọn hơn hẳn người dùng điện thoại di động thông minh dần có xu hướng xứng smartphone nhiều hơn PC. Việc lưu trữ thông tin cá nhân quan trọng trên điện thoại di động như tài khoản ngân hàng, tin nhắn riêng tư, danh bạ cũng thường được lưu trữ trên smartphone hơn là trên PC. Điều này cho thấy cần phải có các giải pháp và phần mềm bảo mật tin cậy cho nền tảng smartphone.

2. Tin nhắn rác và cuộc gọi quấy rối

Nhắn tin và gọi điện là hai trong số những phương thức liên lạc, kết nối phổ biến nhất của thế giới hiện đại. Cùng với chiếc điện thoại di động hai dịch vụ này giúp con người tiết kiệm thời gian, công sức, tiền bạc, của cải trong việc liên lạc, giao tiếp, bày tỏ... giúp khoảng cách không còn là rào cản lớn của sự trao đổi thông tin. Tiện lợi, khả năng đáp ứng mọi lúc mọi nơi, giá thành thấp là những ưu điểm không thể bàn cãi của hai phương thức trên tuy nhiên đi kèm với đó cũng là những vấn nạn lớn mà người dùng

điện thoại phải đối mặt đó là tin nhắn rác và những cuộc gọi quấy rối, làm phiền.

2.1 Tin nhắn rác

Tin nhắn rác là tin nhắn gửi qua các mạng điện thoại di động tới người dùng điện thoại mà đem lại phiền toái, khó chịu, không mang giá trị thông tin với người nhận hoặc gây hoang mang, tác động tiêu cực đến người nhận. Hiện nay cùng với thư điện tử rác tin nhắn rác cũng trở thành một vấn nạn của truyền thông hiện đại và ngày càng có xu hướng nghiêm trọng hơn. Khi thị trường thông tin di động mới hình thành tin nhắn văn bản chỉ là một công cụ giúp con người liên lạc thông thường. Tuy nhiên bằng sự tiện lợi, giá thành rẻ và cùng với công nghệ kinh doanh, sự phức tạp của đời sống xã hội tin nhắn rác càng xuất hiện nhiều hơn.

Ngày nay gần như không một người sử dụng điện thoại nào là không nhận phải tin nhắn rác. Từ tin nhắn quảng cáo, tiếp thị, cho tới tin đánh vào tâm lí người dùng như bối rối, lừa đảo của những số điện thoại không quen biết cho tới tin nhắn từ những người quen biết nhưng với mục đích quấy rối, làm phiền, phá hoại... Tác hại của tin nhắn rác không chỉ dừng ở mức khiến người nhận cảm thấy khó chịu, stress mà đôi khi còn dẫn tới sự mất tiền oan. Tin nhắn rác có thể được chia làm hai loại:

Tin rác vì tiền: là tin nhắn được các cơ quan, tổ chức hay cá nhân sử dụng trong mục đích kinh doanh, kiếm lợi nhuận từ tin nhắn, được gửi hàng loạt với số lượng lớn tới người dùng với nội dung tương tự nhau. Thông thường đây là tin nhắn quảng cáo dịch vụ sản phẩm, tin nhắn tiếp thị gói dịch vụ hoặc những tin nhắn có nội dung hấp dẫn khuyến khích người sử dụng điện thoại tham gia hoặc trả lời tin gửi đến, và chi phí cho tin phản hồi này lại rất tốn kém. Sự phổ biến của điện thoại di động, nhu cầu két nối cao của con người và sự tiện dụng của phương tiện tin nhắn là nguyên nhân chính dẫn tới sự bùng nổ của các dịch vụ qua tin nhắn. Qua đó xuất hiện ngày càng nhiều dự án kinh doanh đem lại lợi nhuận dựa trên khai thác dịch vụ nội dung tin và tất nhiên kèm theo đó là những dịch vụ lừa đảo, làm phiền, hoặc vô cùng tốn kém với người dùng.

Tin rác vì nhận thức: là tin nhắn rác được gửi đến gây ra cảm xúc cực đoan cho người nhận. Thông thường tin nhắn loại này của một nhân vật xác định nào đó gửi tới người dùng với nội dung quấy rối, gây phiền nhiễu, khủng bố tinh thần người nhận...

2.2 Cuộc gọi quấy rối

Cuộc gọi quấy rối là cuộc gọi mà người nhận cuộc gọi không mong muốn thực hiện kết nối. Thông thường cuộc gọi quấy rối là hành vi liên lạc bằng đường viễn thông tới một người nào đó nhằm mục đích trêu trọc, lừa đảo, trả thù, quảng cáo... khiến người dùng dịch vụ (gồm cả người sử dụng máy điện thoại di động lẫn để bàn) bức xúc, phiền toái hoặc bị stress... Tuy nhiên khi người dùng đang thực hiện những công việc riêng hoặc đơn giản là không có tâm trạng liên lạc với bất cứ ai thì cuộc gọi đến từ người thân, bạn bè, khách hàng... cũng được coi là cuộc gọi quấy rối.

Đối tượng của quấy rối điện thoại bao gồm từ người dùng cá nhân cho tới các cơ quan, tổ chức... Hiện nay đối tượng của quấy rối điện thoại nhiều nhất là những người nổi tiếng, những người có nhiều mối quan hệ trong xã hội, sinh viên và các cơ quan nhà nước.

2.3 Các phương pháp ngăn chặn vấn nạn tin nhắn rác và cuộc gọi quấy rối

Sự bùng nổ của nạn tin nhắn rác và việc quấy rối bằng điện thoại di động gia tăng có nguyên nhân chủ yếu là việc phát triển lớn mạnh của ngành thông tin di động, internet, viễn thông. Nhu cầu liên lạc tăng cao, xuất hiện thêm nhiều công cụ liên lạc mới công với đời sống xã hội phức tạp của con người gọi điện và nhắn tin cũng trở thành công cụ thực hiện nhu cầu kinh doanh, nhu cầu trao đổi thông tin chính. Để ngăn chặn vấn nạn tin nhắn và cuộc gọi quấy rối này cần những biện pháp đồng bộ từ:

Phía pháp luật: Các nhà làm luật cần đưa ra những điều luật chống lại việc làm phiền qua di động.

Các nhà cung cấp dịch vụ nội dung: Thực chất ngày nay tin nhắn rác lại chủ yếu được gửi bởi các nhà cung cấp dịch vụ nội dung với mục đích thông báo dịch vụ cho khách hàng, quảng cáo... Tuy nhiên những thông báo đó vô hình chung gây khó cho người dùng điện thoại di động, thậm chí ngay cả các nhà mạng cũng nhảy vào khung bô hòm tin nhắn của người dùng với những thông tin quảng cáo vào những giờ, những thời điểm nhạy cảm khiến người dùng khó chịu. Vì vậy để ngăn chặn tin nhắn rác kiểu này cũng rất cần sự thận trọng, khéo léo của những nhà cung cấp dịch vụ nội dung. Ngoài ra còn một hình thức nữa là các nhà cung cấp dịch vụ nội dung tung tin nhắn lừa đảo để yêu cầu người dùng sử dụng dịch vụ tin nhắn tới các đầu số thu phí cao nhằm thu lợi bất chính. Ngăn chặn hình thức nhắn tin rác này thì cần phải có sự ra tay của luật pháp.

Các nhà mạng: Các nhà mạng hơn ai hết là những người được lợi từ tin nhắn rác. Họ thu phí với mọi tin nhắn, cuộc gọi trong khi chi phí truyền tải tin nhắn, cuộc gọi là rất thấp. Vì vậy cũng cần các nhà mạng chấp nhận bớt đi một phần lợi nhuận ngăn chặn các đầu số phát tán tin rác, cuộc gọi làm phiền.

Người sử dụng điện thoại di động: việc nhận thức một tin nhắn là rác hay không phải là tin nhắn rác thì tùy mỗi người, mỗi thể loại tin nhắn... Một cuộc gọi với người

này có thể hữu ích nhưng với cá nhân khác thì đem lại sự phiền phức khó chịu. Từ phía người sử dụng biện pháp hiệu quả nhất để tránh vấn nạn này là tránh chia sẻ thông tin cá nhân một cách tùy tiện, nhận thức được tầm quan trọng của thông tin cá nhân. Ngoài ra khi bị làm phiền báo cáo số máy điện thoại với cơ quan chức năng khi bị làm phiền. Người dùng cũng nên lựa chọn các gói dịch vụ, phần mềm, công cụ chặn cuộc gọi và tin nhắn hiệu quả.

3. Bảo mật thông tin cá nhân, dữ liệu trong điện thoại di động

Thông tin cá nhân ngày nay là tài sản của xã hội hiện đại. Nhiều công ty tổ chức có được số điện thoại, email của khách hàng từ sự bắt cần trong thói quen trao đổi thông tin của con người. Người dùng điện thoại, máy tính thường vô tình để lộ thông tin của mình trên mạng xã hội, trên diễn đàn vô hình chung lại là điều kiện cho nhiều tổ chức kinh doanh, cá nhân khai thác và làm phiền đôi khi phương hại tới danh dự cá nhân. Thông tin cá nhân cũng có thể bị lộ ra bởi những tổ chức lưu trữ thông tin như các nhà mạng, các công ty có thông tin khách hàng, các tổ chức có thông tin thành viên. Thông tin cá nhân trở thành hàng hóa khi mà các tổ chức này bán chúng đi cho những người quan tâm.

Về phía người dùng điện thoại di động thông tin cá nhân của họ không chỉ là tên tuổi, email, số điện thoại mà còn là thông tin cá nhân của bạn bè, đồng nghiệp, đối tác... Việc khai thác, theo dõi một chiếc điện thoại di động của người dùng có thể giúp cho kẻ xấu biết được quan hệ xã hội, bí mật riêng tư, bí mật công việc của người dùng di động và thông tin cá nhân của những người có liên quan tới chủ máy điện thoại. Ngoài ra khi mà chiếc điện thoại ngày nay càng trở nên mạnh mẽ, đa năng hơn, điện thoại có thể lưu trữ hình ảnh, giải trí với game và nghe nhạc, cũng có thể thực hiện việc kết nối internet, thanh toán điện tử... thì con người ngày càng lưu trữ các thông tin quan trọng của mình trong điện thoại. Những thông tin đó có thể là thông tin về các giao dịch ngân hàng nếu điện thoại có chức năng thanh toán điện tử, có thể là mật khẩu của tài khoản ngân hàng, mật khẩu của website, một công thông tin nào đó được người dùng lưu trữ trên điện thoại. Khi kẻ xấu thu thập được những thông tin đó thì hậu quả là vô cùng khôn lường. Vì vậy điều cần thiết là người dùng điện thoại cũng phải dùng những biện pháp để bảo vệ thông tin quan trọng của mình. Những thông tin đó bao gồm:

- Danh bạ điện thoại
- Tin nhắn, email
- Lịch sử cuộc gọi
- Tài khoản, mật khẩu của các truy cập (trong database của phần mềm điện thoại)

Các biện pháp để bảo vệ thông tin cá nhân trên điện thoại của người dùng:

- Cải thiện nhận thức về tầm quan trọng của thông tin cá nhân cho người dùng điện thoại.
- Lưu trữ, mã hóa, đặt mật khẩu cho thông tin quan trọng.
- Bảo vệ các truy cập không được phép tới điện thoại bằng mật khẩu.
- Sử dụng phần mềm diệt virus để tránh việc điện thoại bị mất thông tin cá nhân, bị theo dõi bởi phần mềm xấu.

4. Virus, phần mềm độc hại trên điện thoại di động

Ngày nay khái niệm về virus không còn là mới mẻ với nhiều người nữa. Hầu như người sử dụng máy tính nào cũng biết và hình dung những khó chịu và tác hại của virus máy tính gây ra cho người dùng. Đa số người sử dụng máy tính đều hiểu rằng virus máy tính có thể làm chậm chạp việc hoạt động của máy tính, gây phiền nhiễu cho người dùng hoặc ăn cắp thông tin trên máy tính... Mặc dù vậy nhiều người vẫn còn thờ ơ, cho rằng điện thoại di động không có nhiều mối lo âu, bận tâm về vấn đề virus trên thiết bị liên lạc này. Trên thực tế hiện tại phần mềm độc hại trên điện thoại di động không phổ biến như trên máy tính bởi vì sự phân mảnh của các nền tảng hệ điều hành cho di động là đa dạng, các hệ điều hành di động ít có tính mở, ít lỗi bảo mật hơn (hoặc ít được tìm ra hơn) so với các hệ điều hành máy tính thông dụng hiện tại, việc phát triển ứng dụng cho điện thoại di động chưa lớn bằng trên máy tính hoặc bị kiểm duyệt nội dung trước khi đăng tải lên kho ứng dụng... Tuy nhiên sự lớn mạnh của điện thoại thông minh gần đây đã dẫn tới việc xuất hiện ngày càng nhiều virus trên điện thoại di động.

Năm 2004, virus đầu tiên trên điện thoại di động được ghi nhận đó là Cabir một virus có khả năng lây lan qua Bluetooth bằng việc tự động gửi chính nó đến các máy điện thoại mở Bluetooth trong khoảng cách 10m gần nó. Tác hại của Cabir không lớn, nó chỉ khiến máy điện thoại của người dùng chậm chạp hơn, khiến người dùng khác khó chịu khi liên tục gửi yêu cầu kết nối Bluetooth. Nhưng Cabir lại là minh chứng cho thấy rằng trên nền tảng điện thoại di động hoàn toàn có khả năng bị virus tấn công. Theo số liệu từ công ty Kaspersky thì cho tới năm 2010 phát hiện tới 153 loại virus tấn công các thiết bị di động với hơn 1000 biến thể, tăng 65% so với năm 2009 cho thấy sự phát triển của virus trên điện thoại là ngày càng nhiều hơn và tốc độ tăng nhanh hơn. Hệ điều hành Android của Google được công bố năm 2008, virus đầu tiên trên hệ điều hành này được phát hiện tháng 8 năm 2010. Cho tới hiện tại tháng 5/2011 trung tâm an ninh mạng Bkis (đại học BKHN) phát hiện tới hơn 30 mẫu virus trên nền tảng hệ điều hành này cho thấy Android cũng là mảnh đất màu mỡ của virus trên điện thoại di động.

4.1 Tác hại của virus trên điện thoại di động

Trong khi virus đầu tiên trên di động chỉ có tác dụng khống định sự tồn tại của nó trên điện thoại thì virus đầu tiên trên điện thoại chạy Android lại khiến người dùng mất tiền oan khi tự động nhắn tin đến một số điện thoại của một công ty khai thác dịch vụ nội dung di động với chi phí cho tin nhắn là không hề nhỏ. Những tác hại mà virus trên điện thoại thường là:

- Ăn cắp thông tin cá nhân, thông tin về máy điện thoại IMEI, số điện thoại...
- Tự động gửi tin nhắn, tạo cuộc gọi khiến người sử dụng tốn kém chi phí viễn thông mà không hề hay biết.
- Trở thành thành viên của mạng Botnet tấn công DDos website.
- Chiếm quyền điều khiển điện thoại, khiến điện thoại của người dùng là mục tiêu theo dõi của kẻ xấu.

4.2 Phân loại virus trên điện thoại di động

Cũng giống như virus trên máy tính virus trên điện thoại di động ngày nay rất đa dạng với đủ loại hành vi có hại. Tuy nhiên do đặc thù của hệ điều hành cho điện thoại di động ngày nay thường là các distro linux khá an toàn trong việc bảo vệ người sử dụng với các tính năng bảo mật của nền tảng hệ điều hành này mà virus trên điện thoại di động hiện tại chưa thể nắm sâu trong hệ thống, chưa phát hiện phần mềm lây nhiễm vào các file thực thi trên di động. Virus trên điện thoại di động chủ yếu là những phần mềm lừa đảo, ăn cắp, hoặc có hành vi phá hoại tập tin, gây hạn chế về mặt hiệu năng hoạt động của điện thoại... và thường được chia làm 3 loại chính là Trojan, Spyware và Adware.

Trojan trên di động: là những ứng dụng di động có các chức năng chạy thầm lặng các thao tác không mong muốn trong máy điện thoại của người dùng. Thông thường một Trojan trên điện thoại di động hay ẩn mình dưới danh nghĩa một ứng dụng hữu ích. Người dùng vẫn có thể sử dụng các chức năng mà người dùng mong muốn với ứng dụng đó, tuy nhiên bên trong ứng dụng Trojan có thể gây thiệt hại về tiền cho tài khoản điện thoại; lắng nghe, ăn cắp thông tin của người dùng trong khi người dùng không hề hay biết. Với việc điện thoại thông minh hiện tại đem lại trải nghiệm truy cập internet nhanh và tiện lợi như việc dùng máy tính cá nhân, virus trên điện thoại di động ngày nay còn có thể tấn phục vụ cho mục đích xấu là tấn công DDos bằng việc lây lan vào điện thoại của người dùng và biến điện thoại thành các zombie của một mạng botnet nào đó.

Spyware trên di động: là ứng dụng di động chuyên thu thập các thông tin từ máy điện thoại của người dùng rồi truyền tải chúng cho kẻ xấu mà không có sự nhận biết, cho phép của chủ điện thoại. Người sử dụng điện thoại thường bị nhiễm spyware khi cài phải những ứng dụng miễn phí qua mạng hoặc trên kho ứng dụng. DroidDream là dòng Spyware phổ biến nhất hiện nay ẩn mình dưới danh nghĩa là

game cho điện thoại di động Android. Người sử dụng các ứng dụng game DroidDream đều chơi game và cảm thấy đây như một ứng dụng giải trí bình thường, tuy nhiên bên trong ứng dụng này lại lắng nghe, thu thập toàn bộ các tin nhắn sms mà người dùng nhận được.

Adware trên di động: là ứng dụng trên điện thoại di động hay đi kèm với những thông tin quảng cáo khiến người dùng khó chịu. Người dùng phải trả tiền thì mới loại bỏ được sự phiền toái của quảng cáo. Loại virus này thường được tải về và cài đặt dưới dạng phần mềm miễn phí hay phiên bản dùng thử. Cần phân biệt rõ giữa Adware và phần mềm miễn phí sử dụng quảng cáo ở chỗ Adware chủ động yêu cầu người sử dụng trả phí để có thể sử dụng phần mềm thoái mái, trong khi đó phần mềm miễn phí trên di động ngày nay thông thường được chèn quảng cáo của nhà sản xuất ứng dụng, của google... Những quảng cáo này thường không gây khó chịu nhiều cho người dùng mà chỉ xuất hiện một cách hạn chế giúp người dùng lưu ý tới thông tin quảng cáo. Đây là phương pháp đem lại lợi nhuận phổ biến nhất của phần mềm miễn phí trên di động hiện tại.

4.3 Các phương pháp phòng tránh rủi ro với virus trên điện thoại

Sự phát triển của mạng viễn thông thế hệ mới với tốc độ cao như 3G, 4G (hay còn gọi là LTE) đã làm gia tăng các nguy cơ tấn công của virus trên điện thoại di động. Tâm lí chung của đa phần người sử dụng, đặc biệt là giới trẻ, giới doanh nhân với công việc bận rộn hay lưu trữ dữ liệu cá nhân quan trọng ngay trên điện thoại. Lợi dụng điều này điện thoại di động hiện tại trở thành con mồi lớn cho giới tội phạm công nghệ. Xét về độ nguy hiểm thì virus trên điện thoại di động nguy hiểm hơn nhiều so với virus trên máy tính bởi vì người sử dụng thường hay không quan tâm tới việc bảo vệ điện thoại di động của mình. Thông tin lưu trữ trên điện thoại di động thường nhạy cảm và rất có giá trị với kẻ xấu.

Để bảo vệ điện thoại, lợi ích của bản thân người sử dụng điện thoại tốt nhất nên tin dùng các phần mềm diệt virus cho điện thoại để tránh các mối nguy hại từ virus. Ngoài ra người sử dụng di động cũng nên trang bị tốt những kiến thức phòng tránh virus như:

- Luôn kiểm tra kĩ lưỡng xuất xứ của phần mềm sử dụng phần mềm của các hãng phần mềm uy tín.
- Chỉ cho nên cài những ứng dụng đòi hỏi sự cho phép truy cập thông tin cá nhân, đọc dữ liệu, truy cập internet... khi đã xem xét kỹ. Ví dụ khi muốn cài một phần mềm chat trên điện thoại thì ứng dụng đó đòi hỏi cấp quyền truy cập internet là điều hợp lý. Nhưng khi muốn cài một ứng dụng đọc sách mà phần mềm đó đòi hỏi người dùng cấp quyền đọc thông tin trong danh bạ thì người dùng nên cân nhắc kỹ.

- Thường xuyên kiểm tra tiền trong tài khoản điện thoại hoặc hóa đơn sử dụng dịch vụ viễn thông khi có điều bất thường có thể nghĩ tới việc điện thoại bị nhiễm virus
- Kiểm tra các tiến trình lạ chạy trong điện thoại.
- Luôn luôn lưu trữ, backup dữ liệu trong điện thoại, có phương án mã hóa, thay thế thông tin để đề phòng việc thông tin cá nhân bị ăn cắp.

5. Các phương thức kiểm soát truy cập trái phép, chống mất trộm cho điện thoại

Ở nước ta hiện tại điện thoại di động là tài sản có giá trị và không thể thiếu với rất nhiều người. Đặc biệt là những chiếc điện thoại thông minh thì ngoài giá trị về thông tin cá nhân là không thể đo đếm được còn là giá trị về kinh tế. Vì vậy ở nước ta việc phòng tránh các rủi ro về mất trộm điện thoại không chỉ bảo vệ người dùng trước việc bị lộ các bí mật riêng tư, thông tin nhạy cảm, thông tin quan trọng mà đôi khi còn là bảo vệ tài sản cá nhân có giá trị.

Thông thường phương pháp hiệu quả nhất để bảo vệ các truy cập trái phép của người khác, tránh sự tò mò, ăn cắp thông tin đó là đặt mật khẩu cho điện thoại. Việc phá mật khẩu của một chiếc điện thoại hiện tại còn khó hơn việc phá mật khẩu của máy tính bởi các phần mềm phá mật khẩu trên điện thoại chưa nhiều và khó khăn trong phương thức phá mật khẩu.

Người dùng điện thoại di động nên mã hóa các thông tin quan trọng trong điện thoại bằng mật khẩu để tránh trường hợp bị ăn cắp thông tin. Chỉ nên lưu trữ mật khẩu đăng nhập vào các giao dịch điện tử như giao dịch ngân hàng điện tử, giao dịch chứng khoán điện tử khi thực sự đảm bảo an toàn. Nếu không đảm bảo được điều này thì nên ghi nhớ các thông tin này hơn là lưu chúng lại dưới dạng file hay dữ liệu trong máy. Kẻ xấu hoàn toàn có thể dịch ngược được ứng dụng của nhà cung cấp dịch vụ giao dịch ngân hàng, giao dịch chứng khoán, mua bán... mà lấy được thông tin vô cùng quý giá này. Người sử dụng cần nâng cao nhận thức về tầm quan trọng của dữ liệu trên điện thoại. Thông thường những người như doanh nhân, người nổi tiếng, người có quan hệ xã hội rộng... thì các thông tin về bản thân họ luôn cần được bảo vệ bí mật. Trong trường hợp máy điện thoại bị phát hiện lấy cắp, việc hủy thông tin trên điện thoại là một điều cần thiết.

Ngoài ra với sự phát triển của khoa học kĩ thuật, ngày nay đa số điện thoại di động thông minh đều có thêm chức năng định vị. Việc cài đặt một phần mềm theo dõi, kiểm tra, điều khiển từ xa cho điện thoại là cần thiết. Trong trường hợp mất trộm điện thoại, việc ra lệnh tìm vị trí điện thoại giúp người dùng xác định và lấy lại điện thoại của mình là vô cùng hữu ích.

6. Phần mềm security cho điện thoại trên nền Android

Android là nền tảng hệ điều hành cho di động mới và nổi tiếng nhất hiện nay. Chính thức ra đời từ năm 2008 với sự xuất hiện của điện thoại thông minh Google Nexus

One, cho tới nay hệ điều hành Android đã trở thành hệ điều hành thông dụng nhất trên điện thoại thông minh. Là hệ điều hành mã nguồn mở, được Google cung cấp dưới bản quyền Apache vì vậy được rất nhiều hãng sản xuất điện thoại trên thế giới sử dụng cho sản phẩm của họ. Với sự lớn mạnh nhanh chóng của mình, Android hiện tại có kho ứng dụng rất lớn và phong phú và đi kèm với đó tất nhiên là cả những phần mềm độc hại. Tuy nhiên do là hệ điều hành non trẻ, những yếu tố rủi ro khi người dùng sử dụng sản phẩm chạy Android chưa được nghiên cứu nhiều trên thế giới, vì vậy việc đưa ra một phần mềm bảo mật cho nền tảng Android là rất cần thiết.

Trong đề tài nghiên cứu này em tập trung xây dựng sản phẩm phần mềm an toàn, bảo mật cho điện thoại di động chạy Android với hi vọng đưa được ra một công cụ hữu ích giúp cho người dùng sản phẩm trên nền tảng này được bảo vệ một cách an toàn hơn. Từ những vấn đề, rủi ro của việc sử dụng điện thoại như nạn tin nhắn rác, các cuộc gọi, tin nhắn quấy rối, virus trên điện thoại... các chức năng của phần mềm này là những công cụ hữu ích giúp giải quyết phần nào những vấn đề đó.

Các chức năng chính của phần mềm:

6.1 Chặn cuộc gọi quấy rối, chặn tin nhắn rác bao gồm:

Chặn các cuộc gọi, tin nhắn đến từ các số điện thoại có hại (chặn theo danh sách đen). Người dùng có thể định nghĩa, danh sách đen này, khi không muốn bị quấy rối bởi một số thuê bao nào đó, hay đưa số thuê bao này vào danh sách đen.

Chặn các cuộc gọi, tin nhắn theo luật: chức năng này thường được sử dụng khi người dùng bận rộn, không có thời gian để trả lời cuộc gọi, tin nhắn. Người dùng có thể định nghĩa các khoảng thời gian, những tình huống bận rộn, những trường hợp chặn cuộc gọi, tin nhắn trong chức năng này.

Chặn tin nhắn rác thông minh: thông thường các tin nhắn rác theo dạng quảng cáo, lừa đảo... có mẫu chung, có sự trùng lặp về nội dung. Phương pháp chặn thông minh dựa trên nội dung của các thông tin trong tin nhắn mà tự động đưa ra quyết định có chặn lại hay không, tránh làm phiền cho người sử dụng.

6.2 Backup, lưu trữ dữ liệu cá nhân của người dùng: đây là chức năng lưu trữ dữ liệu cá nhân của người dùng điện thoại trên thẻ nhớ, hay trên online server. Chức năng này cần thiết cho người dùng trong trường hợp mất máy điện thoại, đồng bộ thông tin cá nhân của người dùng khi sử dụng nhiều thiết bị.

6.3 Diệt virus: chức năng tự động nhận diện và xóa bỏ các phần mềm độc hại (Trojan, Spyware, Adware) trên điện thoại di động. Tự động cập nhật các mẫu nhận diện virus mới nhất từ server của Bkav.

6.4 Chống trộm bao gồm:

Điều khiển điện thoại di động bằng tin nhắn: trong trường hợp phát hiện điện thoại bị mất cắp, người sử dụng có thể nhắn tin tới điện thoại của mình theo cú pháp nhất định nhằm khóa máy tránh các truy cập của người lạ, xóa dữ liệu cá nhân quan trọng khi không muốn chúng lọt vào tay kẻ xâm.

Điều khiển điện thoại di động qua website: người dùng cũng có thể ra lệnh khóa máy, xóa thông tin cá nhân hoặc tìm vị trí điện thoại qua một cổng thông tin điện tử. Khi máy điện thoại của người dùng bị khóa lại, phần mềm sẽ tự động gửi địa chỉ lên server của cổng thông tin điện tử thông qua hệ thống định vị toàn cầu GPS, qua đó người sử dụng có thể xác định được vị trí và lấy lại thiết bị đã mất.

Trong nhiều trường hợp kẻ xâm khi đã lấy cắp được điện thoại của người dùng, chúng thường thay số điện thoại của máy để cắt đứt liên lạc với chủ nhân của thiết bị. Phần mềm cung cấp chức năng tự động khóa máy, và báo số điện thoại mới lên webserver hoặc gửi thông tin qua tin nhắn sms tới một số điện thoại khác mà chủ nhân của máy tin dùng và đăng kí với phần mềm.

7. Lời tri ân

Em xin được tỏ lòng biết ơn tới thầy Lương Mạnh Bá, tới Trung tâm An ninh mạng Bkis đã giúp em hoàn thành đề tài nghiên cứu này.

8. Tài liệu tham khảo

http://en.wikipedia.org/wiki/Mobile_virus

http://en.wikipedia.org/wiki/Sms_spam

<http://en.wikipedia.org/wiki/Antitheft>

Giáo trình An toàn và bảo mật thông tin – thầy Nguyễn Khanh Văn bộ môn Công nghệ phần mềm, viện Công nghệ thông tin và Truyền thông, đại học Bách Khoa Hà Nội.

Botnet Tracking Framework – Framework hỗ trợ theo dõi và giám sát mạng botnet

Triệu Minh Tuân

Tóm tắt - Trong những năm gần đây, hiểm họa botnet đang dần trở lại và ngày càng trở nên nguy hiểm hơn với nhiều công cụ và hình thức tấn công mới, từ ăn cắp thông tin, đến phát tán thư rác, phát tán mã độc hay tấn công từ chối dịch vụ, đã gây ra không ít thiệt hại nặng nề về kinh tế và xã hội. Chính vì vậy, rất nhiều giải pháp phát hiện botnet hay làm giảm nhẹ thiệt hại do botnet gây ra đã được công bố và triển khai trong thực tế. Botnet Tracking (BT) là một trong những giải pháp đó, tư tưởng của Botnet Tracking là giám sát các mạng botnet nhằm thu thập các thông tin hỗ trợ giảm nhẹ hiểm họa này. Báo cáo đề cập tới vấn đề xây dựng một Framework (Botnet Tracking Framework – BNF) hỗ trợ việc theo dõi và giám sát các mạng Botnet. Hiện Framework vẫn đang được phát triển và bước đầu được đưa vào hoạt động tuy nhiên, từ những thông tin thu được ban đầu này đã cho thấy hiệu quả của Framework trong việc hỗ trợ giảm nhẹ hiểm họa Botnet.

Từ khóa - botnet tracking, botnet tracker, honeypot, botnet.

I. Giới thiệu

1. 1 Botnet

Botnet là khái niệm đề cập tới một tập hợp các máy tính bị lợi dụng hay còn gọi là các bots chịu sự kiểm soát của các hackers (hay còn được biết đến như các bot holders và bot masters) phục vụ cho các ý đồ xấu xa của các cá nhân này.

Công trình này được thực hiện dưới sự bảo trợ của trung tâm an ninh mạng BKIS.

Triệu Minh Tuân, sinh viên lớp Công nghệ phần mềm, khóa 51, Viện công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 01687-719-185, E-mail: tuantm@bkav.com.vn).

© Viện Công nghệ thông tin và Truyền thông,
trường Đại học Bách Khoa Hà Nội.

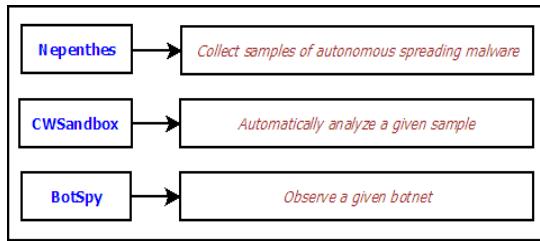
Phân loại botnet : Botnet thường được phân loại theo kiến trúc (tập trung, phân tán) hoặc phân loại theo giao thức mạng mà botnet sử dụng để giao tiếp với nhau (giao thức IRC – Internet Relay Chat, giao thức HTTP – Hyper Text Transfer Protocol, ...).

Các hình thức tấn công của Botnet hiện nay thì rất đa dạng nhưng phổ biến nhất là tấn công từ chối dịch vụ (Ddos), phát tán thư rác số lượng lớn (*spam*), lừa đảo trực tuyến (*phishing*), ăn cắp dữ liệu (*data theft*), ... nguy hiểm nhất có thể còn là phá hoại máy tính người sử dụng.

1.2 Botnet Tracking

Botnet nguy hiểm là vậy, nên tính đến nay, đã có rất nhiều giải pháp được đưa ra nhằm làm giảm các thiệt hại do Botnet gây ra, Botnet tracking là một trong số đó. Botnet Tracking được giới thiệu trong tài liệu Virtual Honeypots: From Botnet Tracking to Instrusion Detection¹, các tác giả đưa ra tư tưởng hệ thống Botnet tracking dựa trên các thành phần của một hệ thống Honeypot để theo dõi botnet, bao gồm:

¹ Chapter 11: Tracking Botnet - Virtual Honeypots: From Botnet Tracking to Instrusion Detection



Hình 1 Mô hình Botnet Tracking trong Virtual Honeypot ...

Trong đó:

- Nepenthes² là hệ thống giả lập các lỗ hổng dịch vụ để thu hút các malware, nhờ đó chúng ta sẽ thu thập được “mẫu” (malware - các chương trình độc hại, trong đó có botnet).
- CWSandbox³ là hệ thống tự động phân tích hành vi của các “mẫu” thu thập được.
- BotSpy là một chương trình IRC client tối ưu cho việc theo dõi các bot với một số hỗ trợ đặc biệt như hỗ trợ theo dõi song song, tự động tải về các biến thể của bot, hỗ trợ cơ sở dữ liệu để lưu trữ thông tin ... Về cơ bản, BotSpy sẽ thực hiện giả lập lại kết nối giao tiếp lên Server điều khiển của các bot để thực hiện thu thập thông tin. Duy trì thu thập thông tin, chúng ta có thể theo dõi được các hoạt động của mạng botnet và từ đó, đưa ra các phương án xử lý phù hợp.

1.3 Bot Milkers

Được sử dụng xuyên suốt nghiên cứu là lý thuyết Bot Milker^[2]. Trong lý thuyết Botnet Milker, người ta tạo ra các “bot giả lập không chứa các thành phần nguy hiểm”⁴, các thành phần này sẽ “tham gia” vào mạng botnet (cụ thể là mạng Botnet Megad-D) để tiến hành thu thập các thông tin, dữ liệu phát tán thư rác của mạng botnet này, từ đó theo dõi hành vi của cả mạng botnet này.

² Chapter 6 - Virtual Honeypots: From Botnet Tracking to Instruction Detection

³ Chapter 12 - Virtual Honeypots: From Botnet Tracking to Instruction Detection

⁴ “Bot emulators without malicious side effects”

Ý nghĩa của cái tên “Bot Milker” bắt nguồn từ việc: Ở đây, các bot giả lập sẽ không giao tiếp với server điều khiển mà chỉ thực hiện “vắt”⁵ thông tin từ các server này, do đó, tính ứng dụng của Bot Milker có thể bị thu hẹp lại trong phạm vi một số mạng botnet hiện nay.

II. Phát triển hệ thống Botnet Tracking

2.1 Phân tích giới hạn

Thực hiện phân tích hai lý thuyết nêu trên, ta nhận thấy một số giới hạn:

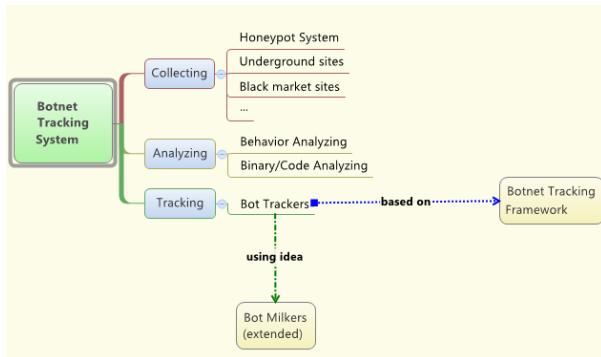
- Về hệ thống Botnet Tracking: sử dụng các thành phần của một hệ thống honeypot, do đó, có phần phụ thuộc vào hệ thống này. Chưa kể đến, lý thuyết về hệ thống này đã được công khai tính tới nay đã được 6 năm, trong thời gian này, Botnet đã phát triển và nguy hiểm hơn rất nhiều: từ phát hiện đến chống phát các hệ thống anti-botnet (ngăn chặn botnet – trong đó có Botnet tracking).
- Về lý thuyết Bot Milkers: Đây có thể coi là một lý thuyết vòi trong việc theo dõi thông tin về các mạng Botnet, đó là sử dụng các thành phần mô phỏng lại các bot (*bot emulator*) để thu thập thông tin trực tiếp từ các server điều khiển. Tuy nhiên, lý thuyết này có một nhược điểm đó là chỉ thực hiện “vắt” thông tin từ phía các server điều khiển mà không kể đến các giao tiếp khác của các bot, do đó, có khả năng sẽ bị các bot master phát hiện và thực hiện điều hướng (trả về thông tin sai), chống phá hay ngăn chặn.

2.2 Thiết kế hệ thống Botnet Tracking

2.2.1 Hệ thống Botnet Tracking

Từ các giới hạn trên kết hợp với tham khảo tài liệu, báo cáo thực hiện mở rộng hệ thống Botnet Tracking ở trên thành một hệ thống Botnet Tracking bao gồm 3 thành phần:

⁵ Ý nghĩa của động từ **milk** trong tiếng Việt.



Hình 2 Hệ thống Botnet Tracking

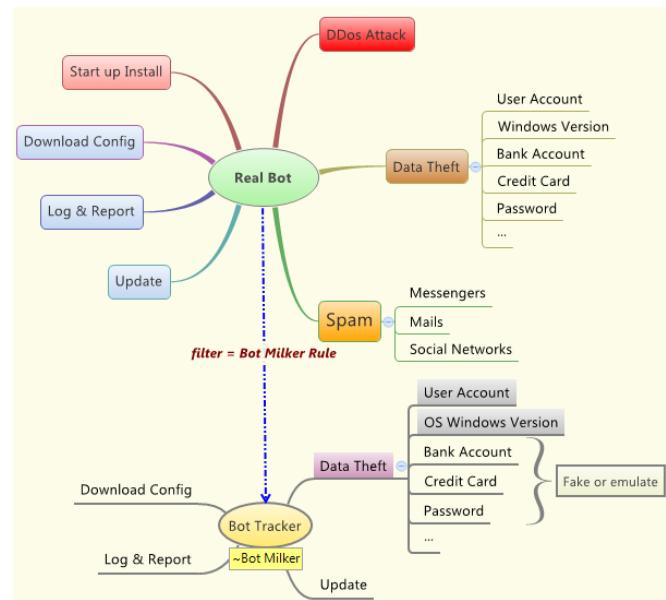
- Thành phần Collecting: chịu trách nhiệm thu thập mẫu, các biến thể, các bộ cài đặt botnet từ nhiều nguồn, không chỉ giới hạn trong phạm vi của một hệ thống honeypot ở trên (Nepenthes), mở rộng tới các phạm vi mà các botnet được rao bán, giới thiệu như một sản phẩm phần mềm đó là các kênh thông tin ngầm, các diễn đàn chợ đen, các kênh thông tin của các hackers, crackers v...v... và tới phạm vi mà các bot lây lan, đó là trên máy tính của người dùng bị lây nhiễm Botnet mà thông qua các cơ chế phát hiện mã độc của các phần mềm Antivirus, chúng ta có thể thu thập được những “mẫu” này.
- Thành phần Analyzing: thực hiện phân tích các mẫu thu được trong đó kết hợp linh động giữa hai phương pháp phân tích là: phân tích theo hành vi (*sử dụng các công cụ giám sát, bắt gói tin, các sandbox ...*) và phân tích theo mã/phân tích nhị phân (*sử dụng các kỹ thuật Reverse engineering, các công cụ hỗ trợ như các disassemblers, debuggers ...*). Trong đó, phương pháp phân tích nhị phân có thể cho chúng ta thông tin đầy đủ và chi tiết về bot hơn rất nhiều là phương pháp chỉ sử dụng CWSandbox ở trên, tuy nhiên, phương pháp này đòi hỏi chi phí và công sức lớn hơn nhiều.
- Thành phần Tracking: là thành phần chính của cả hệ thống, thông qua quá trình phân tích kể trên, kết hợp với lý thuyết Bot Milker mở rộng(*sẽ được trình bày phía dưới*), các thành phần “non malicious” của bot sẽ được trích xuất và tiến hành mô phỏng dựa trên Botnet Tracking Framework, thành phần này được gọi là các Bot Tracker, thực hiện giao tiếp với các server điều khiển mạng botnet và thu thập thông tin hỗ trợ cho việc giảm nhẹ các thiệt hại do Botnet gây ra.

2.2.2 Lý thuyết Bot Milker mở rộng

Để có thể đáp ứng được nhu cầu tracking thực tế, lý thuyết Bot Milker ở trên cần được mở rộng. Giữ nguyên tư tưởng của

Bot Milker là “*Bot giả lập không chứa các thành phần nguy hiểm*”, sau đó, mở rộng ý nghĩa “vắt” ở đây là chúng ta sẽ có giao tiếp với các server điều khiển của botnet để duy trì hoạt động thu thập thông tin. Tuy nhiên, để đảm bảo “*không chứa các thành phần nguy hiểm*”, các thông tin, dữ liệu được sử dụng trong quá trình giao tiếp, cũng được *giả lập* để tránh những thông tin cá nhân, nhạy cảm, những dữ liệu mật có thể bị rò rỉ ...

Hình dưới đây mô tả quá trình “lọc” hành vi của một Bot để xác định các thành phần cần mô phỏng phục vụ theo dõi botnet:



Hình 3 Lọc hành vi của một bot thực bằng lý thuyết Bot Milker mở rộng.

2.3 Bot net Tracking Framework

Nhằm phục vụ cho thành phần Tracking kể trên, Botnet Tracking Framework (BTF) được xây dựng như một thư viện hỗ trợ việc giả lập các bot để đi theo dõi, thu thập thông tin từ các server điều khiển nhằm xác định các thông tin hỗ trợ cho mục tiêu làm giảm thiệt hại do Botnet gây ra. Để làm được như vậy, nghiên cứu đã thực hiện tham khảo các bài báo, tài liệu có liên quan[4][5][6], phân tích một tập các mẫu Botnet, cùng các công cụ có liên quan theo cả hai phương pháp: phân

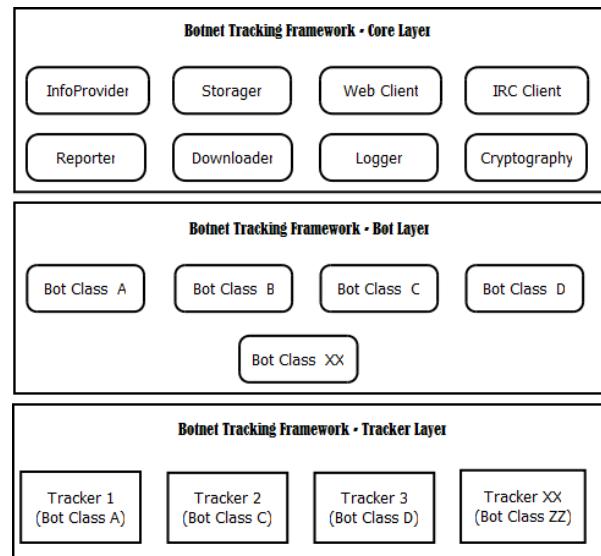
tích theo hành vi và phân tích nhị phân kẽ trên, từ đó xác định ra các đặc điểm chung về Botnet mà Framework cần đáp ứng, đó là:

- Giả lập Web Client: hỗ trợ mô phỏng các Botnet sử dụng giao thức HTTP để giao tiếp.
- Giả lập IRC Client: hỗ trợ mô phỏng các Botnet sử dụng giao thức IRC để giao tiếp.
- Bộ các hình thức mã hóa, giải mã hóa, các hàm băm ... hỗ trợ mã hóa, giải mã dữ liệu theo giao thức mà các Botnet sử dụng.
- Giả lập cung cấp thông tin hệ thống: đảm bảo cung cấp thông tin giao tiếp nhưng không chứa các thành phần nguy hiểm.

Bên cạnh đó là một số nhu cầu về lưu trữ, theo dõi thông tin thu thập được đòi hỏi BTF phải đáp ứng:

- Hỗ trợ tải file: để thu thập những thông tin cấu hình, cập nhật biến thể mới của bot.
- Hỗ trợ ghi log, báo cáo, lưu trữ file: để tiện cho quá trình theo dõi, lập thống kê sau này.

Từ những yêu cầu trên, BTF được thiết kế như sau:



Hình 4 Botnet Tracking Framework Architecture

- Core Layer: là thành phần chính của BTF, làm nền tảng phát triển các layer còn lại.
- Bot Layer: là thành phần thực hiện mô phỏng các bot (*dạng bot milker mờ róng*) của các mạng botnet.
- Tracker Layer: là thành phần được xây dựng từ các layer khác, thực hiện thu thập thông tin các mạng botnet và báo cáo lại cho các mô đun điều khiển của hệ thống theo dõi.

III. CÀI ĐẶT VÀ ĐÁNH GIÁ

3.1 Cài đặt

Từ những phân tích trên, Framework được cài đặt trên Visual C++/MFC, sau hơn một tháng cài đặt đã cơ bản hoàn thành phần Core Layer và hiện đang được sử dụng để di theo dõi một số mạng Botnet như :

- ngrBot: là Botnet IRC, thực hiện ăn cắp thông tin, lừa đảo qua các chương trình chat (*messenger*), liên tục cập nhật các biến thể mới, các server điều khiển mới.
- Botnet tấn công Ddos các trang báo tiếng Việt (trong đó có danlambao, zing ...): Botnet này liên tục cập nhật các cấu hình mới để đổi mục tiêu tấn công, cũng như thay đổi các hình thức tấn công.

- Vector IRC botnet: là module botnet đi kèm theo một dòng virus lây file nguy hiểm, sử dụng kênh thông tin IRC để phát tán rất nhiều loại malware khác nhau.

3.2 Kết quả thử nghiệm

Sau một thời gian theo dõi hoạt động của các tracker kể trên, nghiên cứu đã thu được một số kết quả ban đầu, đặt nền móng cho việc phát triển hoàn thiện hệ thống Botnet Tracking sau này:

- Cập nhật mẫu và các liên kết độc hại vào hệ thống honeypot hiện tại của Bkav.
- Cập nhật cấu hình tấn công, phân tích, giải mã xác định đối tượng tấn công của botnet để hỗ trợ cảnh báo.

IV. LỜI CẢM ƠN

Em xin chân thành cảm ơn Tiến sĩ Nguyễn Khanh Văn, anh Nguyễn Thé Luân cùng tập thể các bạn trong nhóm làm thực tập tốt nghiệp do thầy Văn hướng dẫn, đã giúp đỡ em trong quá trình làm đồ án tốt nghiệp với nghiên cứu này. Và em

cũng xin được gửi lời cảm ơn tới anh Đỗ Mạnh Dũng, anh Nguyễn Công Cường, anh Nguyễn Ngọc Dũng và các bạn cùng nghiên cứu đề tài này thuộc trung tâm an ninh mạng BKIS - Đại học Bách Khoa Hà Nội, đã tận tình chỉ bảo, giúp đỡ, tạo điều kiện cho em trong suốt thời gian nghiên cứu.

V. TÀI LIỆU THAM KHẢO

- [1]. Niels Provos; Thorsten Holz, Virtual HoneyPots: From Botnet Tracking to Intrusion Detection.
- [2]. Chia Yuan Cho, Juan Caballero, Chris Grier, Vern Paxson, Dawn Song (University of California, Berkeley), Insights from the Inside:A View of Botnet Management from Infiltration
- [3]. Wikipedia, một số khái niệm lý thuyết liên quan đến Botnet.
- [4]. Jianwei Zhuge, Thorsten Holz, Xinhui Han, Jinpeng Guo, and Wei Zou, Characterizing the IRC-based Botnet Phenomenon.
- [5]. Mahathi Kiran.Kola, Botnets: Overview and Case Study.
- [6]. Felix C. Freiling, Thorsten Holz Georg Wicherski, Botnet Tracking: Exploring a Root-Cause Methodology to Prevent Distributed Denial-of-Service Attacks

RSED: Môi Trường Giả Lập Mạng Giống Thực Tế Phục Vụ Cho Nghiên Cứu Tấn Công Từ Chối Dịch Vụ (DDoS)

Trương Thảo Nguyên

Tóm tắt - Ngày nay, trong lĩnh vực nghiên cứu các tính chất của cuộc tấn công từ chối dịch vụ cũng như xây dựng các phương pháp phòng tránh và giảm thiểu tác hại của hình thức tấn công này, hầu hết các nghiên cứu sinh đều sử dụng các mô trường mô phỏng để thực nghiệm những nghiên cứu của mình. Kết quả của quá trình giả lập được sử dụng làm cơ sở đánh giá ý nghĩa của vấn đề được đặt ra, tính hiệu quả của những đề xuất, giải pháp cụ thể trong quá trình nghiên cứu. Mô trường mô phỏng càng giống với thực tế càng đảm bảo kết quả của quá trình nghiên cứu là chính xác và có tính thực tiễn. Do đó, vấn đề xây dựng một bộ giả lập mạng phù hợp với tính chất đặc thù của nghiên cứu tấn công từ chối dịch vụ và đảm bảo tính chất gần với thực tế là rất cần thiết. Nghiên cứu này trình bày về RSED (Realistic network Simulation Environment for DDoS attacks), một framework được xây dựng trên nền bộ giả lập OMNET++ [1], các thư viện có sẵn của INET framework [2] và được tham khảo nhất định từ ReaSE framework [3]. RSED cung cấp cho người sử dụng một môi trường giả lập có khả năng tạo được các mạng có cấu trúc lên tới hàng trăm nút và hỗ trợ giả lập đường truyền mạng giống với thực tế. Ngoài ra, RSED cũng hỗ trợ giả lập các mạng bootnet, hay tiến hành các thực nghiệm các bộ phòng chống DDoS một cách nhanh chóng. Nghiên cứu đã cài đặt thử nghiệm phương pháp phòng chống DDoS sử dụng BloomFilter trên môi trường giả lập này.

Từ khóa - DDoS attack, Network simulation environment, Network topology generator, OMNET++

1 GIỚI THIỆU

Ngày nay, trong lĩnh vực nghiên cứu các tính chất của cuộc tấn công từ chối dịch vụ cũng như xây dựng các phương pháp phòng tránh và giảm thiểu tác hại của hình thức tấn công này, hầu hết các nghiên cứu sinh đều sử dụng các mô trường mô phỏng để thực hiện những nghiên cứu của mình. Kết quả từ quá trình giả lập sẽ được thu thập, thống kê để tiếp tục trở thành đầu vào cho việc đánh giá kết quả nghiên cứu, đánh giá những đề xuất, giải pháp mà người nghiên cứu thực hiện, tìm hiểu. Mô trường mô phỏng giống thực sẽ là tiền đề cho việc đánh giá các kết quả

Trương Thảo Nguyên, sinh viên lớp Công Nghệ Phần Mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: (+84)0984-196-715, e-mail: nguyen.88.tt@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

nghiên cứu một cách chính xác, là cơ sở để khẳng định những kết quả này là có khả năng ứng dụng, và phát triển. Vì vậy, yêu cầu xây dựng một bộ giả lập càng giống với thực tế càng tốt để phục vụ cho việc nghiên cứu đã trở thành một vấn đề bức thiết. Đặc biệt, vấn đề này trở nên vô cùng quan trọng trong lĩnh vực nghiên cứu về các cuộc tấn công từ chối dịch vụ (DDoS) vốn khó có thể tiến hành nghiên cứu, thử nghiệm trực tiếp trên mạng thực. Một bộ mô phỏng như vậy không chỉ có thể giả lập các thành phần đặc thù trong nghiên cứu DDoS (các thành phần của mạng botnet, các router định tuyến, các bộ phòng chống, giảm thiểu tác hại, các server chịu tấn công...) mà còn phải có khả năng giả lập được cấu hình mạng (network topology) đủ lớn, mang tính chất, đặc tính của cấu hình mạng trong thực tế. Bên cạnh đó, giả lập giao thông mạng (network traffic) cũng phải giống thực. Một bộ mô phỏng mạng DDoS yêu cầu cần phải giả lập cả giao thông mạng bình thường lẫn giao thông mạng của cuộc tấn công.

Trong phạm vi tìm hiểu của người viết, hiện tại chỉ có một vài bộ giả lập mạng phục vụ cho việc nghiên cứu DDoS như DDoSSim [4], và ReaSE[3][14]. Tuy nhiên các bộ giả lập mạng này đều chưa đáp ứng được các yêu cầu nêu trên. DDoSSim cung cấp một mô hình giả lập dựa trên tác nhân (agent-based simulation) [4] cho phép người dùng có thể thêm mới, thay đổi cơ chế tấn công cũng như cơ chế phòng chống phục vụ cho các thử nghiệm khác nhau. Tuy nhiên bộ giả lập này chỉ hướng tới giao thông mạng, đặc biệt là giao thông mạng của một cuộc tấn công mà chưa chú ý đến khía cạnh cấu hình mạng. Được phát triển từ năm 2008, ReaSE[3] là một bộ giả lập xây dựng trên nền OMNET++ [1]. Bộ giả lập này có khả năng hỗ trợ cả sinh cấu hình mạng và giao thông mạng giống với thực tế. Tuy nhiên, các tính năng này ReaSE được xây dựng tách biệt trong hai gói cài đặt khác nhau và được sử dụng độc lập. Điều này khiến cho quá trình xây dựng các thực nghiệm trở nên kém nhanh chóng và tốn động hơn. Bên cạnh đó, ReaSE cũng chưa cung cấp được mô phỏng hầu hết các thành phần của cuộc tấn công DDoS. Bộ giả lập này mới chỉ dừng lại ở mô hình tấn công trực tiếp (từ các máy tấn công tới server bị tấn công) mà chưa có khả năng mô phỏng các cuộc tấn công phân tán có nhiều cấp độ của kẻ tấn công [3]. Do đó, theo người viết, ReaSE chỉ mới áp dụng cho việc nghiên cứu DoS ở phạm vi mạng nhỏ, chưa thể áp dụng trong nghiên cứu các vấn đề khác nhau của tấn công DDoS như phòng chống tấn công phân tán, dò vết IP (IP traceback)...

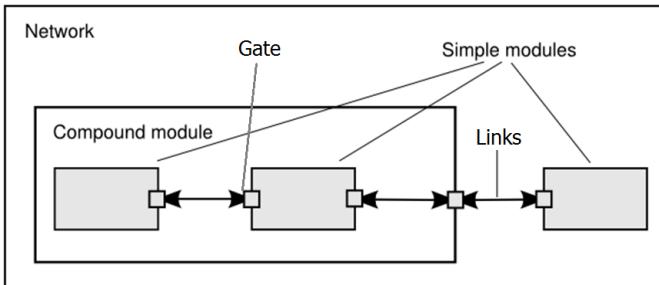
Trong nghiên cứu của mình, người thực hiện đề xuất ra một mô hình giả lập dựa trên nền OMNET++[1], INET framework[2]

và có tham khảo nhất định từ ReaSE framework [3] có tên là RSED (Realistic network Simulation Environment for DDoS attacks). Đây là một bộ giả lập đặc thù cho việc nghiên cứu các cuộc tấn công và phòng chống, giảm thiểu DDoS, đáp ứng được các yêu cầu về cấu hình mạng cũng như giao thông mạng giống thật. Mô hình cho phép mở rộng, thêm mới các thành phần mạng, giải thuật sinh cấu hình mạng, giao thông mạng phù hợp với xu thế phát triển, nghiên cứu trong các lĩnh vực hẹp này. Trong nghiên cứu này, người thực hiện cũng tiến cài đặt thực nghiệm phương pháp phòng chống và giảm thiểu tác hại sử dụng BloomFilter[19] như là một minh chứng cho khả năng của môi trường giả lập.

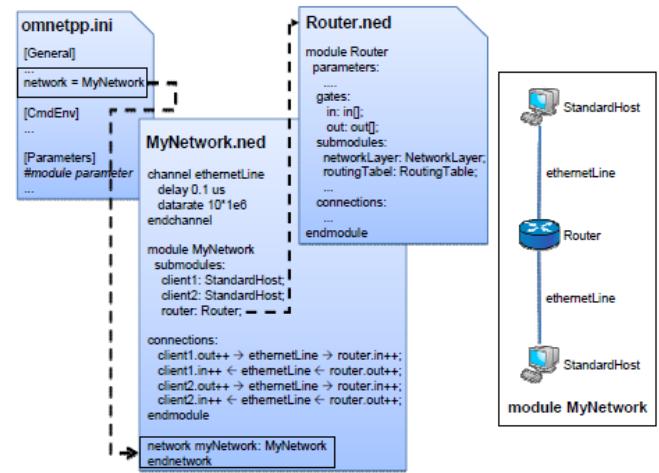
Trong khuôn khổ của báo cáo, cách thức xây dựng bộ giả lập RSED dựa trên OMNET++ và INET framework sẽ được người viết chỉ rõ trong phần 2 dưới đây. Phần 3 nêu chi tiết các thành phần của bộ giả lập cũng như các lựa chọn giải thuật sử dụng trong quá trình sinh cấu hình mạng, giao thông mạng. Người viết cũng trình bày về ví dụ minh họa cài đặt phương pháp phòng chống và giảm thiểu tấn công DDoS sử dụng BloomFilter trong phần 4, từ đó rút ra những đánh giá kết luận về môi trường giả lập RSED ở phần 5.

2 XÂY DỰNG BỘ GIẢ LẬP RSED DỰA TRÊN OMNET++ VÀ INET FRAMEWORK

OMNET++ (Objective Modular Network Testbed in C++) là một bộ mô phỏng các sự kiện rời rạc được ứng dụng rộng rãi trong giả lập mạng lưới truyền thông, mạng internet[1]. OMNET++ cung cấp một kiến trúc các module cho các mô hình giả lập khác nhau. Các module được lập trình bằng C++, sau đó được kết hợp lắp ráp với nhau thành các module lớn hơn bằng cách sử dụng ngôn ngữ bậc cao NED (Network description language). Các module được lập trình bằng C++ gọi là *simple module* bao gồm một hoặc nhiều các lớp diễn tả các chức năng của thành phần đó. Các module được kết hợp từ hai hay nhiều simple module được gọi là *compound module*. Chức năng của các compound module này là sự kết hợp các chức năng của simple module bên trong nó. Do đó compound module không được cài đặt bằng bất kì lớp C++ nào. Các module được kết nối với nhau thông qua các gate bằng links hay channel. Thông qua các kết nối này, các thông điệp, gói tin được truyền qua lại tạo nên sự tương tác giữa các module với nhau.



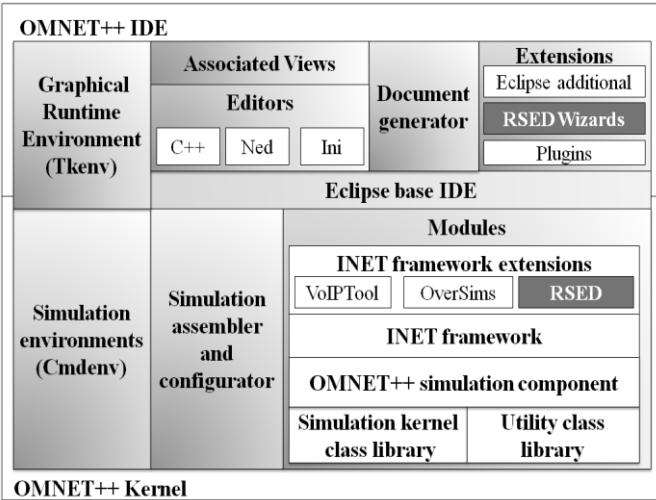
Hình 1: Các thành phần cơ bản của OMNET++



Hình 2: Ví dụ cơ bản về cài đặt mô phỏng bằng OMNET++

Quá trình giả lập mạng bằng OMNET++ cần ít nhất 2 file để có thể thực hiện một mô phỏng: Omnetpp.ini là file chứa cấu hình của mạng được mô phỏng và *.NED là các file chứa thông tin về cấu trúc các modules, định nghĩa các kết nối giữa các modules. Các file này sau đó được kết hợp với các file C++ diễn tả chức năng của các modules, được biên dịch và chuyển sang cho thành phần môi trường giả lập (Cmdenv) nằm trong nhân của OMNET++. Trong khi Cmdenv tiến hành giả lập theo kịch bản được lập trình sẵn, thành phần đồ họa Tkenv sẽ hiển thị quá trình giả lập này cho người sử dụng một cách trực quan.

Bên cạnh các thư viện C++ chuẩn, OMNET++ cũng cung cấp nhiều framework khác nhau phục vụ giả lập mạng trong nhiều lĩnh vực hẹp khác nhau. Trong đó, INET là một framework phục vụ giả lập mạng Internet trong thực tế và được sử dụng nhiều trong các nghiên cứu khoa học. INET hỗ trợ giả lập các mô hình, các giao thức cho cả mạng không dây lẫn mạng có dây như UDP, TCP, SCTP, IP, Ipv6, Ethernet, PPP [2] và nhiều giao thức khác. Nhờ đó, nhiều môi trường giả lập phục vụ nghiên cứu các vấn đề, các giao thức cụ thể của mạng Internet đã được xây dựng như VoIP, OverSim, HTTP [2]. Dựa trên tư tưởng đó, trong nghiên cứu của mình, người viết đã thực hiện xây dựng một môi trường giả lập mạng phục vụ cho việc nghiên cứu DDoS đặt tên là RSED. RSED được phát triển thành hai thành phần hỗ trợ nhau bao gồm: RSEDWizards và RSED framework (Xem hình 3). RSED wizards cung cấp một giao diện đồ họa, cho phép người sử dụng có thể lựa chọn một số tham số đầu vào như tổng số nút trong mạng, các loại nút trong mạng, bộ sinh giả lập mạng... Từ đó sinh ra cấu hình mạng giống thực tế, nhanh chóng và thích hợp với yêu cầu bài toán đặt ra. Ngoài ra, nhờ cài đặt trên OMNET++, các cấu hình mạng được sinh ra cũng đáp ứng được khả năng giả lập cấu hình mạng lên tới hàng trăm, thậm chí hàng nghìn nút. RSED framework là một thư viện bao gồm các mô hình, các module liên quan đến giả lập các thành phần của cuộc tấn công từ chối dịch vụ DDoS cũng như các thành phần tham gia vào quá trình phòng chống và giảm thiểu loại hình tấn công này. Bên cạnh đó, framework này cũng cung cấp các thành phần giả lập giao thông mạng bao gồm giao thông mạng thông thường và mạng có sự tấn công DDoS.

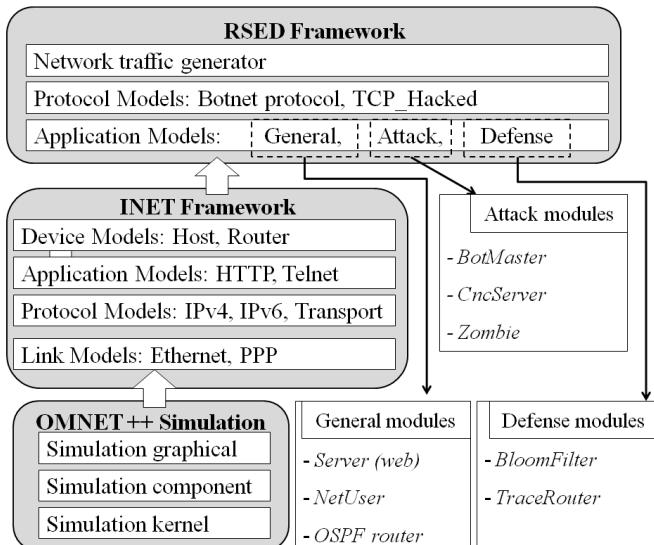


Hình 3: RSED trong kiến trúc tổng thể của OMNET++.
RSED Wizards: thành phần sinh cấu hình mạng. **RSED framework:** các module liên quan đến nghiên cứu DDoS và sinh giao thông mạng

3 MÔI TRƯỜNG GIẢ LẬP RSED

Phần 2 nêu trên đã nêu ra các thành phần của môi trường giả lập RSED cũng như vị trí, vai trò của các thành phần này trong kiến trúc tổng thể của OMNET++ và INET framework. Trong phần 3 này người viết sẽ trình bày cụ thể hơn về các cấu trúc của các thành phần đó. Nội dung của phần 3 bao gồm 3 mục. Mục 3.1 miêu tả về các modules hỗ trợ nghiên cứu DDoS. Mục 3.2 trình bày về các module sinh giao thông mạng nằm trong RSED framework. Cuối cùng mục 3.3 được dành để viết về thành phần sinh cấu hình mạng RSED wizards.

3.1 Các modules hỗ trợ nghiên cứu DDoS



Hình 4: RSED framework

Hình 4 miêu tả một cách tổng quan các thành phần của RSED framework. Framework này bao gồm các module phục vụ cho việc tấn công lẫn phòng chống DDoS như: General modules,

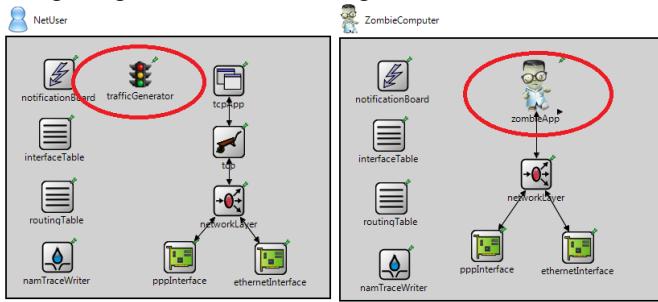
Attack modules và Defense modules. General modules là bộ các module cơ bản, đặc trưng cho các thành phần có mặt trong một mạng internet thông thường như Server, NetUser (đặc trưng cho người dùng mạng thông thường), OSPF router [2] (router định tuyến bằng giải thuật OSPF). Những module này đều đã được phát triển trước đó trong INET framework, tuy nhiên trong RSED đã thêm một số chức năng mới như khả năng bị tràn, khả năng bị đỗ vỡ khi có cuộc tấn công DDoS. Attack modules là bộ các module liên quan đến tấn công DDoS. Cụ thể trong RSED hỗ trợ các module xây dựng một mạng botnet tấn công phân tán: BotMaster (ké phát lệnh tấn công), CncServer (các server trung gian, chuyển tiếp lệnh, điều khiển tấn công) và zombie (các máy tính bị chiếm quyền điều khiển, tham gia trực tiếp vào quá trình tấn công). Cuối cùng Defense modules là tập hợp các module phục vụ cho nghiên cứu phòng chống và giảm thiểu tấn công DDoS. Trong nghiên cứu này, người thực hiện mới chỉ cài đặt hai phương pháp khác nhau đó là BloomFilter [17], [18], [19] và dò vết IP sử dụng phương pháp logging. Bên cạnh các module kể trên, RSED framework còn xây dựng lên một giao thức trao đổi thông tin trong mạng botnet gọi là Botnet protocol. Qua đó hoàn thiện một bộ công cụ hỗ trợ giả lập tấn công từ chối dịch vụ phân tán bằng mạng botnet.

3.2 Bộ sinh giao thông mạng

Quá trình giả lập, sinh giao thông mạng phục vụ cho nghiên cứu DDoS bao gồm hai yêu cầu nhỏ: yêu cầu về việc sinh ra giao thông mạng thông thường (general traffic) và sinh giao thông mạng của một cuộc tấn công (attack traffic). RSED framework đã đáp ứng được cả hai yêu cầu này. Giao thông mạng thông thường được xây dựng cài đặt trong simple module trafficGenerator. Module này được cài đặt không có công và không liên kết với bất kỳ một module nào khác. Trong bất kỳ một compound module nào được cài đặt theo mô hình phân tầng của mạng, có chứa traffic generator, thông tin về các gói tin từ tầng application của compound module được gửi đến các module khác sẽ được lấy ra từ kết quả thực hiện của module này. Hình 5a miêu tả module NetUser (đặc trưng người dùng thông thường) có sử dụng trafficGenerator. Với cách thiết kế như vậy, bộ sinh giao thông mạng thông thường có thể tùy biến sử dụng ở nhiều các module khác nhau, thậm chí cả các module mới được phát triển trong tương lai. Ngoài ra, traffic generator cũng được định nghĩa thông qua khái niệm trừu tượng của C++, cho phép thay đổi giải thuật sinh giao thông mạng một cách linh động. Trong nghiên cứu này, RSED framework đã tham khảo các giải thuật sinh giao thông mạng khác nhau [5] [6] [7] [8] và chọn lựa áp dụng phương pháp được trình bày trong bài báo M. S. Taqqu, W. Willinger, and R. Sherman, “*Proof of a fundamental result in self-similar traffic modeling.*” ACM/SIGCOMM Computer Communication Review, vol. 27, pp. 5-23, 1997 [5].

Trái ngược với giao thông mạng thông thường, giao thông mạng của một cuộc tấn công DDoS được cài đặt trực tiếp trong module zombieApp (đặc trưng cho máy tính bị chiếm quyền điều khiển và trực tiếp thực hiện tấn công). Hiện tại, Zombie mới chỉ sinh được giao thông mạng cho các cuộc tấn công DDoS bằng phương pháp TCP\SYN flood attack [16]. Hình 5b thể hiện cấu

trúc của module ZombieComputer cài đặt giải thuật sinh giao thông mạng của một cuộc tấn công DDoS.



a. General traffic generators b. Attack traffic generators

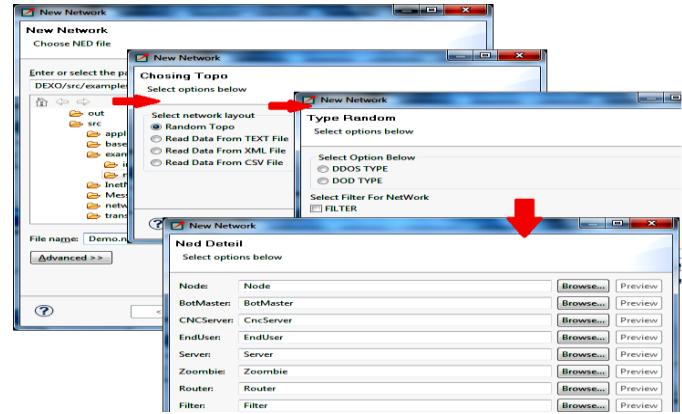
Hình 5: Bộ sinh giao thông mạng

3.3 Bộ sinh cấu hình mạng

Với mục đích xây dựng một bộ sinh cấu hình mạng có tính chất càng giống thực tế càng tốt, theo như Thomas Gamer và Michael Scharf trình bày trong [14], có hai phương pháp tiếp cận khả thi: dựa trên những quan sát thực tế ví dụ như quan sát dữ liệu của bảng định tuyến tại các router biên BGP [12] hoặc sử dụng một bộ sinh ngẫu nhiên như trong BRITE[14], INET [12], ReaSE [13] [14]. Phương pháp tiếp cận thứ nhất có thể sinh ra một cấu hình mạng rất giống với thực tế bằng cách trích xuất ra một phần nhỏ của mạng thực quan sát được tại một khoảng thời gian. Tuy nhiên trong phương pháp này yêu cầu phải thu thập một lượng lớn các dữ liệu, tiến hành các phân tích phức tạp để thu được một cấu hình mạng duy nhất phục vụ việc nghiên cứu (đôi khi chỉ là một bài giả lập đơn giản). Điều này là rất tốn kém tài nguyên cũng như công sức. Bên cạnh đó, do kết quả của quá trình thu thập dữ liệu thực tế chỉ là một cấu hình mạng rất lớn (3027 nút mạng ở mức độ Autonomous system từ năm 1997 đến 2000[12]). Vói mỗi lần giả lập, lại trích rút từ đó ra một cấu hình mạng con để phục vụ yêu cầu giả lập. Do đó, quá trình sinh cấu hình mạng khó có thể được tiến hành tự động và linh động cho từng trường hợp giả lập cụ thể. Phương pháp tiếp cận thứ hai đã ra đời và khắc phục được vấn đề này. Trong phương pháp này, cấu hình mạng được sinh ra một cách ngẫu nhiên có định hướng. Chi tiết hơn, cấu hình mạng được sinh ra dựa trên một số các luật, các tính chất của mạng Internet thực tế được nhiều nhà nghiên cứu bỏ công sức tìm hiểu[14]. Với cách thức như vậy, quá trình sinh cấu hình mạng hoàn toàn có thể tự động hóa (người cần cấu hình mạng chỉ cần nhập các tham số cho bộ sinh giả lập) và linh động đối với từng yêu cầu cấu hình khác nhau. Tuy nhiên, tính chất giống thực của phương pháp này còn phụ thuộc rất lớn vào độ chính xác của các quy luật, tính chất nền trên.

Hiện nay, có nhiều mô hình áp dụng khác nhau dựa trên những quy luật nói trên. Tuy nhiên trong nghiên cứu của mình, người viết chỉ tham khảo và áp dụng mô hình PFP (Positive Feedback Preference Model) do các tác giả S. Zhoua, G. Zhang, G. Zhang, và Z. Zhuge trình bày trong bài báo “*Towards a Precise and Complete Internet Topology Generator*” [13]. Trong mô hình này, cấu hình mạng được xây dựng bắt đầu từ một tập m_0 (m_0 là nhỏ) các nút sinh ra dưới dạng đồ thị ngẫu nhiên, cấu hình mạng sẽ được sinh ra dựa trên hai quy tắc: “*Interactive growth*” và

“*positive-feedback preference*”. Interactive growth là quy tắc dựa trên xác suất cho phép lựa chọn cách thức thêm các liên kết mới khi thêm một nút mới vào cấu hình mạng có sẵn bao gồm cả liên kết trong (liên kết giữa các nút trong cấu hình mạng) và liên kết ngoài (liên kết giữa một nút trong cấu hình mạng với nút mới được thêm vào). Còn positive-feedback preference là cũng quy tắc dựa trên xác suất nhưng cho phép xác định các nút sẽ được ưu tiên lựa chọn để thêm các liên kết new trên. Quá trình áp dụng hai quy tắc này được tiến hành lặp lại cho đến khi thêm đủ số lượng các nút mạng yêu cầu.



Hình 6: Giao diện sử dụng RSED wizards

Bên cạnh áp dụng một giải thuật sinh cấu hình mạng ngẫu nhiên, RSED wizards còn cung cấp cho người sử dụng một cách thức sử dụng lại những cấu hình mạng đã có trước đó được xây dựng dựa trên những phương pháp khác. Người dùng có thể xây dựng cấu hình mạng dựa trên dữ liệu được định nghĩa từ một file có sẵn. Hiện tại, trong bộ giả lập của mình, người viết chỉ hỗ trợ các định dạng file txt, xml và csv.

RSED wizards được thiết kế dựa trên cơ chế wizards, template của OMNET++ [1], do đó, thành phần giả lập này có khả năng mở rộng, thêm mới các tùy chọn cho phù hợp với yêu cầu đặc thù của từng lớp bài toán giả lập. Hình 6 là một số giao diện sử dụng của RSED wizards, cung cấp cho người dùng những tùy chọn khác nhau.

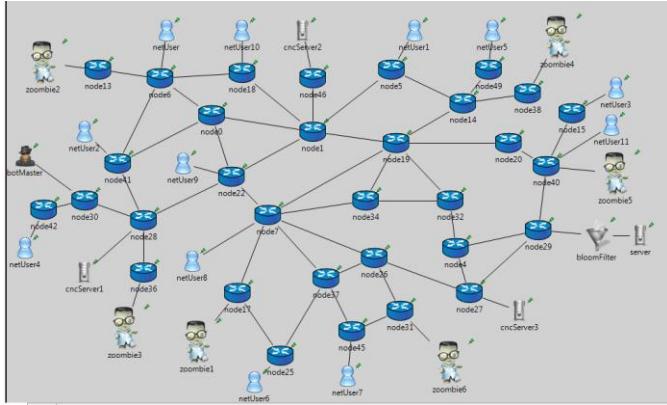
4 VÍ DỤ MINH HỌA (1)

Song song với quá trình phát triển các kiến trúc của RSED, trong nghiên cứu của mình, người viết cũng thực hiện ví dụ minh họa như là chứng minh về khả năng giả lập của môi trường mô phỏng được xây dựng trong nghiên cứu. Một trong những ví dụ đó là cài đặt giả lập bộ phòng chống và giảm thiểu tác hại của phương pháp tấn công TCP/SYN flood bằng phương pháp sử dụng BloomFilter. Chi tiết về các cuộc tấn công TCP/SYN flood và cách giảm thiểu này được trình bày trong [17] [18] [19]. Trong khuôn khổ báo cáo này, người viết xin chỉ đề cập đến mô hình ví dụ giả lập. Mô hình này giả lập 53 nút bao gồm 7 thành phần: BotMaster, CnCServer, Zombie, OSPF router, NetUser, Server và BloomFilter. Các thành phần này được kết hợp giả lập với nhau thông qua kịch bản.

- BotMaster điều khiển 3CnCserver và 6 zombie khác nhau tiến hành cuộc tấn công DDoS vào một Server. Loại tấn công là

TCP/SYN flood .

- Trước server tồn tại một bộ phòng chống và giảm thiểu tấn công BloomFilter.
- Trong mạng cũng tồn tại những người dùng thông thường NetUser để truyền các gói tin hợp lệ đến với server.



Hình 7: Mạng giả lập minh họa

Quá trình giả lập được thiết lập các tham số giống như được trình bày trong phần III.B của [18] với 1500 SYN/sec, thời gian thực hiện tấn công là 300s. Thời gian bắt đầu thực hiện tấn công tại giây thứ 200. Kết quả thực nghiệm nhận được trong hai file BloomFilter.vec và BloomFilter.sna. Phân tích thông tin thu thập được từ hai file này, cho thấy phương pháp phòng chống và hạn chế sử dụng bloom filter đã phát hiện sớm cuộc tấn công DDoS loại bỏ hầu hết các gói tín SYN không hợp lệ từ phía kẻ tấn công (trình bày cụ thể trong [17]).

5 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong nghiên cứu của mình, người thực hiện đã xây dựng, phát triển một bộ giả lập mạng hoàn chỉnh, thực tế phục vụ cho quá trình nghiên cứu tấn công từ chối dịch vụ DDoS và các phương pháp phòng chống, giảm thiểu tác hại của loại hình tấn công này. Nghiên cứu cũng tiếp cận tới các phương pháp giả lập cấu hình mạng và giao thông mạng giống với thực tế và lựa chọn ra một phương pháp để áp dụng phát triển trong bộ giả lập này. Bộ giả lập được xây dựng bằng ngôn ngữ C++ trên nền OMNET++. Điều này không chỉ giúp cho bộ giả lập có tính chất gần với thực tế mà còn có khả năng mở rộng nâng cấp linh động, phù hợp với yêu cầu cụ thể của người sử dụng.

Hiện tại, trong quá trình nghiên cứu, bộ giả lập mới chỉ lựa chọn, hỗ trợ một vài phương pháp tấn công, sinh cấu hình mạng hay sinh giao thông mạng điển hình. Trong tương lai, hướng phát triển của nghiên cứu này là có thể hỗ trợ giả lập nhiều hình thức tấn công cũng như phòng chống DDoS khác nhau, hỗ trợ nhiều hơn các phương pháp giả lập cấu hình mạng và giao thông mạng để người sử dụng có nhiều lựa chọn hơn phù hợp với nghiên cứu của mình.

6 LỜI TRI ÂN

Để thực hiện thành công đề án sinh viên nghiên cứu khoa học này người viết xin gửi lời cảm ơn chân thành đến thầy giáo TS.Nguyễn Khanh Văn, người đã theo sát, động viên và hướng

dẫn trong quá trình nghiên cứu, cũng như viết báo cáo này. Người viết cũng xin gửi lời cảm ơn đến cô giáo ThS. Nguyễn Phi Lê đã góp ý cho thiết kế cũng như giúp đỡ tìm kiếm tài liệu thực hiện nghiên cứu khoa học này.

7 TÀI LIỆU THAM KHẢO

- [1] “OMNET++”, Technical University of Budapest - Department of Telecommunications (BME-HIT), URL: <http://www.omnetpp.org/>, last visited 22/04/2011.
- [2] “INETframework”, URL: <http://inet.omnetpp.org/>, last visited 03/03/2011.
- [3] Thomas Gamer, Christoph P. Mayer, “ReaSE - Realistic Simulation Environments for OMNeT++”, Forschungsbereich Telematik Institut für Telematik Karlsruher Institut für Technologie (KIT), URL: <https://i72projekte.tum.de/trac/ReaSE/wiki/ReaSEFeatures>, last visited 22/04/2011.
- [4] Igor Kotenko and Alexander Ulanov, “Simulation of Internet DDoS Attacks and Defense”, Lecture Notes in Computer Science, 2006, Volume 4176/2006, 327-342, DOI: 10.1007/11836810_24.
- [5] S. Taqqu, W. Willinger, and R. Sherman, “Proof of a fundamental result in self-similar traffic modeling” ACM/SIGCOMM Computer Communication Review, vol. 27, pp. 5-23, 1997.
- [6] S. Avallone, D. Emma, A. Pescap, and G. Ventre. “A Practical Demonstration of Network Traffic Generation”. In Proc. of the 8th IMSA, pages 138{143, Aug. 2004.
- [7] M. E. Crovella and A. Bestavros, “Self-similarity in World Wide Web traffic: evidence and possible causes”, IEEE/ACM Transactions on Networking, 5(6):835{846, Dec. 1997.
- [8] I. Dietrich, “OMNeT++ Traffic Generator”, Sept. 2006, URL: <http://www7.informatik.uni-erlangen.de/~isabel/omnet/modules/TrafGen/>.
- [9] Jared Winick, “Net topology generator”, University of Michigan Technical, URL: <http://topology.eecs.umich.edu/inet/>, last visited 04/06/2002.
- [10] The Internet mapping project, URL: <http://cheswick.com/ches/map/>, last visited 04/05/2011.
- [11] Hamed Haddadi, Andrew Moore, Richard Mortier, Miguel Rio, and Gianluca Iannaccone, “End-to-End Network Topology Generation”, [pdf]: <http://www.ee.ucl.ac.uk/~hamed/docs/sc2k7hamed.pdf>
- [12] C. Jin, Q. Chen, and S. Jamin, “Inet: Internet topology generator”. Tech.rep.cse-tr-433-00, University of Michigan EECS Dept., 2000.
- [13] S. Zhou, G. Zhang, G. Zhang, and Z. Zhuge. “Towards a Precise and Complete Internet Topology Generator”. In Proc. of ICCCAS, volume 3, pages 1830{1834, June 2006.
- [14] Thomas Gamer, Michael Scharf, “Realistic Simulation Environments for IP-based Networks”, Dig. Proceedings of 1st International Workshop on OMNeT++, Marseille, France, Mar 2008.
- [15] Information Sciences Institute (University of Southern California, 4676 Admiralty Way, Marina del Rey, California 90291), “Transmission control protocol - Darpa internet program – protocol specification”, URL: <http://www.faqs.org/rfcs/rfc793.html>, September 1981.
- [16] EC-Council, “Ethical Hacking and Countermeasures version 6 - Module XIV Denial of Service”.
- [17] Changhua Sun, Jindou Fan, Lei Shi, Bin Liu, “A Novel Router-based Scheme to Mitigate SYN Flooding DDoS Attacks”, in Proc. IEEE INFOCOM (Poster), Anchorage, Alaska, USA, May 6-12, 2007.
- [18] Changhua Sun, Jindou Fan, Bin Liu, “A Robust Scheme to Detect SYN Flooding Attacks”, in Proc. International Conference on Communications and Networking in China (ChinaCom), Shanghai, China, August 22-24, 2007.
- [19] Basheer Al-Duwairi and G.Manimaran, “Intentional Dropping: A Novel Scheme for SYN Flooding Mitigation”, INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings, 23-29 April 2006, Barcelona, 0743-166X
- [20] J.Lawrence Carter and Mark N.Wegman, “Universal Classes of Hash Functions”. Proceedings of Ninth Annual SIGACT Conference May ,1977.
- [21] M. V. Ramakrishna, E. Fu, and E. Bahcekapili, “Efficient hardware hashing functions for high performance computers”, IEEE Transactions on Computers, vol. 46, no. 12, pp. 1378–1381, 1997

Ứng dụng công nghệ GPS, GIS xây dựng hệ thống theo dõi và quản lý xe buýt Hà Nội

Vũ Ngọc Thành

Tóm tắt - Hệ thống theo dõi và quản lý xe buýt Hà Nội tích hợp công nghệ GPS, GIS cho phép nhà quản lý theo dõi trực quan lịch trình, hoạt động của xe buýt trên bản đồ theo thời gian thực. Hơn nữa, đây cũng là 1 công thông tin hữu dụng dành cho người sử dụng và các bên liên quan với 1 số chức năng : tra cứu thông tin tuyến buýt, bãi xe, tìm đường. Ngoài ra, Hệ thống còn hỗ trợ đầy đủ nhà quản lý trong việc quản lý hệ thống, báo cáo.

Từ khóa - Buýt Hà Nội, GPS, GIS, quản lý.

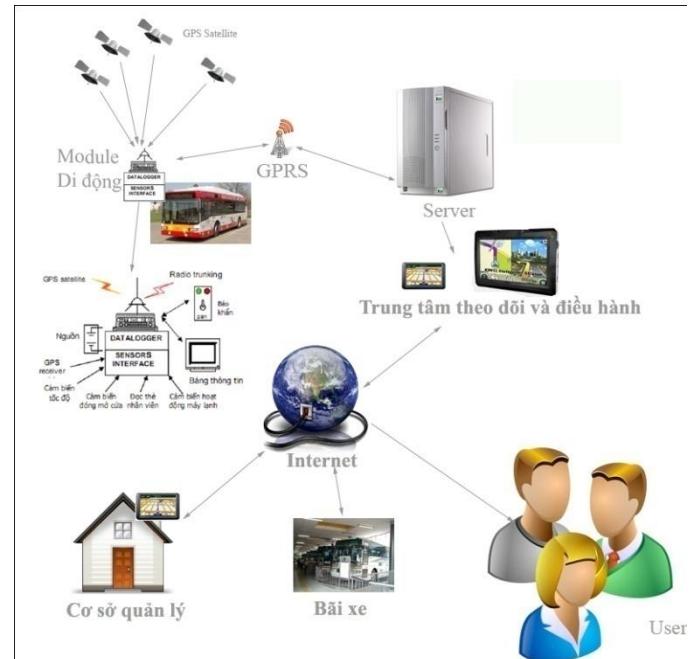
1. GIỚI THIỆU

GPS, GIS là những công nghệ không mới. Tuy nhiên, chỉ trong vài năm gần đây những ứng dụng của nó mới thực sự nở rộ ở Việt Nam. Có thể kể đến là các hệ thống theo dõi xe tải, xe taxi, các dịch vụ dựa vị trí của nhà mạng, và cả 1 loạt các bản đồ trực tuyến ra đời. Một ưu điểm nổi bật của các ứng dụng dựa trên nền tảng GPS, GIS là việc quản lý trực quan vị trí hiện tại của các đối tượng cần theo dõi trong hệ thống. Cùng với hệ thống thông tin sẽ giúp cho việc quản lý hết sức tiện lợi, giảm sức người, súc của.

Trong thực tế, với nhu cầu đi lại ngày càng tăng, xe buýt trở thành phương tiện phổ biến trong cuộc sống của người dân thủ đô. Tuy nhiên, cũng vì thế mà hệ thống quản lý cũ cũ trở nên quá tải, đặc biệt vào các dịp lễ tết. Điều này khiến chất lượng dịch vụ giảm sút, gây nhiều phiền hà với người đi xe. Hơn nữa, hệ thống cũng phải sử dụng rất nhiều nhân viên ghi chép, ghi nhật ký, lịch trình, .. những công việc mà hoàn toàn có thể thực hiện bởi máy tính một cách tự động.

Với mục đích muốn thay đổi, cải thiện, nâng cao hệ thống quản lý xe buýt Hà Nội. Đề tài mạnh dạn đưa ra mô hình quản lý, và xây dựng hệ thống demo áp dụng triệt để nền tảng GPS, GIS.

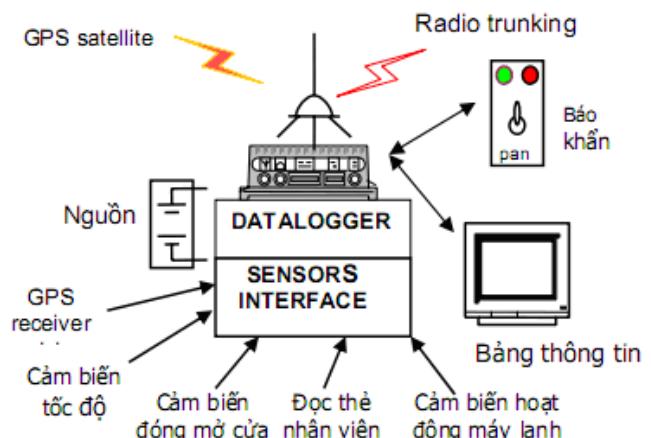
2. MÔ HÌNH THEO ĐỔI QUẢN LÝ



Hình 1. Mô hình theo dõi và quản lý xe buýt

2.1 Module di động

Theo mô hình trên thì trên mỗi xe buýt sẽ được gắn 1 module di động gồm các thành phần khác nhau :



Hình 2. Module di động

- Chip GPS Receiver : Bắt tín hiệu từ GPS Satellite để lấy các thông tin tọa độ hiện tại của xe.
- Cảm biến tốc độ : Lấy thông tin về tốc độ của xe.
- Cảm biến mở cửa.

Công trình này được thực hiện bởi :

Vũ Ngọc Thành, sinh viên lớp Công nghệ phần mềm, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0944-222-012, e-mail: lampard_hut@yahoo.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

- Bộ phận đọc thẻ nhân viên.
- Cảm biến hoạt động của máy lạnh.
- Bảng thông tin : Hiển thị các thông số của xe, thông báo, lịch trình, sơ đồ đường đi, hoặc hiển thị các thông báo cho khách.
- Thiết bị báo khẩn khi có sự cố tới trung tâm điều hành.

Chức năng của module di động :

- Cung cấp các thông tin cho hệ thống : tọa độ(kinh độ, vĩ độ), hướng di chuyển của xe, thông số vận tải của xe.
- Cung cấp các thông tin về tài xế, phụ xe, tình trạng hoạt động của xe(dừng trả khách đúng qui định, mở cửa xe, điều hòa...).
- Cung cấp cho khách hàng lộ trình di chuyển, thông tin về bến đỗ sắp tới, giá vé, các dịch vụ của hệ thống.

2.2 Module kết nối GPSTracker - Server



Hình 2. Module kết nối GPS Tracker - Server

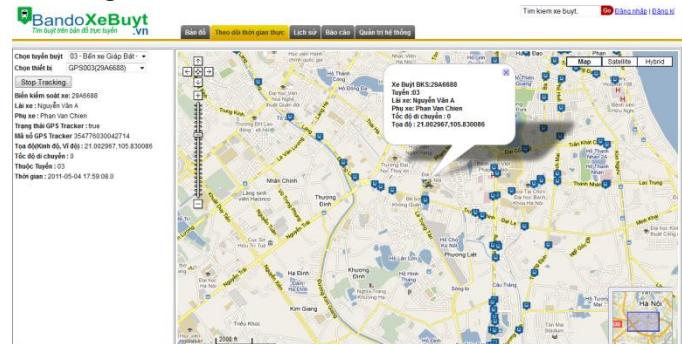
Với mỗi GPS Tracker thì đều được gắn 1 sim card điện thoại. Để có thể truyền dữ liệu lên Server, chúng phải được cấu hình[1] dựa trên sim điện thoại đó. Sau khi thiết bị đã được cấu hình xong, thì cứ sau mỗi khoảng thời gian t giây. dữ liệu sẽ được đẩy 1 cách liên tục lên Server vào một cổng nào đó (đã được mở) thông qua mạng GPRS hoặc 3G. Tùy theo từng thiết bị GPS Tracker khác nhau mà format dữ liệu truyền lên khác nhau. Chính vì thế việc tạo các giao thức nhận dữ liệu trên Server là hết sức quan trọng. Từ đó, Server sẽ lọc, và format lại theo dữ liệu cần hiển thị và cuối cùng đẩy vào cơ sở dữ liệu.

2.3 Server

Server của hệ thống được triển khai trên môi trường Linux hoặc Window. Tuy nhiên, máy đó phải có ip tĩnh. Đây là điều bắt buộc để có thể bắt được dữ liệu từ GPS Tracker đẩy lên.

Nhiệm vụ của Server là xử lý các thông tin nhận được, hiển thị trên nền bản đồ số. Ngoài ra còn thực hiện các công tác nghiệp vụ : Tra cứu, tìm kiếm, xử lý việc theo dõi , ..của hệ thống.

2.4 Trung tâm theo dõi và điều hành



Hình 3. Trung tâm theo dõi và điều hành

Trung tâm theo dõi và điều hành là nơi theo dõi trực tiếp hoạt động vận tải của xe buýt dựa trên nền bản đồ GIS. Cơ sở dữ liệu GIS được tổ chức, lưu trữ, và quản lý trong một hệ cơ sở dữ liệu gồm các thành phần :

- Không gian : sử dụng nền bản đồ địa hình với tỉ lệ 1/1000 tạo các lớp chuyên đề thể hiện : tuyến xe buýt, trạm dừng, bãi xe, trung tâm quản lý.
- Thuộc tính :
 - Hoạt động của tuyến xe gồm : đơn vị quản lý, các loại vé, thời gian bắt đầu, thời gian kết thúc, khoảng cách thời gian giữa 2 xe trong giờ cao điểm, và lúc bình thường. Ngoài ra là các thông tin lộ trình.
 - Thông tin đặc điểm của xe : loại xe, biển kiểm soát, ngày sản xuất, ..
 - Nhân sự vận hành hệ thống : lái xe, phụ xe, ..

Các dữ liệu hoạt động của xe buýt được trung tâm thu thập tự động từ hộp đen của xe buýt, từ thiết bị GPS Receiver, thiết bị cảm biến, ...sau đó được lưu trữ vào cơ sở dữ liệu của hệ thống phục vụ cho việc hiển thị và công tác báo cáo.

Ngoài việc theo dõi hoạt động của xe, thì trung tâm còn có nhiệm vụ cảnh báo tốc độ với các xe đi quá tốc độ, lập báo cáo các sai phạm trong quá trình vận hành và làm báo cáo theo tháng/quý/năm về tình hình hoạt động của hệ thống, lưu giữ hành trình từng xe.

2.5 Cơ sở quản lý, bãi xe

Nhiệm vụ chính của cơ sở quản lý và bãi xe là truy vấn các thông tin về các đối tượng cũng như tình trạng hoạt động của xe buýt trực thuộc cơ sở.

Cập nhật dữ liệu hoạt động vận tải cấp cơ sở, theo chu kỳ ngày, tuần, tháng, năm.

Lập báo cáo, bảng biểu, thống kê, về việc bảo dưỡng phương tiện vận chuyển : Xe buýt trực thuộc cơ sở.

Xây dựng các bên đợi thông minh : Là các bảng điện tử. Ngoài các thông tin về tuyến xe đi qua còn thời gian tương đối xe sắp đến để người đi xe có thể chủ động trong việc bắt xe của mình.

2.6 Người dùng

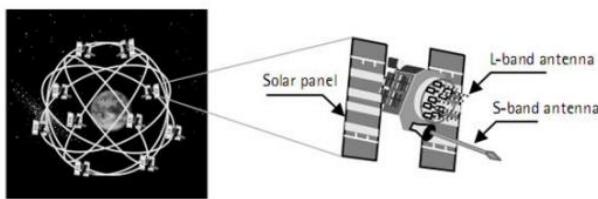
Người dùng hệ thống có thể tra cứu các thông tin liên quan tới xe buýt : tuyến buýt, bãi xe, thời gian hoạt động, lịch

trình cù thể từ tuyến, tìm đường đi trên xe buýt...
Ngoài ra người dùng có thể mua vé tháng trực tuyến qua hệ thống bán trực tuyến của trung tâm quản lý.

3. GIS, GPS VÀ XÂY DỰNG BẢN ĐỒ SỐ

3.1 Lý thuyết về GPS, GIS

GPS(Global Positioning System) gồm một chòm sao 24 vệ tinh. Để đảm bảo hệ thống vệ tinh này phủ khắp toàn bộ trái đất một cách liên tục, những vệ tinh này được sắp xếp sao cho mỗi 4 vệ tinh được đặt trong mỗi 6 mặt phẳng quỹ đạo. Như vậy sẽ có 4 đến 10 vệ tinh sẽ hiện hữu tại bất kỳ nơi đâu trên trái đất nếu góc ngang khoảng 10 độ. Để cung cấp cho việc định vị ta chỉ cần 4 trong số đó.



Hình 4. Hệ thống vệ tinh GPS Satellite

Hoạt động cơ bản của GPS :

- Các vệ tinh GPS bay vòng quanh trái đất hai lần trong một ngày theo một quỹ đạo rất chính xác và phát tín hiệu xuống trái đất. Các máy thu GPS nhận thông tin này và bằng phép tính lượng giác sẽ tính được chính xác vị trí người dùng. Về bản chất thì GPS so sánh thời gian tín hiệu phát đi từ vệ tinh với thời gian tín hiệu nhận được ở bộ thu. Sai lệch về thời gian cho biết máy thu GPS đang ở cách vệ tinh bao xa.
- Máy thu GPS phải khóa được với tín hiệu của ít nhất 3 vệ tinh để tính ra được kinh độ, vĩ độ. Với 4 hay nhiều hơn thì có thể tính được : kinh độ, vĩ độ, độ cao. Ngoài ra máy thu có thể tính thêm các thông tin khác như : tốc độ, hướng chuyển động, khoảng cách hành trình...

GIS(Geographic Information System) - Hệ thống thông tin địa lý là hệ thống quản lý, phân tích và hiển thị tri thức địa lý, tri thức này được hiển thị qua các tập thông tin :

- Các bản đồ : giao diện trực tuyến với dữ liệu địa lý để tra cứu, trình bày kết quả, và sử dụng như 1 nền thao tác với thế giới thực.
- Các tập thông tin địa lý : thông tin địa lý dạng file, dạng cơ sở dữ liệu gồm các yếu tố, mạng lưới, topology, địa hình, thuộc tính.
- Các mô hình xử lý : tập hợp các quy trình xử lý để phân tích tự động.
- Các mô hình dữ liệu : GIS cung cấp công cụ mạnh hơn cơ sở dữ liệu thông thường bao gồm quy tắc và sự kiện toàn bộ giống các hệ thống tin khác. Trong đó lược đồ, quy tắc, và sự kiện đóng vai trò quan trọng.
- Metadata : tài liệu miêu tả dữ liệu , cho phép người sử dụng tổ chức, tìm hiểu, và truy nhập tới tri thức địa lý.

Các cách nhìn với hệ thống GIS :

- Cơ sở dữ liệu địa lý (Geodatabase) : GIS là một cơ sở dữ liệu không gian truyền tải thông tin địa lý theo quan điểm gốc của mô hình dữ liệu(yếu tố, topology, mạng lưới, raster...).
- Hiện tượng hóa(Geovisualization) : GIS là tập các bản đồ thông minh thể hiện các yếu tố và quan hệ giữa các yếu tố trên mặt đất.
- Xử lý(Geoprocessing) : GIS là công cụ xử lý thông tin cho phép tạo ra các thông tin mới từ thông tin đã có. Chức năng xử lý thông tin được lấy từ các thông tin đã có, áp dụng các chức năng phân tích và ghi kết quả mới.

Cơ sở dữ liệu địa lý : Hệ thống GIS sử dụng cơ sở dữ liệu địa lý với các thành phần, bao gồm :

- Tập hợp các dữ liệu dạng Vector(điểm, đường, vùng).
- Tập hợp dữ liệu dạng Raster(dạng mô hình DEM hoặc ảnh)
- Tập hợp dữ liệu dạng lưới(ví dụ : đường giao thông, lưới cấp thoát nước, lưới điện).
- Tập hợp dữ liệu địa hình 3 chiều và bề mặt khác.
- Dữ liệu đo đạc
- Dữ liệu dạng địa chỉ.

3.2 Xây dựng bản đồ nền dành cho hệ thống quản lý xe buýt sử dụng nền bản đồ Google Map.

Việc xây dựng bản đồ nền , và bộ công cụ tích hợp thao tác trên nó thực sự là một công việc hết sức khó khăn mà không phải một cá nhân nào có thể làm được và trong một thời gian ngắn. Chưa nói đến việc triển khai các dịch vụ, tính năng trong môi trường trực tuyến .

Do đó, lựa chọn bản đồ nền trực tuyến có sẵn là ưu tiên số một. Hiện nay có khá nhiều bản đồ trực tuyến : google map, diadiem, vietbando, yahoo,... Tuy nhiên trong hệ thống này chúng ta tích hợp với Google Map bởi 1 số lý do sau :

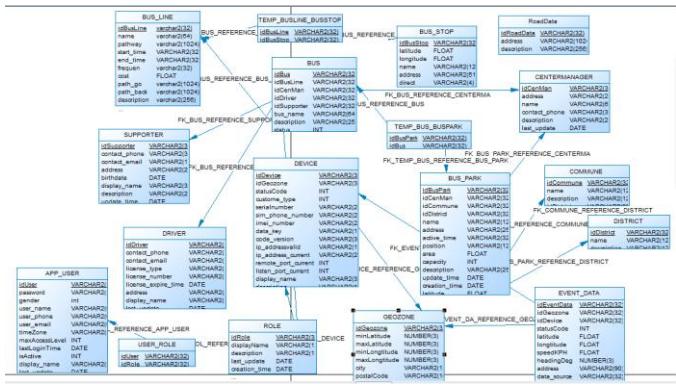
- Google map là bản đồ trực tuyến phổ biến nhất hiện nay.
- Việc tích hợp và sử dụng bản đồ sử dụng các API là không khó.
- Kế thừa các framework có sẵn tích hợp thao tác với googleMap cho phép ta triển khai dễ dàng hơn các ứng dụng mang tính nghiệp vụ của mình.

Để xây dựng bản đồ nền dựa trên google map, công việc đầu tiên là sử dụng GPS Tracker ,đi đến từng điểm trạm dừng xe buýt ,lấy được tọa độ rồi ghi lại. Bằng cách này chúng ta sẽ có được 1 bản đồ chính xác nhất về hệ thống đường đi của xe buýt hiện tại.

4. THIẾT KẾ HỆ THỐNG VÀ CÀI ĐẶT

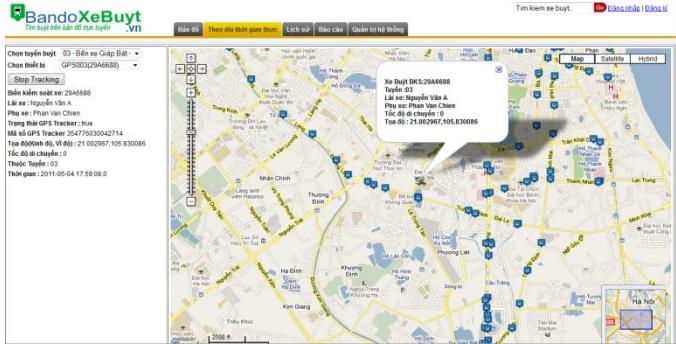
4.1. Xây dựng cơ sở dữ liệu hệ thống

CSDL của hệ thống là 1 tập hợp các đối tượng mô tả thông tin xe buýt, thiết bị, tuyến đường, bãi xe, người sử dụng...



Hình 5. Thiết kế cơ sở dữ liệu hệ thống

4.2. Xây dựng hệ thống theo dõi thời gian thực



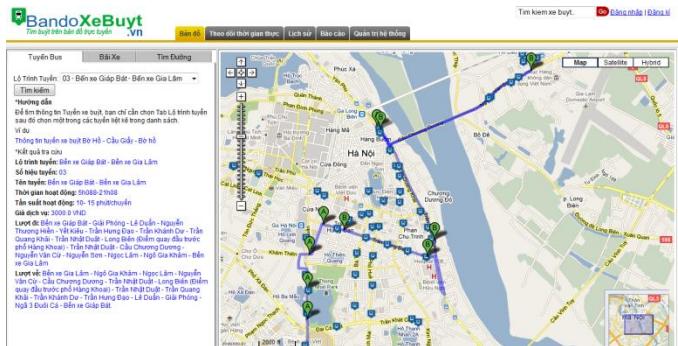
Hình 6. Chức năng theo dõi thời gian thực

Dựa vào cấu hình truyền nhận từ GPS Tracker - Server. Cứ $t(s)$ GPS Tracker sẽ gửi dữ liệu lên Server theo 1 cổng nào đó, ví dụ : Server có địa chỉ IP : 210.211.77.11:xxxx. Trên Server, chúng ta cài đặt 1 tiến trình liên tục đọc dữ liệu từ cổng xxxx, sau đó lọc, format và insert dữ liệu vào DataBase.

Ở một mặt khác, khi người quản trị bắt đầu "Start Tracking" (sau khi đã chọn thiết bị/xe cần theo dõi), nó sẽ liên tục truy cập vào DB để lấy dữ liệu mới nhất : các thông tin bao gồm : tọa độ(kinh độ, vĩ độ, thời gian, thông số thiết bị, thông tin về lái xe, phụ xe, ...) Dữ liệu lấy về liên tục được hiển thị lên bản đồ.

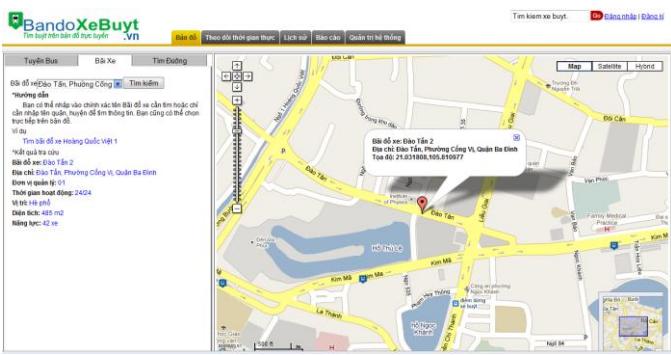
Trong thực tế thì cứ khoảng 10s chúng ta sẽ update vị trí của xe một lần vì ngoài thời gian truyền dữ liệu, server còn mất thời gian phân tích, dística dữ liệu vào Database, và thời gian lấy dữ liệu đó ra.

4.3. Xây dựng các chức năng tra cứu, tìm đường, lịch sử

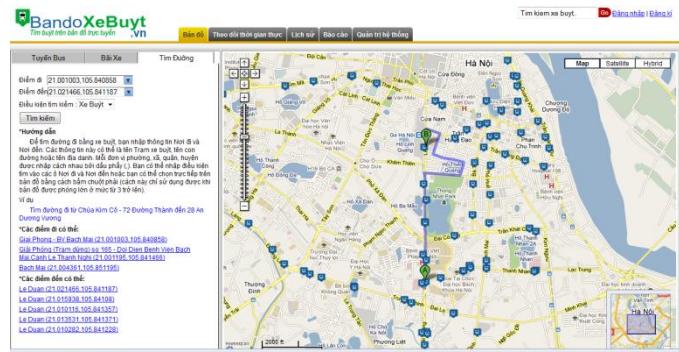


Hình 7. Chức năng tra cứu tuyến buýt

Chức năng tra cứu bãi xe trên bản đồ



Hình 8. Chức năng tra cứu bãi xe

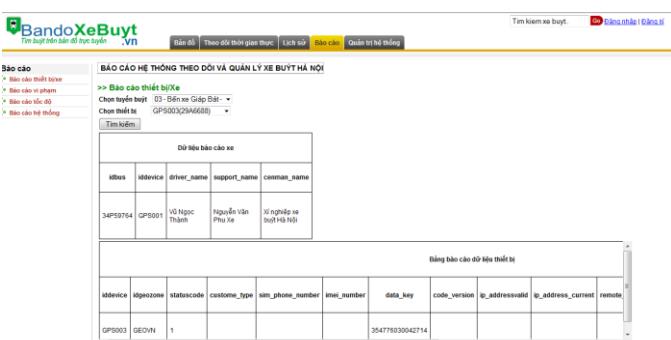


Hình 9. Chức năng tìm đường



Hình 10. Chức năng xem lịch sử đường đi

4.4 Báo cáo hệ thống



Hình 11. Báo cáo hệ thống

4.4. Các chức năng nghiệp vụ hệ thống

The screenshot shows a web-based vehicle management system. At the top, there's a navigation bar with links like 'Tìm kiếm xe buýt', 'Bán vé', 'Thống kê thời gian thực', 'Lịch sự', 'Báo cáo', and 'Quản trị hệ thống'. The main content area is titled 'QUẢN TRỊ HỆ THỐNG THEO DỜI VÀ QUẢN LÝ XE BUýt' and shows a form for managing routes. The form includes fields for 'Mã số tuyến(*)', 'Tên tuyến(*)', 'Bến xe Giáp Bát - Bến xe...', 'Thời gian bắt đầu(*)', 'Thời gian kết thúc(*)', 'Tần suất(*)', 'Giá vé(*)', 'Chiều dài(*)', and 'Chiều rộng(*)'. Below these are dropdown menus for 'Bến xe Giáp Bát - Bến xe...', 'Giá vé', and 'Tần suất'. At the bottom of the form are buttons for 'Mở tệp tin ảnh' (Open file), 'Cập nhật' (Update), 'Thêm tuyến' (Add route), and 'Xóa tuyến' (Delete route).

Hình 12. Quản trị hệ thống

5. PHỤ LỤC

5.1 [1] Cấu hình truyền dữ liệu GPS Track - Server

Chuẩn bị :

- 1 Sim card (ví dụ 1 sim Viettel : 0944.111.222)
- 1 Server với địa chỉ IP tĩnh (ví dụ : 192.168.9.1, port :6688).
- Qui trình cấu hình như sau :
- "format123456"
- "begin123456"
- "adminip123456 192.168.9.1 6688"
- "apn123456 v-internet"
- "time zone123456 7"
- "monitor123456"
- "t060s***n1234567"
- Tất cả cú pháp trên 0944.111.222

6. LỜI TRI ÂN

Tôi xin gửi lời cảm ơn sâu sắc

- Thầy - Th.s. - GVC Lương Mạnh Bá đã tận tình giúp đỡ, chỉ bảo, và góp ý trong suốt quá trình thực hiện đề tài này.
- Anh Trịnh Văn Tú - Giám đốc phòng VAS - công ty viễn thông và dịch vụ phần mềm TELSOFT đã nhiệt tình giúp đỡ về mặt cơ sở vật chất và thiết bị.
- Anh Trần Văn Toàn - Phòng GIS - viện khoa học nông nghiệp Việt Nam.

9. TÀI LIỆU THAM KHẢO

- [1] Phạm Duy Thành, đồ án tốt nghiệp đại học - lớp INPG-04, Đại học bách khoa Hà Nội.
- [2] Open GTS - Open GTS Tracking System.<http://opengts.sourceforge.net>
- [3] Framework GMap4JSF.
- [4] JSF In Action - Kito Mann.

Xây dựng hệ mờ nhận dạng biển số xe

Đoàn Hồng Quân

Tóm tắt - Nhận dạng biển số xe ô tô tự động (Automatic License Plate Recognition) được ứng dụng trong rất nhiều hệ thống giao thông (bãi đỗ xe tự động, xác định xe vi phạm luật giao thông, tìm kiếm xe mất cắp...). Trong bài báo này, một thuật toán đơn giản và hiệu quả được trình bày để xây dựng một hệ thống nhận dạng biển số xe. Thuật toán đề xuất bao gồm 3 phần chính: Tách vùng biển số, trích các ký tự và nhận dạng các ký tự biển số. Đôi với tách vùng biển số, áp dụng giải thuật phát hiện biên (edge) và giải thuật làm nhòe (smearing). Trong phần phân đoạn, giải thuật lọc và một số phép biến đổi hình thái (morphological) được sử dụng. Cuối cùng, các ký tự biển số kết quả của quá trình phân đoạn được trích chọn đặc trưng và đưa vào mô hình mờ để nhận dạng. Hiệu quả của thuật toán đã được thử nghiệm qua các hình ảnh thực tế. Dựa trên kết quả thực nghiệm, thuật toán đề xuất ước đạt độ chính xác nhận dạng trên 60% và vẫn tiếp tục được cải tiến.

Từ khóa - Nhận dạng ký tự, nhận dạng biển số xe, mô hình hóa mờ, hệ mờ nhận dạng biển số xe.

1. GIỚI THIỆU

Cùng với sự phát triển kinh tế, gia tăng dân số và nhu cầu đi lại, số lượng các phương tiện tham gia giao thông nói chung và ô tô nói riêng xuất hiện ngày càng nhiều. Điều này đặt ra một yêu cầu lớn trong việc kiểm soát và quản lý loại phương tiện này. Thực trạng tại các điểm trông giữ ô tô ở các đô thị lớn tại Việt Nam hiện nay cho thấy một vấn đề bất cập: việc xử lý thủ công (ghi biển số xe) gây tốn kém tiền thuê nhân công và không hiệu quả, nhất là ở những bãi đậu xe lớn. Yêu cầu đặt ra là cần xây dựng các hệ thống tự động. Một trong những hệ thống tự động như vậy là hệ thống tự động nhận dạng biển số xe, đặc trưng của nó là có khả năng thu nhận hình ảnh cũng như "đọc" và "hiểu" các biển số xe một cách tự động.

Một hệ thống như vậy có thể được sử dụng trong rất nhiều ứng dụng như: trạm cân và rửa xe tự động, bãi giữ xe tự động, các hệ thống kiểm soát lưu lượng giao thông hay trong các ứng dụng về an ninh như tìm kiếm xe mất cắp...

Với tính thực tiễn của mình, nhận dạng biển số xe tự động đã được nghiên cứu rộng rãi với nhiều cách tiếp cận khác nhau. Có thể kể tên một vài công trình nghiên cứu mà người viết dùng để tham khảo như sau. Lotufo, Morgan và

Đoàn Hồng Quân, sinh viên lớp Hệ thống thông tin, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0167-331-7934, e-mail: hongquan2512@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

Johnson[3] đề xuất nhận dạng biển số xe sử dụng kỹ thuật nhận dạng ký tự quang học. Johnson và Bird[5] đề xuất phương pháp phân tích dựa trên đặc trưng của biển số và so khớp mẫu cho việc nhận dạng biển số xe tự động. Fahmy[6] đề xuất mạng nơron nhân tạo hai chiều nhớ trong đọc biển số. Đó là cách tiếp cận cho một số lượng nhỏ các mẫu. Nijhuis, Ter Brugge, Helmholz J.P.W. Pluim, L. Spaanenburg, R.S. Venema and M.A. Westen đề xuất dùng logic mờ và mạng nơron nhân tạo để nhận dạng biển số. Phương pháp này sử dụng logic mờ cho việc phân đoạn và mạng nơron nhân tạo cho trích chọn đặc trưng...

Trong nghiên cứu này, thuật toán đề xuất dựa trên việc tách vùng biển số, trích các ký tự biển số và nhận dạng ký tự.

Bài báo được tổ chức theo cấu trúc sau: Chương 2 cung cấp một cái nhìn tổng quan về hệ thống. Tách vùng biển số được trình bày trong Chương 3. Chương 4 đề cập đến việc trích ký tự vùng biển số. Chương 5 giới thiệu kỹ thuật trích chọn đặc trưng từ các ảnh ký tự. Chương 6 trình bày phương pháp nhận dạng ký tự dựa trên mô hình hóa mờ. Chương 7 đề cập đến kết quả thực nghiệm và Chương 8 là phần kết luận.

2. CẤU TRÚC CỦA MỘT HỆ THỐNG NHẬN DẠNG BIỂN SỐ XE TỰ ĐỘNG

Thuật toán đề xuất trong bài báo này được thiết kế để nhận dạng biển số xe (ô tô) tự động. Đầu vào của hệ thống là ảnh xe chụp bởi camera. Ảnh chụp này sẽ được xử lý thông qua việc tách vùng ảnh chứa biển số từ ảnh xe. Vùng biển số này sẽ là đầu vào cho chức năng phân đoạn để trích các ký tự biển số riêng lẻ. Tiếp sau các ký tự này sẽ được trích chọn đặc trưng. Các đặc trưng ký tự này sẽ là đầu vào (các biến mờ) của hệ mờ nhận dạng. Đầu ra cuối cùng của hệ thống là các biển số xe.

3. TÁCH VÙNG BIỂN SỐ

Tách vùng biển số là bước đầu tiên trong thuật toán này. Ảnh thu được ban đầu từ camera sẽ được lọc nhiễu (lọc trung vị), sau đó chuyển sang dạng nhị phân chỉ gồm các bit 0 và 1 (chỉ màu đen và trắng) bởi 1 giá trị ngưỡng. Các điểm ảnh kém sáng hơn ngưỡng sẽ nhận giá trị 0 (đen) và những điểm còn lại có giá trị 1 (trắng). Hình ảnh thu được từ camera (ảnh gốc) và ảnh nhị phân được thể hiện dưới hình 1(a) và 1(b) tương ứng.



(a) Ảnh gốc



(b) Ảnh nhị phân

Hình 1 (a) Ảnh gốc – (b) Ảnh nhị phân

Ảnh nhị phân sau đó được xử lý bằng một số phương pháp. Để tìm vùng biển số, đầu tiên áp dụng giải thuật làm nhòa ảnh. Làm nhòa là phương pháp để tách ra vùng text từ một ảnh hỗn hợp. Với giải thuật làm nhòa, ảnh được xử lý theo chiều dọc và chiều ngang (quét dòng). Nếu số lượng các pixel trắng là nhỏ hơn một ngưỡng cho trước hoặc lớn hơn bất kỳ một ngưỡng cho trước nào khác, thì các pixel trắng này sẽ được chuyển thành màu đen. Trong hệ thống này, giá trị ngưỡng được chọn là 10 và 100 cho cả làm nhòa theo chiều ngang và chiều dọc.

If số lượng các pixel 'trắng' < 10; các pixel này chuyển thành 'đen'

Else: không làm gì

If số lượng các pixel 'trắng' > 100; các pixel này chuyển thành 'đen'

Else: không làm gì

Sau khi làm nhòa, một thao tác hình thái học, làm trương ảnh (dilation) sẽ được áp dụng cho ảnh để xác định vị trí biển. Tuy nhiên, có thể có nhiều hơn một ứng viên cho khu vực vị trí biển số. Để tìm được vùng chính xác và loại bỏ các vùng còn lại, một số hàm lượng giá sẽ được áp dụng. Ở đây dùng số lượng pixel trong từng vùng ứng viên làm căn cứ. Ảnh thu được sau bước xử lý này được trình bày trong hình 2 (a) và ảnh biển số trong ảnh ban đầu được đánh dấu bằng vùng biên xanh cho trong hình 2 (b)



(a) Vùng biển số



(b) Vùng biển số được đánh dấu trên ảnh gốc

Hình 2 (a) Vùng biển số - (b) Vùng biển số được đánh dấu trên ảnh gốc

Sau khi xác định được vị trí vùng biển số, vùng này sẽ được cắt ra khỏi ảnh và thu được ảnh biển số như Hình 3



Hình 3 - Ảnh biển số

4. TRÍCH KÝ TỰ VÙNG BIỂN SỐ

Trong trích vùng biển số, ảnh biển số sẽ được chia nhỏ thành các phần tử cấu thành nó là các ký tự biển số riêng lẻ. Đầu tiên ảnh được lọc để nâng cao chất lượng ảnh và loại bỏ nhiễu cùng với các thành phần không mong muốn. Bước tiếp theo là nhị phân ảnh theo ngưỡng. Tiếp theo sẽ tiến hành đảo bit để đảm bảo rằng các ký tự sẽ là màu trắng (giá trị pixel = 1). Các vùng ảnh có dung lượng nhỏ hơn 50 pixel sẽ bị loại (để bỏ đi ký tự - trong biển số). Cuối cùng đánh dấu các "đoạn" ký tự dựa vào điểm bắt đầu và kết thúc của vùng bit 1 (theo cả chiều ngang và chiều dọc). Các ký tự đánh dấu sẽ được tách ra khỏi ảnh sau khi đưa về cùng kích thước chuẩn là 42×24 pixels. Kết quả thu được từ việc trích các ký tự biển số từ hình 3 được cho trong hình bên dưới



Hình 4 – Các ký tự đơn lẻ thu được sau quá trình trích ký tự

5. TRÍCH CHỌN ĐẶC TRƯNG

Kỹ thuật trích chọn đặc trưng của nghiên cứu này dựa trên sự phân tích của Jesse Hansen trong **Error! Reference source not found.**

Các đặc trưng hữu ích trong quá trình nhận dạng được trích chọn dựa vào sự phân tích hình thái và moment 2-D của ký tự:

$$m_{pq} = \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} x^p y^q f(x, y)$$

Từ các moment trên, chúng ta có thể tính các đặc trưng như:

1. Tổng số pixel (số lượng pixel trên một ký tự ảnh nhị phân)

2. Trọng tâm (trung tâm của vùng ảnh ký tự)

3. Các tham số Elliptical:

- Eccentricity (tỷ lệ của trục lớn trên trục nhỏ)
- Orientation (góc của trục chính)

4. Độ đo Skewness

5. Độ đo Kurtosis

6. Các moment mức cao hơn.

7. Extend (khoảng rộng của ảnh)

8. Perimeter (Chu vi vùng ảnh)

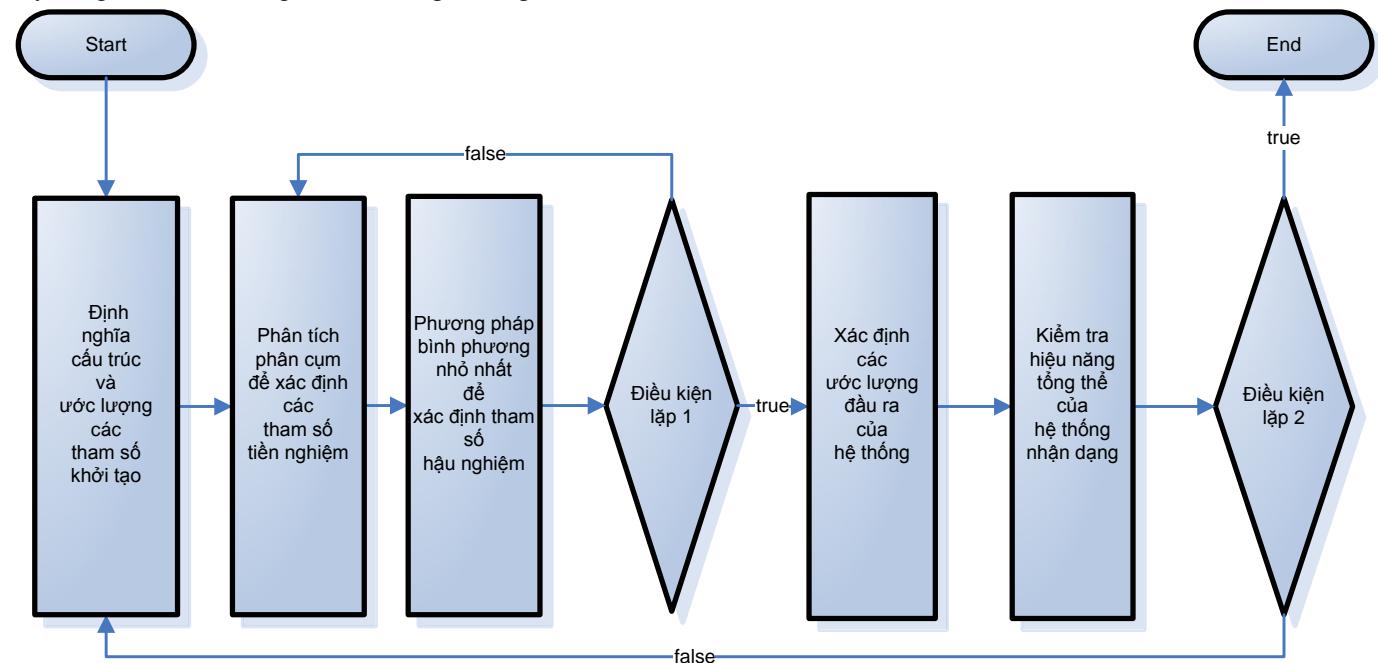
Cuối cùng, mỗi ký tự biến số sẽ được đại diện bởi 9 đặc trưng.

Hình 5 minh họa các đặc trưng thu được từ sự phân tích ký tự A

	0.869923	-88.296029	0.512897
	-0.010447	0.005537	0.003899
	0.003983	131.941125	7.000000

Hình 5 – Các đặc trưng trích chọn từ ký tự A

Các đặc trưng này sẽ được sử dụng làm đầu vào cho việc xây dựng hệ mờ sẽ được giới thiệu trong chương 6.



Hình 6 – Sơ đồ cho quá trình nhận dạng hệ thống (sinh luật mờ và tối ưu hóa tham số).

6. NHẬN DẠNG KÝ TỰ

6.1. Cơ sở lý thuyết

Kỹ thuật được sử dụng để nhận dạng ký tự là kỹ thuật mô hình hóa mờ.

Các bước chính để xây dựng một hệ mờ bao gồm: mờ hóa (định nghĩa biến mờ và hàm thuộc), sinh luật, suy diễn mờ và khử mờ.

Quá trình mờ hóa và sinh luật trình bày trong bài báo này dựa trên kỹ thuật trong [1].

Mờ hóa:

Có 13 đầu vào (tương ứng với 13 đặc trưng ảnh ký tự) và 1 đầu ra. Vì thế mô hình sẽ có 13 biến mờ đầu vào và 1 biến mờ đầu ra.

Hàm thuộc của các biến mờ vào ra có dạng tam giác cân.

$$\mu = \begin{cases} 1 - \frac{|x - a|}{\delta a}, & \text{if } x \in [a - \delta a, a + \delta a] \\ 0, & \text{otherwise} \end{cases}$$

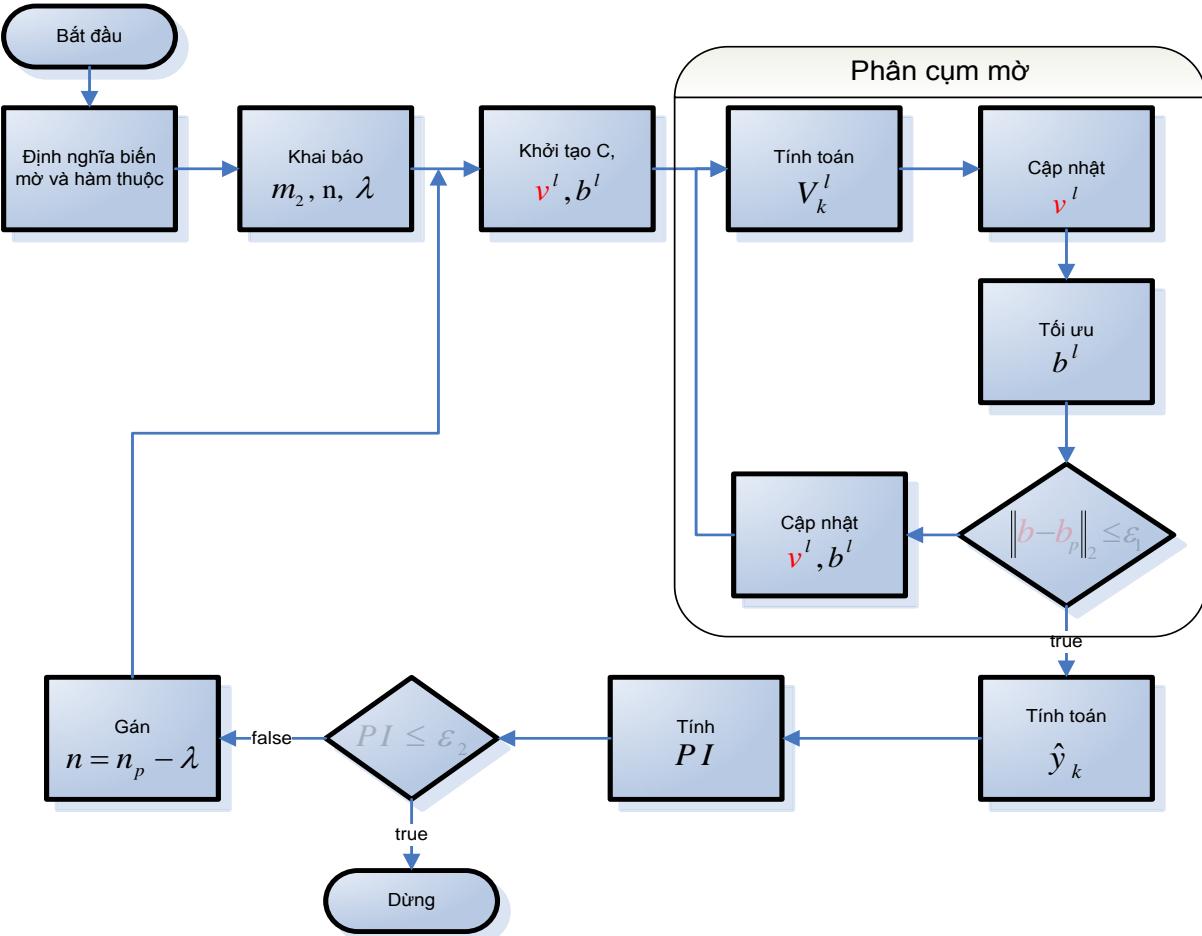
Mỗi biến mờ đầu vào đều có 5 giá trị ngôn ngữ với các khoảng cách đều.

T = {Very Small, Small, Medium, Large, Very Large}

Biến mờ đầu ra chứa 35 giá trị ngôn ngữ (26 chữ cái + 9 số).

Sinh luật: Phương pháp phân tích phân cụm dữ liệu được dùng để sinh luật mờ, một cải tiến so với fuzzy c-mean truyền thống.

Một cách tổng quát, quá trình sinh luật mờ và tối ưu hóa các tham số hệ thống được thể hiện trong hai vòng lặp ở sơ đồ bên dưới



Hình 7: Lưu đồ giải thuật nhận dạng

Trong đó:

- m_2, n, λ là các tham số được khai báo. Giá trị m_2 ảnh hưởng đến V_k^l và là tham số cố định. n thay đổi theo từng vòng lặp một bước λ . Giá trị của n quyết định số luật được sinh. Với $n=1$, có duy nhất 1 luật, n càng nhỏ số luật càng nhiều, và $n=0$ thì số luật đúng bằng số bộ dữ liệu.

- C: số lượng luật sinh ra.

- v^l : ma trận chứa tâm biến vào (hình chiếu của đỉnh hàm thuộc tam giác lên trực).

- b^l : vector chứa tâm biến ra.

- V_k^l : ma trận thuộc được tính theo công thức

$$V_k^l = \frac{1}{\sum_{j=1}^C (d^l(x_k) / d^j(x_k))^{2/(m_2-1)}},$$

$$l = 1, 2, \dots, C; k = 1, 2, \dots, N$$

với $d^j(x_k) = \|x_k - v^j\|_2$, khoảng cách của ma trận đầu vào $x_k = [x_{k1}, x_{k2}, x_{k3}, \dots, x_{kn}]^T$ với ma trận tâm biến vào v^j .

- b_p, n_p lần lượt là giá trị ở vòng lặp trước của b, n .

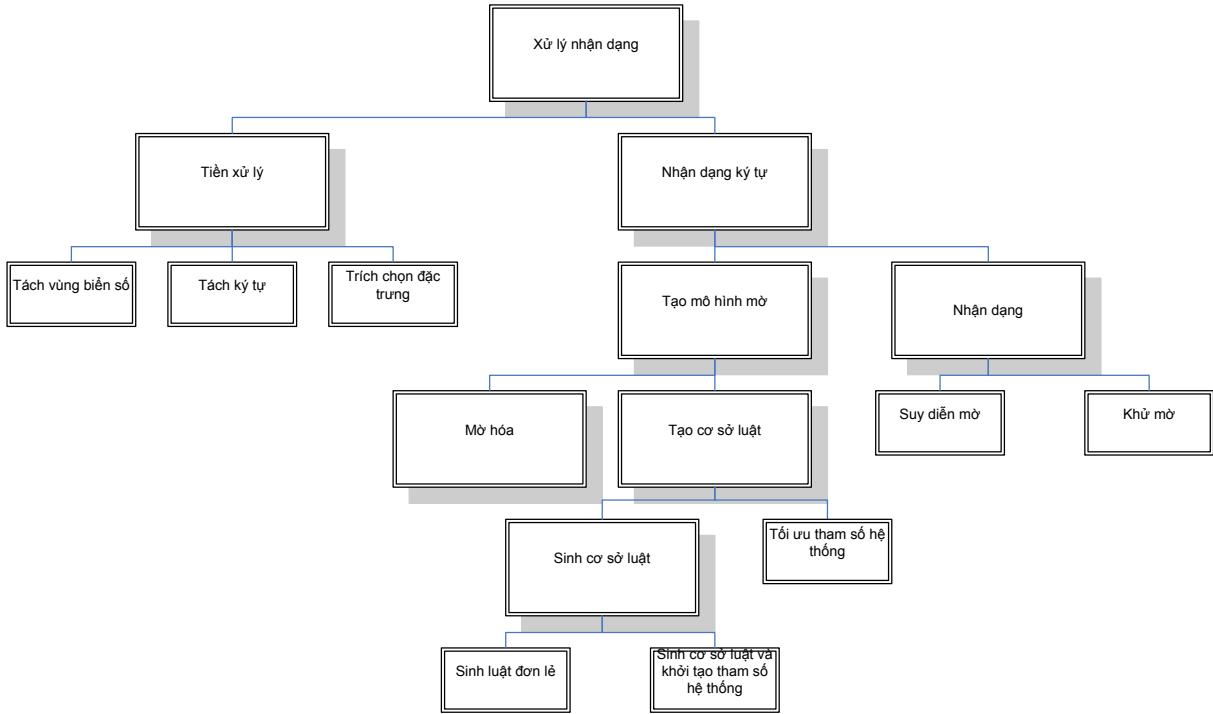
- ϵ_1, ϵ_2 lần lượt là hai tham số quyết định điều kiện dừng của hai vòng lặp. ϵ_1 là tham số đánh giá độ chính xác, trong khi ϵ_2 là tham số đánh giá hiệu năng.

- $\hat{y}_k = \sum_{l=1}^C V_k^l b^l$: kết quả suy diễn của hệ thống

- $PI = \frac{\sum_{k=1}^N (y_k - \hat{y}_k)^2}{N}$, hiệu năng của hệ thống.

6.2. Phân tích, thiết kế hệ thống

a. Biểu đồ phân cấp chức năng (BPC)



Hình 8: Biểu đồ phân cấp chức năng của hệ thống.

Như trình bày trong Hình 8, hệ thống xây dựng gồm hai chức năng lớn là chức năng tiền xử lý và chức năng nhận dạng ký tự.

Trong chức năng tiền xử lý có ba chức năng nhỏ là tách vùng biển số, tách ký tự và trích chọn đặc trưng.

Chức năng nhận dạng ký tự gồm hai bước: Tạo mô hình và nhận dạng. Quá trình tạo mô hình kết thúc bằng việc định nghĩa hàm thuộc, biến mờ, sinh cơ sở luật và tối ưu hóa các tham số hệ thống. Quá trình nhận dạng lựa chọn cơ chế suy diễn mờ, một phương pháp khử mờ và trả về kết quả đầu ra là biến số được nhận dạng.

b. Lược đồ cấu trúc của hệ thống (LCT)

Hình 9 thể hiện LCT của hệ thống. Các nút lá chính là các chức năng cơ bản (tương ứng với một modul trong chương trình). Có 9 modul tổng thể, trong đó 2 modul Suy diễn mờ và Khử mờ là các modul đã được xây dựng sẵn với sự trợ giúp của Fuzzy Logic Toolbox trong Matlab. 7 modul còn lại được xây dựng thành các hàm để giải quyết từng nhiệm vụ đơn lẻ.

6.3. Cài đặt hệ thống

a. Công cụ

Matlab version 7.5.0 release 2007b

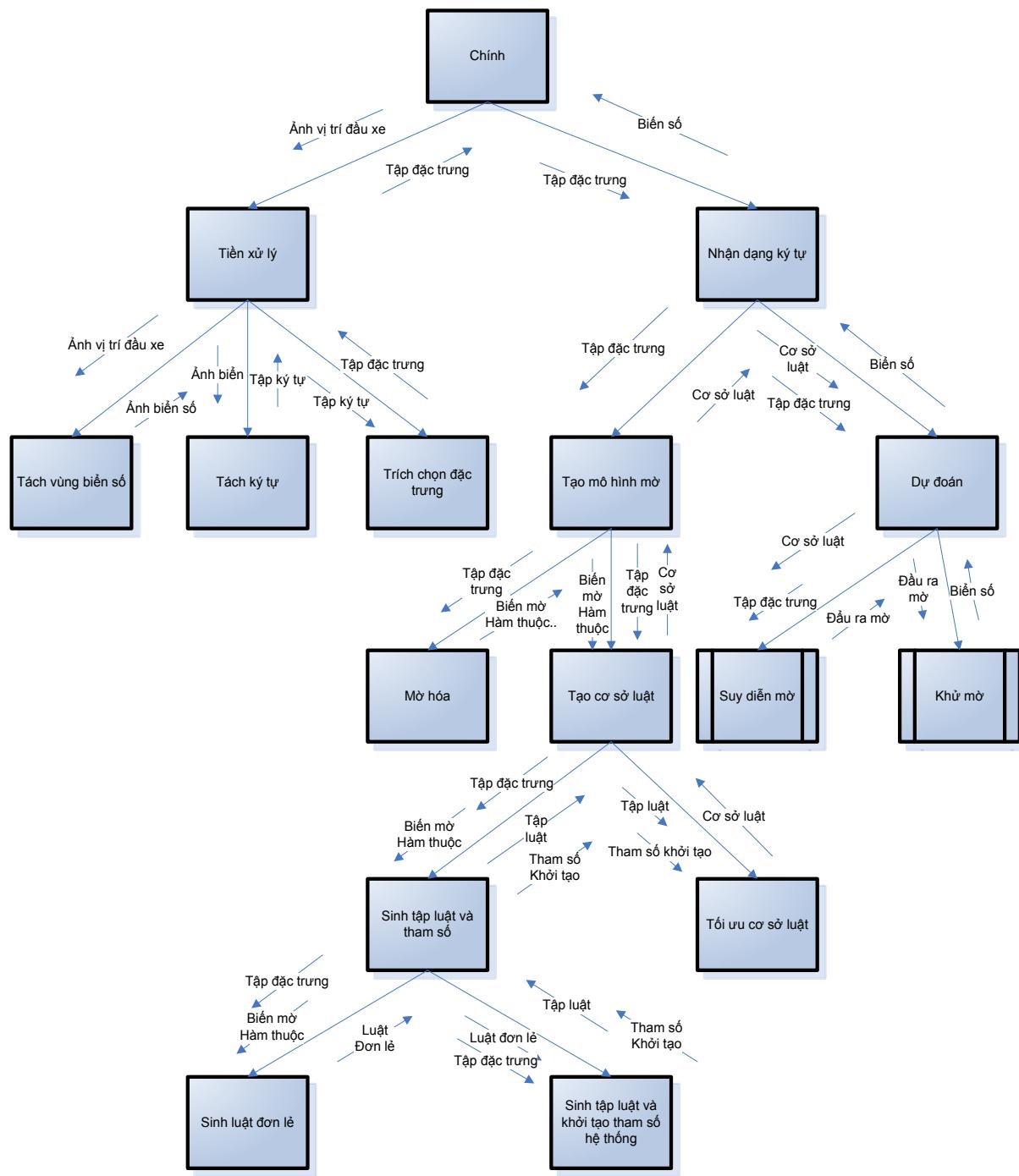
- Xử lý ảnh: Image Processing Toolbox
- Xây dựng hệ mờ: Fuzzy Logic Toolbox, trong đó đặc biệt là các hàm hỗ trợ xây dựng một Fuzzy Inference System (FIS).

b. Giao diện chương trình

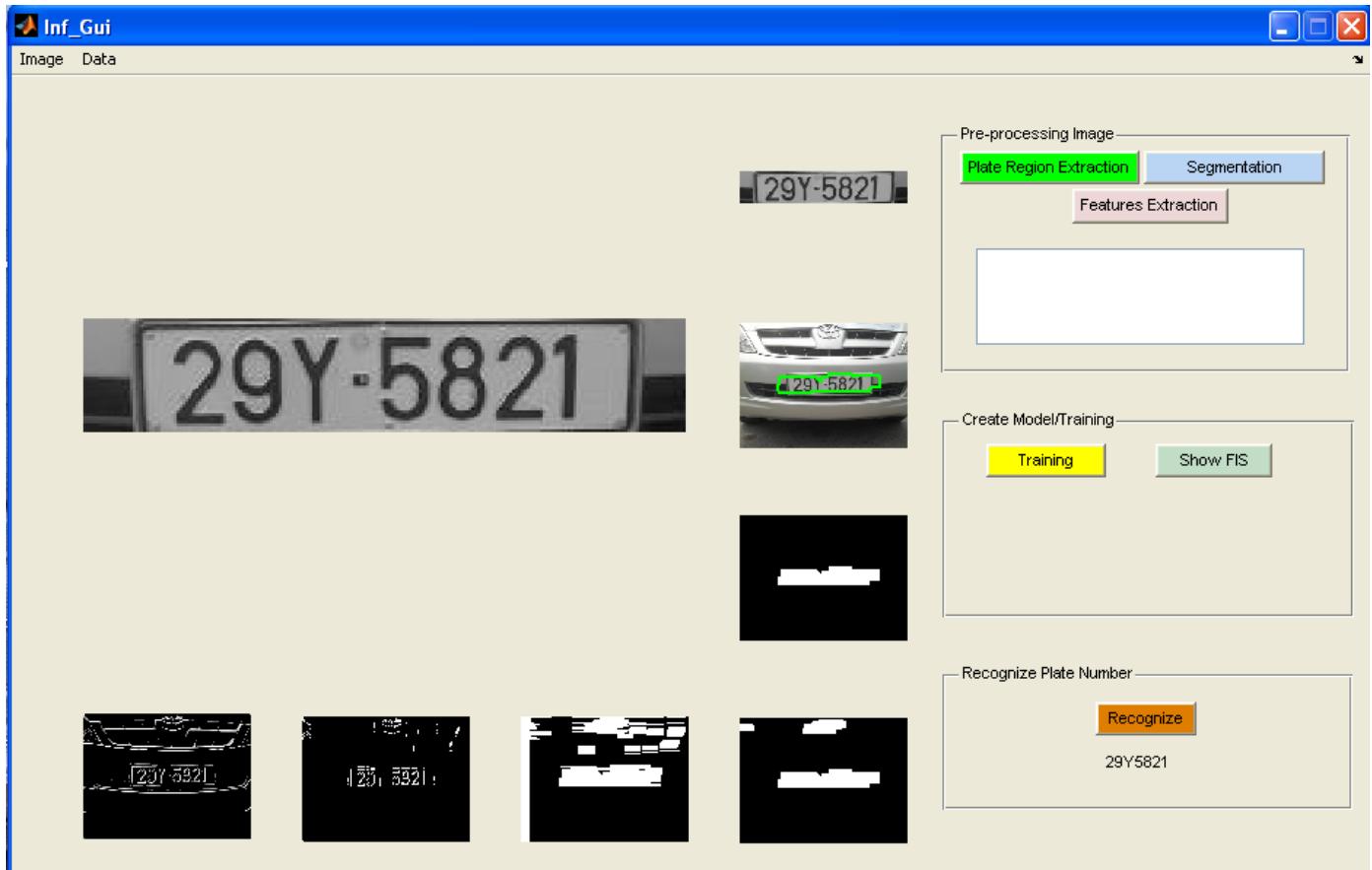
Giao diện chương trình được thể hiện trong Hình 10 bên dưới.

Ba vùng chức năng được phân định rõ ràng trong ba vị trí phía bên phải của Form. Khu vực tiền xử lý ảnh có ba chức năng: tách vùng biển số, trích ký tự biển số và trích chọn đặc trưng. Vùng tạo mô hình gồm chức năng training để sinh một FIS, chức năng Show FIS để hiện các thông số của FIS đã xây dựng, bao gồm hàm thuộc, biến mờ, cơ chế suy diễn, phương pháp khử mờ, cơ sở luật... Vùng nhận dạng ký tự bao gồm 1 button cho chức năng nhận dạng biển số và 1 output text cho việc hiện thị kết quả.

Vào ra của hệ thống được thực hiện trong các chức năng đọc ghi file, chức năng save và load dữ liệu, open và save ảnh trên menu nằm ngang Image và Data.



Hình 9: Lược đồ cấu trúc của hệ thống



Hình 12: Giao diện của hệ thống được xây dựng bằng Matlab

7. KẾT QUẢ THỰC NGHIỆM VÀ ĐỀ XUẤT CẢI TIẾN

7.1. Kết quả thực nghiệm:

Các thực nghiệm đã được tiến hành để kiểm tra hệ thống đề xuất. Hệ thống được thiết kế trên Matlab version 7.5.0 để tiến hành nhận dạng các biển số xe ô tô ở Việt Nam. Ảnh đầu vào của hệ thống là ảnh màu với kích thước 359×269 . Các ảnh thực nghiệm đã được chụp trong những điều kiện ánh sáng khác nhau. Kết quả của quá trình thực nghiệm được cho bởi

Bảng 1

Bảng 1: Kết quả thử nghiệm hệ thống

Các công đoạn	Số mẫu chính xác	Tỉ lệ mẫu chính xác
Tách vùng biển số	128/144	88.89%
Trích ký tự	120/144	83.33%
Nhận dạng	127/144	88.19%

Độ chính xác chung của toàn hệ thống:

$$\text{Độ chính xác hệ thống} = \prod_{i=1}^3 r_i, \text{ trong đó } r_i \text{ là tỉ lệ mẫu}$$

chính xác của 1 trong 3 công đoạn.

Độ chính xác hệ thống $\approx 65.32\%$.

7.2. Đề xuất cải tiến

Vì một số ký tự khá giống nhau, nên dẫn đến xảy ra các nhầm lẫn trong quá trình nhận dạng. Các cặp ký tự giống nhau có thể kể đến như B và 8, S và 5, Z và 2... Bằng cách cải tiến phương pháp trích chọn đặc trưng, bổ sung thêm tri thức về đặc thù của biển số xe (các chữ cái nằm ở vị trí thứ 3 trong biển số), độ chính xác sẽ được cải thiện.

8. KẾT LUẬN

Trong bài báo này, chúng tôi trình bày một ứng dụng phần mềm được thiết kế cho việc nhận dạng tự động biển số xe. Đầu tiên, chúng tôi tiến hành tách vùng biển số, sau đó chúng tôi trích từng ký tự biển đơn lẻ trong vùng biển số. Cuối cùng, một hệ mờ được xây dựng để nhận dạng các ký tự biển số. Hệ thống này được xây dựng cho mục đích nhận dạng biển số xe ô tô tại Việt Nam và đã được kiểm thử trên một số lượng lớn ảnh. Các kết quả ban đầu thu được cho thấy độ chính xác ở mức chấp nhận được. Các nghiên cứu tiếp sau dựa trên nghiên cứu này về việc nâng cao độ chính xác và mở rộng cho việc nhận dạng biển số xe đa quốc gia có thể được phát triển trong tương lai.

9. LỜI CẢM ƠN

Cuối cùng, tôi xin gửi lời cảm ơn chân thành và sâu sắc nhất đến PGS. TS Trần Đình Khang. Xin cảm ơn thày trong

thời gian qua đã nhiệt tình hướng dẫn và gợi ý nhiều kỹ thuật quan trọng để tôi có thể hoàn thành đề tài nghiên cứu này.

10. TÀI LIỆU THAM KHẢO

- [1] George Tsekouras, Haralambos Sarimveis and George Bafas "A method for fuzzy system identification based on clustering analysis", SAMS 2002, Vol. 42(6), pp. 797-823J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp. 68-73.
- [2] Serkan Ozbay, and Ergun Ercelebi, "Automatic Vehicle Identification by Plate Recognition", World Academy of Science, Engineering and Technology 9 2005
- [3] <http://www.ele.uri.edu/~hansenj/projects/ele585/OCR/>
- [4] R.A. Lotufo, A.D. Morgan, and AS. Johnson, 1990, "Automatic Number-Plate Recognition," Proceedings of the IEE Colloquium on Image analysis for Transport Applications, V01.035, pp.6/1-6/6, February 16, 1990.
- [5] A.S. Johnson, B.M. Bird, 1990, "Number-plateMatching for Automatic Vehicle Identification," IEE Colloquium on Electronic Image and Image Processing in Security and Forensic, April, 1990.
- [6] M.M.M. Fahmy, 1994, "Automatic Number-plate Recognition : Neural Network Approach," Proceedings of VNIS'94 Vehicle Navigation and Information System Conference, 31 Aug-2 Sept, 1994
- [7] J.A.G. Nijhuis, M.H. Ter Brugge, K.A. Helmholt, J.P.W. Pluim, L. Spaanenburg, R.S. Venema, M.A. Westenberg, 1995, "Car License Plate Recognition with Neural Networks and Fuzzy Logic," IEEE International Conference on Neural Networks, 1995
- [8] [MathWorks - MATLAB and Simulink for Technical Computing](#)

Các gợi ý cá nhân hóa được gửi tự động cho người dùng di động

Hoàng Minh Thuấn, Tạ Thị Quỳnh Lan

Tóm tắt - **Đối với những lĩnh vực ứng dụng mà trong đó các sản phẩm, dịch vụ được cung cấp bởi hệ thống thay đổi thường xuyên** (ví dụ: các chương trình khuyến mãi, những sự kiện, giải trí...), người sử dụng thường gặp khó khăn trong việc tìm kiếm các sản phẩm, dịch vụ mà họ mong muốn. Các hệ thống gợi ý là những công cụ trợ giúp quyết định nhằm giải quyết vấn đề tràn ngập thông tin bằng cách gợi ý những sản phẩm, dịch vụ phù hợp nhất với nhu cầu và sở thích của mỗi người dùng. Trong bài báo này, chúng tôi đề xuất một phương pháp cho phép hệ thống tự động gửi các gợi ý cá nhân hóa phù hợp với nhu cầu và sở thích của mỗi người dùng tại những ngữ cảnh thích hợp. Đồng thời, chúng tôi cũng giới thiệu hệ thống Prom4U, được cài đặt theo phương pháp được đề xuất, nhằm hỗ trợ người dùng di động kịp thời nhận được các khuyến mãi sản phẩm mà họ mong muốn.

Từ khóa - hệ thống gợi ý, gợi ý tự động, cá nhân hóa, ứng dụng cho người dùng di động.

1. GIỚI THIỆU

Các hệ thống thương mại điện tử thường cung cấp một số lượng rất lớn các sản phẩm và dịch vụ khác nhau. Vì vậy, nếu không có hỗ trợ của hệ thống, người dùng sẽ gặp khó khăn trong việc quyết định lựa chọn các sản phẩm, dịch vụ phù hợp nhất với nhu cầu và sở thích của họ. Các hệ thống gợi ý (recommender systems) được xây dựng để nhằm giải quyết vấn đề tràn ngập thông tin, bằng cách cung cấp các gợi ý về sản phẩm, dịch vụ được cá nhân hóa (làm cho phù hợp) đối với nhu cầu và sở thích của mỗi người dùng [1], [2]. Hầu hết các hệ thống gợi ý hiện nay đều hoạt động theo mô hình gợi ý theo yêu cầu (the pull-delivery recommendation approach), trong đó người dùng phải chủ động đưa ra yêu cầu gợi ý về sản phẩm, dịch vụ, sau đó hệ thống mới cung cấp các gợi ý. Tuy nhiên, trong một số lĩnh vực ứng dụng thực tế (ví dụ: bài toán cung cấp các khuyến mãi về sản phẩm thương mại phù hợp cho mỗi người dùng), các sản phẩm, dịch vụ thay đổi nhanh chóng và thường xuyên (các khuyến mãi thường tồn tại trong một khoảng thời gian ngắn và danh sách khuyến mãi thay đổi liên tục). Với các lĩnh vực ứng dụng như vậy thì phương pháp gợi ý theo yêu cầu tỏ ra phù hợp.

Một hệ thống gửi thông tin tự động (a push-delivery information system) là hệ thống có thể gửi các thông tin tới người dùng mà không cần người dùng đó phải đưa ra yêu cầu. Mô hình hệ thống tự động gửi thông tin rất hiệu quả trong các lĩnh vực ứng dụng mà thông tin thay đổi thường xuyên, bởi vì nó giúp người dùng nhận được thông tin họ quan tâm một cách kịp thời. Tuy nhiên, nếu hệ thống gửi những thông tin không phù hợp hoặc thông tin phù hợp nhưng không đúng hoàn cảnh (thời gian, địa điểm) thì sẽ gây ra sự khó chịu cho

người dùng và họ sẽ có xu hướng từ chối nhận thông tin đó. Do đó, hệ thống phải đảm bảo là các gợi ý được cá nhân hóa phù hợp với người dùng và được gửi vào đúng ngữ cảnh (thời gian, địa điểm). Trong một số phương pháp được đề xuất bởi các tác giả trước đây, hệ thống tự động gửi các thông tin (sản phẩm, dịch vụ) liên quan đến (gần) vị trí hiện tại của một người dùng, mà không quan tâm đến sở thích của họ [5], [6]. Trong một số phương pháp khác, mặc dù hệ thống có tính đến sở thích của người dùng, nhưng không xác định ngữ cảnh phù hợp để gửi thông tin (hệ thống luôn luôn gửi các thông tin quảng cáo mỗi khi người dùng ở gần hoặc bên trong một cửa hàng) [7], [8].

Trong bài báo này, chúng tôi trình bày phương pháp được đề xuất đối với bài toán cung cấp các gợi ý tự động (the *push-delivery recommendation methodology*). Phương pháp gợi ý được đề xuất cho phép hệ thống chủ động (tự động) gửi các gợi ý phù hợp cho người dùng tại những ngữ cảnh thích hợp. Để cung cấp các gợi ý tự động cho người dùng, hệ thống cần phải xác định: *những gợi ý nào sẽ được gửi đến cho một người dùng xác định, và khi nào* thì hệ thống nên (tự động) gửi các gợi ý này đến cho anh ta. Để giải quyết vấn đề thứ nhất, phương pháp gợi ý được đề xuất khai thác cả sở thích dài hạn và sở thích trong phiên gợi ý hiện thời của người dùng, đồng thời cũng sử dụng phương pháp gợi ý hội thoại dựa trên đánh giá (critique-based conversational recommendation) [3]. Để giải quyết vấn đề thứ hai, hệ thống mô hình hóa mỗi tình huống gợi ý tự động thành một trường hợp, và áp dụng chiến lược giải quyết vấn đề dựa trên suy diễn theo trường hợp (case-based reasoning) [4] – là một phương pháp học máy. Cũng trong bài báo này, chúng tôi xin giới thiệu hệ thống gợi ý **Prom4U**, được cài đặt theo phương pháp gợi ý tự động được đề xuất, nhằm giúp người dùng di động có thể kịp thời nhận được các khuyến mãi (về sản phẩm, dịch vụ) mà họ quan tâm.

Các nội dung tiếp theo của bài báo được trình bày như sau. Phần 2 giới thiệu biểu diễn hình thức của các khuyến mãi sản phẩm, sở thích người dùng và câu tìm kiếm của người dùng. Phần 3, chúng tôi trình bày đề xuất phương pháp cung cấp các gợi ý cá nhân hóa được gửi tự động cho người dùng di động. Cuối cùng phần 4 trình bày cài đặt hệ thống và kết luận.

2. BIỂU DIỄN HÌNH THỨC

2.1 Biểu diễn khuyến mãi sản phẩm

Trong bài toán ứng dụng minh họa cho phương pháp của chúng tôi, mục tiêu gợi ý của hệ thống là các khuyến mãi, trong khi các thông tin về sản phẩm được khuyến mãi và quà tặng là các thông tin bổ sung. Các khuyến mãi được biểu diễn có cấu trúc (structuredly represented), bao gồm ba thành phần chính: thông tin về khuyến mãi (*PROMOTION_INFO*),

những sản phẩm được khuyến mãi (*PROMOTED_PRODUCTS*), những quà tặng đi kèm (*GIFTS*).

Mỗi thành phần chính này được biểu diễn bởi các thành phần và các thuộc tính của nó. Cụ thể, thành phần *PROMOTION_INFO* lưu trữ thông tin về khuyến mãi, bao gồm thuộc tính *Prom_Type* và hai thành phần: *DURATION* và *PROVIDER*:

PROMOTION_INFO = (*Prom_Type, DURATION, PROVIDER*)

- Thuộc tính *Prom_Type* chứa thông tin về kiểu của khuyến mãi.

- Thành phần *DURATION* lưu trữ thông tin về thời gian của khuyến mãi *DURATION* = (*Begin_Time, End_Time*); trong đó *Begin_Time* và *End_Time* là thời gian bắt đầu và kết thúc của khuyến mãi.

- Thành phần *PROVIDER* chứa thông tin về nhà cung cấp khuyến mãi (bao gồm số hiệu của nhà cung cấp, khoảng cách đến vị trí hiện tại của người dùng): *PROVIDER* = (*Provider_Id, Distance*).

Thành phần *PROMOTED_PRODUCTS* chứa thông tin về tập các sản phẩm được khuyến mãi: số hiệu sản phẩm, loại sản phẩm và giá của nó:

PROMOTED_PRODUCTS = {(*Product_Id, Category, Price*)};

Thành phần *GIFTS* chứa thông tin về tập các quà tặng của khuyến mãi kiểu quà tặng và mô tả:

GIFTS = {(*Gift_Type, Description*)};

2.2 Biểu diễn hồ sơ người dùng

Hồ sơ người dùng (User Profile) chứa sở thích dài hạn của người dùng, sở thích này sẽ được khai thác bởi hệ thống để xây dựng biểu diễn câu tìm kiếm của người dùng (sẽ được đề cập trong Mục 3.3). Tương tự như đối với khuyến mãi sản phẩm, hồ sơ người dùng được biểu diễn một cách có cấu trúc như sau:

U = (*PROMOTION_PREF, PRODUCT_PREF, GIFT_PREF*)

Thành phần *PROMOTION_PREF* chứa các sở thích dài hạn của người dùng về các khuyến mãi, và được biểu diễn như sau:

PROMOTION_PREF = (*PROM_TYPES_PREF, PROVIDERS_PREF*)

Trong đó:

- *PROM_TYPES_PREF* biểu diễn sở thích của người dùng đối với các kiểu khuyến mãi: *PROM_TYPE_PREF* = {(*Prom_Type, Score*)}, trong đó *Prom_Type* là kiểu khuyến mãi, và *Score* là chỉ số thể hiện mức độ ưa thích của người dùng đối với kiểu khuyến mãi đó;

- *PROVIDERS_PREF* biểu diễn sở thích của người dùng về các nhà cung cấp khuyến mãi:

PROVIDERS_PREF = {(*Provider_Id, Score*)}, trong đó *Provider_Id* là số hiệu của nhà cung cấp khuyến mãi, và *Score* là chỉ số thể hiện mức độ ưa thích của người dùng đối với nhà cung cấp khuyến mãi đó.

Thành phần *PRODUCT_PREF* biểu diễn sở thích của người dùng đối với các loại sản phẩm:

PRODUCT_PREF = {(*Prd_Category, Price, Score*)}, trong đó *Prd_Category* là loại sản phẩm, *Price* là mức giá ưa thích của người dùng đối với loại sản phẩm đó, *Score* là thể hiện mức độ ưa thích của người dùng đối với loại sản phẩm đó. Thành phần *GIFT_PREF* biểu diễn sở thích của người dùng đối với các kiểu quà tặng.

GIFT_PREF = {(*Gift_Type, Score*)}; trong đó *Gift_Type* là kiểu của quà tặng mà người dùng thích, *Score* thể hiện mức độ ưa thích của người dùng đối với kiểu quà tặng đó.

2.3 Biểu diễn câu truy vấn tìm kiếm của người dùng trong phiên gọi ý

Biểu diễn câu tìm kiếm của người dùng (User Query) lưu giữ các thông tin về sự hiểu biết của hệ thống đối với các yêu cầu và sở thích của người dùng trong phiên gọi ý hiện thời (session-specific user preferences). Trong một phiên gọi ý, tại mỗi chu kỳ (bước) gọi ý, hệ thống sử dụng biểu diễn câu tìm kiếm *Q* để tính toán danh sách gợi ý cho người dùng.

Trong phương pháp được đề xuất, câu tìm kiếm của người dùng bao gồm hai thành phần: mẫu sở thích FP (favorite pattern) và trọng số (mức độ quan trọng) của các thuộc tính *W*.

$$Q = (FP, W)$$

Thành phần *FP*, được biểu diễn có cấu trúc, bao gồm ba thành phần liên quan đến sở thích ngắn hạn của người dùng về: khuyến mãi, sản phẩm được khuyến mãi và quà tặng.

$$FP = (PROMOTION_QUERY, PRODUCT_PREF, GIFT_PREF)$$

trong đó hai thành phần *PRODUCT_PREF* và *GIFT_PREF* biểu diễn sở thích hiện thời của người dùng đối với sản phẩm được khuyến mãi và đối với quà tặng, được biểu diễn tương tự như trong biểu diễn hồ sơ người dùng (xem mục 2.2).

Thành phần *PROMOTION_QUERY* xác định sở thích hiện thời của người dùng đối với khuyến mãi. Ngoài những thành phần và thuộc tính như trong biểu diễn của *PROMOTION_PREF* (xem mục 2.2.), trong biểu diễn của *PROMOTION_QUERY* có thêm thành phần *DURATION* và thuộc tính *Distance*, để biểu diễn sở thích hiện thời của người dùng đối với khoảng thời gian của khuyến mãi và đối với khoảng cách đến nhà cung cấp khuyến mãi.

Vector trọng số *W* được biểu diễn có cấu trúc, phù hợp với biểu diễn của *FP*. Đối với mỗi mức biểu diễn, giá trị trọng số của từng thành phần (hoặc thuộc tính) thể hiện mức độ quan trọng của thành phần (hoặc thuộc tính) đó đối với người dùng.

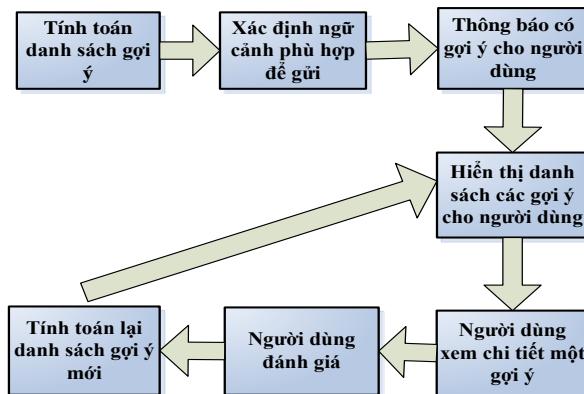
3. PHƯƠNG PHÁP GỌI Ý ĐƯỢC ĐỀ XUẤT

3.1. Tổng quan quá trình gợi ý

Trong phương pháp gợi ý được đề xuất, một phiên gọi ý bắt đầu khi hệ thống (tự động) tính toán tập gợi ý phù hợp cho người dùng và hiển thị thông báo có khuyến mãi mới cho anh ta. Quá trình gợi ý tổng thể được thể hiện trong hình 3.1.

Hệ thống khai thác sở thích dài hạn của người dùng (được lưu trong hồ sơ người dùng User Profile) để khởi tạo biểu diễn

ban đầu về câu tìm kiếm của người dùng Q^0 . Sau đó, hệ thống sử dụng biểu diễn Q^0 này để tính toán danh sách gợi ý ban đầu cho người dùng. Trong quá trình tính toán danh sách gợi ý, hệ thống sắp xếp thứ tự các khuyến mãi theo mức độ tương tự của chúng đối với (FP^0, W^0) , nếu một khuyến mãi có độ tương tự càng cao với (FP^0, W^0) thì nó được xếp càng cao trong danh sách gợi ý (Chi tiết xem tại mục 3.2). Sau khi tính toán danh sách gợi ý ban đầu, hệ thống phải xác định thời điểm thích hợp để tự động gửi danh sách này cho người dùng. Ngữ cảnh gợi ý tự động được tính toán dựa trên phương pháp học máy Case-Based Reasoning (CBR). Ý tưởng của chúng tôi là sử dụng chiến lược giải quyết vấn đề của phương pháp CBR để khai thác tri thức chứa trong các trường hợp gợi ý tự động trong quá khứ (past push cases) (Xem chi tiết tại mục 3.3)



Hình 3.1. Quá trình gợi ý

Tại thời điểm gửi thích hợp đã xác định, hệ thống gửi đến người dùng một thông báo có gợi ý (trong bài toán minh họa, là màn hình thông báo có khuyến mãi mới Hình 3.2-a). Tại thời điểm này, người dùng có thể đưa ra một trong 3 quyết định: đồng ý xem danh sách gợi ý, từ chối xem danh sách gợi ý hoặc trì hoãn việc xem danh sách gợi ý (người dùng sẽ nhận danh sách sau một khoảng thời gian). Sau khi người dùng đồng ý xem danh sách gợi ý, hệ thống sẽ lưu lại trường hợp gợi ý tự động này để sử dụng trong tương lai, và hiển thị danh sách các khuyến mãi cho người dùng (Hình 3.2-b). Nếu người dùng từ chối xem danh sách gợi ý thì hệ thống chỉ lưu lại trường hợp gợi ý tự động này và kết thúc phiên làm việc. Nếu người dùng lựa chọn trì hoãn việc xem danh sách gợi ý, hệ thống sẽ lưu lại danh sách gợi ý này và đợi đến sau khoảng thời gian đã được chỉ định bởi người dùng thì sẽ gửi lại thông báo có gợi ý mới này.

Khi danh sách các khuyến mãi được hiển thị trên màn hình điện thoại của người dùng (Hình 3.2-b), người dùng có thể lựa chọn một khuyến mãi để xem các thông tin chi tiết (Hình 3.2-c). Sau khi người dùng xem các thông tin chi tiết của một khuyến mãi, có ba tình huống có thể xảy ra. Nếu người dùng thỏa mãn với khuyến mãi, thì khuyến mãi này được lưu vào danh sách lựa chọn của người dùng, và người dùng có thể tiếp tục xem các khuyến mãi khác trong danh sách. Nếu người dùng không thích khuyến mãi này, thì anh ta có thể xem các khuyến mãi khác trong danh sách hoặc thoát khỏi phiên gợi ý hiện thời. Nếu người dùng quan tâm (thích) khuyến mãi, nhưng có một số thuộc tính của khuyến mãi này

không hoàn toàn thỏa mãn anh ta. Vì thế, người dùng đánh giá (critique) khuyến mãi này để chỉ ra các thuộc tính anh ta không thỏa mãn (Hình 3.2-d). Những đánh giá này giúp hệ thống cập nhật lại (điều chỉnh) biểu diễn câu tìm kiếm của người dùng (Q) cho chính xác với yêu cầu và sở thích của người dùng, và tính toán danh sách gợi ý mới dựa trên biểu diễn câu tìm kiếm được cập nhật này. Nhờ việc đánh giá khuyến mãi, người dùng tại bước gợi ý tiếp theo (của phiên gợi ý hiện tại) sẽ nhận được danh sách khuyến mãi phù hợp hơn với yêu cầu và sở thích của anh ta. Danh sách gợi ý mới sau đó được hiển thị cho người dùng, và hệ thống tiếp tục tiến triển tới bước gợi ý tiếp theo (the next recommendation cycle).

Khi phiên gợi ý kết thúc, hệ thống khai thác các thông tin về các khuyến mãi được chọn bởi người dùng và những đánh giá của người dùng để cập nhật hồ sơ người dùng. Việc cập nhật này cho phép hệ thống “hiểu được” về sở thích dài hạn của người dùng, và nhờ vậy có thể đáp ứng nhu cầu của người dùng tốt hơn trong tương lai.



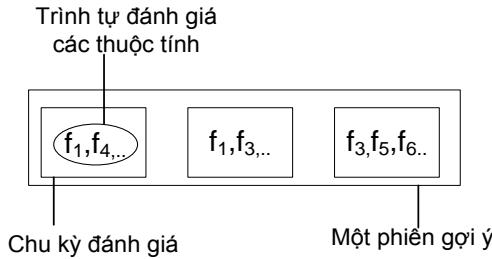
Hình 3.2. Giao diện hệ thống *Prom4U*

Khởi tạo câu tìm kiếm của người dùng Q^0

Tại thời điểm bắt đầu của phiên gợi ý, hệ thống khai thác sở thích dài hạn của người dùng (được lưu trong hồ sơ người dùng) để khởi tạo biểu diễn câu tìm kiếm (Q^0). Trong bước khởi tạo này, các giá trị của các thành phần và thuộc tính của FP^0 được xác định bởi các giá trị của các thành phần và thuộc tính tương ứng trong hồ sơ người dùng (User Profile). Ngoài ra, giá trị của thành phần *DURATION* và thuộc tính *Distance* được thiết lập là không xác định - vì tại thời điểm bắt đầu của phiên gợi ý, hệ thống không biết về các sở thích của người

dùng đối với khoảng thời gian của khuyến mãi (*DURATION*) và đối với khoảng cách đến nhà cung cấp (*Distance*).

Hệ thống khởi tạo các giá trị trọng số W bằng cách khai thác lịch sử tương tác của người dùng với hệ thống, cụ thể là thông tin về các đánh giá của người dùng đối với các khuyến mãi được gọi ý. Ý tưởng cơ bản là giá trị trọng số của một thành phần (hoặc thuộc tính) tỷ lệ với tần suất mà người dùng đánh giá thành phần (hoặc thuộc tính) đó [3]. Hình 3.3 mô tả ví dụ về trình tự người dùng đánh giá các gợi ý trong một phiên gọi ý. Một phiên gọi ý bao gồm nhiều bước gọi ý. Tại mỗi bước gọi ý, các gợi ý được hiển thị cho người dùng (Hình 3.2-b), sau đó người dùng xem các thông tin chi tiết của một gợi ý, và đánh giá (critique) gợi ý đó (Hình 3.2-d).



Hình 3.3. Quá trình đánh giá trong một phiên gọi ý

Ở mỗi mức biểu diễn của W , hệ thống tính toán giá trị trọng số của các thành phần và thuộc tính thuộc vào mức đó. Ví dụ, đối với mức biểu diễn *PROMOTION_INFO*, hệ thống tính toán giá trị trọng số cho thuộc tính *Prom_Type* và các giá trị trọng số cho hai thành phần *DURATION* và *PROVIDER*.

Trước tiên, hệ thống tính toán giá trị trọng số của các thuộc tính (hoặc thành phần) f_i trong phiên gọi ý s_k của người dùng u_j :

$$w_i(u_j, s_k) = \frac{1}{\lambda_k} \cdot \sum_{l=1}^{\lambda_k} \frac{Ctz(f_i, u_j, c_l)}{\alpha^{(\lambda_k-l)}}$$

Trong đó:

- c_l : bước gọi ý của phiên gọi ý s_k ;
- λ_k : độ dài (số bước gọi ý) của phiên gọi ý s_k ;
- $Ctz(f_i, u_j, c_l) = 1$, nếu tại bước gọi ý c_l người dùng u_j đưa ra một đánh giá đối với thuộc tính (thành phần) f_i ;
= 0, nếu ngược lại;
- $\alpha (>=1)$: là tham số được sử dụng nhằm tăng độ quan trọng của các đánh giá gần đây nhất (những đánh giá xuất hiện cuối cùng trong phiên gọi ý s_k). Trong hệ thống thử nghiệm *Prom4U* chúng tôi sử dụng giá trị $\alpha = 1$.

Sau đó, hệ thống tính toán giá trị trọng số của thuộc tính (hoặc thành phần) f_i đối với tất cả các phiên gọi ý của người dùng u_j :

$$w_i(u_j) = \frac{1}{\|S(u_j)\|} \cdot \sum_{k=1}^{\|S(u_j)\|} \frac{w_i(u_j, s_k)}{\beta^{\|S(u_j)\|-k}}$$

Trong đó :

- $S(u_j)$: tập tất cả (được sắp theo thứ tự thời gian) các phiên gọi ý của người dùng u_j .
- $\beta (>1)$: là tham số xác định mức độ quan trọng của các phiên gọi ý theo thời gian. Trong hệ thống thử nghiệm *Prom4U*, chúng tôi sử dụng giá trị $\beta=1,2$ (một phiên gọi ý gần thời điểm hiện tại hơn sẽ có mức độ ảnh hưởng lớn hơn).

Dự đoán ngữ cảnh gửi tự động

Ngữ cảnh gửi tự động được tính toán dựa trên phương pháp học máy Case-Based Reasoning nhằm khai thác tri thức trong các trường hợp gợi ý tự động trong quá khứ (past push cases). Cụ thể, trong bài toán gợi ý khuyến mãi sản phẩm,, mỗi trường hợp gửi tự động (a push case) được biểu diễn bởi hai thành phần: *vấn đề* (problem) và *giải pháp* (solution):

Vấn đề	Giải pháp
<ul style="list-style-type: none"> • Thời gian gửi tự động (the push time) • Khoảng cách của người dùng đến các nhà cung cấp • Các nhà cung cấp các khuyến mãi có trong danh sách gợi ý • Sở thích của người dùng đối với những nhà cung cấp đó 	<p>Lựa chọn (quyết định) của người dùng đối với các gợi ý tự động của hệ thống:</p> <ul style="list-style-type: none"> • Đồng ý nhận; hoặc • Từ chối nhận

Bảng 3.1 Biểu diễn của một trường hợp gửi tự động

Để xác định ngữ cảnh thích hợp cho việc tự động gửi các gợi ý, hệ thống xác định:

1. Ngữ cảnh của trường hợp hiện tại now_case
2. Một tập gồm k các trường hợp gửi tự động trong quá khứ gần giống nhất với trường hợp hiện thời mà có giải pháp là người dùng **đồng ý** nhận danh sách gợi ý (ký hiệu là $C^{Accepted}$).
3. Một tập gồm k các trường hợp gửi tự động trong quá khứ gần giống nhất với trường hợp hiện thời mà có giải pháp là người dùng **từ chối** nhận danh sách gợi ý (ký hiệu là $C^{Rejected}$).

Tiếp theo, hệ thống tính toán mức độ phù hợp để gửi (*push_degree*):

$$\text{push_degree} = \sum_{i=1}^k 1 - \text{distance}(\text{now_case}, C_i^{Accepted})$$

và mức độ không phù hợp để gửi (*not_push_degree*):

$$\text{not_push_degree} = \sum_{i=1}^k 1 - \text{distance}(\text{now_case}, C_i^{Rejected})$$

Trong đó:

- $C_i^{Accepted}$ là trường hợp của ngữ cảnh trong quá khứ nằm trong tập $C^{Accepted}$

- C_i^{Rejected} là trường hợp của ngũ cảnh trong quá khứ nằm trong tập C^{Rejected}
- distance(): là hàm tính khoảng cách (mức độ khác nhau) của hai trường hợp ngũ cảnh. Hàm này được định nghĩa như dưới đây:

$$\text{distance} = 0.5 * \text{distance_time} + 0.5 * \text{distance_provider}$$

(trong đó các hàm khoảng cách của các thành phần sẽ được định nghĩa riêng)

Nếu $\text{push_degree} \geq \text{not_push_degree}$, thì hệ thống gửi thông báo có gợi ý mới cho người dùng (xem Hình 3.2-a). Ngược lại, hệ thống lưu lại danh sách gợi ý này trong hàng đợi; và hệ thống sẽ đợi đến ngũ cảnh (khoảng thời gian) tiếp theo để tính toán lại xem có phù hợp để gửi danh sách gợi ý cho người dùng.

4. CÀI ĐẶT HỆ THỐNG VÀ KẾT LUẬN

Dựa trên phương pháp được đề xuất ở Phần 3, chúng tôi đã xây dựng hệ thống **Prom4U** nhằm cung cấp kịp thời các khuyến mại sản phẩm phù hợp cho mỗi người dùng di động. Hệ thống **Prom4U** được xây dựng sử dụng công nghệ Java và hệ quản trị cơ sở dữ liệu MySQL. Hệ thống được chia ra làm 2 phần. Phần Server thực hiện tính toán danh sách gợi ý và xác định ngũ cảnh phù hợp để (tự động) gửi các khuyến mãi đến cho người dùng, lưu trữ và quản lý dữ liệu về các khuyến mại, quản lý hồ sơ người dùng (sở thích dài hạn của người dùng). Phần Client cung cấp giao diện để người dùng xem thông tin các khuyến mại, đưa ra các đánh giá (critique) đối với các khuyến mãi đó, và lựa chọn các khuyến mãi ưa thích.

Chúng tôi đã bắt đầu quá trình thử nghiệm hệ thống **Prom4U** với các người dùng (test with real users). Trong thời gian tới, chúng tôi sẽ tiến hành thu thập và phân tích các kết quả thí nghiệm để đánh giá hiệu quả của phương pháp gợi ý và sự hữu ích của hệ thống **Prom4U**.

5. LỜI CẢM ƠN

Chúng tôi xin chân thành cảm ơn sự hướng dẫn tận tình của TS. Nguyễn Nhật Quang đã giúp đỡ, chỉ bảo để chúng tôi có thể hoàn thành công trình nghiên cứu này. Đồng thời, chúng tôi cũng xin cảm ơn sự hỗ trợ tài chính của Quỹ phát triển khoa học và công nghệ quốc gia (NAFOSTED) để thực hiện thí nghiệm nghiên cứu này.

6. TÀI LIỆU THAM KHẢO

- [1] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor. *Recommender Systems Handbook*. Springer, 2011.
- [2] R. Burke. Hybrid web recommender systems. In P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.): *The Adaptive Web: Methods and Strategies of Web Personalization*, pp. 377–408. Springer, 2007.
- [3] F. Ricci and Q. N. Nguyen. Acquiring and revising preferences in a critique-based mobile

recommender system. *IEEE Intelligent Systems*, vol. 22, n. 3, pp. 22–29, 2007.

- [4] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, vol. 7, n. 1, pp. 39–59, 1994.
- [5] L. Aalto, N. Göthlin, J. Korhonen, and T. Ojala. Bluetooth and WAP push-based location-aware mobile advertising system. In *Proc. 2nd Int. Conf. Mobile Systems, Application, and Services*, pp. 49–58, 2004.
- [6] S. Kurkovsky and K. Harihar. Using ubiquitous computing in interactive mobile marketing. *J. Personal and Ubiquitous Computing*, vol. 10, n. 4, pp. 227–240, 2006.
- [7] J. E. de Castro and H. Shimakawa. Mobile advertisement system utilizing user's contextual information. In *Proc. 7th Int. Conf. Mobile Data Management*, pp. 91, 2006.
- [8] A. Ciaramella, M. G. C. A. Cimino, B. Lazzerini, and F. Marcelloni. Situation-aware mobile service recommendation with fuzzy logic and semantic Web. In *Proc. 9th Int. Conf. Intelligent Systems Design and Applications*, pp. 1037–1042, 2009.

Hệ thống lưu trữ và chia sẻ dữ liệu Lindax

Nguyễn Đức Huy, Nguyễn Thị Khen, Phạm Việt Linh

Tóm tắt - Ngày nay khi tin học hóa ngày càng phát triển cả về phần cứng lẫn phần mềm, khi xu hướng điện toán đám mây đã ngày càng trở nên quen thuộc với người dùng, thì nhu cầu làm việc mọi lúc mọi nơi, không phụ thuộc vào không gian lưu trữ, trở nên cần thiết và thực sự hữu ích với người dùng máy tính, di động hay các thiết bị ICT khác. Thực tế hiện nay cũng đã xuất hiện nhiều hệ thống lưu trữ, chia sẻ dữ liệu nổi tiếng như: RapidShare, MediaFire, MegaUpload... Tuy nhiên các tính năng mà hệ thống trên đem lại chưa đủ để người dùng có thể dễ dàng làm việc trực tuyến với kho dữ liệu cá nhân của mình. Điều này đã thúc đẩy chúng tôi cùng nghiên cứu và phát triển một hệ thống đa tiện ích trong việc lưu trữ, chia sẻ, đồng bộ dữ liệu. Với lợi ích thiết thực mang lại từ công nghệ lưu trữ lưới – công nghệ nền tảng của điện toán đám mây, cùng với việc phát triển các phần mềm nguồn mở, phát triển hệ thống theo kiến trúc hướng dịch vụ, chúng tôi đã giải quyết được bài toán mở rộng các tính năng cho một hệ thống lưu trữ và chia sẻ dữ liệu thông thường.

Từ khóa - Grid Computing, Service - Oriented Architecture (SOA), Upload, Download, Synchronization.

1. DẪN NHẬP

Hệ thống lindax gồm 3 tầng chính:

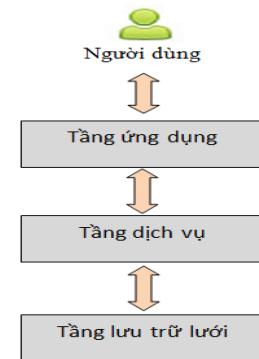
Tầng ứng dụng, tầng dịch vụ và tầng lưu trữ lưới.

- Tầng ứng dụng: đây là tầng giao tiếp với phía người sử dụng. Nó cung cấp các tính năng như quản lý người dùng, quản lý cây thư mục của người dùng trong hệ thống.
- Tầng dịch vụ: đây là tầng cung cấp các dịch vụ phục vụ cho quá trình upload, download, đồng bộ dữ liệu hay tạo bản sao,...

Thông tin về tác giả:

1. Nguyễn Đức Huy, Sinh viên lớp Hệ thống thông tin A, Khóa 51 (điện thoại: 0164.961.2486, E-mail: huyitbk@gmail.com).
2. Nguyễn Thị Khen, Sinh viên lớp Hệ thống thông tin A, Khóa 51 (điện thoại: 0074.749.060, E-mail: khennt@gmail.com).
3. Phạm Việt Linh, Sinh viên lớp AS1 Việt nhật, Khóa 51 (điện thoại: 0984.454.500, E-mail: vietlinhbka@gmail.com).

- Tầng lưu trữ lưới: đây là tầng lưu trữ dữ liệu người dùng tải lên, tại đây cung cấp hai module cơ bản đó là module upload và download. Ngoài ra các module sao lưu, chuyển đổi dữ liệu cũng được thực hiện tại đây.

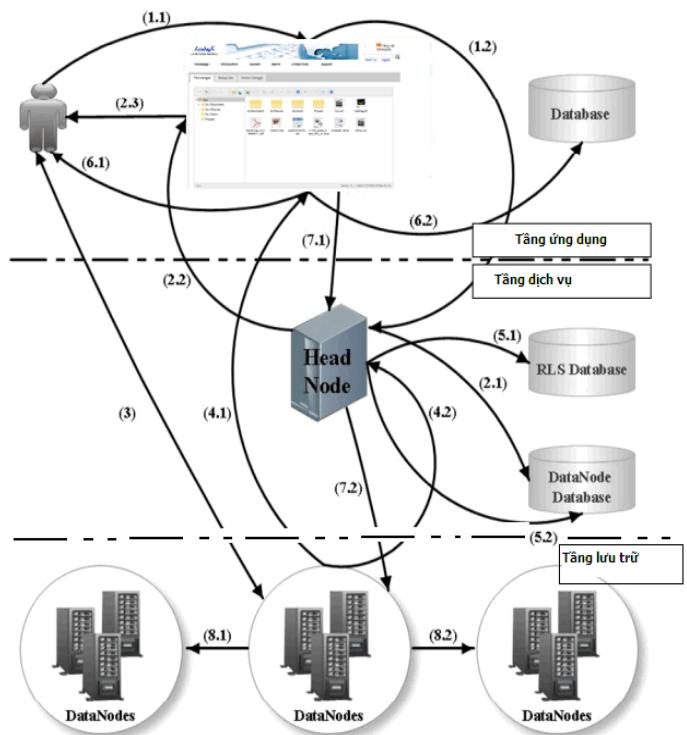


Trong bài báo này, chúng tôi sử dụng các thuật ngữ như: Web Application để chỉ máy chủ chạy ứng dụng web của hệ thống, HeadNode để chỉ máy chủ quản lý các máy lưu trữ, DataNode để chỉ máy lưu trữ.

2. Mô tả chức năng:

2.1 Chức năng upload file

Mô hình kiến trúc :



- 1.1 Client yêu cầu upload tệp
- 1.2 Ứng dụng gọi dịch vụ lấy về module upload
- 2.1 HeadNode chọn nút lưu trữ để thực hiện upload
- 2.2 Gửi địa chỉ nút lưu trữ về ứng dụng
- 2.3 Ứng dụng gọi module upload trên datanode
- 3 Quá trình upload giữa người dùng và datanode sao
- 4.1 DataNode trả lại kết quả upload cho ứng dụng
- 4.2 DataNode trả lại kết quả upload cho headnode

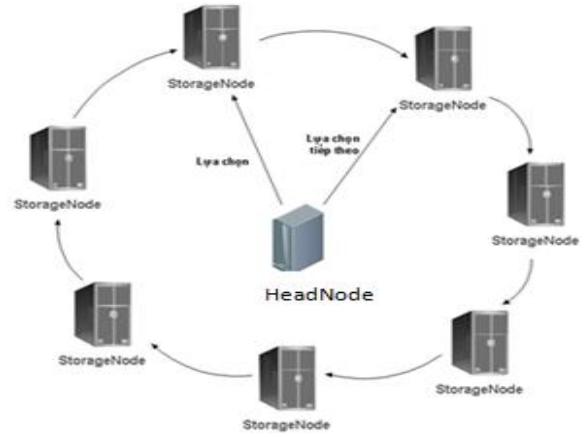
- 5.1 HeadNode cập nhật RLS db
- 5.2 HeadNode cập nhật grid db
- 6.1 Ứng dụng trả kết quả cho client
- 6.2 Ứng dụng cập nhật db
- 7.1 Gửi yêu cầu tạo bản sao
- 7.2 HeadNode chọn nút tạo bản sao
- 8.1 Tạo bản sao
- 8.2 Tạo bản sao

Upload là một trong những pha quan trọng nhất của hệ thống lưu trữ và chia sẻ dữ liệu nói chung và hệ thống LINDAX nói riêng. Khi người dùng Upload dữ liệu lên lưới, nếu không có một cơ chế lưu trữ tệp trên DataNode một cách hợp lý thì dữ liệu lưu trữ trên các máy DataNode sẽ không có được sự “đồng đều”, sẽ có máy lưu trữ quá nhiều tệp, trong khi đó lại có máy lưu trữ rất ít tệp. Điều này cũng sẽ ảnh hưởng đến khả năng download trong pha Download hay cả tốc độ truyền tệp khi tạo bản sao giữa các DataNode. Dù cho tốc độ hay cơ chế download có được cải thiện đến mấy, nhưng nếu tệp phân bố trên DataNode không đồng đều thì pha Download cũng sẽ không đạt được hiệu năng cao nhất. Có thể hiểu ở đây là, sẽ có những khả năng xảy ra như, các DataNode sẽ bị đầy nhanh chóng so với các DataNode khác, tài nguyên mạng có thể bị quá tải dẫn đến hiệu năng truyền và tải tệp giảm, máy lưu trữ cũng sẽ phải chịu tải lớn ảnh hưởng đến khả năng đáp ứng.

Theo đánh giá của chúng tôi, quá trình upload tốt là quá trình luôn tạo được sự cân bằng giữa các nút lưu trữ sau quá trình upload đó. Sự cân bằng ở đây được thể hiện ở:

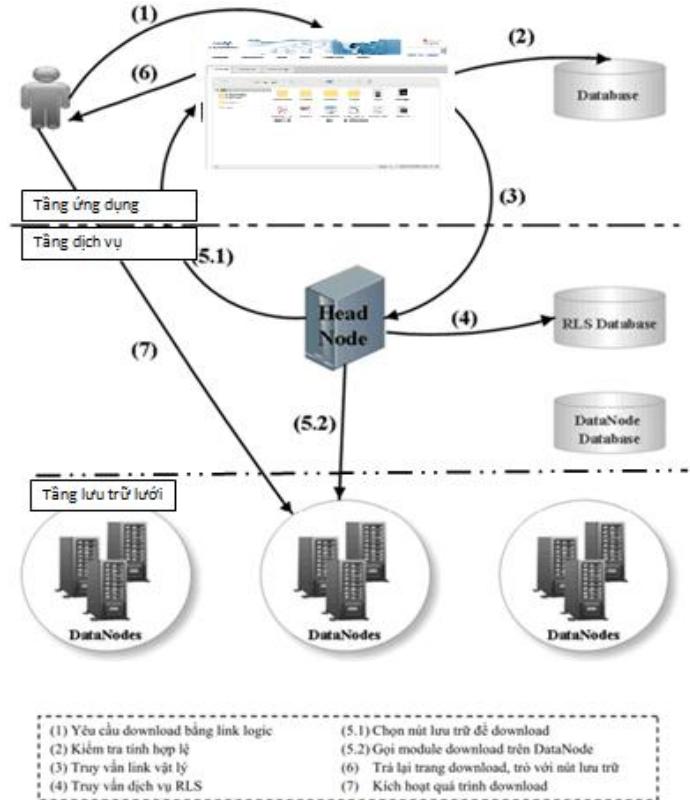
- Quá trình download các tệp sau khi được upload lên sẽ thực hiện một cách cân bằng trên các DataNode.
- Các DataNode luôn có sự cân bằng về dung lượng lưu trữ.

Để chọn được nút lưu trữ trong quá trình upload, chúng tôi cũng đã thử nghiệm các giải thuật như: upload tệp lên máy có dung lượng lưu trữ hiện thời cao nhất, hay upload tệp lên máy có % cpu nhàn rỗi là cao nhất, hay giải thuật RoundRobin. Kết quả thực nghiệm cho thấy áp dụng giải thuật RoundRobin trong quá trình chọn nút lưu trữ đáp ứng được cả 2 tiêu chí ở trên. Sơ đồ của giải thuật RoundRobin được minh họa trên hình vẽ.



2.2 Chức năng download file

Mô hình kiến trúc:



Khi có một yêu cầu download tới 1 tệp trên hệ thống. HeadNode sẽ chọn máy nào để trả về cho người dùng download? Trong khi tệp tin của người dùng có thể được sao lưu trên nhiều nút lưu trữ khác nhau. Vấn đề làm sao để người dùng có thể download tệp tin với tốc độ tốt nhất có thể. Khi mà yêu tố tiên quyết ảnh hưởng tới tốc độ download có thể kể tới là băng thông của các máy lưu trữ, chính vì vậy khi thiết kế giải thuật chọn nút cho pha Download chúng tôi đã quyết định dựa

trên dung lượng băng thông đã được sử dụng trên DataNode để đánh giá DataNode có thích hợp cho việc download dữ liệu hay không.

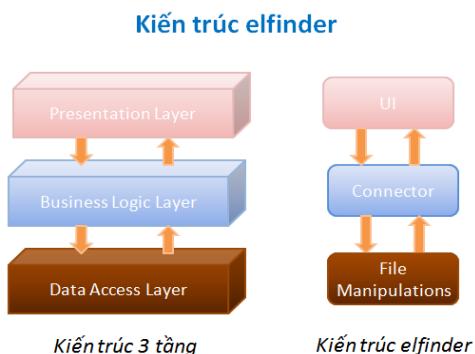
Giải thuật lập lịch cho pha download được thực hiện như sau:

- Cứ sau một khoảng thời gian nhất định, các DataNode sẽ trả về cho HeadNode lượng băng thông mà DataNode đó đã sử dụng.
- Khi có yêu cầu download từ tầng ứng dụng, HeadNode sẽ dựa trên lượng băng thông mà DataNode đã sử dụng để chọn ra DataNode có băng thông sử dụng ít nhất trả về cho client.

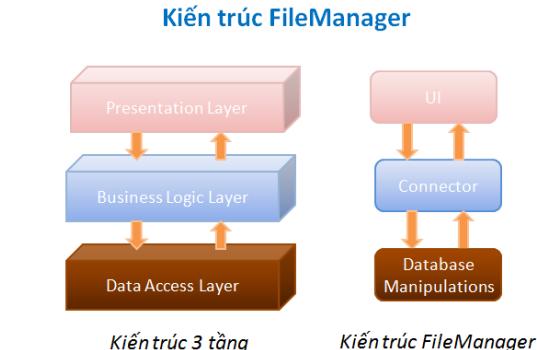
2.3 Chức năng quản lý tệp tin

Bất kỳ một hệ thống lưu trữ và chia sẻ dữ liệu nào cũng đều cần có một công cụ để quản lý dữ liệu. Và lindax cũng vậy, xuất phát từ bài toán quản lý file của người dùng, phải đảm bảo kiến trúc độc lập giữa tầng ứng dụng và tầng lưu trữ dưới, chúng tôi đã tìm hiểu một số công cụ quản lý file mã nguồn mở, nhưng kết quả cho thấy tất cả các công cụ đó đều tương tác với dữ liệu thật trên server, server lưu trữ đồng thời là server ứng dụng, do vậy gặp phải nhiều hạn chế. Vì chúng tôi quyết định viết lại trình quản lý file dựa trên một mã nguồn mở đã có đó là ElFinder.

Kiến trúc của Elfinder được thể hiện như hình vẽ:



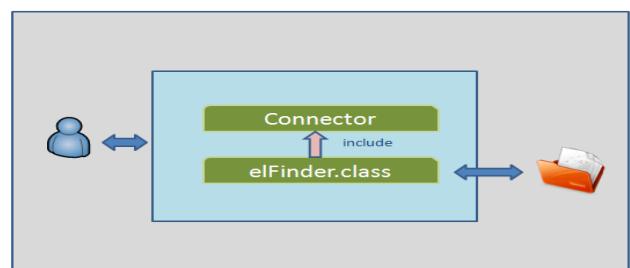
Để phù hợp với bài toán của mình, kiến trúc của trình quản lý file mà chúng tôi xây dựng như sau:



Tầng Presentation là tầng chứa giao diện tương tác trực tiếp với người dùng, thông qua Connector, giao tiếp với tầng Database phía dưới nhờ Database Manipulations.

Với Elfinder, tầng connector bao gồm class connector class elfinder, làm nhiệm vụ xử lý tương tác của người dùng trên hệ thống file thật.

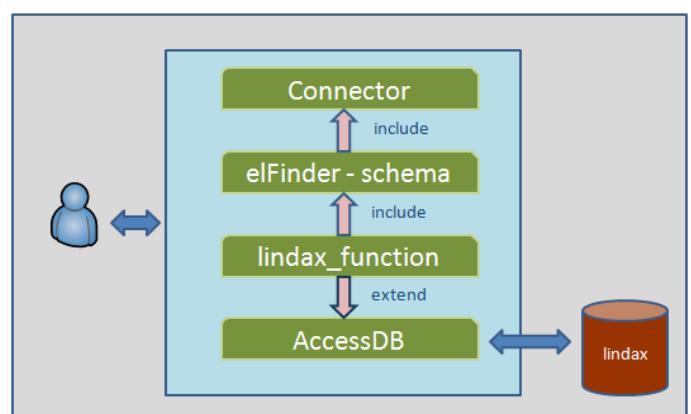
elFinder – Connector



Với filemanager, tầng connector bao gồm 4 layer: connector, elfinder-schema,lindax_function và accessdb. Trong đó layer

Connector làm nhiệm vụ cung cấp giao diện tương tác người dùng và layer elFinder – schema. Layer elFinder – schema là khung mô tả tương ứng các thao tác trên giao diện người dùng. Layer này triết lý gọi đến layer lindax_function để lấy về thông tin về file khi có thao tác truy vấn đến tầng dưới. Layer lindax_function truy vấn đến cơ sở dữ liệu nhờ kế thừa từ layer AccessDB. Trong đó, layer AccessDB là layer truy vấn trực tiếp đến cơ sở dữ liệu trên tầng ứng dụng.

FileManager – Connector



2.4 Chức năng đồng bộ dữ liệu từ xa

Đây là chức năng cho phép người dùng có thể đồng bộ dữ liệu trên máy tính cục bộ với thư mục trên phía ứng dụng web. Chức năng này được cài đặt trên các máy end-user.

Xây dựng được tính năng này phải đảm bảo:

- khi người dùng có bất kỳ sự thay đổi nào ở thư mục được đồng bộ desktop client như tạo mới thư mục hay tạo mới file, sự thay đổi tương ứng xảy ra ở ngay trên tầng ứng dụng và tầng lưu trữ.
- Và đồng thời khi người dùng có bất kỳ sự thay đổi nào khi tương tác trên trình quản lý file ở nền web-based của hệ thống, sự thay đổi tương ứng cũng được đáp ứng ở phía client.

Để xây dựng tính năng này, chúng tôi đề ra các bước thực hiện như sau:

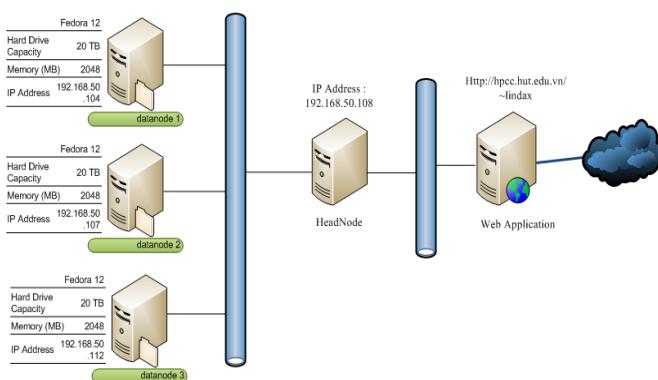
- Tầng dịch vụ cài đặt dịch vụ trả về file lưu cấu trúc thư mục của từng người dùng theo định dạng xml.
- Phía client gọi dịch vụ trên lấy về file xml, duyệt file và tạo lại cây thư mục trên máy client tương ứng. Khi duyệt file xml:
 - nếu nút đang được duyệt là dir: chương trình tiến hành tạo thư mục.
 - nếu nút đang được duyệt là file: chương trình tiến hành gọi dịch vụ download file về.
- Phía client xây dựng chương trình chạy liên tục sau một khoảng thời gian nhất định, so sánh file lưu cấu trúc thư mục phía client và file lưu cấu trúc thư mục trên webserver, cập nhật sự thay đổi tương ứng. Chương trình phía client luôn lắng nghe sự thay đổi trên thư mục được đồng bộ. Cụ thể khi người dùng

tạo mới thư mục, chương trình gọi dịch vụ cập nhật cơ sở dữ liệu logic trên tầng ứng dụng, khi người dùng tạo mới file, chương trình gọi dịch vụ để upload file.

Ứng dụng đồng bộ dữ liệu từ xa được cài đặt và chạy dưới dạng system-tray, giúp người sử dụng dễ dàng thao tác với các tính năng của hệ thống.

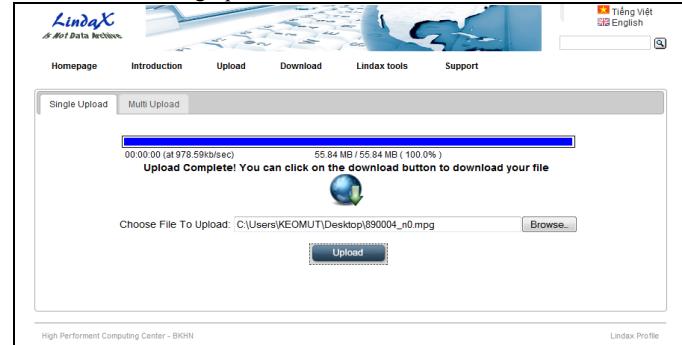
3. Kết quả thực nghiệm

Hệ thống của chúng tôi được triển khai với sơ đồ cụ thể như sau :

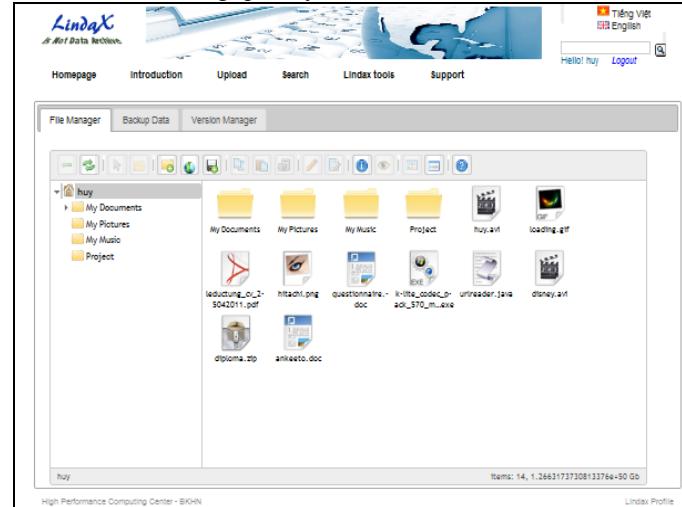


Đây là mô hình triển khai hệ thống với một máy web application, 1 máy headnode, và 3 máy datanode.

Giao diện tính năng upload dữ liệu:



Giao diện tính năng quản lý file:



4. Kết luận

Trong bài báo này, chúng tôi đã trình bày hệ thống lưu trữ và chia sẻ dữ liệu được xây dựng trên kiến trúc mở dựa trên nền tảng lưu trữ lưới. Với các tính năng đã xây dựng được như: trình quản lý file, trình đồng bộ dữ liệu từ xa, giúp thao tác thuận tiện dễ dàng cho người dùng. Đồng thời chúng tôi cũng đã xây dựng được một tập các api nâng cao khả năng đáp ứng cho hệ thống như các tiện ích sao lưu, chuyển đổi dữ liệu,...hứa hẹn hệ thống trong thời gian tới sẽ mang lại nhiều hữu ích và thiết thực hơn nữa.

5. Lời tri ân

Tập thể nhóm chúng tôi xin bày tỏ lời cảm ơn chân thành đến giáo sư Nguyễn Thanh Thủy, tiến sĩ Nguyễn Hữu Đức, thạc sĩ Lê Đức Tùng, kỹ sư Lê Đức Hùng, kỹ sư Đào Quang Minh đã giúp đỡ chúng tôi trong quá trình thực hiện đề tài này.

9. TÀI LIỆU THAM KHẢO

- [1] Distributed Systems: Principles and Paradigms. Andrew S. Tanenbaum
Maarten van Steen.
- [2] Enabling Applications for Grid Computing with Globus , IBM Redbooks,
6/2003
- [3] Michael Di Stefano, “Distributed Data Management for Grid Computing”,
John Wiley & Sons, Inc. 2005.
- [4] GridFTP develop guide at
<http://www.globus.org/toolkit/docs/latest-stable/data/gridftp/>
- [5] GridFTP: <http://www.globus.org/toolkit/docs/4.0/data/gridftp/>.
- [6] Ann Chervenak, “RLS Java Client API Documentation”
<http://www.isi.edu/~annc/rls/doc/client-java/index.html>.
- [7] Addison Wesley, A Brief Guide to the Standard Object Modeling Language,
Third Edition.
- [8] Leszek Maciaszek, "Requirements Analysis and System Design: Developing
Information Systems with UML 3rd ed", Addison Wesley (2007); ISBN:
978-0-321-44036-5.
- [9] Công nghệ phát triển Web: HTML, JavaScript, Ajax, trang Web:
<http://www.w3schools.com/>.
- [10] Michael Glass, Yann Le Scouarnec, Elizabeth Naramore, Gary Mailer,
Jeremy Stolz, Jason Gerner, “Beginning PHP, Apache, MySQL Web
Development”.
- [11] Công nghệ j-Query: <http://api.jquery.com/>
- [12] How to design a good API and why it matters-PDF, Joshua Bloch.

Chương trình tạo video 3D từ mô hình 3D sử dụng công nghệ GPGPU

Trịnh Quốc Việt, Nguyễn Hữu Dũng

Tóm tắt - Mô hình 3 chiều xây dựng bởi các chương trình đồ họa như 3ds Max chỉ hiển thị qua mắt người dưới dạng hình ảnh 2 chiều mà không có bề nổi hay chiều sâu. Hiện nay vẫn chưa có 1 công cụ cho phép render trực tiếp từ mô hình 3D ra ảnh hay video nổi do đó mục tiêu mà chúng tôi hướng tới ở đề tài này là xây dựng một chương trình cho phép render trực tiếp từ một mô hình 3 chiều thành video 3D có hiệu ứng nổi.

Để thực hiện được công việc đó, trước hết cần hiểu nguyên lý tạo hiệu ứng ảnh nổi. Có rất nhiều cách để ta thu nhận được hình ảnh 3 chiều. Tuy nhiên tất cả đều dựa trên nguyên lý cơ bản là sự mô phỏng thị giác hai mắt đối với đối tượng sự vật. Nói cách khác, hiệu ứng 3D ở các loại ảnh nổi hay phim nổi đều giống nhau ở bản chất: nhằm gửi đến mắt trái và mắt phải người quan sát một cách tách biệt hai hình ảnh tương ứng với góc lệch bên trái và bên phải của đối tượng (nếu tách biệt không tốt sẽ có hiện tượng nhòe hình). Sự chập ảnh vô thức của não bộ sẽ gây nên ảo tượng chìm hay nổi của đối tượng sự vật. Có thể kể đến các loại ảnh nổi và phim nổi 3D như ảnh anaglyph, ảnh autostereogram, phim 3D dùng công nghệ phân cực, màn hình 3D dùng công nghệ chớp tắt điện tử theo thời gian... Trong phạm vi nghiên cứu này, chúng tôi lựa chọn công nghệ anaglyph với những ưu điểm là tạo hiệu ứng nổi rõ rệt, có thể hiển thị hình ảnh nổi trên bất cứ loại màn hình nào với chi phí trang thiết bị rẻ. Trong khi đó, các rạp phim hiện nay phải đầu tư tới hàng trăm triệu để xây dựng hệ thống chiếu phim nổi 3D bằng công nghệ phân cực, đó là điều không tưởng với các hộ gia đình. Tuy nhiên công nghệ anaglyph cũng có một nhược điểm khá lớn là bị mất bớt màu sắc khi quan sát qua kính lọc màu.

Như đã nêu ở trên, mục tiêu của nghiên cứu là tạo video nổi 3 chiều từ mô hình 3D. Chúng tôi đã đi từng bước trong việc nghiên cứu công nghệ này, bắt đầu từ việc tạo ra ảnh anaglyph từ hai ảnh đầu vào. Tiếp theo chúng tôi nghiên cứu và thực hiện tạo ra hai ảnh ứng với hai góc nhìn của mắt từ một mô hình 3D để tạo ra ảnh anaglyph sử dụng thuật toán render raytracing.

Sau đó, chúng tôi đặt một hoạt cảnh cho camera chuyển động và render liên tục thành các frame để tạo ra video. Và cuối cùng là sử dụng công nghệ tính toán đa dụng trên GPU để song song hóa bài toán, giúp cải thiện thời gian chạy của chương trình.

Công trình này được thực hiện dưới sự bảo trợ của Trung tâm tính toán hiệu năng cao – trường Đại học Bách Khoa Hà Nội

Trịnh Quốc Việt, sinh viên lớp Hệ thống thông tin, khóa 51, Trung tâm đào tạo tài năng, trường Đại học Bách Khoa Hà Nội (điện thoại: 01696.984.004, e-mail: rockman88v@gmail.com).

Nguyễn Hữu Dũng, sinh viên lớp Tin Pháp, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0988.678.689 e-mail: dungnh3388@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

Từ khóa - anaglyph, CUDA, GPGPU, 3D.

1. TẠO ẢNH ANAGLYPH TỪ HAI ẢNH ĐẦU VÀO

Ảnh anaglyph là loại ảnh được tạo bằng cách trộn màu của hai ảnh đầu vào tương ứng với hai góc nhìn của mắt người. Ảnh anaglyph khi đi qua kính anaglyph tương ứng sẽ bị lọc màu và tạo cho ta thấy hiệu ứng nổi. Ví dụ, ảnh anaglyph red-cyan là sự kết hợp của màu đỏ của ảnh trái với màu xanh lá và xanh dương của ảnh trái. Có nhiều loại ảnh anaglyph tương ứng với công thức trộn màu, trong đó ảnh anaglyph loại red-cyan là phổ biến nhất.

Dưới đây là một số loại ảnh anaglyph và công thức trộn màu:

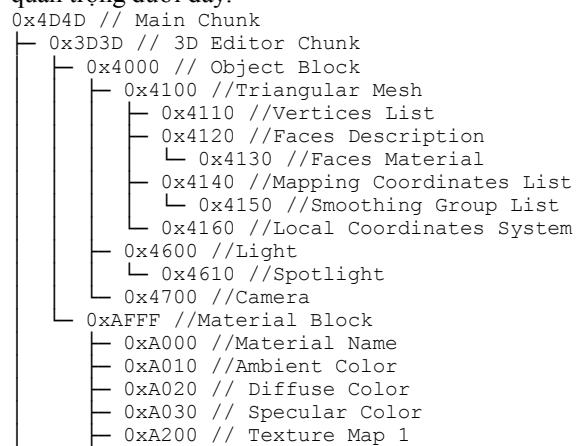
Loại ảnh	Ảnh trái	Ảnh phải
Red-green	Pure red	Pure green
Red – blue	Pure red	Pure blue
Red – cyan	Pure red	Pure cyan (blue + green)
Maganta cyan	Maganta (red + blue)	Cyan (blue + green)

2. TẠO ẢNH ANAGLYPH TỪ MÔ HÌNH 3D

2.1 Định dạng file .3ds

File .3ds chứa một chuỗi thông tin được dùng để mô tả chi tiết khung 3D bao gồm một hay nhiều vật thể. File .3ds chứa một chuỗi các khối được gọi là Chunks. Những gì được chứa trong các khối này? Tất cả đều cần thiết để mô tả khung cảnh: tên của mỗi đối tượng, các tọa độ đỉnh, tọa độ ánh xạ, danh sách các đa giác, các mặt màu và cả các frame hoạt cảnh.

Các chunks này không có cấu trúc tuyến tính. Điều này có nghĩa là một vài chunks sẽ phụ thuộc vào các chunks khác và chỉ được đọc nếu chunk cha của nó được đọc trước. Không cần thiết phải đọc hết tất cả các chunk mà ta chỉ cần xét tới một vài chunk quan trọng dưới đây.



```

    └── 0xA230 // Bump Map
    └── 0xA220 // Reflection Map
        /* Sub Chunks For Each Map */
        └── 0xA300 // Mapping Filename
        └── 0xA351 // Mapping Parameters
    0xB000 // Keyframer Chunk
    └── 0xB002 // Mesh Information Block
    └── 0xB007 //Spot Light Information Block
    └── 0xB008 // Frames (Start and End)
        └── 0xB010 // Object Name
        └── 0xB013 // Object Pivot Point
        └── 0xB020 // Position Track
        └── 0xB021 // Rotation Track
        └── 0xB022 // Scale Track
    └── 0xB030 // Hierarchy Position

```

Trong nghiên cứu này, chúng tôi tập trung vào Object block với 3 chunk quan trọng là Triangular Mesh (các lưỡi tam giác biểu diễn vật thể), ánh sáng, camera. Các thành phần của 1 tam giác bao gồm tọa độ các đỉnh và 2 mặt của tam giác. Ánh sáng có tọa độ nguồn sáng và cường độ sáng. Camera có tọa độ và hướng nhìn.

Như đã nêu trên, nếu ta muốn đọc một chunk nào đấy thì trước hết phải đọc chunk cha của nó. Tưởng tượng như file .3ds là một cây và chunk cần đọc là lá. Theo thứ tự để đến được lá, ta cần bắt đầu từ thân cây và đi qua các nhánh để tới lá. Ví dụ ta cần đi đến chunk Vertices List, ta phải đọc Main Chunk trước tiên, sau đó tới 3D Editor Chunk, Object Block và cuối cùng là chunk Triangular Mesh. Các chunk còn lại có thể được bỏ qua một cách an toàn.

2.2 Đối tượng camera trong mô hình 3D

Một đối tượng quan trọng trong nghiên cứu của chúng tôi là camera trong mô hình 3D. Camera thể hiện góc nhìn của mắt người tới mô hình. Trong đối tượng camera có hai thuộc tính quan trọng là vị trí đặt camera và vị trí mà camera nhìn vào. Để tạo ra ảnh nổi, ta cần tạo ra được hai ảnh đầu vào ứng với hai góc nhìn của mắt người. Từ đó, chúng tôi đã tạo ra hai đối tượng camera ứng với vị trí hai mắt người nhìn vào mô hình.

Ban đầu chúng tôi đặt camera ở vị trí mặc định, sau đó người dùng có thể dịch chuyển camera theo ý muốn theo chiều ngang, dọc hay thay đổi góc nhìn vật và chương trình sẽ hiển thị mô hình theo hướng nhìn của camera đó. Sau khi người dùng đã chọn xong vị trí nhìn của camera và chọn chức năng render, chương trình sẽ tính toán vị trí đặt camera thứ hai có độ lệch tương đối so với camera thứ nhất để tạo góc nhìn thứ hai.

Người dùng có thể dịch camera sang trái, phải, lên xuống và thay đổi góc nhìn của camera. Việc dịch chuyển camera là theo mặt phẳng song song với chiều ngang nếu dịch ngang và dịch theo phuong thẳng đứng nếu dịch theo chiều dọc, do đó ta sẽ thấy vật dịch chuyển sang ngang, hoặc dịch chuyển lên.

Khi người dùng thay đổi góc nhìn vào vật thể, tọa độ của camera sẽ được tính toán lại và hiển thị lên màn hình. Góc nhìn này chính là góc tạo bởi tia điểm đầu là vị trí đặt camera và điểm cuối là điểm camera nhìn vào so với mặt phẳng nằm ngang Oxy.

Người dùng có thể chọn chế độ mặc định là để camera tự quay, khi đó góc nhìn sẽ được tự động thay đổi, camera sẽ quay xung quanh vật thể.

2.3 Thuật toán render raytracing

Ý tưởng cơ bản của thuật toán ray tracing là dò theo tia sáng từ nguồn sáng. Đường đi và cường độ, màu sắc của chúng được tính toán cho độ phản xạ và khúc xạ.

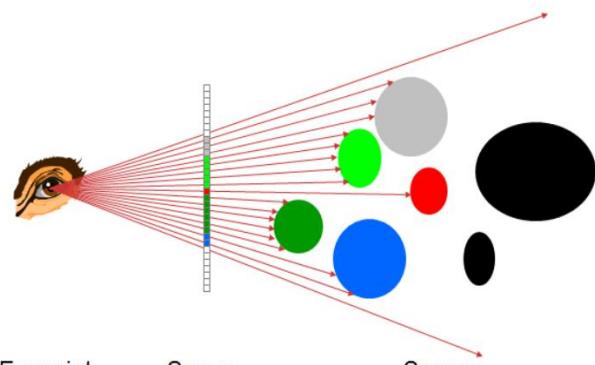
Tưởng tượng rằng có một camera đặt trong một căn phòng với các vật thể khác nhau. Căn phòng được chiếu sáng bằng một bóng đèn. Các tia sáng từ bóng đèn phản xạ vào các vật thể. Chỉ có 1 vài tia sáng đi qua camera và tạo ra ảnh. Đây chính là quá trình mà thuật toán ray tracing cố gắng mô phỏng.

Mọi thứ trong căn phòng đều có thể được biểu diễn dưới dạng toán học. Bóng đèn có tọa độ x, y, z và cường độ sáng. Các vật thể được mô hình hóa bằng các hình học nguyên thủy như là hình cầu hay hình đa giác. Camera được biểu diễn bằng tọa độ, hướng và khoảng không gian thấy được của thấu kính.

Trong chương trình, các tia sáng được đưa tới nguồn sáng từ mọi hướng. Góc phản xạ và khúc xạ của tia sáng được tính toán khi nó đập vào các vật thể. Các thuộc tính bề mặt (màu sắc, độ bóng...) của vật thể làm thay đổi màu sắc và cường độ tia sáng. Một số tia sáng đi tới camera sẽ được lưu lại các thông số về màu sắc và cường độ.

Chỉ có một số lượng nhỏ các tia là đi tới camera trong một lượng cực lớn các tia được phát ra từ nguồn sáng. Điều này làm cho khối lượng tính toán trở nên rất tốn kém.

Như vậy chìa khóa của thuật toán ray tracing là điều chỉnh một chút quá trình nêu trên. Thay vì chiếu các tia sáng từ nguồn sáng với hầu hết các tia sáng bị đi ra ngoài camera, các tia sáng bây giờ sẽ được chiếu từ camera tới căn phòng. Mỗi tia sẽ được chiếu qua mỗi điểm ảnh trên màn ảnh để lưu lại màu sắc của tia. Đây chính là màu của từng điểm ảnh thu nhận được.



Hình 1. Biểu diễn các tia từ mắt người đến vật thể

Thuật toán Ray Tracing tổng quát

Với mỗi điểm ảnh trên buffer,

Xác định điểm P tương ứng trên viewplane;

Xây dựng tia Ray đi từ mắt qua P;

Xác định giao điểm gần nhất giữa tia Ray và các vật thể Objects trong khung cảnh;

Nếu không tồn tại giao điểm thì điểm ảnh có màu là màu nền; Ngược lại,

Tính vector trực chuẩn của vật thể tại giao điểm;

Màu điểm ảnh là màu ambient;

Với mỗi nguồn sáng,

Tạo tia đi từ giao điểm đến nguồn sáng;

Tính sự phân phối ánh sáng diffuse tới vật thể và cộng dồn vào màu của điểm ảnh;

Nếu vật thể phản xạ,

Tính toán tia phản xạ;

Xác định màu của tia phản xạ;

Cộng dồn vào màu của điểm ảnh;

Nếu vật thể trong suốt,

Tính toán tia khúc xạ;

Xác định màu của tia khúc xạ;

Cộng dồn vào màu của điểm ảnh;

Tuy nhiên phương pháp ray tracing có 1 nhược điểm là việc tính toán giao điểm gần nhất giữa tia và vật thể khá phức tạp. Đây là thuật toán tổng quát của thủ tục trên:

```
MinT = INFINITY
forall Object ObjectList {
    // Tính khoảng cách từ gốc của tia và giao điểm với vật
    // thể
    T = Object.Intersection(Ray)
    if (T > 0 && T < MinT) {
        MinT = T
        MinObject = Object
    }
}
Point = Ray.Origin + MinT * Ray.Direction
Normal = MinObject.Normal(Point)
```

2.4 Tạo ảnh anaglyph

Từ những trình bày ở trên về việc đặt camera thứ hai và thuật toán render raytracer, chúng tôi đã áp dụng để tạo ra hai ảnh với hai góc nhìn đối với vật thể. Việc tính toán màu cho mỗi điểm ảnh được thực hiện đồng thời với cả hai camera. Việc tính giá trị ảnh của mỗi pixel được tính toán theo giải thuật raytracer. Kết quả thu được cho mỗi điểm ảnh là 3 byte màu RGB. Tại đây chúng tôi sẽ kết hợp màu của hai điểm ảnh lại theo công thức của trộn màu của ảnh anaglyph red-cyan. Kết quả sau khi trộn màu ta sẽ thu được một điểm ảnh trong ảnh anaglyph và tiến hành ghi vào file ảnh bitmap.

Có thể mô tả giải thuật như sau:

```
for each_pixel
{
    Calculate_color(pixel, camera1);
    Calculate_color(pixel, camera2);
    Final_color = Mix(pixel1, pixel2);
    Write(output.bmp, final_color);
}
```

3. CUDA VÀ CÔNG NGHỆ TÍNH TOÁN SONG SONG TRÊN CÁC BỘ XỬ LÝ ĐỒ HỌA

Trong vài năm gần đây, năng lực tính toán của các bộ xử lý đồ họa (GPU) đã tăng lên với tốc độ đáng kể so với CPU. Tính đến tháng 6/2008, GPU thế hệ GT200 của NVIDIA đã đạt tới ngưỡng 933GFlops gấp hơn 10 lần so với bộ xử lý hai lõi Intel Xeon 3.2 GHz tại cùng thời điểm. Hình 2 thể hiện sự tăng tốc về năng lực tính toán của các bộ xử lý đồ họa NVIDIA so với bộ xử lý Intel.



Hình 2. Biểu đồ năng lực tính toán GPU-CPU.

Tuy vậy, sự vượt trội về hiệu năng này không đồng nghĩa với sự vượt trội về công nghệ. GPU và CPU được phát triển theo hai hướng khác biệt: trong khi công nghệ CPU cố gắng tăng tốc cho một nhiệm vụ đơn lẻ thì công nghệ GPU lại tìm cách tăng số lượng nhiệm vụ có thể thực hiện song hành. Chính vì vậy, trong khi số lượng lõi tính toán trong CPU chưa đạt đến con số 8 lõi thì số lượng lõi xử lý GPU đã đạt đến 240 và còn hứa hẹn tiếp tục tăng tới trên 500 lõi trong năm 2010.

Để trả giá cho năng lực tính toán, GPU hy sinh tính linh động của các lõi xử lý. Hiện tại các lõi xử lý của GPU tại một thời điểm chỉ thực hiện được một đoạn mã duy nhất trên nhiều luồng, do vậy GPU chỉ thích hợp với những bài toán song song dữ liệu trong đó cùng một đoạn mã chương trình được thực thi song song cho nhiều bộ dữ liệu khác nhau. Rất may là đa số các bài toán yêu cầu năng lực tính toán lớn đều có thể quy về dạng song song dữ liệu này.

Bên cạnh việc phát triển các bộ xử lý đồ họa có năng lực tính toán lớn, các hãng sản xuất cũng quan tâm tới môi trường phát triển ứng dụng cho các bộ xử lý đồ họa này. CUDA[7] là môi trường phát triển ứng dụng cho các bộ xử lý đồ họa của NVIDIA, bao gồm một ngôn ngữ lập trình song song dữ liệu cùng với các công cụ biên dịch, gỡ rối, và giám sát thực thi cho các ứng dụng trên các bộ xử lý này. Dưới đây là một số đặc điểm chính của ngôn ngữ lập trình do CUDA hỗ trợ (gọi tắt là ngôn ngữ CUDA).

– Ngôn ngữ CUDA là mở rộng của ngôn ngữ C, do vậy quen thuộc với đa số người phát triển ứng dụng.

– Mã CUDA chia làm 2 phần: phần thực thi trên CPU và phần thực thi trên GPU. Phần thực thi trên GPU, còn gọi là nhân song song (*kernel*), khi được gọi có thể thực hiện song song trên hàng ngàn luồng (*thread*) riêng biệt. Mỗi luồng có một định danh riêng dùng để xác định nhiệm vụ của luồng đó.

– CUDA cho phép người lập trình tùy ý xác định số lượng luồng song song, tuy nhiên để tránh sự phụ thuộc vào phần cứng, các luồng này cần được phân theo từng khối luồng (*block*) với số lượng không quá 512. Cách phân này giúp người lập trình không cần quan tâm tới năng lực của phần cứng, đồng thời giúp việc tổ chức thực thi được hiệu quả thậm chí là trên các GPU khác nhau.

– Bộ nhớ được tổ chức phân cấp bao gồm các lớp sau:

- Bộ nhớ chính: là vùng bộ nhớ dành cho phần mã CPU. Chỉ có phần mã này có thể truy nhập và sửa đổi thông tin trên đó.

- Bộ nhớ toàn cục GPU: là vùng bộ nhớ mà tất cả các tiến trình của GPU có thể truy nhập. Người lập trình có thể chuyển

dữ liệu từ bộ nhớ chính sang bộ nhớ này thông qua một số hàm thư viện của CUDA. Bộ nhớ này thông thường được sử dụng để lưu trữ các dữ liệu đầu vào và đầu ra cho các luồng song song trên GPU.

- Bộ nhớ chia sẻ: là vùng bộ nhớ mà chỉ các luồng trong cùng một khối luồng mới có thể truy nhập được. Đây là bộ nhớ tích hợp ngay trên chip xử lý nên tốc độ truy nhập dữ liệu cao hơn rất nhiều so với bộ nhớ toàn cục. Bộ nhớ này thường được sử dụng để lưu trữ các dữ liệu chia sẻ tạm thời nhằm tăng tốc quá trình sử dụng bộ nhớ.

- Bộ nhớ cục bộ GPU: Là vùng bộ nhớ được cấp phát cho các biến cục bộ của từng luồng GPU và không thể truy nhập được từ các luồng khác.

4. ỨNG DỤNG GPU TẠO VIDEO 3D ANAGLYPH

4.1. Ý tưởng

Từ bài toán tạo ảnh anaglyph từ mô hình 3D, một vấn đề đặt ra là làm sao để tăng tốc độ render ra ảnh. Như chúng ta đều biết, mỗi ảnh đều được cấu thành từ tập rất lớn các điểm ảnh. Với loại ảnh kích thước trung bình 800x600 cũng có đến 480.000 điểm ảnh. Để render một ảnh cỡ như vậy không tốn quá nhiều thời gian nhưng khi tạo ra một đoạn video thì sẽ tốn một thời gian rất lớn bởi ta sẽ phải render rất nhiều frame ảnh. Có thể tính toán sơ bộ một video dài 5 phút, với tốc độ 25frame/giây, thì ta sẽ phải render số lượng frame ảnh là 7.500 frame. Do đó, chúng tôi tiếp cận theo hướng song song hóa việc render cho từng điểm ảnh, từ đó giảm được thời gian render từng frame ảnh cũng như thời gian render video.

4.2 Giải thuật

Mỗi một frame ảnh sẽ có n điểm ảnh, tùy vào độ phân giải của ảnh mà n lớn hay nhỏ. Với ảnh độ phân giải 800x600 thì số n sẽ là 480.000 điểm ảnh. Mỗi điểm ảnh đều cần tính toán giá trị màu của nó để lưu vào ảnh. Gọi p là số luồng có thể thực thi song song của GPU, số p này bị hạn chế bởi phần cứng. Do đó để render một frame, ta cần gọi (n/p) lần, mỗi lần thực hiện p luồng song song để tính toán giá trị màu của điểm ảnh. Giá trị này sẽ được lưu vào mảng, sau khi hoàn thành tính giá trị màu cho cả một frame thì mảng chứa thông tin màu đó sẽ được ghi vào file ảnh.

5. THỬ NGHIỆM VÀ ĐÁNH GIÁ

Chúng tôi đã thử nghiệm với một số mô hình 3D và kết quả cho thấy có hiệu ứng nổi 3D khi đeo kính 3D.

Thử nghiệm chương trình tuần tự và chương trình song song để so sánh thời gian render ra các frame chúng tôi đã thu được kết quả như sau:

Số frame	Thời gian chạy trên CPU	Thời gian chạy trên GPU
1	12 giây	1.8 giây
15	3 phút	24.6 giây
50	10 phút	1 phút 17 giây
100	20 phút	2 phút 20 giây

Có thể thấy khi số frame render càng nhiều thì thời gian render

trung bình của một frame của chương trình song song sẽ giảm dần.

6. LỜI TRI ÂN

Chúng tôi xin gửi lời cảm ơn đến TS. Nguyễn Hữu Đức và KS. Dương Nhật Tân đã luôn tạo điều kiện và giúp đỡ chúng tôi thực hiện bài báo này.

7. TÀI LIỆU THAM KHẢO

- [1] Paul Bourke, Creating and viewing anaglyph, URL: http://paulbourke.net/texture_colour/anaglyph/, last visited April 2011.
- [2] Mike Kohn, 3D Images, URL: <http://www.mikekohn.net/3dimages.php>, last visited April 2011
- [3] Thanassis Tsiodras, Basic 3D Algorithms URL: <http://users.softlab.ece.ntua.gr/~ttsiod/renderer.html#intro>, last visited April 2011

Mô hình dịch vụ điện toán đám mây BKloud

Lê Quang Hiếu, Hoàng Quốc Nam, Lưu Thị Thùy Nhung

Tóm tắt - Điện toán đám mây là một mô hình điện toán sử dụng các công nghệ máy tính và phát triển dựa vào mạng Internet. Ở mô hình điện toán này, mọi khả năng liên quan đến công nghệ thông tin đều được cung cấp dưới dạng các dịch vụ, cho phép người dùng truy cập, sử dụng các dịch vụ công nghệ mà không cần quan tâm đến cơ sở hạ tầng phục vụ công nghệ đó. Bằng cách chia sẻ sức mạnh điện toán ảo, các mức độ tiện ích được nâng cao vì tận dụng tối đa thời gian nhàn rỗi của máy chủ, và do đó sẽ giảm chi phí đáng kể trong khi tốc độ phát triển của ứng dụng được gia tăng. Dựa trên tiêu chí đó, trong bài báo này, chúng tôi đề xuất một mô hình cung cấp các máy chủ ảo thuộc mức hạ tầng như một dịch vụ (IaaS) của điện toán đám mây dựa trên các gói phần mềm nguồn mở, đưa ra những chiến lược, mô hình phù hợp với nhu cầu của người dùng và kiến tạo một công thông tin điện toán đám mây phục vụ cộng đồng. Chúng tôi đã xây dựng thành công hạ tầng điện toán đám mây với cơ sở hạ tầng gồm một máy chủ (head node) và hai máy dịch vụ (service node), từ đó cung cấp các máy chủ ảo dịch vụ tùy theo nhu cầu người dùng một cách nhanh chóng, tiện lợi nhất và có thể đảm bảo tận dụng tối đa sức mạnh của hệ thống tính toán. Kết quả này phần nào đã mở ra được những hướng tiếp cận rộng lớn hơn với điện toán đám mây tại Việt Nam.

Từ khóa - BKloud, Cloud Computing, IaaS, Cloud

1. GIỚI THIỆU

Từ những năm 1980, khi tính toán hiệu năng cao có được sự quan tâm và bắt đầu phát triển, cho tới những năm gần đây với tốc độ phát triển chóng mặt của Internet, điện toán đám mây đã có những bước tiến vượt bậc.[1] Thuật ngữ điện toán đám mây ra đời bắt nguồn từ nhu cầu sở hữu tài nguyên tính toán cao mà chi phí đầu tư, chi phí bảo trì được tối thiểu hóa ở mức tối đa nhất có thể. So với các công nghệ tính toán hiệu năng cao khác như tính toán phân tán (Distributed Computing), tính toán lưới (Grid Computing), điện toán theo nhu cầu (Utility Computing) thì điện toán đám mây đang dần trở thành xu hướng công nghệ.

Công trình này được thực hiện tại Trung tâm tính toán hiệu năng cao – trường Đại học Bách Khoa Hà Nội

Lê Quang Hiếu, sinh viên lớp Hệ Thống Thông Tin, khóa 52, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (Điện thoại: 0974.616.850, e-mail: hielq89@gmail.com).

Hoàng Quốc Nam, sinh viên lớp Hệ thống thông tin, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (Điện thoại: 0975.308.547, e-mail: hqnam1988@gmail.com).

Lưu Thị Thùy Nhung, sinh viên lớp Hệ thống thông tin, khóa 52, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (Điện thoại: 0982.921.190, e-mail: nhung.luu@gmail.com).

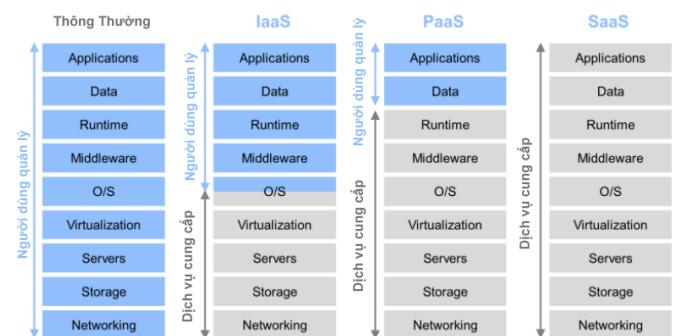
© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

Hình vẽ dưới đây tổng kết xu hướng tìm kiếm trên Internet về các công nghệ hiệu năng cao sử dụng máy tìm kiếm Google.



Hình 1. Bảng so sánh xu hướng công nghệ hiệu năng cao: parallel programming – grid computing – cloud computing sử dụng máy tìm kiếm Google.

Trong điện toán đám mây, mọi khả năng liên quan đến công nghệ thông tin đều được cung cấp dưới dạng các dịch vụ, cho phép người dùng truy cập sử dụng các dịch vụ công nghệ mà không cần phải quan tâm tới cơ sở hạ tầng phục vụ công nghệ đó. Có ba mô hình dịch vụ công nghệ trong điện toán đám mây phổ biến nhất, đó là: mô hình phần mềm như một dịch vụ (Software as a Service - SaaS), mô hình nền tảng như một dịch vụ (Platform as a Service - PaaS) và mô hình hạ tầng như một dịch vụ (Infrastructure as a Service - IaaS). Sử dụng các dịch vụ này, với các cá nhân, tổ chức, chi phí cài đặt, bảo dưỡng, xây dựng cơ sở hạ tầng phục vụ nhu cầu tính toán, nghiên cứu sẽ được giảm tới mức tối thiểu trong khi về phía nhà cung cấp dịch vụ sẽ bảo đảm có thể sử dụng tối đa hiệu năng của hệ thống, không gây lãng phí, mất mát tài nguyên.



Hình 2. Các mô hình dịch vụ điện toán đám mây.

Tuy nhiên, cho tới thời điểm này, vẫn chưa có một định nghĩa chính xác về điện toán đám mây; thông thường điện toán đám mây bị nhập nhằng với tính toán lưới hoặc tính toán phân tán. Trong bài báo này, tiếp cận của chúng tôi đưa ra là một mô hình cung cấp dịch vụ điện toán đám mây ở mức hạ tầng như một dịch vụ - IaaS (Infrastructure as a Service), do đó các định

nghĩa, khái niệm của điện toán đám mây chỉ tập trung quanh mức IaaS, hay còn được hiểu đơn giản hơn là cung cấp các máy chủ ảo dành cho người dùng cuối.

Với hướng tiếp cận như vậy, trong bài báo này, một đám mây được định nghĩa là một khối các máy tính được cấu hình theo cách sao cho bất kỳ người dùng cuối nào cũng có thể tạo một hoặc nhiều máy ảo trên đó với cấu hình thiết lập theo nhu cầu. Đám mây sẽ phân tán các máy ảo đó tới những máy tính thật ở trong nó. Thuật ngữ “đám mây” ở đây một mặt được hiểu bởi tính mờ trong các dịch vụ mà nó cung cấp thông qua Internet, mặt khác bộc lộ được việc người dùng không biết cũng như không quan tâm tới chính xác máy ảo của họ đang nằm ở đâu hay phần cứng của hạ tầng bên dưới được thiết lập như thế nào.

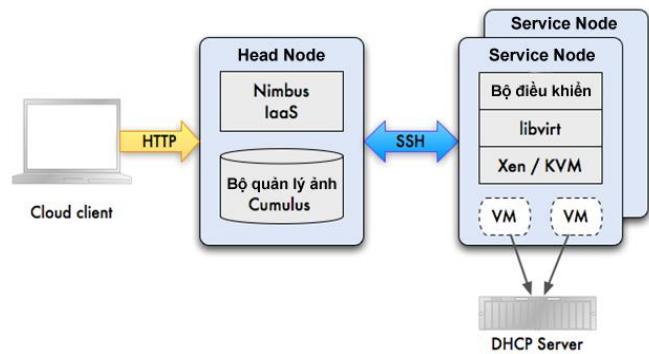
Tuy nhiên, một vấn đề được đặt ra đối với các nhà cung cấp dịch vụ điện toán đám mây là chi phí bỏ ra để xây dựng phần cứng, hạ tầng bên dưới và chi phí duy trì, quản lý dịch vụ bên trên. Trên thế giới có rất nhiều công nghệ giúp quản lý tài nguyên, dịch vụ bên trên như VMWare, Amazon EC2, ... tuy nhiên chi phí bỏ ra đối với các công nghệ trên khá cao đi kèm với khả năng có thể tùy biến còn rất thấp. Để giải quyết vấn đề trên, chúng tôi sử dụng các công nghệ nguồn mở với khả năng tùy biến cao, dễ dàng khi sử dụng.

Các phần tiếp theo của bài báo giới thiệu sơ bộ về mô hình cung cấp dịch vụ điện toán đám mây mà nhóm tiến hành thiết lập, chúng tôi gọi hệ thống đó là Bkloud, và kết quả nghiên cứu của chúng tôi trong việc xây dựng cổng thông tin cung cấp các dịch vụ này.

2. NIMBUS VÀ ĐIỆN TOÁN ĐÁM MÂY

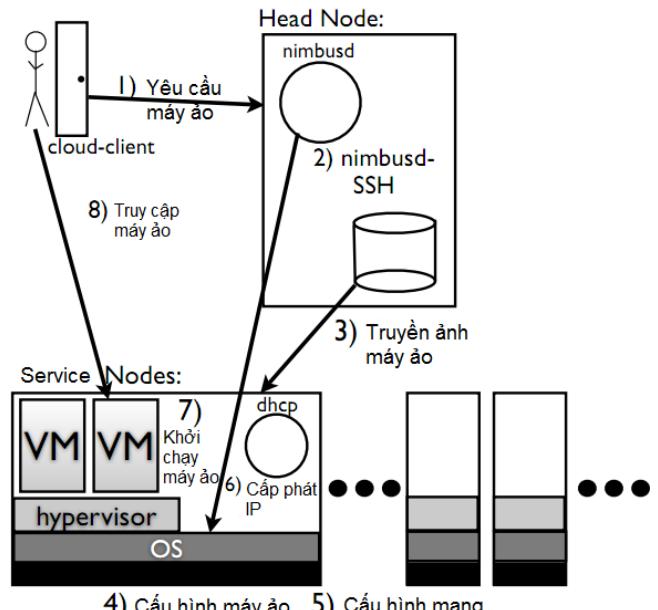
Cùng với sự phát triển nhanh chóng của điện toán đám mây, các dự án, phần mềm nguồn mở phục vụ cho công tác nghiên cứu, phát triển ứng dụng điện toán đám mây cũng liên tục xuất hiện. Với mục đích xây dựng mô hình cung cấp hạ tầng như một dịch vụ, chúng tôi lựa chọn dự án Nimbus bởi khả năng mở rộng cao, dễ tùy biến và dễ tích hợp với các dự án nguồn mở cũng như nguồn đóng khác như: Amazon EC2[10], OpenNebula[9], Eucalyptus[8].. Trên thế giới, Nimbus[7] đã góp mặt trong nhiều dự án khoa học như: SkyComputing[5], OOI (Ocean Observation Initiative)[6],..

Nimbus là một dự án mã nguồn mở được phát triển với tiêu chí xây dựng điện toán đám mây dành cho khoa học[2][3]. Nimbus bao gồm một Head Node có nhiệm vụ tương tác với người dùng, quản lý máy ảo và các Service Node có nhiệm vụ cung cấp tài nguyên cho máy ảo.



Hình 3. Hệ thống Nimbus.

Nimbus có được sự hỗ trợ từ dự án Globus (một dự án mã nguồn mở cho tính toán lưới), do đó Nimbus sử dụng các chứng thực số của Globus (chuẩn X.509) cho xác thực người dùng và hệ thống Cumulus – được cài tiền từ GridFTP của Globus để quản lý các ảnh máy ảo và dễ dàng tương thích với các hệ thống khác như Amazon S3. Một ảnh của máy ảo bao gồm các thiết lập của hệ thống bên trong máy ảo cũng như các dữ liệu cá nhân của người dùng. Khi có yêu cầu từ phía người dùng, Nimbus sẽ sử dụng các thiết lập bên trong để khởi chạy máy ảo. Hình 4 mô tả cách thức hoạt động của Nimbus khi người dùng yêu cầu tạo máy ảo với thiết lập đơn giản nhất.



Hình 4. Cách thức hoạt động của Nimbus.

Các bước khởi tạo một máy ảo của Nimbus bao gồm:

1. Người dùng sử dụng Nimbus cloud-client yêu cầu máy ảo.
2. Nimbus headnode sẽ truy cập tới các node tính toán thông qua phương thức ssh.
3. Tệp ảnh (bao gồm đĩa ảo và các thiết lập bên trong) của máy ảo được truyền tới các node tính toán qua Cumulus.
4. Thiết lập các cấu hình phần cứng cho máy ảo (CPU, RAM,...) theo yêu cầu người dùng.
5. Thiết lập mạng bằng cách gán một địa chỉ MAC ảo cho máy ảo.
6. Yêu cầu cấp phát IP tới DHCP Server.
7. Khởi chạy máy ảo.

8. Người dùng truy cập tới máy ảo thông qua ssh hoặc các phần mềm remote desktop.

Ngoài khả năng cung cấp các máy ảo cho người dùng, Nimbus còn có thể tùy biến các cấu hình bên trong để cung cấp cho người dùng các cụm máy tính cluster ảo, phục vụ cho việc nghiên cứu, lập trình song song. Bởi vậy, với khả năng tùy biến cao, dễ dàng mở rộng cũng như tích hợp với các hệ thống khác, khả năng bảo mật chặt chẽ thông qua chuẩn X.509, Nimbus là một lựa chọn tối ưu để làm nền tảng phát triển cho dịch vụ điện toán đám mây.

Tuy nhiên, một vấn đề với Nimbus là phía người dùng (client) còn rất thụ động khi sử dụng Nimbus do phải sử dụng một phần mềm client của Nimbus.Thêm nữa, việc sử dụng hệ thống hiện nay toàn bộ phải thao tác qua các dòng lệnh, phía người dùng và nhà cung cấp dịch vụ có cái nhìn rất hạn chế về hệ thống mình sử dụng. Để giải quyết các vấn đề đó, bài báo đã đưa ra một mô hình khác, sử dụng những chiến lược tối ưu hơn để có thể đem lại sự tiện dụng, dễ dàng sử dụng, quản lý cho người dùng và nhà cung cấp dịch vụ.

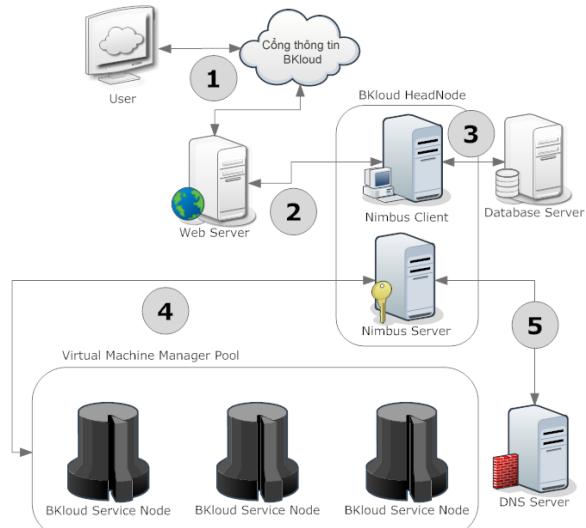
3. CHIẾN LƯỢC CUNG CẤP DỊCH VỤ ĐIỆN TOÁN ĐÁM MÂY VÀ CÔNG THÔNG TIN BKLOUD

Như đã giới thiệu ở phần trước, cách tiếp cận của chúng tôi trong việc xây dựng một mô hình cung cấp dịch vụ điện toán đám mây hạ tầng như một dịch vụ IaaS là tùy biến dự án nguồn mở Nimbus theo những chiến lược do chúng tôi đề xuất, với mục tiêu đem lại cho người dùng sự tiện dụng, dễ sử dụng và các nhà quản lý, cung cấp dịch vụ có một cái nhìn tổng quan nhất về hệ thống.

Từ mục tiêu đó, chúng tôi đã xây dựng công thông tin cung cấp các dịch vụ điện toán đám mây BKloud, giúp người dùng có thể thao tác và quan sát tình trạng máy ảo với các thao tác đơn giản nhất mà vẫn đảm bảo tính mềm dẻo, bảo mật của hệ thống.

3.1. CHIẾN LƯỢC CUNG CẤP DỊCH VỤ

Chiến lược của hệ thống BKloud là thay vì bắt buộc mỗi người dùng phải mang theo gói phần mềm Nimbus cloud-client cùng những thành phần bắt buộc như khóa xác thực chuẩn X.509, các thuộc tính cấu hình khởi động máy ảo, ... thì tất cả được gói gọn vào trong headnode của hệ thống. Mỗi khi người dùng đăng nhập vào hệ thống, BKloud sẽ kiểm tra tài khoản và sẽ lấy chứng thực số chuẩn X.509 tương ứng và cài đặt vào Nimbus client (được đặt ngay tại headnode), do đó Nimbus client có thể xác thực người dùng, nhận biết các máy ảo của người dùng trên hệ thống và tiến hành xử lý các yêu cầu từ phía người dùng.Thêm nữa, khi các nhà quản lý dịch vụ muốn cài đặt thêm số lượng các node cung cấp tài nguyên tính toán (Service node), việc cài đặt và quản lý tài nguyên hạ tầng bên dưới sẽ dễ dàng và đơn giản hơn rất nhiều. Hình 5 mô tả mô hình logic của hệ thống BKloud:



Hình 5. Mô hình logic BKloud

Trong đó, các bước lần lượt khi người dùng sử dụng hệ thống:

- Người dùng đăng nhập vào cổng thông tin BKloud portal và gửi yêu cầu tạo máy ảo tới web server, yêu cầu bao gồm: cấu hình máy ảo (CPU, RAM), hệ điều hành và tổng thời gian thuê máy ảo. Từ các thông tin yêu cầu này, hệ thống sẽ tính toán và đưa ra chi phí phù hợp.
- Web server kiểm tra yêu cầu và chuyển các yêu cầu đến Nimbus client.
- Nimbus client xác thực người dùng bằng cách cài đặt các chứng thực số chuẩn X.509 và các thông số cấu hình cài đặt máy ảo (CPU, RAM,...)
- Sau khi cài đặt thành công, lúc này Nimbus client sẽ nhận biết được các máy ảo của người dùng trên hệ thống BKloud và thông báo cho Nimbus server gửi yêu cầu tới các cụm tài nguyên tính toán quản lý máy ảo phía dưới (Virtual machine manager pool) bao gồm các service node.
- Máy ảo sau khi đã được thiết lập cấu hình sẽ được Nimbus server cấp phát IP thông qua một DHCP-DNS Server.

Cuối cùng, người dùng có thể truy nhập vào máy ảo thông qua địa chỉ IP hoặc tên miền do hệ thống BKloud cung cấp qua giao diện web của cổng thông tin. Trong khi đang sử dụng, người dùng có thể tùy chỉnh lại cấu hình hệ thống như nâng cấp CPU, RAM bằng cách đăng ký và trả thêm phí với hệ thống. Sau khi đã hết thời gian sử dụng, người dùng có thể gia hạn thêm thời gian hoặc ngừng sử dụng.

Như vậy, tổng kết lại, với chiến lược đề ra, mô hình dịch vụ BKloud sẽ có thể:

- Cung cấp cái nhìn trực quan nhất cho người dùng trong việc quản lý máy ảo.
- Thao tác đơn giản, bật tắt máy bằng giao diện web, không phải luôn mang theo các gói phần mềm để có thể sử dụng.
- Cung cấp máy ảo dựa theo nhu cầu và chi phí người sử dụng bỏ ra.

- Một tính năng đặc biệt khác là có thể cung cấp một cluster ảo, bao gồm nhiều máy ảo liên kết, chia sẻ công việc qua lại với nhau. Do đó, có thể cải thiện được hiệu suất công việc, hoặc tạo môi trường nghiên cứu các lĩnh vực khoa học yêu cầu hiệu năng cao như lập trình song song, dự báo bão, dự báo động đất, các bài toán thiên văn ...
- Quản lý tài nguyên tính toán ở hạ tầng phía dưới một cách đơn giản, thuận tiện nhất. Thêm mới hoặc tạm ngắt các node tính toán một cách linh hoạt, không gây ảnh hưởng tới hệ thống.

Tuy nhiên, vì toàn bộ mô hình được triển khai dựa trên tiêu chí mã nguồn mở nên tới thời điểm này, BKloud mới chỉ cung cấp cho người dùng các máy ảo và cách thức sử dụng hoàn toàn trên nền Linux. Các máy ảo cài đặt hệ điều hành Windows có cơ chế hoạt động khác so với Linux nhưng vẫn có thể tích hợp vào mô hình dù tính ổn định chưa cao.

Trong tương lai, có rất nhiều hướng phát triển cho hệ thống BKloud như mở rộng cung cấp các máy ảo nền hệ điều hành Windows, tích hợp với các hệ thống điện toán đám mây khác như OpenNebula, Amazon EC2, tích hợp thêm các thành phần hỗ trợ quản lý điện toán đám mây khác như Amazon S3, ... hoặc làm cơ sở để phát triển mô hình nền tảng như một dịch vụ PaaS (Platform as a Service).

3.2. CÔNG THÔNG TIN BKLOUD

Công thông tin BKloud là một ứng dụng chạy trên nền web mà có khả năng tích hợp và cá nhân hóa các ứng dụng, thông tin và các dịch vụ cộng tác. Công thông tin BKloud cung cấp cho người dùng một điểm truy cập đơn giản tới các nguồn dữ liệu, nội dung và các dịch vụ đa dạng của một đơn vị nghiệp vụ hay các tài nguyên trên mạng internet. Mô hình công thông tin BKloud được xây dựng dựa trên ngôn ngữ PHP, hệ quản trị cơ sở dữ liệu MySQL, tích hợp công nghệ AJAX với tiêu chí có thể phản hồi thông tin theo yêu cầu người dùng một cách nhanh nhất. Công thông tin BKloud được xây dựng dựa trên mẫu kiến trúc kinh điển MVC (Model – View – Controller); tận dụng được tính mềm dẻo và thiết kế logic của MVC, công thông tin BKloud được thiết kế để thuận lợi cho việc phát triển, mở rộng các tính năng và dễ dàng cài đặt.

Các chức năng chính mà công thông tin cung cấp:

- Giúp người dùng dễ dàng đăng ký tài khoản và đăng nhập vào hệ thống, lựa chọn các thông số cho máy ảo, quản lý các máy ảo của mình bằng cách tùy chỉnh các chế độ hiển thị hay bật/tắt các máy ảo theo ý muốn.
- Giúp quản trị viên theo dõi các máy ảo có trong hệ thống, danh sách các người dùng đã đăng ký sử dụng, bật/tắt các máy ảo, thêm/bớt các tài khoản.
- Giúp quản trị viên quản lý tài nguyên, theo dõi tình trạng các node tài nguyên ở hạ tầng bên dưới.

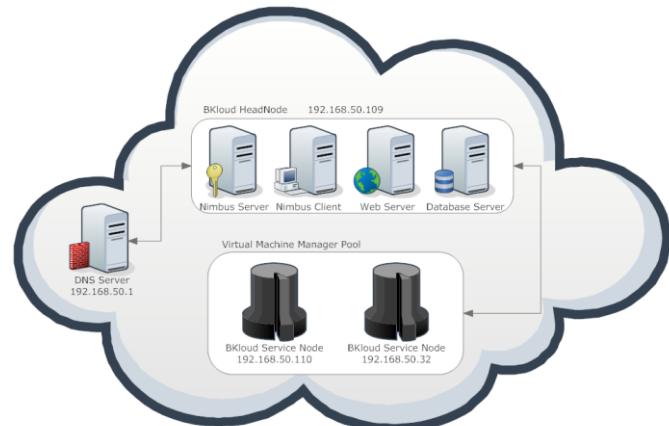
3.3. MÔ HÌNH TRIỂN KHAI - KẾT QUẢ THỰC NGHIỆM

Với mô hình logic được miêu tả ở phần trên, chúng tôi đã tiến hành cài đặt và thử nghiệm trên cơ sở hạ tầng của Trung tâm tính toán hiệu năng cao, trường Đại học Bách Khoa Hà Nội

và đã thu được những kết quả bước đầu đáng khích lệ. Tài nguyên tính toán của hệ thống BKloud bao gồm 3 node liên kết với nhau qua mạng LAN, đường truyền Ethernet 100Mbps, cấu hình chi tiết:

- Headnode:**
CPU: Intel Dual Core 1.80 MHz – không hỗ trợ ảo hóa.
RAM: 1GB DDRII bus 667 MHz
HDD: 80GB
Hệ điều hành: CentOS 5.5 x64
IP: 192.168.50.109
Các thành phần cài đặt: Nimbus Server, Nimbus Client, Apache Webserver, MySQL Database Server.
- Service node 1:**
CPU: Intel Dual Core 1.80 MHz – không hỗ trợ ảo hóa.
RAM: 2GB DDRII bus 667 MHz
HDD: 120GB
Hệ điều hành: CentOS 5.5 x64
IP: 192.168.50.110
Các thành phần cài đặt: Nimbus workspace service, Xen Hypervisor.
- Service node 2:**
CPU: Intel Core 2 Duo 2.80 MHz - hỗ trợ ảo hóa.
RAM: 2GB DDRII bus 667 MHz
HDD: 160GB
Hệ điều hành: CentOS 5.5 x64
IP: 192.168.50.32
Các thành phần cài đặt: Nimbus workspace service, QEMU/KVM Hypervisor.
- DHCP – DNS Server:** Sử dụng máy chủ của HPC – IP: 192.168.50.1.

Do hạn chế trong cơ sở vật chất nên BKloud headnode bao gồm cả Web Server, Database Server, Nimbus Server và Nimbus Client, thêm nữa khi người dùng tạo và sử dụng máy ảo bị giới hạn trong cấu hình CPU tối đa 2 nhân, RAM tối đa 2GB và thời gian sử dụng tối đa là 5 tiếng. Tuy nhiên do đây chỉ là mô hình triển khai thử nghiệm của BKloud nên những kết quả đạt được từ mô hình thử nghiệm này đạt ở mức khả quan. Hình 6 mô tả mô hình triển khai của BKloud:



Hình 6. Mô hình triển khai của BKloud

Với mô hình triển khai như trên, hệ thống BKloud đã vận hành và đạt được một số kết quả khả quan như:

- Khởi tạo và duy trì hoạt động đồng thời 4 máy ảo ở chế độ ổn định, trong đó có 3 máy ở chế độ không có đồ họa (no graphic), chỉ cung cấp dịch vụ ra bên ngoài (SSH, ..) và 1 máy ở chế độ đồ họa, có thể tương tác qua lại với người dùng thông qua các phần mềm như VNCViewer hoặc Vinagre, nhìn vào hình 10 có thể thấy máy ảo Ubuntu đang chạy ở chế độ đồ họa với CPU ảo hóa của QEMU (QEMU Virtual CPU).
- Thử nghiệm tách rời service node 2 có IP 192.168.50.32 khỏi mô hình và bổ sung ngược trở lại một cách dễ dàng.



Hình 7. Trang chủ công thông tin BKloud.

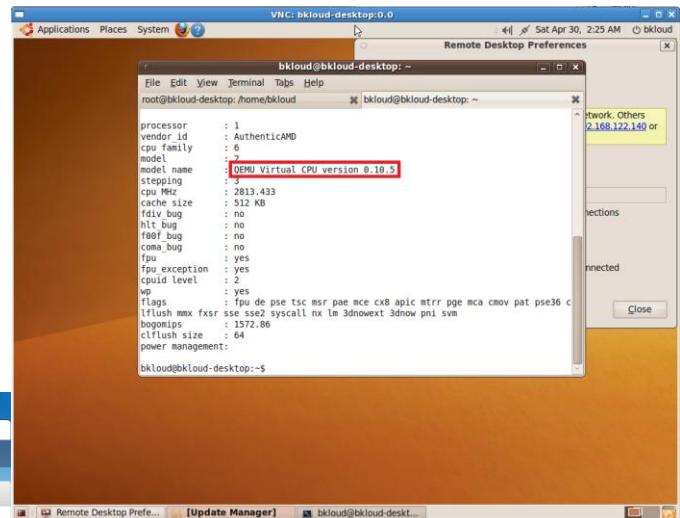
The screenshot shows a 'Khởi tạo Máy Ảo' (Create Virtual Machine) form. It includes fields for 'VPS Name' (Name), 'Chọn Template' (Select Template) set to 'Ubuntu 9.10', 'CPU' (1 Core), 'RAM' (64 MB), and a 'Thời gian sử dụng' (Usage time) field set to '1 / giờ'. There is also a 'Thêm mới' (New) button at the bottom.

Hình 8. Tham số khởi tạo máy ảo.

The screenshot shows a list of managed virtual machines. The table includes columns for ID, Tên miền (Domain name), Hệ điều hành (Operating System), Ngày tạo (Created Date), Thời gian sử dụng (Usage time), and Tinh trạng (Status). The listed machines are:

ID	Tên miền	Hệ điều hành	Ngày tạo	Thời gian sử dụng	Tinh trạng
26	pub02 IP: 192.168.0.2	Bộ vi xử lý 2 Core	05-05-2011	2.00 giờ	Tắt máy
25	Tên mặc định: vm-017 Tên: test	Bộ nhớ trong 512 MB	04-05-2011	0 days 01:57:22	Đang chạy
24	Tên miền: IP:	Bộ vi xử lý 1 Core	04-05-2011	1.00 giờ	Tắt máy
23	Tên miền: IP:	Bộ nhớ trong 64 MB	04-05-2011	0 days 01:57:22	Đang dùng
22	Tên miền: IP:	Bộ vi xử lý 4 Core	04-05-2011	1.00 giờ	Bật máy
21	Tên miền: IP:	Bộ nhớ trong 64 MB	04-05-2011	0.00 giờ	Bật máy

Hình 9. Quản lý máy ảo trên công thông tin BKloud.



Hình 10. Truy cập từ xa vào máy ảo Ubuntu bằng phần mềm VNCViewer

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Với mục tiêu ban đầu đặt ra là xây dựng và triển khai một mô hình BKloud cung cấp hạ tầng như một dịch vụ, các kết quả nghiên cứu đã phần nào chứng minh được tính mềm dẻo, tiềm năng triển khai thực tế của mô hình, đồng thời cho thấy khả năng ổn định và cân bằng của hệ thống BKloud. Với tốc độ phát triển chóng mặt của công nghệ thông tin, khi hạ tầng phần cứng ngày càng được nâng cao về hiệu năng và tốc độ, việc mở rộng và nâng cấp hệ thống cũng không gặp phải khó khăn.

Trong xu thế thị trường hiện nay, nhu cầu ảo hóa và điện toán đám mây là hai vấn đề được quan tâm hàng đầu của các doanh nghiệp và các tổ chức, bởi vậy mô hình BKloud phần nào đã tiếp cận được với xu hướng của thế giới, bắt đầu đặt nền móng phát triển điện toán đám mây tại Việt Nam.

Trong tương lai, ngoài việc mở rộng, nâng cấp dịch vụ như cung cấp thêm các máy ảo chạy đa hệ điều hành, các cluster ảo, hệ thống có thể tích hợp thêm với các dịch vụ điện toán đám mây khác như Amazon EC2, OpenNebula, Eucalyptus hoặc tích hợp các thành phần mở rộng như Amazon S3, Walrus. Ngoài ra, đi kèm với những mở rộng ấy là việc nâng cấp đường truyền mạng nhanh hơn, ổn định hơn, giúp người dùng có thể truy cập sử dụng các dịch vụ điện toán đám mây mọi nơi mọi lúc, nhanh chóng và đơn giản nhất, đối với các nhà cung cấp dịch vụ điện toán đám mây, hệ thống sẽ có thể giao tiếp với nhau nhanh chóng hơn, đem lại hiệu suất công việc cao hơn.

5. LỜI CẢM ƠN

Chúng tôi xin chân thành cảm ơn TS. Nguyễn Hữu Đức, ThS. Lê Đức Tùng cùng các anh, các bạn trên trung tâm tính toán hiệu năng cao trường Đại học Bách Khoa Hà Nội đã tận tình chỉ dẫn, giúp đỡ nhóm trong quá trình hoàn thiện sản phẩm.

6. TÀI LIỆU THAM KHẢO

- [1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brand. Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for

- Delivering Computing as the 5th Utility. Future Generation Computer Systems, 25(6):599–616, June 2009.
- [2] C. Hoffa, G. Mehta, T. Freeman, E. Deelman, K. Keahey, B. Berriman, and J. Good. On the Use of Cloud Computing for Scientific Workflows. SWBES 2008, December 2008.
 - [3] K. Keahey, R. Figueiredo, J. Fortes, T. Freeman, and M. Tsugawa. Science Clouds: Early Experiences in Cloud Computing for Scientific Applications, August 2008.
 - [4] L. Wang, J. Tao, M. Kunze, D. Rattu, and A. C. Castel-lanos. The Cumulus Project: Build a Scientific Cloud for a Data Center. In proceedings of International Conference of Cloud Computing and Applications, October 2008.
 - [5] Sky Computing, Keahey, K., Tsugawa, M., Matsunaga, A., Fortes, J. IEEE Internet Computing, vol. 13, no. 5, September/October 2009.
 - [6] OCEANS 2009, MTS/IEEE Biloxi - Marine Technology for Our Future: Global and Local Challenges.
 - [7] Trang chủ Nimbus: <http://www.nimbusproject.org/>
 - [8] Eucalyptus: <http://open.eucalyptus.com/>
 - [9] OpenNebula: <http://www.opennebula.org/>
 - [10] Amazon EC2: <http://aws.amazon.com/ec2/>

Hệ điều hành hiệu năng cao HPOS

Cao Minh Quỳnh, Nguyễn Đức Minh, Ngô Văn Vĩ

Tóm tắt - Hiện nay nhu cầu cần giải quyết các bài toán phức tạp và xử lý lượng dữ liệu ngày càng lớn, tính toán hiệu năng cao hiện nay là xu hướng mới trên thế giới. Để giải quyết nhu cầu tính toán, việc xây dựng hệ thống hiệu năng cao dần chuyển dịch sang hướng phát triển các hệ thống lớn đều được kết hợp từ các máy tính cỡ nhỏ thông qua mạng và các dịch vụ quản trị hệ thống. Qua phân tích trên, chúng tôi đưa ra một giải pháp xây dựng một phiên bản hệ điều hành hiệu năng cao, mang tên HPOS, nhằm tiết kiệm chi phí, tăng hiệu năng và chất lượng công việc, phục vụ cả nhu cầu thực tế và nhu cầu nghiên cứu. Hệ điều hành của chúng tôi được kế thừa từ hệ thống LiveParaSystem, năm nay chúng tôi đã tiếp tục nghiên cứu phát triển tiếp hệ thống theo hướng tiếp cận gần hơn với các nhu cầu thực tế như xử lý bài toán dữ liệu lớn, cân bằng tài và xây dựng hệ thống theo mô hình Cloud Computing với các giải pháp cụ thể như: mềm dẻo, dễ triển khai, dễ dàng mở rộng; có thể đa khởi động qua mạng, CD, USB ...; thiết lập hệ thống Web Server và Database Server cân bằng tài với độ tin cậy cao, lập lịch xử lý tác vụ; xử lý dữ liệu phân tán với MapReduce; hỗ trợ thêm các ứng dụng nền tảng để triển khai mô hình Cloud Computing.

Ngoài ra nhằm hỗ trợ người dùng tương tác và sử dụng hệ thống, công thông tin (môi trường) thông qua web cũng đã được chúng tôi xây dựng, được cập nhật và theo dõi qua trang web: <http://www.hpos.com.vn>. Hệ thống của chúng tôi đã hoàn thiện và triển khai chạy thử nghiệm tại một số nơi như: Dự án đào tạo về CNTT&TT Việt Nhật, Bộ môn Khoa học máy tính – Viện công nghệ thông tin và truyền thông – ĐH Bách Khoa Hà Nội, Trung tâm tính toán hiệu năng cao của Trường Đại Học Khoa học tự nhiên...Hiệu năng hệ thống đã đạt được những kết quả đáng khích lệ và rất mong được sự hỗ trợ về cơ sở hạ tầng, máy móc để thử nghiệm, cải tiến nâng cao hiệu năng hơn nữa.

Từ khóa: parallel computing, cloud computing, hpc, linux cluster, opensource, ide

I. GIỚI THIỆU

Với nhu cầu cần giải quyết các bài toán phức tạp và xử lý lượng dữ liệu ngày càng lớn, tính toán hiệu năng cao hiện nay là xu hướng mới trên thế giới. Để giải quyết nhu cầu tính toán, các nhà sản xuất lớn trên thế giới không ngừng đầu tư nhằm tăng tốc

Công trình này được thực hiện bởi các tác giả:

Cao Minh Quỳnh, sinh viên lớp Khoa học máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (e-mail: quynhcm@gmail.com)

Nguyễn Đức Minh, sinh viên lớp Khoa học máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (e-mail: quicksort88@gmail.com).

Ngô Văn Vĩ, sinh viên lớp Công nghệ phần mềm, khóa 52, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (e-mail: ngovi.se.fit@gmail.com)

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

độ xử lý (tăng mật độ tích hợp transistor trên chip). Tuy nhiên cách phát triển theo chiều sâu này gấp phải một trở ngại lớn là chi phí đầu tư quá lớn. Chính vì vậy mà xu hướng của thế giới hiện nay trong việc xây dựng hệ thống hiệu năng cao dần chuyển dịch sang hướng phát triển theo chiều rộng. Các hệ thống lớn hiện nay đều được kết hợp từ các máy tính cỡ nhỏ thông qua mạng và các dịch vụ quản trị hệ thống.

Ở Việt Nam, lập trình song song đã từng bước được đưa vào giảng dạy tại các trường đại học và bước đầu có ứng dụng thực tế. Tuy nhiên, đây là một hướng đi mới, tài liệu về lập trình song song chưa nhiều và không tập trung. Bên cạnh đó, môi trường để trao đổi kiến thức cũng chưa có, mới chỉ có một số website nhỏ nhưng không mang tính mở cho người dùng trao đổi. Các sinh viên còn gặp nhiều khó khăn với môi trường thực hành vì xây dựng một hệ thống song song không hề đơn giản về mặt kỹ thuật và tốn kém về mặt chi phí. Một sản phẩm giải quyết được vấn đề này và có thể áp dụng vào thực tế là một nhu cầu tất yếu.

Hiện nay trên thế giới có 2 dòng sản phẩm chính là miễn phí và thu phí. Nổi bật trong các giải pháp thu phí là: Windows HPC Server, Platform Solution, Redhat solution,... Bên phía sản phẩm miễn phí có: Ganglia, Rock Cluster, Pelicanhpc. Mỗi dòng sản phẩm lại có những điểm mạnh và điểm yếu riêng. Dòng sản phẩm thu phí có chất lượng dịch vụ cao, sản phẩm tốt và ổn định nhưng giá thành lại rất đắt, còn dòng sản phẩm miễn phí khó sử dụng hơn và triển khai phức tạp với người sử dụng.

Qua phân tích đặc tính kỹ thuật của 2 dòng sản phẩm, chúng tôi đưa ra một giải pháp tốt hơn nhằm tiết kiệm chi phí, tăng hiệu năng và chất lượng công việc, phục vụ cả nhu cầu thực tế và nhu cầu nghiên cứu. Trong báo cáo này các giải pháp của chúng tôi đưa ra bao gồm:

- Xây dựng một hệ điều hành nền mã nguồn mở
- Thiết lập nhanh, ổn định hệ thống tính toán hiệu năng cao
- Thiết lập hệ thống Web Server và Database Server cân bằng tài với độ tin cậy cao
- Quản trị tập trung
- Lập lịch xử lý tác vụ
- Mềm dẻo, dễ triển khai, dễ dàng mở rộng
- An ninh cao
- Trong suốt với người dùng
- Xử lý dữ liệu phân tán với MapReduce



Figure 1. Màn hình khởi động HPOS

Trong báo cáo này, chúng tôi đề xuất HPOS, một phiên bản hệ điều hành Linux, có khả năng xây dựng các hệ thống tính toán hiệu năng cao một cách nhanh chóng, tin cậy, ổn định và hiệu quả nhất. Bên cạnh đó chúng tôi muốn xây dựng một hệ thống tính toán hiệu năng cao phù hợp với tình hình tại Việt Nam như:

- Miễn phí hoặc giá rẻ bao gồm các chi phí tối thiểu.
- Hướng đến giải quyết hiệu quả các bài toán thực tế cấp thiết hiện nay như: cân bằng tải, xử lý dữ liệu lớn...
- Dễ tiếp cận, thân thiện với người dùng
- Cộng đồng một cộng đồng phát triển lớn mạnh (wiki, blog).

Phản còn lại của bài báo được tổ chức như sau. Thiết kế hệ thống được đề cập trọng phần II. Các chi tiết khi cài đặt được mô tả trong phần III. Trong phần IV, chúng tôi trình bày kết quả thu được khi chạy thử hệ thống tín toán trên nền HPOS và hướng phát triển. Cuối cùng, kết luận được đặt ở phần V.

II. THIẾT KẾ HỆ THỐNG

Hệ điều hành HPOS được xây dựng sử dụng các giải pháp opensource nhằm giảm thiểu thời gian phát triển và đón nhận sự trợ giúp từ cộng đồng phát triển.

- Hệ điều hành HPOS xây dựng trên nền LFS [3, 6]
- Cụm máy tính tự động cấu hình HPCluster.
- Xử lý dữ liệu phân tán với MapReduce trên Hệ điều hành HPOS
- Kho chứa mã nguồn được đặt tại Google Code.

A. Hệ điều hành mã nguồn mở HPOS

HPOS là một hệ điều hành được thiết kế để làm việc với các cụm máy tính (cluster) kết nối với nhau thông qua mạng nội bộ (LAN). Thêm vào đó, HPOS hỗ trợ đa khởi động hệ thống từ USB/CD, đặc biệt có thể điều khiển khởi động các máy tính trạm trong hệ thống qua mạng LAN từ phần mềm quản trị máy chủ, nên cluster có thể dựng rất nhanh và không cần ổ cứng.

HPOS sử dụng cách tiếp cận của Linux From Scratch, xây dựng Linux dựa trên nhiều lớp phần mềm mã nguồn mở thành công và ổn định, bao gồm nhân Linux, các công cụ của GNU và giao diện Gnome.

Một tính năng khác giúp nâng cao tính mềm dẻo của HPOS là khả năng kết nối với các máy tính không sử dụng HPOS. Các tính năng cốt lõi của HPOS có thể đóng gói thành HPOS Pack giúp các hệ thống đã có trên nền Linux như Red Hat Enterprise Linux, CentOS, ... kết nối được với máy chủ HPOS và vận hành như một máy trạm HPOS.

Ngoài chức năng là một hệ điều hành thông thường giống như Windows của Microsoft, Mac OS của Apple và một số bản phân phối Linux khác, HPOS tập trung vào việc hỗ trợ giải quyết các bài toán tính toán xử lý dữ liệu lớn, cân bằng tải trong suốt với người dùng cũng như các dịch vụ quản trị thông tin của cụm máy tính. Chính vì thế HPOS là một hệ điều hành hỗ trợ rất mạnh cho người dùng có nhu cầu tính toán hiệu năng cao, xu hướng chủ chốt trên thế giới.



Figure 2. Mô hình xây dựng hệ điều hành HPOS

Kết hợp với đặc thù bảo mật tốt hơn, qua đó ổn định hơn, của các phiên bản Linux so với Windows, HPOS là một lựa chọn đúng đắn và tiết kiệm cho các doanh nghiệp cũng như các tổ chức có nhu cầu xử lý, tính toán thông tin khối lượng lớn.

HPOS được phát hành với giấy phép công cộng GNU/GPL version 2.

B. Cụm máy tính HPCluster

Qua hơn 2 năm nghiên cứu và thử nghiệm công nghệ, cluster trên nền HPOS đã hoàn thiện phiên bản 1.0 thành công trong việc cấu hình tự động cluster, làm việc tập trung tại máy chủ và quản trị một phần tài nguyên của cụm máy tính.

Phiên bản hiện tại HPOS hỗ trợ các máy tính kiến trúc Intel x86. Cấu hình tối thiểu để chạy HPOS là:

Kiểu cài đặt	RAM	HDD	CPU
Thực thi trực tiếp	256 MB	0 GB	> 1 GHz, x86, 32-bit
Cài vào ổ cứng	128 MB	2 GB	> 1 GHz, x86, 32-bit

Để vận hành cụm máy HPOS cần

- Một (hoặc hai) máy chủ.
- Nhiều hơn một máy trạm tham gia tính toán.
- Switch kết nối các máy trạm và máy chủ
- Switch kết nối máy chủ và mạng Internet.

1) Mô hình mạng

HPOS được xây dựng trên nền mạng LAN, trong đó các máy trạm tham gia tính toán được kết nối trực tiếp với nhau giúp giảm thiểu thời gian trao đổi dữ liệu giữa các máy.

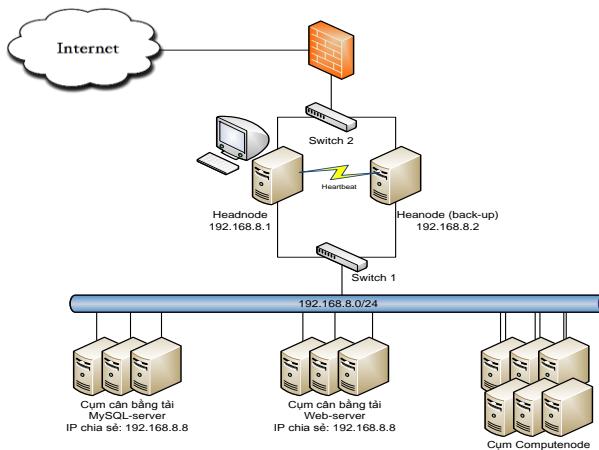


Figure 3. Mô hình cluster dựng trên nền HPOS

a) Switch 1

Kết nối nội bộ cụm máy, không kết nối với các router khác.

b) Switch 2

Kết nối nội bộ máy chủ và đường ra Internet.

c) Tường lửa (Firewall)

Tường lửa trong cụm máy HPOS được đặt trên máy chủ chỉ cho phép truy cập hạn chế từ bên ngoài để quản trị cụm máy:

- Cổng 22 Secure Shell (SSH) quản trị cụm máy qua giao diện dòng lệnh
- Cổng 10000 Quản trị cấu hình cụm máy thông qua giao diện web

2) Triển khai cụm máy tính HPOS

1. Kết nối máy chủ và các máy trạm vào Switch_1
2. Nếu máy chủ muốn ra Internet thì sử dụng thêm một card mạng và kết nối vào Switch_2.
3. Khởi động máy chủ từ đĩa CD/USB, chọn chế độ server.
4. Khởi động máy trạm từ đĩa CD/USB, chế độ client được chọn theo mặc định.
5. Quản trị cụm máy tính tại máy chủ
6. Thông qua Switch_2, quản trị mạng có thể thao tác với máy chủ bằng SSH thông qua LAN hoặc Internet.

Chú ý: Khi chọn chế độ boot qua mạng các máy thợ cần được ngắt kết nối với các thiết bị có cung cấp dịch vụ DHCP.

3) Công cụ quản trị cụm máy tính

Toàn bộ hoạt động của cụm máy tính được điều khiển thông qua công cụ quản trị “HPOS Cluster Management”

Công cụ quản trị sẽ chia thành 3 mục chính:

- Node Management: quản trị toàn bộ máy tính trong cụm máy
- Job Management: cập nhật thông tin, bổ sung công việc cho cụm máy
- Report: thống kê hoạt động của hệ thống

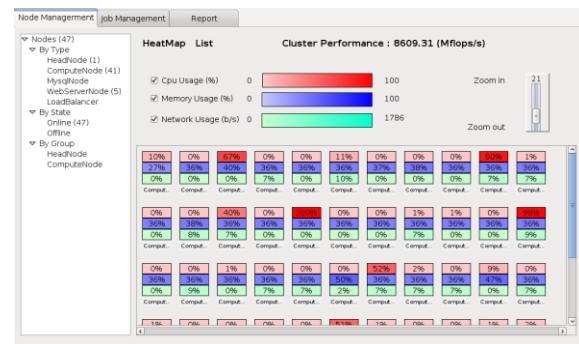


Figure 4: Heatmap

Công cụ này là phần mềm do nhóm tự phát triển và đã tích hợp vào hệ điều hành HPOS (trên bản dành cho máy chủ Headnode) giúp biểu diễn các node dưới dạng đồ họa, giúp người quản trị có cái nhìn sinh động, tổng quan về tình trạng hoạt động của cụm máy tính ví dụ như: có thể chọn xem mức độ sử dụng của cpu, ram, network của từng máy tính trong cụm máy tính, phân chia công việc cho từng máy trong cụm máy tính, báo cáo thống kê tình trạng hoạt động cả cụm máy tính.

4) Cài đặt

Bên cạnh việc chạy trực tiếp từ CD/USB, HPOS cho phép người dùng cài đặt hệ điều hành vào ổ cứng để thuận tiện cho việc sử dụng cũng như xây dựng chương trình giải các bài toán.

Chạy trực tiếp HPOS từ ổ cứng không tốn nhiều RAM như khi chạy trực tiếp từ CD/USB và giúp tăng hiệu năng hệ thống.

a) Cài đặt vào đĩa cứng

Quản trị viên cài đặt HPOS từ trình cài đặt trên Desktop. Cũng có thể tiến hành cài đặt từ dòng lệnh như sau:

```
#Remastersys-installer
```

b) Backup hệ thống

Quá trình xây dựng và vận hành hệ thống tốn rất nhiều thời gian và công sức do vậy cần thiết phải backup đảm bảo an toàn dữ liệu hệ thống trong trường hợp xảy ra sự cố.

HPOS phiên bản server hỗ trợ hình thức backup toàn bộ (full backup) hệ thống tới thiết bị lưu trữ trên mạng, CD, hoặc USB. Quá trình backup và restore chỉ diễn ra trong vòng 5-10 phút.

Phiên bản HPOS-server có thể được backup với kích thước vừa đủ một đĩa CD. Công cụ backup trong HPOS-server cho phép tạo file .iso là file ảnh của HPOS-server.

```
#remastersys backup /home/hpos/custom.iso
```

Khi xảy ra sự cố, quản trị viên chỉ cần khởi động từ đĩa CD backup và kích hoạt chương trình phục hồi HPOS từ Desktop.

Quản trị viên cũng có thể chạy lệnh

```
#remastersys-installer
```

5) Các công nghệ sử dụng

Các công nghệ sử dụng trong quá trình xây dựng cluster là:

a) Hệ thống SSH và SSHFS

SSH và SSHFS giúp các máy tính sử dụng HPOS giao tiếp một cách dễ dàng và bảo mật. Ngoài ra, người quản trị có thể điều khiển cụm máy được đặt ở xa bằng việc truy nhập vào máy chủ mà không lo lắng về tính bảo mật của thông tin trên đường truyền Internet.

b) Dynamic host configuration protocol (DHCP)

Dhcp là một giao thức tầng ứng dụng trong mô hình mạng được sử dụng phục vụ việc thu thập thông tin cấu hình trong mạng. Giao thức này giảm tải cho máy chủ, cho phép thiết lập cấu hình một cách tự động cho các thiết bị mạng mới được kết nối.

HPOS sử dụng Dhcp cấp IP và đánh số cho các máy thợ làm việc trong cụm máy tính. Qua đó quản lý toàn bộ tài nguyên của cụm máy theo các chỉ số này.

Phiên bản DHCP được HPOS sử dụng là udhcp, với ưu điểm nhỏ gọn và dễ cấu hình. Cấu hình hệ thống như sau:

- Máy chủ được cấp IP tĩnh là: 192.168.8.1
- Máy thợ được cấp IP động trong dải: 192.168.8.2-254

c) MPI & OpenMPI

Message passing interface (MPI) [8] là chuẩn chung để xây dựng API cho phép nhiều máy tính trao đổi các thông điệp với nhau. MPI được sử dụng rộng rãi trong các cụm máy tính và các siêu máy tính.

OpenMPI là một cài đặt mã nguồn mở cho MPI được phát triển bởi cộng đồng nghiên cứu, hàn lâm và các doanh nghiệp. Chính vì vậy mà OpenMPI được xây dựng bởi công nghệ và tài nguyên của tất cả các lĩnh vực trong cộng đồng tính toán hiệu năng cao.

HPOS lựa chọn MPI và OpenMPI bởi tính phổ dụng và ổn định.

d) SquashFS

Là một hệ thống file chỉ đọc cho Linux. SquashFS nén file và thư mục cùng các thành phần khác với khả năng nén rất tốt.

Các hệ điều hành thông thường chiếm dụng nhiều tài nguyên ổ cứng, để có thể nén lại trên một đĩa CD, cần thiết phải sử dụng SquashFS. Và HPOS cũng không phải ngoại lệ.

e) UnionFS

Là một hệ thống file, giành cho Linux, cho phép các thư mục và tập tin khác của các hệ thống file khác nhau có thể làm việc được với nhau.

Khi giải nén HPOS từ LiveCD, cần thiết phải sử dụng UnionFS để giả lập việc ghi xóa dữ liệu (nếu chạy trực tiếp trên LiveCD).

f) PXE Booting with LiveCD

PXE (Preboot eXecution Environment hoặc Pre-eXecution Environment) là một môi trường cho phép khởi động máy tính bằng việc sử dụng card mạng cùng với RAM. Việc khởi động đó sẽ không phụ thuộc vào những thiết bị của máy tính như CD, harddisk và các hệ điều hành đã được cài đặt

HPOS đã sử dụng PXE để khởi động qua mạng, trước hết cần khởi động HPOS-server. Sau đó bật chế độ PXE-boot trong BIOS các máy trạm và khởi động, máy trạm sẽ tự động khởi động HPOS-client. Kỹ thuật này cho phép người quản trị hệ thống có thể điều khiển bật tắt, khởi động các máy client trong cluster từ server chỉ bằng những thao tác click chuột đơn giản mà không phải thực hiện thủ công như thông thường.

C. Xử lý dữ liệu phân tán với MapReduce trên hệ điều hành HPOS

MapReduce là một “mô hình lập trình” (programming model), lần đầu được giới thiệu trong bài báo của Jefferey Dean và Sanjay Ghemawat ở hội nghị OSDI 2004[9]. Đó là mô hình giúp xử lý tập hợp dữ liệu siêu lớn đặt tại các máy tính phân tán, có thể xử lý được cả dữ liệu không cấu trúc (dữ liệu lưu trữ dạng tệp tin hệ thống) và dữ liệu cấu trúc (dữ liệu quan hệ 2 chiều), cho phép xử lý song song trên các tập dữ liệu lớn giúp tiết kiệm chi phí xây dựng máy chủ lưu trữ dữ liệu; phát triển điện toán mây.... Trong MapReduce, các máy tính chứa dữ liệu đơn lẻ được gọi là các nút (node), định nghĩa dữ liệu (cấu trúc và không cấu trúc) dưới dạng cặp khóa/giá trị (key/value). Ví dụ, key có thể là tên của tập tin (file) và value nội dung của tập tin, hoặc key là địa chỉ URL và value là nội dung của URL,... Việc định nghĩa dữ liệu thành cặp key/value này linh hoạt hơn các bảng dữ liệu quan hệ 2 chiều truyền thống (quan hệ cha – con hay còn gọi là khóa chính – khóa phụ).

Để xử lý khối dữ liệu bao gồm rất nhiều cặp (key, value), lập trình viên viết hai hàm map và reduce. Hàm map có đầu vào là một cặp (k1, v1) và đầu ra là một danh sách các cặp (k2, v2). Như vậy hàm Map có thể được viết theo dạng:

map(k1,v1) => list(k2,v2).

Và hàm reduce có dạng:

reduce(k2, list (v2)) => list(v3).

MapReduce cho phép lập trình viên dễ dàng sử dụng thư viện định tuyến để lập trình song song chính xác và hiệu quả, không phải bận tâm đến việc trao đổi dữ liệu giữa các cluster khác nhau vì sự độc lập dữ liệu khá cao; không phải theo dõi xử lý lỗi, các tác vụ...

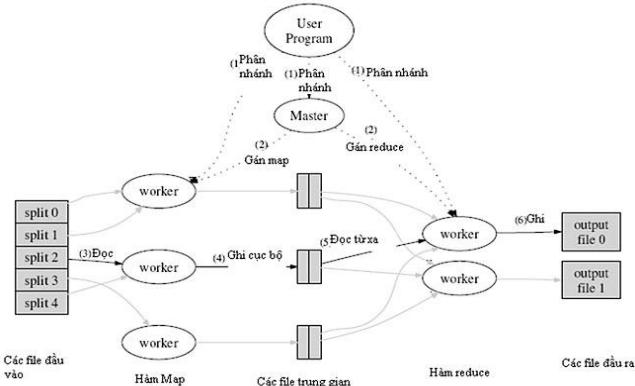


Figure 5: Mô hình MapReduce

(1): Thư viện MapReduce mà chương trình người dùng (User Program) sử dụng chia các tập tin đầu vào (dữ liệu cần xử lý) thành các phần nhỏ. Dung lượng mỗi phần từ 16 megabytes đến 64 megabytes (MB). Và sau đó sao chép chương trình thành các tiến trình song song chạy trên các máy tính phân tán chứa dữ liệu.

(2): Chương trình điều khiển Master sẽ gán mỗi phần dữ liệu cho một hàm Map và một hàm Reduce.

(3) – (4): worker là phần được gán một hàm Map và Reduce để xử lý, nó sẽ đọc dữ liệu, phân tích cặp key/value ở đầu vào và phân tích thành các cặp trung gian khác được lưu tại vùng nhớ đệm.

(5): Định kỳ, các cặp dữ liệu trung gian sẽ được đẩy đến các worker tương ứng (do master điều khiển) để hàm reduce xử lý. Các thuật toán sắp xếp, so sánh, phân vùng dữ liệu sẽ được sử dụng tại giai đoạn này. Các tập dữ liệu trung gian có cùng key sẽ được sắp xếp cùng một nhóm.

(6): Khi tất cả các tác vụ Map và Reduce đã hoàn tất thì sẽ cho ra kết quả cuối cùng của quy trình MapReduce.

Để sử dụng được mô hình MapReduce thì chúng tôi đã tiến hành cài đặt gói phần mềm mã nguồn mở Hadoop trên hệ điều hành HPOS, đây là một giải pháp triển khai MapReduce do Apache xây dựng. Hiện nay, rất nhiều trang web lớn trên thế giới đang sử dụng Hadoop để xử lý khối dữ liệu đồ sộ của họ như Google, Amazon, Yahoo, Facebook...

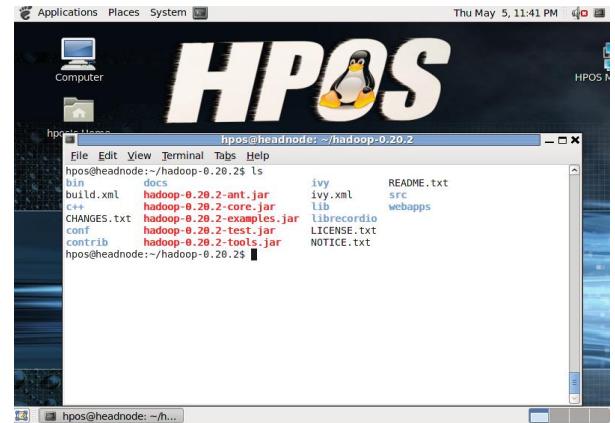


Figure 6: Cài đặt Hadoop trên HPOS

III. KẾT QUẢ THỰC NGHIỆM VÀ HƯỚNG PHÁT TRIỂN

Hệ điều hành HPO đã được chúng tôi đưa vào thử nghiệm triển khai thực tế tại một số nơi và có kết quả như sau:

Tại Trung tâm máy tính của dự án Đào tạo Việt Nhật: nhóm đã triển khai HPOS trên 40 máy tính bộ HP, với cấu hình Core 2 Duo 2Ghz, 1 Gb Ram và ổ cứng 80Gb, với giải pháp là khởi động HPOS và thực thi trực tiếp trên đĩa CD, ổ USB và khởi động qua mạng. Có 20 máy tính chạy từ đĩa CD, 10 máy chạy từ USB và 10 máy khởi động qua mạng. đã được nhóm chạy kiểm thử các tính năng: quản trị từ xa, quản trị công việc, bật tắt từ xa, xử lý bài toán word count với file dữ liệu là 2GB. Các tính năng trên đều vận hành tốt và hiệu năng tổng của cả hệ thống đạt 8 Giga FLOPS.

Tại phòng lab bộ môn Khoa Học Máy Tính, nhóm phát triển sau khi tiến hành cấu hình Hadoop hoàn thiện trên hệ điều hành HPOS, chúng tôi bắt đầu tiến hành cài đặt bài toán Cây khung nhỏ nhất với đường kính bị chặn (Bounded Diameter Minimum Spanning Tree - BDMST) và chạy thử nghiệm trên cụm máy tính để giải bài toán này. Với hướng tiếp cận mô hình hóa giải thuật di truyền bền vững theo mô hình MapReduce, số lượng cá thể trong một thế hệ của một quần thể có thể tăng lên 10^7 cá thể. Trải qua nhiều thế hệ, số cá thể mới được sinh ra có thể lên tới hàng trăm triệu, từ đó thu được các kết quả rất khả quan. Chúng tôi đã tiến hành chạy trên các bộ test chuẩn và sẽ thống kê, đánh giá kết quả trong thời gian sớm nhất.

Trong thời gian tới, nhóm phát triển sẽ tiếp tục nghiên cứu, phát triển hệ điều hành HPOS để nâng cao hiệu năng của hệ thống hơn nữa. Bên cạnh đó chúng tôi cũng sẽ cài đặt nền tảng tính toán đám mây mã nguồn mở Eucalyptus (Elastic Utility Computing Architecture for Linking Your Programs To Useful Systems) Public Cloud, là một bản thực thi mã nguồn mở của Amazon Elastic Compute Cloud (EC2), trên HPOS để có thể phát triển theo hệ điều hành hướng Cloud Computing trong tương lai.

IV. KẾT LUẬN

Báo cáo trên đây là cái nhìn tổng quan về hệ điều hành tính toán hiệu năng cao HPOS, các điểm mạnh và yếu của HPOS. Với nhiệt huyết của mình cùng với sự giúp sức của cộng đồng, nhóm sẽ cố gắng xây dựng một sản phẩm tốt, hoàn thiện và phù hợp với người sử dụng.

V. PHỤ LỤC

- VNOFOSS – Vietnam Free Open Source Software
- IDE – Integrated Development Environment
- HPC – High Performance Computing
- MPI - Message Passing Interface

VI. LỜI TRI ÂN

Chúng tôi xin chân thành cảm ơn:

Thầy giáo Ngô Duy Hòa, người hướng dẫn dự án từ những ngày đầu tiên, đã định hướng và khuyên bảo nhóm nhiều điều có ích.

Thầy giáo Nguyễn Việt Huy, người đã hướng dẫn nhóm tham dự cuộc thi VNFOSS.

Cô giáo Huỳnh Thị Thanh Bình và thầy giáo Nguyễn Đức Nghĩa đã hướng dẫn và tạo điều kiện cho nhóm xây dựng cluster tại bộ môn Khoa Học Máy Tính, ĐHBK Hà Nội.

Các bạn Nguyễn Thành Trung, Phan Đình Phúc, Cao Minh Phương và nhóm PCGHUT đã tham gia trong suốt quá trình phát triển dự án.

Các bạn Nguyễn Toàn Thắng, Trần Văn Luân đã cùng cộng tác, tham gia phát triển và duy trì cổng thông tin.

Cuối cùng là cộng đồng mã nguồn mở tại Việt Nam đã đóng góp nhiều ý kiến quý báu cho nhóm trong quá trình hoàn thiện sản phẩm.

VII. TÀI LIỆU THAM KHẢO

- [1] “An open-source Linux Cluster distribution” <http://www.rocksclusters.org>.
- [2] Các công cụ hỗ trợ lập trình C/C++ và phát triển plugin cho eclipse.
<http://eclipse.org/>
- [3] Gerard Beekmans (2007): “Linux From Scratch, version 6.3”, “Beyond Linux From Scratch, version 6.2”, US: LFS, <http://linuxfromscratch.org>
- [4] Negus Christopher (2005) “Live Linux CDs: Building and Customizing Bootables”
- [5] Tomas M., “The ultimate way to bring your linux to life”, Linux Live for CD&USB,
- [6] Marcelvdboer, “Linux From Script”, <http://www.marcelweb.nl/lfscrip>
- [7] Gatech, “Travelling Sales Man Problem Test Data”,
<http://www.tsp.gatech.edu/data/index.html>
- [8] Wikipedia, “Message Passing Interface”,
http://en.wikipedia.org/wiki/Message_Passing_Interface
- [9] MapReduce: Simplified Data Processing on Large Clusters ,
<http://labs.google.com/papers/mapreduce.html>

Giải thuật di truyền lai giải bài toán phủ đỉnh

Nguyễn Hữu Phước

Tóm tắt - Bài toán phủ đỉnh là một bài toán thuộc lớp NP-Khó, có rất nhiều phương pháp khác nhau đã được áp dụng để giải quyết bài toán phủ đỉnh, một trong các hướng đó là sử dụng giải thuật di truyền. Các kĩ thuật tối ưu hóa cục bộ (Local Optimize) thường được áp dụng vào giải thuật di truyền nguyên thủy để tạo ra giải thuật di truyền lai. Các giải thuật di truyền lai đa phần cho kết quả tốt hơn và có tốc độ hội tụ nhanh hơn so với giải thuật di truyền nguyên thủy. Đề tài nghiên cứu này đề xuất một phương pháp tối ưu mới One Point Optimize (OPO) áp dụng hiệu quả vào giải thuật di truyền lai trong bài toán phủ đỉnh. Thực nghiệm trên bộ dữ liệu BHOSLIB (Benchmark with Hidden Optimum Solutions) cho kết quả tốt hơn khi so sánh với các giải thuật di truyền lai đã được công bố trong các bài báo của Huo Hongwei [1] và Keta Kotechan [2].

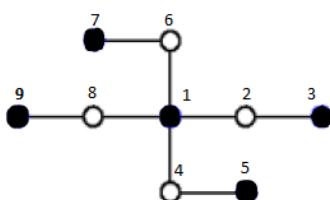
Từ khóa—hybrid genetic algorithm, minimum vertex cover, one point optimize, local optimize

1. GIỚI THIỆU

Gọi một phủ đỉnh của đồ thị vô hướng $G = (V, E)$ là một tập con các đỉnh của đồ thị $S \subseteq V$ sao cho mỗi cạnh của đồ thị có ít nhất một đầu mút trong S .

Bài toán yêu cầu tìm ra phủ đỉnh có số đỉnh nhỏ nhất của một đồ thị cho trước, vì là một bài toán NP-đầy đủ nên không tồn tại thuật toán để giải nó trong thời gian đa thức. Có khá nhiều thuật toán xấp xỉ đã được áp dụng, ví dụ giải thuật tham lam cơ bản sau:

Chọn đỉnh có bậc cao nhất và loại bỏ đỉnh đó và tất cả các cạnh kề với nó ra khỏi đồ thị, lặp lại thao tác trên đến khi không còn cạnh nào trên đồ thị. Có thể dễ dàng chỉ ra một trường hợp thuật toán trên cho kết quả không tối ưu. Với đồ thị trong hình 1, đầu tiên chọn đỉnh 1 là đỉnh có bậc cao nhất, loại 1 và các cạnh $(1, 6)$, $(1, 2)$, $(1, 4)$ và $(1, 8)$ ra khỏi đồ thị. Sau đó tiếp tục chọn các đỉnh 7, 3, 5 và 9. Như vậy phủ đỉnh tìm được là $(1, 3, 5, 7, 9)$ trong khi phủ đỉnh tối ưu là $(2, 4, 6, 8)$.



Hình 1 – Giải thuật tham lam

Nguyễn Hữu Phước, sinh viên lớp Khoa học máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0947874117, e-mail: phuocnh88@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

Có khá nhiều phương pháp xấp xỉ khác như ưu tiên chọn cạnh, tìm cặp ghép cực đại... Những phương pháp này tuy có ưu điểm là có độ phức tạp thấp, thời gian đáp ứng nhanh nhưng kết quả đa phần là không tối ưu, và có thể rất tồi trong một số trường hợp đồ thị được xây dựng đặc biệt.

Một hướng tiếp cận khác với bài toán phủ đỉnh là sử dụng giải thuật di truyền. Giải thuật di truyền (GA-Genetic Algorithms) là giải thuật tìm kiếm, chọn lựa các giải pháp tối ưu để giải quyết các bài toán thực tế khác nhau, dựa trên cơ chế chọn lọc của tự nhiên: từ tập lời giải ban đầu, thông qua nhiều bước tiến hoá, hình thành tập lời giải mới phù hợp hơn, và cuối cùng dẫn đến lời giải tối ưu toàn cục.

Sơ đồ của giải thuật di truyền:

- Khởi tạo một quần thể ban đầu (các đáp án ban đầu của bài toán).
- Xác định giá trị hàm mục tiêu (*fitness*) cho mỗi cá thể trong quần thể.
- Tạo ra quần thể mới bằng cách lai ghép chéo (*crossover*) từ các cá thể hiện tại có chọn lọc (*selection*), đồng thời tạo ra các đột biến (*mutation*) trong quần thể mới theo một xác suất nhất định. Các cá thể tốt nhất có thể được bảo toàn sang thế hệ tiếp theo.
- Các cá thể trong quần thể mới sinh ra được thay thế cho các cá thể trong quần thể cũ.
- Nếu điều kiện dừng thỏa thì giải thuật dừng lại và trả về cá thể tốt nhất cùng với giá trị hàm mục tiêu của nó, nếu không thì quay lại bước 2.

Khác biệt quan trọng giữa tìm kiếm của GA và các phương pháp tìm kiếm khác là các phương pháp khác chỉ xử lý một điểm trong không gian tìm kiếm, còn GA duy trì và xử lý một tập các lời giải (một quần thể), các toán tử di truyền sẽ giúp trao đổi thông tin giữa các vùng (lai ghép) hoặc chuyển sang tìm kiếm ở vùng khác (đột biến) do đó giảm khả năng kết thúc tại một điểm tối ưu cục bộ mà không thấy tối ưu toàn cục.

Các các kĩ thuật tìm kiếm và tối ưu thường được áp dụng vào giải thuật di truyền cổ điển để tạo thành giải thuật di truyền lai. Trong phần tiếp theo của báo cáo này, chúng tôi sẽ từng bước xây dựng giải thuật di truyền lai để giải bài toán phủ đỉnh.

2. XÂY DỰNG GIẢI THUẬT DI TRUYỀN LAI

2.1 Mã hóa lời giải

Mỗi cá thể sẽ có bộ nhiễm sắc thể được biểu diễn bằng một dãy N bit (với N là số đỉnh của đồ thị). Bit thứ i có giá trị là 0 và

1 tương ứng đỉnh thứ i không được chọn hay được chọn vào tập kết quả.

Ví dụ một phương án sẽ được mã hóa như sau:

$$X = [X_1, X_2, \dots, X_n] = [0 1 1 0 1]$$

Nghĩa là các đỉnh thứ 2, 3 và 5 được chọn. Để thấy một mã hóa với dãy bit bất kì có thể là một phương án sai, nghĩa là các đỉnh được chọn không phủ hết toàn bộ các cạnh của đồ thị, yêu cầu đặt ra phải xây dựng một phương thức điều chỉnh các phương án sai.

Với cách mã hóa như trên ta có hàm thích nghi sau:

$$f(x) = N - \sum_{i=1}^n x_i$$

Hàm thích nghi tính bằng số đỉnh không được lựa chọn vào tập kết quả, như vậy các phương án tốt hơn có số đỉnh sử dụng ít hơn sẽ có hàm thích nghi lớn hơn.

2.2 Lựa chọn các toán tử di truyền

Ở trong cài đặt này, chúng tôi sử dụng một phương pháp lai ghép riêng áp dụng cho bài toán phủ đỉnh do Keta Kotachan đề xuất [2], đó là lai ghép HVX. Phương pháp này có đặc điểm cho tốc độ hội tụ nhanh do cá thể con sẽ tận dụng các gen tốt của cá thể cha mẹ, chi tiết phép lai ghép này sẽ được trình bày ở phần sau của báo cáo.

Với toán tử chọn lọc, chúng tôi lựa chọn cách chọn lọc bằng quay bánh xe (Roulette wheel), phương pháp này cho phép một cá thể có thể được lựa chọn nhiều lần, và các cá thể tốt hơn sẽ có xác suất lựa chọn lớn hơn, để tận dụng được các gen tốt của cá thể đó.

Các cá thể sau lai ghép sẽ được đột biến với xác suất quy định trước. Vì lựa chọn cách đột biến bằng đảo bít nên xảy ra trường hợp cá thể sau đột biến không phải là một phủ đỉnh của đồ thị, vì thế ta cần phải xây dựng một hàm phụ trợ để sửa chữa các cá thể sau đột biến, đó là kỹ thuật Scan and Repair.

3. SCAN AND REPAIR

Bởi vì phương án đang xét chưa phải là một phủ của đồ thị nên sẽ còn các cạnh chưa được phủ. Xét tất cả các cạnh này, với mỗi cạnh ta chọn một trong hai đỉnh để cho vào phương án đang xét, như vậy thỏa mãn phủ được tất cả các cạnh của đồ thị

Bước 1: Kiểm tra nếu phương án đang xét là một phủ của đồ thị thì kết thúc, nếu không chuyển sang bước 2.

Bước 2: Với mỗi đỉnh P_i được chọn trong phương án đang xét, ta xóa tất cả các cạnh được phủ bởi P_i ra khỏi đồ thị. $A_{i,j} = 0$ & $A_{j,i} = 0$ với $1 \leq j \leq N$

Bước 3: Lần lượt duyệt các cạnh chưa được phủ, lựa chọn một trong hai đỉnh của cạnh đó và thêm vào phương án đang xét. Nếu tồn tại $A_{i,j} = 1$, ta lựa chọn đỉnh có bậc cao hơn trong 2 đỉnh P_i và P_j .

4. LAI GHÉP HVX

Các kiểu lai ghép One Point, Two Point hay Uniform thường được sử dụng rộng rãi trong giải thuật di truyền, thích hợp khi

chúng ta sử dụng mã hóa nhị phân và không có các thông tin cụ thể. Nhưng trong bài toán phủ đỉnh Ketan Kotachan đã đề xuất một phép lai ghép riêng biệt là HVX, sử dụng heuristic để sinh ra một nhiễm sắc thể con từ một cặp cha mẹ:

Procedure HVX

Begin

$$V' = \{ \}$$

Khởi tạo bảng VT và ET

VT = (F(v), N(v)), với F(v) là tần số của đỉnh v trong P1 và P2, N(v) là bậc của v, chỉ xét các đỉnh v thuộc ít nhất P1 hoặc P2.

ET = Tập cạnh của đồ thị

while ET $\neq \{ \}$ do

Chọn đỉnh v1 có N(v1) lớn nhất

Nếu có nhiều đỉnh cùng có bậc cao nhất thì chọn đỉnh có tần số F(v1) lớn hơn

Nếu vẫn còn nhiều lựa chọn thì lấy ngẫu nhiên.

Loại bỏ v1 và các cạnh kề nó ra khỏi đồ thị

$$ET = ET - \{E(x, y) \text{ với } x = v1 \text{ hoặc } y = v1\}$$

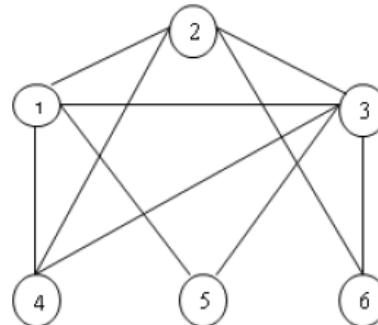
$$V' = V' - \{v1\}$$

end while

return V'

end

Ví dụ xét đồ thị sau trong hình 2:



Hình 2 – Ví dụ HVX

Đồ thị G gồm 6 đỉnh và 9 cạnh

2 cá thể cha mẹ được chọn để tiến hành lai ghép là:

$$X1 = [0, 0, 1, 1, 0, 1]$$

$$X2 = [1, 1, 1, 1, 0, 0]$$

Xây dựng bảng F(v) là tần số xuất hiện của các đỉnh trong X1 và X2.

N(v) là bậc của các đỉnh trong đồ thị hiện tại

Tại lần lặp đầu tiên, đỉnh có bậc cao nhất là đỉnh 3 $\Rightarrow V' = \{3\}$

Sau khi loại bỏ cách cạnh kề 3 ra khỏi đồ thị ta có đỉnh 1 và 2 cùng có bậc là 3 nhưng đỉnh 1 có tần số cao hơn (xuất hiện trong cả 2 cá thể cha mẹ) $\Rightarrow V' = \{3, 1\}$

Cuối cùng đỉnh 2 được chọn và tất cả các cạnh đều đã được xóa khỏi đồ thị, thuật toán kết thúc và ta có $V' = \{3, 1, 2\}$.

Vertex (V)	F(v)	N(v)	Ran 1 v1=3 V={3} Ni(v1)	Ran 2 v1=1 V={3,1} Ni(v1)	Ran 3 v1=2 V={3,1,2} Ni(v1)
1	2	4	3	0	0
2	1	4	3	2	0
3	2	5	0	0	0
4	2	3	2	1	0
6	1	2	1	1	0

Hình 3 – HVX Crossover

Có thể thấy phép lai ghép HVX ưu tiên lựa chọn các gen tốt của cha mẹ, qua đó cho ra cá thể con có độ thích nghi lớn hơn.

5. TỐI ƯU HÓA ONE POINT OPTIMIZE

Các cá thể sau khi áp dụng các toán tử di truyền có thể chứa các đỉnh dư thừa, một đỉnh gọi là dư thừa nếu ta có thể loại bỏ nó ra khỏi phủ đỉnh hiện tại mà không làm mất tính phủ của đồ thị.

One Point Optimize (OPP) là một kỹ thuật tối ưu các cá thể sau lai ghép bằng cách loại bỏ đỉnh dư thừa và tìm kiếm trong các lân cận xung quanh nó. Kỹ thuật này dựa trên 2 nhận xét sau:

- Từ một phủ đỉnh của đồ thị đang có, nếu ta xóa một đỉnh đi thì các đỉnh kề với đỉnh vừa xóa sẽ phải thêm vào để bảo đảm tính phủ của đồ thị
- Ngược lại, khi thêm một đỉnh vào một phủ đỉnh, thì có thể kéo theo một số đỉnh kề với đỉnh vừa thêm trở nên dư thừa và có thể loại bỏ khỏi đồ thị

Kỹ thuật này có thể mô tả như sau, xuất phát từ một đỉnh x bắt kì của phủ đỉnh X, ta loại bỏ x và thêm tất cả các đỉnh kề x (mà chưa xuất hiện trong phủ đỉnh)

Với mỗi đỉnh thêm vào, ta lần lượt kiểm tra các đỉnh kề với nó, nếu là đỉnh dư thừa thì loại bỏ.

Nếu sau các bước trên, số đỉnh dư thừa bị loại bỏ nhiều hơn số đỉnh thêm vào thì ta có một phủ đỉnh tối ưu hơn.

Ví dụ:



Hình 4

Ta có phủ đỉnh hiện tại $\{2, 3, 6, 7, 8\}$

Chọn đỉnh gốc là đỉnh 3, xóa 3 ra khỏi phủ đỉnh, như vậy các đỉnh kề với 3 có đỉnh 4 và 5 phải thêm vào.

Lần lượt thêm đỉnh 4, kiểm tra các đỉnh kề với 4, có đỉnh 8 trở nên dư thừa (vì 4 và 7 đã được chọn) \Rightarrow xóa đỉnh 8

Thêm đỉnh 5 \Rightarrow đỉnh 6 trở nên dư thừa, xóa đỉnh 6

Kết thúc thuật toán, như vậy ta đã loại các đỉnh 3, 6, 8 và thêm vào đỉnh 4, 5.

Phủ đỉnh mới có kích thước giảm đi một đỉnh là $\{2, 4, 5, 7\}$

6. DI TRUYỀN LAI VỚI HVX VÀ OPO

Gài thuật di truyền lai kết hợp HVX và OPO do chúng tôi đề xuất như sau:

Procedure OPO

Begin

 t = 0

 Khởi tạo quần thể ban đầu P(t)

 while (t < số quần thể tối đa) do

 Sắp xếp P(t) theo hàm thích nghi

 Lựa chọn 50% cá thể tốt nhất từ P(t) sang

 P(t+1)

 m = POP_SIZE / 2

 while (m < POP_SIZE) do

 Chọn 2 cá thể P1, P2 từ P(t) bằng quay bánh xe

 P' = HVX(P1, P2)

 P = đột biến(P')

 P = Scan and Repair(P)

 P = One Point Optimize(P)

 Thêm P vào P(t+1)

 m = m + 1

 end while

 Áp dụng OPO với cá thể tốt nhất và cập nhật kết quả

 t = t + 1 //Chuyển sang thế hệ tiếp theo

 End while

 Return cá thể tốt nhất

End

7. KẾT QUẢ THỰC NGHIỆM

7.1 Bộ dữ liệu thực nghiệm

Kết quả thử nghiệm sử dụng bộ dữ liệu BHOSLIB (Benchmark with Hidden Optimum Solutions) được sinh ngẫu nhiên dựa trên mô hình RB. Bộ dữ liệu này đã được sử dụng rộng rãi trong các báo cáo khoa học như một cơ sở để đánh giá, so sánh các phương pháp heuristics áp dụng vào bài toán phủ đỉnh.

Test	Số đỉnh	Phủ đỉnh nhỏ nhất
FrB30-15-mis	450	420
FrB35-17-mis	595	560
FrB40-19-mis	760	720
FrB45-21-mis	945	900
FrB50-23-mis	1150	1100
FrB53-24-mis	1272	1219
FrB56-25-mis	1400	1344
FrB59-26-mis	1534	1475

Hình 5 – BHOSLIB

Mỗi bộ đều bao gồm 5 test cùng kích thước đồ thị và lời giải tối ưu.

7.2 Cấu hình thực nghiệm

Chương trình thực nghiệm được viết bằng ngôn ngữ C++,

trình biên dịch MSVC 2008. Chạy trên máy tính cá nhân với cấu hình: Intel Core i7 2.66 GHz, 4GB RAM, Windows 7 OS.

Với mỗi bộ test chúng tôi cho chạy 10 lần và ghi lại giá trị tốt nhất tìm được, giá trị trung bình và thời gian chạy trung bình.

Do không tìm được các bộ dữ liệu mà các bài báo [1] và [2] sử dụng nên chúng tôi đã cài đặt lại hai giải thuật đó để so sánh.

7.3 Kết quả thực nghiệm:

Input		Ketan Kotecha			Huo Hongwei			HGA OPO		
		Min	Best	Average	Time	Best	Average	Time	Best	Average
frb30-15-1.mis	420	422	423.2	115	423	424.4	105	420	420	204
frb30-15-2.mis	420	421	421.8	120	424	426.2	115	420	420	196
frb30-15-3.mis	420	422	422.6	121	423	424.2	109	420	420.2	201
frb30-15-4.mis	420	422	422.8	117	424	425.2	112	420	420	203
frb30-15-5.mis	420	423	423.4	119	424	425	110	420	420	198

Input		Ketan Kotecha			Huo Hongwei			HGA OPO		
		Min	Best	Average	Time	Best	Average	Time	Best	Average
frb35-17-1.mis	560	563	564.4	185	564	565.4	162	561	561	413
frb35-17-2.mis	560	563	563.8	184	565	565.2	165	560	560.4	405
frb35-17-3.mis	560	562	563.3	187	563	564.5	167	560	560.8	395
frb35-17-4.mis	560	562	563	177	563	565.2	164	561	561.4	401
frb35-17-5.mis	560	562	563.5	183	564	565.1	171	560	560.7	398

Input		Ketan Kotecha			Huo Hongwei			HGA OPO		
		Min	Best	Average	Time	Best	Average	Time	Best	Average
frb40-19-1.mis	720	724	725.2	371	725	725.4	353	721	721	718
frb40-19-2.mis	720	725	725	374	726	726	347	721	721.4	744
frb40-19-3.mis	720	723	723.4	377	725	725.8	358	720	721.3	727
frb40-19-4.mis	720	723	723.8	377	725	725.5	352	722	722	712
frb40-19-5.mis	720	724	724.3	373	725	725.8	355	720	721.2	698

Giải thuật HGA OPO cho kết quả tốt hơn nhiều so với các giải thuật di truyền do Keta và Huo đề xuất. Mặc dù có thời gian chạy nhiều hơn, do hàm tối ưu hóa OPO có độ phức tạp cao. Nhưng trong đa số các bài test, HGA OPO có tốc độ hội tụ rất nhanh và thường kết quả tốt nhất xuất hiện ở khoảng 50 thế hệ đầu tiên, thời gian trong bảng dữ liệu thực nghiệm là thời gian chạy đủ số thế hệ quy định là 200. Chúng tôi cũng tiến hành chạy thử với bài test frb100-40.mis là bài test có đồ thị gồm 4000 đỉnh và phủ đỉnh tối ưu bao gồm 3900 đỉnh. Đến thời điểm hiện tại lời giải tốt nhất ghi nhận được là của Shaowei Cai và Kaile Su với giải thuật Local Search – EWLS, tìm ra phủ đỉnh gồm 3902 đỉnh. Với bộ dữ liệu trên HGA OPO cho kết quả là phủ đỉnh gồm 3909 đỉnh.

8. LỜI TRI ÂN

Em xin gửi lời cảm ơn sâu sắc đến giáo viên hướng dẫn, cô Nguyễn Thị Hải Yến, người đã giúp đỡ em rất nhiều trong quá trình thực hiện đề tài này. Sự chỉ bảo và theo sát của cô đã giúp em có cách làm việc khoa học và định hướng tốt trong quá trình nghiên cứu.

9. TÀI LIỆU THAM KHẢO

- [1] Huo Hongwei, Xu Xuezhou, XuJin and Bao - Solving Vertex Covering Problems Using Hybrid Genetic Algorithms
- [2] ProblemKetan Kotecha and Nilesh Gambhava - A Hybrid Genetic Algorithm for Minimum Vertex Cover.
- [3] Isaac K. Evans - Evolutionary Algorithms for Vertex Cover.
- [4] CoverSilvia Richter, Malte Helmert, and Charles Gretton - A Stochastic Local Search Approach to Vertex.
- [5] Melanie Mitchell - An introduction to genetic algorithms, 1999.
- [6] Tobias Friedrich , Jun He , Nils Hebbinghaus , Frank Neumann , Carsten Witt - Analyses of Simple Hybrid Algorithms for the Vertex Cover Problem.
- [7] Rajiv Kalapala, Martin Pelikan and Alexander K. Hartmann - Hybrid Evolutionary Algorithms on Minimum Vertex Cover for Random Graphs.
- [8] Shaowei Cai, Kaile Su, Qingliang Chen - EWLS: A New Local Search for Minimum Vertex Cover.
- [9] BHOSLIB: Benchmarks with Hidden Optimum Solutions for Graph Problems (Maximum Clique, Maximum Independent Set, Minimum Vertex Cover and Vertex Coloring). Available at <http://www.nlsde.buaa.edu.cn/~kexu/benchmarks/graph-benchmarks.htm>

Hệ thống nhận diện Virus máy tính theo hành vi

Trần Minh Quảng

Tóm tắt - Hiện nay, với việc virus máy tính xuất hiện ngày càng nhiều, các phương pháp nhận diện truyền thống đã và đang bộc lộ các điểm yếu của mình: cơ sở dữ liệu cồng kềnh, tỉ lệ phát hiện virus mới thấp,... Để giải quyết bài toán thực tế trên, đề tài nghiên cứu này đã đi sâu nghiên cứu, tìm hiểu các công nghệ nhận diện virus máy tính mới nhất, nhằm thiết lập nên một hệ thống nhận diện virus thông minh, có cơ chế nhận diện mới – theo hành vi, đáp ứng được các nhu cầu hiện có: cơ sở dữ liệu nhỏ gọn, tỉ lệ phát hiện virus mới cao, tài nguyên chiếm dụng ít.

Từ khóa - virus, máy tính, nhận diện, hành vi.

1. VIRUS MÁY TÍNH

Virus máy tính là các phần mềm độc hại được lập trình với mục đích phá hoại hệ thống máy tính, thu thập các thông tin của nạn nhân (mã thẻ tín dụng, tài khoản ngân hàng, thông tin mật, v.v..), nắm quyền truy cập trái phép đến tài nguyên hệ thống, và các hành vi lừa đảo khác[1].

Việc phân loại virus máy tính là không nhất quán đối với từng hằng phân mềm, từng tổ chức,... Trong đó có thể kể ra một số loại sau đây: virus lây file[2], worm[3], trojan[4], spyware[5], adware[6], scareware[7], crimeware[8], rootkit[9], và các chương trình, phần mềm độc hại khác.

Trong những năm gần đây, virus máy tính có sự gia tăng vượt bậc về số lượng. Theo thống kê sơ bộ của Symantec trong năm 2008, tỉ lệ xuất bản của các loại phần mềm độc hại còn lớn hơn các loại phần mềm hợp pháp khác[10]. Còn theo F-Secure, số lượng virus máy tính sinh ra trong năm 2007 bằng tổng số lượng virus máy tính trong 20 năm trước đó[11]. Thiệt hại do virus máy tính gây ra cũng ngày càng lớn. Theo thống kê của Bkav, năm 2010, người sử dụng tại Việt Nam đã phải chịu tổn thất lên tới 5.900 tỷ VNĐ vì virus máy tính[12].

Đề tài nghiên cứu này tập trung nghiên cứu các loại virus máy tính thực thi trên các hệ điều hành Windows, hệ điều hành chiếm tới 90% số lượng người sử dụng[13].

2. CÁC PHƯƠNG PHÁP NHẬN DIỆN VIRUS MÁY TÍNH TRUYỀN THÔNG

Lĩnh vực an ninh mạng nói chung và phòng chống virus máy tính nói riêng đã được các công ty an ninh mạng khai thác và phát

Công trình này được thực hiện dưới sự bảo trợ của Công ty TNHH An ninh mạng BKAV.

Trần Minh Quảng, sinh viên lớp Khoa học máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: (84-4)3 8692463, e-mail: quangtm@bkav.com.vn).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

triển từ khoảng 20 năm trở lại đây. Trong quá trình đó đã có nhiều phương pháp nhận diện virus được phát triển, mỗi phương pháp lại có các ưu nhược điểm khác nhau. Trong khuôn khổ nghiên cứu này, đề tài tìm hiểu hai phương pháp nhận diện truyền thống phổ biến nhất: nhận diện dựa theo mã hash và nhận diện dựa theo chữ ký.

2.1. Phương pháp nhận diện theo mã hash

Phương pháp này tính giá trị hash (MD5, SHA,...) của các mẫu virus đã được chuyên gia phân tích. Sau đó khi gặp một mẫu dữ liệu mới cần kiểm tra, phương pháp này tính giá trị hash của mẫu dữ liệu đó, sau đó so sánh với cơ sở dữ liệu các mã hash đã biết, từ đó đưa ra kết luận và diệt nếu mẫu dữ liệu được xác định là nhiễm virus.

Ưu điểm của phương pháp này là cho kết quả chính xác gần như tuyệt đối, bởi xác suất hai mẫu dữ liệu khác nhau có cùng giá trị hash là vô cùng nhỏ (thực tế chưa bao giờ xuất hiện).

Tuy nhiên, phương pháp nhận diện này có một số mặt hạn chế như: không nhận diện được các virus mới xuất hiện, chưa có trong cơ sở dữ liệu, và khi số lượng virus tăng đột biến như hiện nay, việc bùng nổ kích thước cơ sở dữ liệu là điều không tránh khỏi.

2.2. Phương pháp nhận diện theo chữ ký

Chữ ký là một chuỗi các byte, đặc trưng cho một loại virus. Mỗi virus đã biết đã được các chuyên gia phân tích để biết chính xác các hành vi phá hoại cũng như cách khắc phục đối với nó. Phương pháp nhận diện theo chữ ký tiến hành so khớp chữ ký của các virus đã biết đối với một mẫu dữ liệu đầu vào. Sau đó dựa trên các kết quả phân tích đã có để tiến hành diệt nếu kết quả so khớp phù hợp với một loại virus nào đó.

Ưu điểm của phương pháp này là cho kết quả chính xác và tốc độ nhận diện nhanh. Bởi các đoạn mã đặc trưng của từng virus được lấy sau khi đã được các chuyên gia phân tích và kết luận, còn việc nhận diện chỉ là bài toán so khớp chuỗi byte.

Mặc dù phương pháp này có thể nhận diện được một số biến thể virus mới của một số dòng, nhưng cũng giống như phương pháp nhận diện theo mã hash, phương pháp này hầu như không nhận diện được các loại virus mới xuất hiện, chưa có trong tập mẫu đã có, và phương pháp này cũng gặp tình trạng bùng phát kích thước cơ sở dữ liệu khi số lượng virus tăng đột biến như hiện nay.

3. PHƯƠNG PHÁP NHẬN DIỆN THEO HÀNH VI

Với mong muốn áp dụng các công nghệ nhận diện thông minh trong lĩnh vực virus máy tính để giải quyết bài toán thực tế với hiệu năng tốt nhất, đề tài nghiên cứu này đã đi sâu nghiên cứu và phát triển một phương pháp nhận diện mới, khắc phục được các

nhiệt điểm của phương pháp truyền thống, nhưng vẫn đảm bảo độ chính xác, tin cậy cần thiết – phương pháp nhận diện virus máy tính theo hành vi.

Công nghệ nền tảng được áp dụng trong phương pháp này dựa trên việc nhận diện virus máy tính theo hành vi[14], đây chính là công nghệ hiện đang được các nhà phát triển phần mềm diệt virus trên toàn thế giới tập trung phát triển bởi tính khả thi và hiệu quả của nó. Công nghệ này đánh giá một mẫu dữ liệu là virus dựa trên các hành vi phá hoại mà mẫu dữ liệu đó thể hiện.

Để phát hiện các hành vi của một mẫu dữ liệu, phương pháp này sử dụng bộ giả lập hành vi trên nền 32-bits. Bộ giả lập này sẽ làm thay nhiệm vụ của bộ xử lý, thực thi các lệnh của chương trình đầu vào trên một bộ nhớ ảo. Các hành vi của chương trình đầu vào sẽ được bộ giả lập nhận biết và đưa ra.

Khi tiến hành nhận diện một mẫu dữ liệu mới, phương pháp nhận diện theo hành vi đưa mẫu dữ liệu qua bộ giả lập này, thu được các hành vi của mẫu dữ liệu. Sau đó, các hành vi này được phân tích. Dựa trên các hành vi xấu của virus máy tính và các hành vi đặc trưng của từng dòng virus khác nhau đã biết, hệ thống nhận diện đưa ra kết luận và phương án xử lý tương ứng.

4. XÂY DỰNG TẬP LUẬT CÁC HÀNH VI XẤU

Để có thể đưa ra kết luận dựa trên việc phân tích các hành vi của một mẫu dữ liệu, cần phải có một cơ sở dữ liệu các hành vi xấu của virus máy tính để tiến hành so sánh. Các hành vi xấu là các hành vi tác động đến hệ thống máy tính, thay đổi tùy chọn, tính năng hệ thống, nhằm mục đích phá hoại, trực lợi, v.v... Do đó, đề tài nghiên cứu này đã tiến hành thống kê, hệ thống lại các hành vi của virus máy tính, cũng như của các phần mềm chuẩn, các file chuẩn của Windows, từ đó rút ra được các hành vi xấu.

Trên hệ điều hành Windows, việc tác động vào hệ thống gần như chỉ có thể được thực hiện bằng cách tác động vào hệ thống lưu trữ file và hệ thống các khóa registry. Bởi các tác động được thực hiện trên bộ nhớ trong sẽ không được lưu lại khi tắt máy, còn việc tác động lên phần cứng máy tính là điều hết sức khó khăn. Cá biệt có một số loại virus có thể tác động lên một số phần của thiết bị lưu trữ như Master Boot Record (MBR), Boot Sector,... nhưng số lượng rất nhỏ và phải có phương pháp xử lý đặc biệt. Do đó trong khuôn khổ nghiên cứu này, đề tài tập trung nghiên cứu các hành vi phá hoại của virus máy tính, được thực hiện bằng các tác động lên file và registry.

Dựa trên kinh nghiệm của các chính bản thân cùng các chuyên gia của Công ty Trách nhiệm hữu hạn (TNHH) An ninh mạng Bkav trong quá trình nghiên cứu và phân tích virus máy tính, đề tài đã thống kê lại các hành vi phá hoại của virus, từ đó lựa chọn các hành vi phù hợp đưa vào tập dữ liệu các hành vi xấu. Trong phạm vi nghiên cứu của mình, đề tài kiểm soát các hành vi sau của virus máy tính:

- Tự sao chép bản thân vào thư mục tạm của Windows.
- Tự sao chép bản thân vào các thư mục nhạy cảm (Thư mục Windows, thư mục hệ thống, thư mục cài đặt chương trình,...).
- Tự sao chép bản thân vào các thư mục khác.
- Thiết lập các khóa registry để virus tự động thực thi mỗi khi

Windows khởi động.

- Thay đổi tùy chỉnh FolderOptions của Windows (không hiện file ẩn, không hiện phần mở rộng, v.v...)

- Thiếp lập chế độ tự động chạy cho các ổ đĩa máy tính và các thiết bị lưu trữ di động.

- Vô hiệu hóa các tính năng của Windows (TaskManager, Regedit,...)

- Thay đổi nội dung file HOSTS của Windows.

- Thay đổi thiết lập Internet Explorer (đặt lại trang chủ,...)

- Thay đổi tùy chỉnh về mở file với một số phần mở rộng (mở các file exe, txt,... bằng các chương trình không chính thức,...)

- Thay đổi các thiết lập SecurityOptions (tắt tường lửa, tắt AutoUpdate của Windows,...)

- Thay đổi màn hình nền máy tính.

- Inject mã thực thi vào các tiến trình khác

- Hook một số hàm, chức năng của hệ thống (keylogger,...)

- Ngừng các tiến trình, dịch vụ của các chương trình diệt virus.

- Truy nhập vào các file password của hệ thống.

- Download các virus khác từ Internet.

- Mở cổng sau.

- Các hành vi tác động đến file, registry khác.

- V.v...

Từ cơ sở dữ liệu tập các luật trên, đề tài nghiên cứu này xây dựng nên các nhóm luật đặc trưng, ứng với từng loại virus khác nhau.

Ví dụ: Đối với loại virus: W32.Generic.Worm

- Hành vi đặc trưng:

- + Tự sao chép bản thân vào các thư mục mạng.

- + Tự sao chép bản thân vào các ổ đĩa, các thiết bị di động

- + Ghi file Autorun.inf vào các ổ đĩa.

5. HỆ THỐNG NHẬN DIỆN VIRUS MÁY TÍNH THEO HÀNH VI

Hệ thống nhận diện virus máy tính theo hành vi hoạt động với đầu vào là một mẫu dữ liệu (một file thực thi trên môi trường Windows 32-bits), và đầu ra là kết luận mẫu dữ liệu đó không là virus hoặc nếu là virus thì thuộc loại virus nào.

Mẫu dữ liệu đầu vào được kiểm tra một số thông tin cơ bản như: kích thước mẫu dữ liệu, chữ ký điện tử, thông tin về header của file,... để quyết định liệu có tiếp tục sử dụng bộ giả lập đối với mẫu dữ liệu hay không. Nếu mẫu dữ liệu có kích thước quá lớn (cụ thể trong đề tài là lớn hơn 20MB), hệ thống sẽ không tiến hành kiểm tra tiếp với mẫu dữ liệu, bởi sẽ gây ra sự tổn kém về mặt tài nguyên hệ thống trong khi với kích thước lớn như vậy, khả năng mẫu dữ liệu là virus là rất nhỏ. Nếu mẫu dữ liệu có chữ ký điện tử của các công ty, nhà sản xuất uy tín, hệ thống cũng không tiến hành kiểm tra tiếp với mẫu đó. Ngoài ra, sau khi kiểm tra các thông tin khác về header của mẫu dữ liệu, hệ thống cũng sẽ quyết định có tiếp tục kiểm tra với mẫu hay không, chẳng hạn như sẽ dừng nếu phát hiện mẫu dữ liệu không phải là file thực thi, hoặc là file thực thi lỗi, v.v...

Tiếp theo, bộ giả lập của hệ thống sẽ tiến hành phân tích các hành vi của mẫu dữ liệu nếu mẫu dữ liệu được thực thi. Bộ giả

lập sẽ cho biết tất cả các hành vi tác động đến file, registry, tiến trình,... của mẫu dữ liệu. Sau khi thu nhận được các thông tin đó, hệ thống sẽ kiểm tra các hành vi của mẫu dữ liệu, với các bộ hành vi được định nghĩa trước đối với từng loại virus và đưa ra kết luận.

Trong phạm vi nghiên cứu của đề tài, virus máy tính được chia thành các loại sau đây:

* Worm:

- Mô tả: Virus là file thực thi, có các hành vi phá hoại và có thể lây lan từ máy tính này sang máy tính khác.
- Hành vi đặc trưng:
 - + Tự sao chép bản thân vào các thư mục mạng.
 - + Tự sao chép bản thân vào các ổ đĩa, các thiết bị lưu trữ di động.
 - + Tạo file autorun.inf ở các ổ đĩa.

* Trojan:

- Mô tả: Virus là file thực thi, có các hành vi phá hoại nhưng không có khả năng lây lan.
- Hành vi đặc trưng:
 - + Tự sao chép bản thân vào các thư mục trên máy tính.
 - + Ghi các khóa registry để virus tự thực thi mỗi khi Windows khởi động.
 - + Download, thực thi các virus khác.

* Backdoor:

- Mô tả: Virus là file thực thi, mở cổng sau để tin tặc có thể nắm quyền truy nhập máy tính.
- Hành vi đặc trưng:
 - + Mở cổng trên máy tính.

* Rootkit:

- Mô tả: Virus là file thực thi ở chế độ nhân, thao tác với hệ thống ở mức nhân, có thể dùng để che giấu các file, tiến trình khác.
- Hành vi đặc trưng:
 - + Thao tác với hệ thống ở mức nhân.
 - + Hook một số hàm của Windows.

* Lây file:

- Mô tả: Virus là các đoạn mã lây nhiễm vào các file thực thi trên máy tính, tự động lây nhiễm sang các file khác khi một file bị nhiễm được thực thi.
- Hành vi đặc trưng:
 - + Tìm kiếm các file thực thi trên ổ đĩa.
 - + Ghi dữ liệu vào các file thực thi trên ổ đĩa.

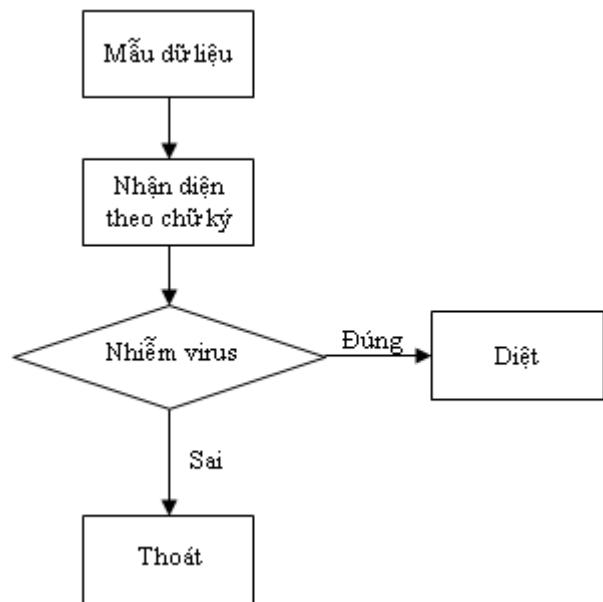
* Các loại virus cụ thể theo từng dòng:

- Mô tả: Là các dòng virus đã được các chuyên gia phân tích và nghiên cứu các hành vi đặc trưng của dòng đó. Việc biết rõ từng dòng virus giúp cho việc diệt virus và khôi phục hệ thống dễ dàng hơn.
- Hành vi đặc trưng: Các hành vi đặc trưng tùy thuộc từng

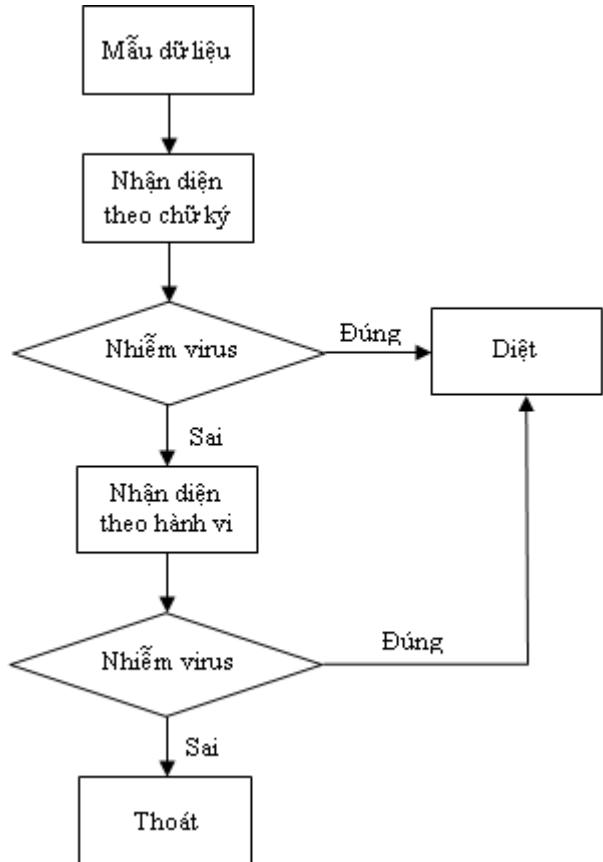
dòng virus.

6. ỨNG DỤNG TRÊN THỰC TẾ

Trên thực tế, hệ thống nhận diện virus máy tính theo hành vi không hoạt động độc lập mà đóng vai trò như một mắt xích trong quá trình kiểm tra, nhận diện các mẫu virus.



Hình 1. Sơ đồ hệ thống nhận diện truyền thống



Hình 2. Sơ đồ hệ thống có áp dụng công nghệ

nhận diện theo hành vi.

Mặc dù phương pháp nhận diện theo chữ ký có những hạn chế nhất định, nhưng không thể phủ nhận rằng, phương pháp này có tốc độ nhận diện rất nhanh, độ chính xác cao. Ngược lại, phương pháp nhận diện theo hành vi tuy có những ưu điểm nổi bật, nhưng thời gian nhận diện chậm hơn, độ chính xác tương đối. Do đó, để phát huy tối đa hiệu năng, hiệu quả của quá trình nhận diện virus máy tính, hệ thống nhận diện virus được thiết kế với sự tham gia của cả hai phương pháp nhận diện kể trên.

Dữ liệu dành cho việc nhận diện theo chữ ký được tối ưu hóa, loại bỏ các loại virus đã xuất hiện từ rất lâu và hiện tại không còn tồn tại nữa. Ngoài ra còn có thể áp dụng công nghệ điện toán đám mây để giảm dung lượng lưu trữ cơ sở dữ liệu chữ ký ở phía người dùng.

Một mẫu dữ liệu sau khi đã được kiểm tra bằng phương pháp nhận diện theo chữ ký, nếu không phát hiện thấy dấu hiệu của virus sẽ được kiểm tra bằng phương pháp nhận diện theo hành vi. Điều này vừa giúp cải thiện tốc độ nhận diện của hệ thống, vừa giúp tăng độ chính xác khi xử lý virus.

Hệ thống nhận diện virus máy tính theo hành vi đã và đang được phát triển tại Công ty TNHH An ninh mạng Bkav, dưới dạng một công nghệ được trang bị trong phần mềm diệt virus Bkav 2011 Beta và đã được đưa vào thử nghiệm nội bộ trong 3 tháng. Trong thời gian thử nghiệm, hệ thống đã cho thấy sự hiệu quả trong quá trình nhận diện các mẫu virus mới, chưa có trong tập mẫu đã biết.

Trên thế giới hiện nay đã có một số công ty an ninh mạng phát triển công nghệ nhận diện này, tuy nhiên phương pháp cụ thể không được tiết lộ. Ở Việt Nam, hiện chưa có công trình nào nghiên cứu về phương pháp nhận diện mới này.

Trong một thử nghiệm với một mẫu virus mới, có hành vi giống hành vi của dòng virus W32.Generic.Trojan. Kết quả nhận diện của một số hãng bảo mật như sau:

Bảng 1. Kết quả thử nghiệm các phần mềm diệt virus

Phần mềm	Phiên bản	Cập nhật	Kết quả
Antiy-AVL	2.0.3.7	2011.05.05	-
Avast	4.8.1351.0	2011.05.04	Win32:Malware-gen
AVG	10.0.0.1190	2011.05.04	-
BitDefender	7.2	2011.05.05	Gen:Trojan.FWDisable.cmW@ae6KMEe
ClamAV	0.97.0.0	2011.05.05	-
DrWeb	5.0.2.03300	2011.05.05	-
F-Secure	9.0.16440.0	2011.05.04	-
GData	22	2011.05.05	Gen:Trojan.FWDisable.cmW@ae6KMEe
Kaspersky	9.0.0.837	2011.05.05	HEUR:Trojan.Win32.Generic
McAfee	5.400.0.1158	2011.05.05	-
Microsoft	1.6802	2011.05.04	-
Norman	6.07.07	2011.05.04	W32/Malware.OFRY
Panda	10.0.3.5	2011.05.04	-

Sophos	4.64.0	2011.05.05	-
Symantec	20101.3.2.89	2011.05.05	-
TrendMicro	9.200.0.1012	2011.05.04	-
Bkav 2010	3295	2011.05.03	-
Bkav 2011 Beta	3295	2011.05.03	- W32.Generic.Trojan

Nhìn vào bảng trên ta có thể thấy, một số hãng phần mềm như Avast, BitDefender, Kaspersky,... có trang bị công nghệ nhận diện virus theo hành vi, với việc nhận diện mẫu dữ liệu là virus với tên gọi có chứa từ viết tắt “gen” – viết tắt của từ “generic”, một tên gọi đặc trưng cho các loại virus được phát hiện bởi phương pháp nhận diện theo hành vi.

7. LỜI TRI ÂN

Để có được những kết quả này, em xin bày tỏ lòng biết ơn sâu sắc nhất tới thầy giáo, TS. Phạm Đăng Hải - người đã tận tâm hướng dẫn em trong suốt quá trình nghiên cứu.

Em cũng xin gửi lời cảm ơn chân thành tới công ty TNHH An ninh mạng Bkav đã tạo điều kiện thuận lợi giúp cho em có một môi trường tốt nhất để thực tập và làm việc.

8. TÀI LIỆU THAM KHẢO

- [1] Troy Nash, Vulnerability & Risk Assessment Program (VRAP), Lawrence Livermore National Laboratory, “An Undirected Attack Against Critical Infrastructure”, 2005. Available: http://www.us-cert.gov/control_systems/pdf/undirected_attack0905.pdf
- [2] Wikipedia, “Computer virus”, URL: http://en.wikipedia.org/wiki/Computer_virus, last visited May 2011.
- [3] Wikipedia, “Computer worm”, URL: http://en.wikipedia.org/wiki/Computer_worm, last visited May 2011.
- [4] Wikipedia, “Trojan horse (computing)”, URL: http://en.wikipedia.org/wiki/Trojan_horse_%28computing%29, last visited May 2011.
- [5] Wikipedia, “Spyware”, URL: <http://en.wikipedia.org/wiki/Spyware>, last visited May 2011.
- [6] Wikipedia, “Adware”, URL: <http://en.wikipedia.org/wiki/Adware>, last visited May 2011.
- [7] Wikipedia, “Scareware”, URL: <http://en.wikipedia.org/wiki/Scareware>, last visited May 2011.
- [8] Wikipedia, “Crimeware”, URL: <http://en.wikipedia.org/wiki/Crimeware>, last visited May 2011.
- [9] Wikipedia, “Rootkit”, URL: <http://en.wikipedia.org/wiki/Rootkit>, last visited May 2011.
- [10] Symantec Corp., “Symantec Internet Security Threat Report: Trends for July–December 2007 (Executive Summary)”. April 2008. Available: http://eval.symantec.com/mktginfo/enterprise/white_papers/b-whitepaper_exec_summary_internet_security_threat_report_xiii_04-2008.en-us.pdf
- [11] F-Secure Corporation, “F-Secure Reports Amount of Malware Grew by 100% during 2007”, December 4, 2007. Available: http://www.f-secure.com/f-secure/pressroom/news/fs_news_20071204_1_eng.html
- [12] Bkav Corporation, “Việt Nam thiệt hại 5.900 tỷ đồng trong năm 2010 vì virus máy tính”, URL: http://www.bkav.com.vn/tin_tuc_noi_bat/viet-nam-thiet-hai-5900-ty-dong-trong-nam-2010-vi-virus-may-tinh-3205/, last visited May 2011.
- [13] Wikipedia, Usage share of operating systems, URL: http://en.wikipedia.org/wiki/Usage_share_of_desktop_operating_systems, last visited April 2011.
- [14] Tristan Aubrey-Jones, School of Electronics and Computer Science, University of Southampton, UK, “Behaviour Based Malware Detection”, 2007. Available: [http://jones.com/papers/info3005_jan2008_behaviour_based_malware_deetection.pdf](http://jones.com/papers/info3005_jan2008_behaviour_based_malware_detection.pdf)

Phân cụm tài liệu sử dụng độ tương đồng dựa trên cơ sở các cụm từ

Nguyễn Kim Thuật-Cao Mạnh Đạt

Tóm tắt - Hiện nay, hầu hết các thuật toán phân cụm tài liệu đều được xây dựng trên cơ sở không gian vecto tài liệu, trong đó, mỗi tài liệu được biểu diễn dưới dạng một vecto trên các từ mà không quan tâm đến thứ tự của các từ. Vì vậy, kết quả phân cụm của các thuật toán phân cụm tài liệu thường rất hạn chế [2, 3].

Cụm từ là một dãy các từ có thứ tự thường xuất hiện trong một hay nhiều tài liệu của tập tài liệu đang xét. Thực tế, cụm từ được xem như là một thuật ngữ có nghĩa, nhờ đó có thể cải thiện rất nhiều kết quả của bài toán phân cụm các tài liệu. Công trình nghiên cứu này sử dụng một độ đo tương tự dựa trên cơ sở các cụm từ để tính toán độ tương đồng giữa các cặp tài liệu bất kỳ trong mô hình cây hậu tố tài liệu. Ánh xạ các nút trong cây hậu tố của tập các tài liệu thành các chiều để xây dựng một không gian vecto và, từ đó, kế thừa độ đo tương tự *tf-idf* để tính độ tương đồng các tài liệu dựa trên cơ sở các cụm từ [1]. Từ không gian vecto xây dựng được, áp dụng thuật toán phân cụm phân cấp GAHC để phân cụm tập các tài liệu đang xét. Kết quả thử nghiệm kĩ thuật này đối với bài toán phân cụm các tài liệu tiếng Việt đã khẳng định kĩ thuật phân cụm dựa trên cụm từ là rất hiệu quả đối với các tài liệu tiếng Việt so với các thuật toán phân cụm cổ điển trước đây.

Từ khóa - Cây hậu tố, độ đo tương đồng, phân cụm tài liệu.

1. GIỚI THIỆU

Phân cụm các tài liệu là kỹ thuật tổ chức lại một tập các tài liệu thành các nhóm, trong đó mỗi nhóm mang một chủ đề riêng biệt. Hầu hết các phương pháp phân cụm tài liệu hiện nay đều dựa trên mô hình không gian vecto của các tài liệu (VSD)

Trong mô hình này, mỗi tài liệu sẽ được biểu diễn dưới dạng một vecot nhiều chiều, mỗi chiều đặc trưng cho một từ khóa xuất hiện trong tài liệu. Các từ khóa là các từ phân biệt trong tập tài liệu đang xét. Sau đó, độ tương đồng giữa các tài liệu sẽ được tính dựa trên các độ đo *Cosine*, *Jaccard*, hoặc khoảng cách Euclidean... Cuối cùng, tập các tài liệu này sẽ là đầu vào của thuật toán phân cụm K-Mean, AHC,... để trả lại các cụm tài liệu.

Với mục đích phân cụm các tài liệu một cách hiệu quả và chính xác hơn, người ta đã nghiên cứu phương pháp phân cụm các tài liệu dựa trên cơ sở các cụm từ. Cụm từ là một chuỗi gồm một hay nhiều từ đơn ghép lại. Hung Chim và Xiaotie Deng[1] đã tiếp cận cơ sở này để phân cụm trên một

tập các tài liệu rất lớn. Kết quả là, chất lượng các cụm tài liệu trả về vượt trội các các cụm thu được từ các phương pháp phân cụm các tài liệu trên cơ sở VSD.

Đặc biệt, một trong những cấu trúc tài liệu mà có khả năng lưu giữ được thứ tự các từ trong tài liệu, ta không thể không nói đến, là mô hình cây hậu tố các tài liệu (STD). Mô hình này biểu diễn các tài liệu dưới dạng các chuỗi từ chứ không phải là các kí tự. Mỗi tài liệu sẽ được đại diện bởi tập chuỗi con các hậu tố của tài liệu đó. Có rất nhiều công trình liên quan tới mô hình STD, tuy nhiên không có nhiều công trình đánh giá được ảnh hưởng của thứ tự các cụm từ tới việc phân cụm các tài liệu. Bằng việc sử dụng cây hậu tố, việc trích rút thông tin cũng như tính độ tương đồng giữa các tài liệu dễ dàng hơn rất nhiều, điều này mang lại hiệu quả cho quá trình phân cụm các tài liệu. Ngược lại, với việc biểu diễn các tài liệu bằng mô hình không gian vec tơ (VSD), độ tương đồng các tài liệu chủ yếu được tính dựa trên trọng số của các từ rời rạc trong tập tài liệu, chủ yếu được tính bằng công thức *tf-idf*, mà không hề xét tới thứ tự của các từ trong trong tập tài liệu đang xét.

Chính vì vậy, một cách tiếp cận mới, để kết hợp cả hai mô hình lại đã được đề ra trong [1]. Trong công trình nghiên cứu này, chúng tôi tập trung vào tìm hiểu phân cụm các tài liệu tiếng Việt, sử dụng cấu trúc cây hậu tố các tài liệu, dựa trên cơ sở các cụm từ. Ngoài ra, chúng tôi còn so sánh việc phân cụm các tài liệu dựa trên mô hình VSD (dựa trên cơ sở các từ đơn) với cùng phương pháp phân cụm, từ đó rút ra kết luận về hiệu quả vượt trội của cách tiếp cận mới này khi xử lý các tài liệu tiếng Việt, điều mà ít được đề cập trong các phương pháp phân cụm truyền thống.

Phần còn lại của bài viết sẽ bao gồm những phần chính sau: phần 2 giới thiệu qua về công trình liên quan, phần 3 là một cái nhìn sơ lược về mô hình cây hậu tố của các tài liệu, phần 4,5 sẽ nói về độ đo tương đồng được sử dụng và thuật toán phân cụm các tài liệu. Phần 6 sẽ trình bày chi tiết những thử nghiệm mà chúng tôi đã tiến hành. Cuối cùng phần 7,8 là những đánh giá và kết luận về những thử nghiệm này.

2. CÔNG TRÌNH LIÊN QUAN

Phân cụm các tài liệu luôn được coi là một pha rất quan trọng để cải thiện hiệu năng của các bộ máy tìm kiếm bằng cách hậu xử lý các kết quả trả về. Thuật toán phân cụm phân cấp (HAC) là thuật toán phổ biến nhất được sử dụng để sắp xếp lại tập các tài liệu này. Thuật toán này có 3 biến thể chính : *single-link*, *complete-link*, and *group-average*. Thực tế, thuật toán HAC trả ra các cụm với chất lượng rất tốt với độ phức tạp tính toán chấp nhận được.

Mô hình thường được dùng để biểu diễn tập các tài liệu là mô hình không gian vecto các tài liệu (VSD model) [2]. Trong mô hình này, việc phân cụm các tài liệu chỉ được tính chủ yếu dựa trên trọng số các từ trong tập tài liệu. Tuy nhiên,

để nâng cao chất lượng của các cụm trả về, việc phát triển một mô hình mới, cho phép biểu diễn được các tập tài liệu dưới dạng một không gian vecto, mà vẫn bảo đảm về mặt ngữ nghĩa là rất cần thiết.

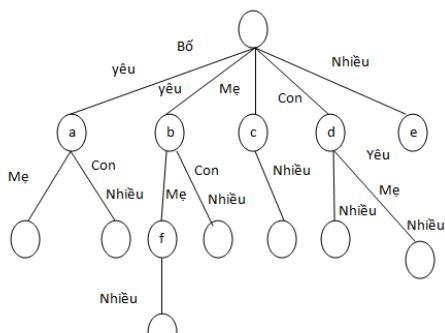
Mô hình cây hậu tố của các tài liệu là một cấu trúc đã thỏa mãn được việc lưu giữ được thứ tự của các từ trong một tài liệu. Mô hình này được đề xuất lần đầu tiên năm 1997 [3]. Khác với việc biểu diễn tập tài liệu dựa trên tập các từ đơn và không quan tâm tới thứ tự của các từ trong tài liệu, mô hình STD biểu diễn mỗi một tài liệu dưới dạng các chuỗi hậu tố con của tài liệu đó. Sau đó, mỗi tiền tố chung của các chuỗi con này được gán thành các nhãn cho các cạnh của cây hậu tố [1]. Bên cạnh đó, cách biểu diễn mới này mang lại hiệu quả cao trong việc trích rút thông tin cũng như trích các cụm từ trong một tài liệu. Mô hình này cũng đã được sử dụng trong rất nhiều trong các công trình nghiên cứu [1][3].

3. ĐẶC ĐIỂM CỦA MÔ HÌNH CÂY HẬU TỐ CÁC TÀI LIỆU (STD)

Cây hậu tố là một cấu trúc dữ liệu được biết đến trong việc lưu giữ thứ tự của các từ trong một tài liệu. Nó đã được nghiên cứu trong nhiều thập kỷ và được sử dụng trong rất nhiều các thuật toán và các ứng dụng thực tế (bài toán tìm xâu con lớn nhất, rút trích ý chính trong văn bản, bài toán sắp xếp chuỗi gen,...). Tuy nhiên, mô hình biểu diễn các tài liệu dựa trên cây hậu tố thì chưa được nghiên cứu nhiều kể từ khi nó ra đời.

Biểu diễn các tài liệu bằng cây hậu tố để bảo lưu thứ tự của tài liệu đã được nói đến từ rất lâu. Nhưng việc xây dựng một không gian vecto đặc tả các tài liệu dựa trên mô hình cây hậu tố là một hướng tiếp cận mới. Mỗi một nút trong cây hậu tố (trừ nút gốc và các nút kết thúc) sẽ đặc trưng cho một chiều trong không gian vecto được xây dựng. Mỗi nút sẽ đặc tả cho một cụm từ mà xuất hiện ít nhất trong một tài liệu trong tập đang xét.

Trong ví dụ bên dưới, ta xây dựng một cây hậu tố cho một tập gồm 3 tài liệu tiếng việt, mỗi tài liệu là một câu ngắn, tạm thời ta chỉ chia tài liệu ra thành các từ dựa trên dấu trắng.



H1.Cây hậu tố của tập 3 tài liệu:
“Bố yêu me”, “Con yêu mẹ Nhiều”, “Bố yêu con nhiều”.

Ta sẽ đi tìm hiểu các tính chất của cây hậu tố. Xét cây hậu tố T là một cây có hướng có gốc, biểu diễn một tập tài liệu

D . Mỗi nút trong v có ít nhất hai con và mỗi cạnh được gắn nhãn bằng một chuỗi con khác rỗng của một tài liệu trong D . Các nhãn của hai cạnh bất kỳ xuất phát từ một nút chung phải bắt đầu bằng các ký tự khác nhau. Đối với nút lá của cây hậu tố, việc kết các nhãn của các nút nằm trên con đường đi từ gốc đến nút lá đó sẽ tạo thành cụm từ P_v . Cụm từ này là một chuỗi con hậu tố của ít nhất một tài liệu trong tập D . Cây hậu tố có ba loại nút: nút gốc, nút trong, và nút kết thúc. Nút trong là nút có ít nhất một con. Nút kết thúc là nút được dùng để phân biệt các tài liệu phân biệt trong D khi ta biểu diễn chúng trên cùng một cây hậu tố. Nút kết thúc này được khởi tạo bởi thuật toán Ukkonen[4]

Trong bài báo cáo này, ta coi các nút là con của nút gốc sẽ có độ sâu là một, tương tự như vậy, con của các nút có độ sâu là một sẽ có độ sâu là hai,...

Tính chất 1: Mỗi nút trong của cây hậu tố T thể hiện các cụm từ chung trong các tài liệu, và mỗi nút lá, việc kết nhãn của các nút trên đường đi từ nút gốc tới nó sẽ tạo thành một chuỗi con hậu tố của một tài liệu trong tập D .

Tính chất 2: Mỗi nút có độ sâu là một sẽ được gắn nhãn bởi các cụm từ xuất hiện ít nhất một lần trong tập các tài liệu của tập D . Số nút có độ sâu là một sẽ bằng số từ khóa (số các từ đơn phân biệt trong không gian vecto các tài liệu) trong tập tài liệu D .

Tính chất 3: Với mỗi cụm từ P_v , nhãn của một nút trong v sẽ chứa ít nhất hai từ. Hay độ dài của cụm từ $|P_v| \geq 2$.

4. ĐỘ ĐO TƯƠNG ĐỒNG GIỮA CÁC TÀI LIỆU DỰA TRÊN MÔ HÌNH STD

Nhu đã đề cập ở trong phần trên, cây hậu tố có ba loại nút chính. Trong đó những nút kết thúc là những nút không mang nhãn. Chúng được sinh ra từ thuật toán xây dựng cây hậu tố Ukkonen. Các nút này được dùng để đánh dấu kết thúc cho mỗi tài liệu trong tập đang xét. Với việc loại bỏ nút gốc và các nút kết thúc, mỗi nút còn lại trên cây hậu tố đều được đại diện bởi một cụm từ khác rỗng, mỗi cụm từ này xuất hiện trong ít nhất một tài liệu trong tập đang xét. Các cạnh khác nhau có thể có cùng nhãn. Ví dụ như trong H1, ta thấy cụm từ “Con Nhiều” là nhãn của 2 cạnh trong cây hậu tố.

Tùy mô hình biểu diễn các tài liệu như vậy, độ đo tương đồng giữa các tài liệu được định nghĩa rất đơn giản và dễ hiểu. Bằng việc coi mỗi nút v (trừ nút gốc và các nút kết thúc) là một đặc trưng, ta sẽ xây dựng một không gian vecto để biểu diễn các tài liệu. Trong bài báo cáo này, M được định nghĩa là số chiều của không gian vecto. M bằng số nút của cây hậu tố trừ đi nút gốc và số nút kết thúc. N là số tài liệu trong tập D , k là số cụm (kết quả trả về sau khi phân cụm tập D).

Mỗi tài liệu d sẽ được biểu diễn dưới dạng một vecto M chiều.

$$d = \{w(1,d), w(2,d), w(3,d), \dots, w(M,d)\}$$

trong đó $w(i,d)$ là trọng số tương ứng của nút i đối với tài liệu d . Giá trị $w(i,d)$ được tính dễ dàng từ công thức

$$tf-idf : w(i,d) = (1 + \log(tf(i,d))).\log(1 + N/df(i)) \quad (1),$$

trong đó $tf(i,d)$ là số lần tài liệu d duyệt qua nút i , $df(i)$ là số tài liệu đã

duyệt qua nút i. Đối với các tài liệu mà không đi qua một nút nào đó, giá trị log của tf được gán bằng 0. Xét ví dụ trong hình 1, giá trị df của nút b là df(3), giá trị tf của nút b đối với tài liệu 1 là tf(b,1)=1. Từ đó, ta có thể tính được trọng số của nút b đối với tài liệu 1:

$$w(b,1) = (1+0).\log(1+3/3) = 1.$$

Sau khi xây dựng được không gian vecto cho tập các tài liệu, ta sẽ dùng độ đo cosine để tính toán độ tương đồng giữa hai tài liệu bất kỳ trong tập D.

Xét hai tài liệu, được biểu diễn dạng hai vecto M chiều, $d_x = \{x_1, x_2, \dots, x_M\}$; $d_y = \{y_1, y_2, \dots, y_M\}$, trong đó x_i, y_i là các trọng số của nút i đối với hai tài liệu. Và công thức tính toán độ tương đồng giữa hai tài liệu sẽ được tính bởi công thức:

$$\text{sim}_{x,y} = \frac{d_x \bullet d_y}{|d_x| \times |d_y|} = \frac{\sum_{i=1}^M x_i y_i}{\sqrt{\sum_{i=1}^M x_i^2 \cdot \sum_{i=1}^M y_i^2}} \quad (2)$$

5. THUẬT TOÁN PHÂN CỤM PHÂN CẤP GAHC

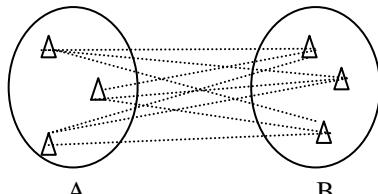
Trong thực tế, có rất nhiều loại thuật toán phân cụm tài liệu như K-MEAN, DBSCAN,... Tuy nhiên, thuật toán phân cụm phân cấp HAC là một hướng tiếp cận rất phổ biến. Các tiếp cận này trả ra các cụm với kết quả rất tốt, đơn giản trong cài đặt. Trong phạm vi bài báo cáo này, chúng tôi sử dụng thuật toán GAHC- một biến thể của thuật toán HAC để phân cụm tập các tài liệu. Đầu vào của thuật toán này là tập tài liệu D bao gồm N tài liệu. Đầu ra là một tập bao gồm k cụm, mỗi cụm bao gồm một số tài liệu trong D, các tài liệu này có nội dung khá tương đồng với nhau.

Ý tưởng cơ bản của thuật toán HAC này khá đơn giản. Ban đầu, ta xem mỗi tài liệu là một cụm phân biệt. Sau đó, ta nhóm 2 cụm gần nhất thành một cụm. Quá trình này lặp lại cho đến khi tất cả các tài liệu được nhóm thành một cụm.

Các bước chi tiết trong kỹ thuật phân cụm này:

1. Xây dựng ma trận độ tương đồng, ma trận này sẽ lưu lại độ tương đồng của các cặp tài liệu bất kì, sử dụng công thức cosine để tính toán.
2. Xem mỗi tài liệu là một cụm (chẳng hạn nếu tập D có 4 tài liệu, ban đầu ta sẽ có 4 cụm).
3. Lặp lại 2 bước sau cho tới khi số cluster là 1
 - a. Gộp 2 cụm gần nhất
 - b. Cập nhật ma trận khoảng cách.

Thuật toán GAHC chỉ là một biến thể của thuật toán trên, với việc gộp các cụm trên cơ sở độ tương đồng trung bình của các cặp tài liệu trong hai cụm.



A, B là hai cụm cá tài liệu.

Độ tương đồng giữa hai cụm A, B được tính theo công thức:

$$d_{A \rightarrow B} = \text{average}(d_{ij})$$

trong đó: d_{ij} là độ tương đồng cosine giữa tài liệu i thuộc cụm A và tài liệu j thuộc cụm B.

6. THỦ NGHIỆM

Để thấy được sự hiệu quả của thuật toán phân cụm sử dụng độ tương đồng dựa trên cụm từ trong cách tiếp cận mới này, chúng tôi đã tiến hành 4 thử nghiệm để so sánh với cách phân cụm tài liệu truyền thống sử dụng độ đo tương tự tf-idf trên cơ sở các từ. Chúng tôi sử dụng đồng bộ thuật toán phân cụm là GAHC để đảm bảo tính đúng đắn trong phép so sánh này. Ngoài ra, để thử nghiệm được chính xác, tập tài liệu đầu vào đã được phân cụm rất cẩn thận bằng tay.

Bốn thử nghiệm cụ thể là:

1. Phân cụm các tài liệu dựa trên mô hình VSD bằng thuật toán GAHC, các từ được phân biệt bằng dấu trắng.
2. Phân cụm các tài liệu dựa trên mô hình cây hậu tố (STD) bằng thuật toán GAHC, các từ được cách nhau bằng dấu trắng.
3. Phân cụm các tài liệu dựa trên mô hình VSD bằng thuật toán GAHC, tài liệu sẽ được tách thành các từ có nghĩa bằng cách so khớp trong từ điển.
4. Phân cụm các tài liệu dựa trên mô hình cây hậu tố (STD) bằng thuật toán GAHC, tài liệu sẽ được tách thành các từ có nghĩa bằng cách so khớp trong từ điển.

Chúng tôi đã tiến hành mỗi thử nghiệm trên theo ba bước chính. Bước 1: Xây dựng không gian vecto cho tập các tài liệu (theo hai cách biểu diễn VSD và STD). Bước 2 : Sử dụng thuật toán phân cụm GAHC để phân cụm tài liệu. Bước 3: Đánh giá kết quả.

Thử nghiệm một mặt tập trung vào việc đánh giá các cụm trả về trong hai cách biểu diễn: theo không gian vecto cổ điển hoặc theo mô hình cây hậu tố (trong thử nghiệm 1 và 2). Mặt khác, chúng tôi còn cố gắng phân cụm các tài liệu tiếng Việt (trong hai thử nghiệm 3 và 4), bằng cách tách các tài liệu này thành các từ tiếng Việt có nghĩa thông qua một xử lý khá đơn giản. Trong hướng tiếp cận này, mỗi tài liệu sẽ được tách thành tập hợp các từ có nghĩa, các từ này được so khớp trong từ điển, điều này sẽ đảm bảo ngữ nghĩa của các tài liệu.

Tập tài liệu thử nghiệm của chúng tôi bao gồm 69 tài liệu ngắn. Mỗi tài liệu là tên một đề tài đồ án của khoa công nghệ thông tin đại học bách khoa năm 2010 và nó chỉ là một chuỗi từ có độ dài tương đối ngắn. Chúng tôi đã tiến hành phân cụm bằng tay tập tài liệu thành 10 cụm. Mỗi cụm chứa các tài liệu mang chủ đề khá tương tự nhau.

Cả bốn thử nghiệm đều được viết bằng ngôn ngữ Java. Riêng với thuật toán xây dựng cây hậu tố, chúng tôi kế thừa từ mã nguồn mở. Thuật toán này chỉ cho phép tạo cây hậu tố với đầu vào là một chuỗi ký tự liền nhau. Vì vậy, chúng tôi đã cải tiến để có thể xây dựng cây hậu tố với đầu vào là tập các tài liệu. Thuật toán GAHC và việc xây dựng cây hậu tố sau khi xử lý tài liệu tiếng Việt chưa hề được đề cập đến trong các công trình gần đây, vì vậy chúng tôi đã tự cài đặt nó để có thể so sánh kết quả của các thử nghiệm.

7. ĐÁNH GIÁ KẾT QUẢ

Để đánh giá được chất lượng của các cụm trả về, chúng tôi sử dụng một tập tài liệu mẫu, các tài liệu này đã được phân cụm trước bằng tay. Sau đó, tiến hành so sánh với các kết quả trả về từ 4 thử nghiệm trên. Trong bài báo cáo này, chúng tôi sử dụng 2 độ đo đánh giá kết quả của thuật toán phân cụm.

Đầu tiên, chúng tôi sử dụng độ đo *F-mesure*. Độ đo này đã được sử dụng rất nhiều trong việc đánh giá chất lượng của các cụm trả về. Xét tập $C = \{C_1, C_2, \dots, C_k\}$ là tập các cụm trả về từ thuật toán GAHC của tập tài liệu D. Bên cạnh đó, tập $C^* = \{C_1^*, C_2^*, \dots, C_k^*\}$ là tập các cụm “mẫu” của tập D do ta xử lý trước bằng tay. Xét 2 đại lượng, $rec(i,j)$ là đặc trưng cho độ tin cậy của cụm i với cụm mẫu j; $prec(i,j)$ đặc trưng cho độ chính xác của cụm j so với cụm mẫu i.

$$\text{Trong đó: } rec(i, j) = |C_i \cap C_j^*| / |C_j^*| \quad (3)$$

$$prec(i, j) = |C_i \cap C_j^*| / |C_i| \quad (4)$$

$i = 1..k; j = 1..l$

Công thức tính F-mesure

$$F(i, j) = \frac{2 \cdot rec(i, j) \cdot prec(i, j)}{rec(i, j) + prec(i, j)} \quad (5)$$

Dựa vào công thức trên, độ đo F-mesure được để đánh giá chất lượng của tập C so với tập mẫu C^* được tính như sau:

$$F = \sum_{i=1}^l \frac{|C_i^*|}{N} \cdot \max_{j=1..k} \{F(i, j)\}. \quad (6)$$

Công thức đánh giá thử nghiệm thứ 2 được là độ đo

Purity.

$$Purity = \sum_{j=1}^k \frac{|C_j|}{N} \cdot \max_{i=1..l} \{prec(i, j)\} \quad (7)$$

Hai công thức này đã được sử dụng trong [1], và trả lại kết quả khá chính xác. Thuật toán phân cụm là càng chính xác khi giá trị của *F-mesure* là càng lớn và giá trị của *Purity* là càng nhỏ. Chúng tôi đã tiến hành so sánh kết quả của thử nghiệm 1,2 với nhau và so sánh thử nghiệm 3,4 với nhau (do cách tách từ khác nhau). Để phép so sánh được chính xác hơn, số cụm trả ra từ thuật toán GAHC được cố định là 10.

Bảng 1: F-mesure của hai trường hợp đầu

Thử nghiệm	I	II
F-Mesure	0.37	0.71
Purity	0.472	0.102
Số chiều	312	991
Số cụm	10	10

Bảng 2: F-mesure và Purity của thử nghiệm 3, 4

Thử nghiệm	IV	III
F-mesure	0.81	0.62
Purity	0.091	0.115
Số chiều	634	258
Số cụm	10	10

Từ bảng 1-2, ta có thể thấy giá trị độ đo F- mesure của thử nghiệm 2,4 cao hơn hẳn so với thử nghiệm 1,3; bên cạnh đó giá trị Purity của nó lại nhỏ hơn đáng kể so với Purity của thử nghiệm 1,3. Điều đó thể hiện tính ưu việt của mô hình STD. Thật vậy, các cụm trả về trong thử nghiệm 2,4 là khá tương đồng với tập các cụm mẫu. Ngược lại, kết quả của thử nghiệm 1,3 lại khiêm tốn hơn rất nhiều. Ngoài ra, ta có thể thấy rõ ràng, số nút trong trường hợp biểu diễn tập tài liệu theo cây hậu tố có sử dụng tách từ tiếng việt bao giờ cũng ít hơn so với trường hợp biểu diễn tập tài liệu theo mô hình STD tách từ bằng dấu trắng (số nút trong trường hợp IV chỉ là 634 trong đó số nút trong trường hợp II là 991). Do vậy, độ phức tạp tính toán trong trường hợp sử dụng xử lý tiếng việt được giảm đi đáng kể (số chiều của không gian vecto giảm).

8. KẾT LUẬN

Mô hình không gian vecto có điểm và cây hậu tố đều giữ vai trò quan trọng trong việc phân cụm tập các tài liệu. Tuy nhiên, hai mô hình được tiếp cận theo hai cách hoàn toàn khác nhau. Hầu hết các thuộc toán phân cụm tài liệu dựa trên mô hình VSD đều không tính đến vị trí của các từ trong một tài liệu. Do đó, việc nghĩa của các từ thay đổi trong từng thứ tự khác nhau đã bị bỏ qua. Mô hình biểu diễn cây hậu tố của các tài liệu đã lưu giữ hoàn toàn thứ tự của các từ trong một tài liệu, đồng nghĩa với việc lưu giữ được ngữ nghĩa của các tài liệu. Tuy nhiên, mô hình cây hậu tố gặp rất nhiều khó khăn trong việc phân cụm cũng như đánh giá chất lượng các cụm trả về. Kết hợp hai mô hình này đã được đề xuất trong [1]. Công trình này sử dụng ý tưởng trong [1] để phân cụm các tài liệu tiếng Việt. Bên cạnh đó, tiếng Việt là một ngôn ngữ đơn âm tiết, nếu chỉ xử lý theo các từ đơn thì sẽ làm mất đi ngữ nghĩa của tài liệu. Vì vậy, trước khi tiến hành phân cụm, chúng tôi đã cố gắng tách các tài liệu thành tập các từ có nghĩa, nhằm giữ được ngữ nghĩa vốn có của chúng.

Định nghĩa của cây hậu tố rất đơn giản nhưng việc xây dựng nó là hoàn toàn không dễ. Trong công trình này, chúng tôi chỉ tập trung vào việc đo độ tương đồng giữa các tài liệu, mà chưa để ý tới việc tối ưu hóa trong quá trình xây dựng cây hậu tố. Chúng tôi sẽ tiến hành công việc này trong tương lai. Ngoài ra, chúng tôi có thể phát triển ý tưởng trong bài nghiên cứu này trong việc phân cụm và đánh nhận các tập các tài liệu trả về từ một bộ máy tìm kiếm.

9. TÀI LIỆU THAM KHẢO

- [1] Efficient Phrase-Based Document Similarity for Clustering - Hung Chim and Xiaotie Deng, Senior Member, IEEE.
- [2] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," Comm. ACM, vol. 18, no. 11, pp. 613-620, 1975
- [3] C.J. van Rijsbergen, Information Retrieval. Butterworths, 1997.

==== RESULT ====

Cluster 0 bao gồm các doc sau :

- Thiết kế cơ sở dữ liệu đã chiết
- Thiết kế công dân tộc cho doanh nghiệp

Cluster 1 bao gồm các doc sau :

- Ông đang công nghệ USD trong bài toán mobilebanking
- Ông đang công nghệ USD vào bài toán Mobile Payment
- Kế hoạch kinh doanh và thông tin khách quan vào bài toán ra quyết định đã mục tiêu
- Bài toán ra quyết định đã thực hiện phục vụ cho công tác tổ chức tuyển sinh đại học
- Bài toán ra quyết định đã tiêu chuẩn sử dụng phương pháp trao đổi ngang hàng

Cluster 2 bao gồm các doc sau :

- Trích rút mifl quan hệ giữa các thư thi trong văn bản tiếng việt sử dụng phương pháp kernel
- Tìm và sao lặp sai chính tả trong văn bản tiếng Việt
- Ông đang họ m/s sử dụng phương pháp phân cụm dữ liệu
- Ông đang họ m/s sử dụng phương pháp lan truyền ngược sai số

Cluster 3 bao gồm các doc sau :

- Nghiên cứu thời hàn ngữ có nghĩa cho liên lạc VoIP
- Nghiên cứu giải pháp ghi âm cho liên lạc VoIP
- Nghiên cứu và IVR cho giao tiếp VoIP
- Xây dựng kich bìn cảng giao tiếp thời trung Đại học Bách Khoa Hà Nội
- Xây dựng CSDL thông tin cung cấp qua cảng giao tiếp thời trung Đại học Bách Khoa Hà Nội

Cluster 4 bao gồm các doc sau :

- Xây dựng website nhà thông sản giao dịch chứng khoán và trên môi trường .NET
- Xây dựng website giới thiệu tour du lịch Campuchia.
- Hệ thống đánh giá là hàng an ninh website
- Hệ thống đánh giá là hàng an ninh website
- Xây dựng website quản lý bán hàng điện thoại di động bằng ngôn ngữ lập trình PHP và MySQL
- Xây dựng trang web quản lý bán hàng siêu thị
- Xây dựng trang web quản lý bán hàng online

Cluster 5 bao gồm các doc sau :

- Logic ngôn ngữ và suy diễn từ động
- Bài toán du doan nops ceph phù hợp để giải các bài giải ý tư động
- Bài toán giải ý tư động cho người dùng di động
- Bài toán giải ý chẩn đoán động giải trí có ràng buộc thời gian

Cluster 6 bao gồm các doc sau :

- Tìm hiểu ngôn ngữ lập trình và cơ sở dữ liệu
- Tìm hiểu về cơ sở dữ liệu và xây dựng website
- Tìm hiểu công nghệ two way anaglyph 3D và hai khung 2D
- Khai phá các truy vấn trong mặt cơ sở dữ liệu
- Xây dựng kiến trúc kim kim
- Phát hiện luật kê khai trong cơ sở dữ liệu

Cluster 7 bao gồm các doc sau :

- Xây dựng công thông tin việc làm
- Xây dựng công thông tin cho các ứng dụng tin sinh sản sử dụng công nghệ SOA
- Xây dựng công thông tin cho công ty Nam Thành của nước Cộng Hòa Dân Chủ Nhân Dân Lào
- Xây dựng công thông tin và chứng khoán cho phép người dùng theo dõi cập nhật, tăng kết các thông tin và chứng khoán
- Nghiên cứu các giải pháp tích hợp dịch vụ trong công thông tin điện tử
- Nghiên cứu triết khai các dịch xác thực trong công thông tin điện tử
- Xây dựng công thông tin quản lý và khai thác mỏng cấp Hà nội, da trên kiến trúc SOA
- Xây dựng ứng dụng trên nền tảng SOA

Cluster 8 bao gồm các doc sau :

- Nghiên cứu và hi CSDL Oracle và ứng dụng quản lý
- Nghiên cứu và hi CSDL Oracle và ứng dụng quản lý thông tin Trang thiết bị trong bệnh viện
- Kiểm tra ứng dụng
- Thiết kế hệ thống datacenter chuẩn Tier 3 và cải đặt hệ thống giám sát
- Nghiên cứu hệ quản trị cơ sở dữ liệu SQL và cải đặt ứng dụng thử nghiệm
- Nghiên cứu hệ thống thời
- Tìm hiểu và xây dựng các công cụ hỗ trợ quản lý hệ thống máy tính của ba môn
- Xây dựng hệ thống tin tức và truy vấn quản lý hoạt động của ba môn HTTT
- Nghiên cứu công nghệ mã hóa và ứng dụng cho hệ thống cao thông minh
- Xây dựng hệ thống quản lý phản công thực tiễn
- Phân tích thiết kế hệ thống quản lý thi-viết
- Hệ thống quản lý chấm công trong các doanh nghiệp

Cluster 9 bao gồm các doc sau :

- Hệ thống điều hành tác nghiệp, xử lý công văn trên nền Web
- Phát triển ứng dụng web cho hệ thống lưu trữ và chia sẻ file dữ liệu công nghệ lưu trữ dữ liệu
- Phát triển ứng dụng cho hệ thống lưu trữ, chia sẻ, quản lý và sao lưu tài liệu trên nền tảng lưu trữ dữ liệu
- Nghiên cứu và áp dụng giải pháp cho ứng dụng Web
- Hệ thống định danh và xác thực dữ liệu vào PKI và Biometric
- Nghiên cứu và triển khai giải pháp dịch vụ tính toán đám mây dữ liệu trên nền tảng UEC
- Hệ thống tv đóng trích lục thông tin ngành điện da trên nền tảng web ngữ nghĩa

H2- Kết quả phân cụm của tập tài liệu trong thử nghiệm IV, đã qua bước tiền xử lý tách từ tiếng việt, độ tương đồng của các tài liệu được tính dựa trên cơ sở các cụm từ. Các tài liệu trong mỗi cụm rất tương đồng so với các cụm mẫu.

Hệ thống trích rút thông tin cho việc xây dựng cơ sở tri thức từ văn bản tiếng Việt

Nguyễn Hữu Thiện, Nguyễn Quang Vinh, Nguyễn Thị Minh Ngọc

Viện Công Nghệ Thông Tin và Truyền Thông

Đại học Bách Khoa Hà Nội

nguyenhuuhien88bk@yahoo.com, vinhnq2112@gmail.com, ngocnguyen1802@gmail.com

Tóm tắt - Sự bùng nổ thông tin trên internet hiện nay làm nảy sinh nhu cầu xây dựng các cơ sở tri thức từ nguồn dữ liệu này. Các cơ sở tri thức sẽ cho phép chúng ta quản lý, truy nhập, trao đổi thông tin một cách dễ dàng hiệu quả hơn. Bên cạnh đó, các cơ sở tri thức cũng cho phép máy móc thực hiện những suy diễn trên đó, từ đó tạo ra những tri thức mới phục vụ con người. Để xây dựng các cơ sở tri thức từ khối dữ liệu khổng lồ trên internet hiện nay, vấn đề trích rút thông tin (thực thể, quan hệ...) từ các tài liệu là một vấn đề then chốt. Trong bài báo này, chúng tôi đề xuất một cách tiếp cận học máy để giải quyết bài toán trích rút thông tin cho việc xây dựng cơ sở tri thức từ văn bản tiếng Việt. Chúng tôi tích hợp các công nghệ học máy tiên tiến: Conditional Random Fields, Support Vector Machine để giải quyết các khía cạnh khác nhau của bài toán trích rút thông tin cho tiếng Việt. Hệ thống là một môi trường tương đối tổng quát, có khả năng thích nghi nhanh chóng với các miền ứng dụng mới, cũng như các ngôn ngữ mới. Kết quả thực nghiệm cho thấy hệ thống của chúng tôi đạt được hiệu năng tốt trên tập văn bản đầu vào và có nhiều triển vọng để tích hợp vào các hệ thống xây dựng cơ sở tri thức thuộc nhiều lĩnh vực khác nhau, qua đó làm cho các hệ thống xây dựng cơ sở tri thức trở nên mạnh mẽ và khả chuyên hơn.

Từ khóa - Conditional Random Fields (CRFs), Coreference Resolution, Information Extraction (IE), Knowledge Base, Named Entity Recognition (NER), Relation Extraction, Support Vector Machines (SVMs).

1. DẪN NHẬP

Sự phát triển mạnh mẽ của World Wide Web hiện nay đã dẫn đến sự bùng nổ của khối lượng thông tin do con người tạo ra. Tuy nhiên, những thông tin này lại thường chỉ hiểu được bởi con người. Máy móc không có hoặc có rất ít hiểu biết về các thông tin này. Điều này kết hợp với hạn chế của con người trong việc xử lý những khối dữ liệu lớn dẫn chúng ta đến một tình huống trong đó chúng ta tràn ngập trong một lượng khổng lồ các thông tin nhưng khai thác được rất ít tri thức hữu ích từ chúng. Một giải pháp cho vấn đề này là biến đổi các thông tin sang một định dạng mà cả con người và máy móc có thể hiểu và trao đổi được với nhau. Từ đó, máy móc với khả năng xử lý lớn của mình sẽ giúp chúng ta tìm ra những tri thức hữu ích. Ý tưởng đó thúc đẩy sự ra đời của web ngữ nghĩa, với trọng tâm là xây dựng nền một cơ sở tri thức làm nền tảng cho các ứng dụng tri thức bên trên.

Cấu trúc tổng quát của một hệ thống xây dựng cơ sở tri thức bao gồm một ontology đặc trưng cho lĩnh vực quan tâm, một

module tìm ra những tài liệu liên quan đến một lĩnh vực đó trên mạng Internet, một module trích rút những thông tin quan tâm (thực thể, quan hệ...) từ các tài liệu thu được, một module để triển khai các thông tin này vào cơ sở tri thức và một module để triển khai các ứng dụng trên nền các cơ sở tri thức này. Trong đó module trích rút thông tin đóng một vai trò quan trọng vì các thực thể, đối tượng cũng như quan hệ là những thành phần cốt lõi bao chất tạo nên một cơ sở tri thức. Các hệ thống tự động xây dựng cơ sở tri thức trước đây thường giải quyết vấn đề này bằng cách xây dựng các tập luật (thủ công hay học máy) để định hướng cho quá trình trích rút. Đặc điểm chung của chúng là thường phụ thuộc lớn vào miền ứng dụng và đòi hỏi nhiều chỉnh sửa nền tảng nếu như muốn chuyển sang các miền ứng dụng mới.

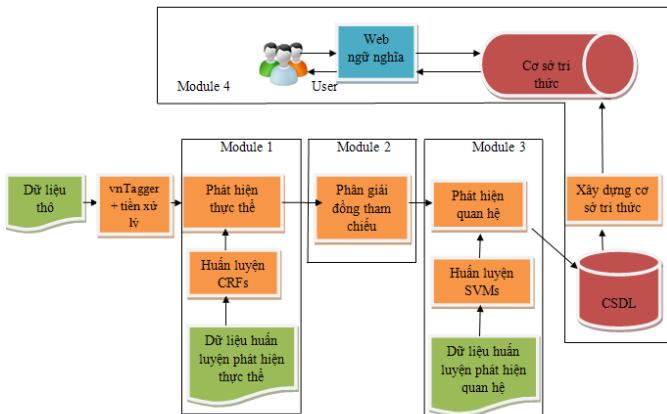
Alani H. et al [1] đề xuất một hệ thống tự động xây dựng cơ sở tri thức về các họa sĩ và các tác phẩm nghệ thuật. Để giải quyết vấn đề trích rút, họ sử dụng môi trường GATE (A Framework and Graphical Development Environment) [4] để thực hiện quá trình trích rút thông tin. Tư tưởng chủ đạo của GATE là cho phép người dùng thêm vào các luật thủ công hướng miền ứng dụng để làm cơ sở cho quá trình trích rút thực thể cũng như phân giải đồng tham chiếu. Sau đó, các thực thể sẽ được kết hợp với một ontology đặc trưng cho miền và từ điển ngữ nghĩa WordNet để trích rút ra các quan hệ giữa các thực thể. Mặc dù cách tiếp cận này có thể đạt hiệu quả cao, nhưng việc thêm các luật thủ công đặc trưng cho mỗi miền ứng dụng vào GATE làm cho hệ thống không có tính khả chuyên và đòi hỏi nhiều công sức, thời gian. Một cách tiếp cận học máy sẽ phù hợp trong tình huống này. Craven M. et al [3] đề xuất một cách tiếp cận học máy để xây dựng cơ sở tri thức. Tư tưởng cơ bản của họ là coi mỗi trang web là một biểu diễn của một thực thể nào đó trong một ontology về các khoa, viện khoa học máy tính cho trước. Công việc chính của hệ thống là phân loại các trang web vào một trong các thực thể (hoặc không thực thể nào) của ontology. Để trích rút thông tin từ các văn bản, Craven M. et al [3] sử dụng thuật toán SRV (Sequence Rules with Validation) [5]. SRV là một thuật toán suy diễn logic học ra các luật first-order phục vụ cho bài toán trích rút thông tin. SRV là một thuật toán trích rút có hiệu quả cao trên những tài liệu có cấu trúc hoặc bán cấu trúc (các trang web cá nhân...). Tuy nhiên, khi được áp dụng cho các văn bản giàu ngữ nghĩa hơn (các bài báo...) thì SRV không nắm bắt được đầy đủ ngữ cảnh cần thiết và hoạt động với một hiệu năng thấp. Để xây dựng nên một hệ thống trích rút có tính khả chuyên cao, chúng ta cần những phương pháp trích rút có thể nắm bắt được các ngữ cảnh và tính chất của văn bản một cách hiệu quả hơn.

Vargas-Vera, M. et al [18] cũng đề xuất một hệ thống trích rút thông tin sử dụng công nghệ gán nhãn dựa trên ontology. Tuy nhiên việc xây dựng các luật trích rút của họ là dựa trên những tri thức chuyên gia về một lĩnh vực cho trước (các trận bóng đá...). Điều này làm việc chuyển đổi giữa các miền ứng dụng của hệ thống gặp nhiều khó khăn.

Ở Việt Nam, hệ thống xây dựng cơ sở tri thức của VN-KIM [19] cũng giải quyết vấn đề trích rút thông tin bằng cách sử dụng GATE và do đó có những nhược điểm như đã đề cập bên trên. Nhận thức được những hạn chế trên của các hệ thống đã có, chúng tôi đề xuất một cách tiếp cận học máy, có tính khả chuyên cao để giải quyết vấn đề trích rút thông tin cho các hệ cơ sở tri thức. Hệ thống tích hợp những công nghệ học máy tiên tiến nhất để giải quyết các công đoạn của vấn đề trích rút một cách triệt để. Phần còn lại của bài báo được tổ chức như sau: phần 2 sẽ trình bày tổng quan về hệ thống trích rút thông tin cho cơ sở tri thức từ văn bản tiếng Việt, các phần 3, 4, 5 trình bày chi tiết các công nghệ để giải quyết bài toán, phần 6 trình bày các kết quả thực nghiệm của chúng tôi đối với hệ thống, cuối cùng phần 7 đưa ra một số đánh giá, kết luận cũng như định hướng tương lai.

2. Hệ thống trích rút thông tin cho cơ sở tri thức

Hệ thống của chúng tôi bao gồm 4 module chính như trong hình vẽ 1.

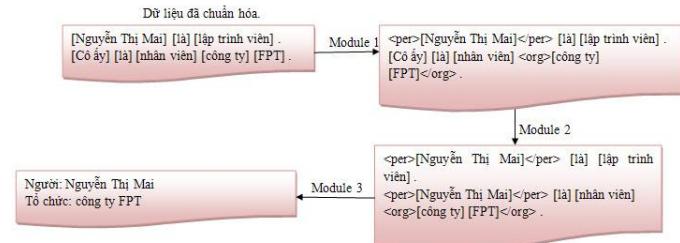


Hình 1: Hệ thống trích rút thông tin

Dữ liệu thô đầu vào của hệ thống là những văn bản có chứa các thông tin (ví dụ thông tin về nơi làm việc, chỗ ở, hướng nghiên cứu... của một nhà khoa học). Dữ liệu thô trước tiên sẽ được đưa qua bộ tách từ vnTagger [10] để thực hiện việc tách từ cho văn bản đầu vào. Các văn bản đã được tách từ sau đó sẽ được xử lý ban đầu để đưa về một dạng chuẩn chung. Các văn bản đã được chuẩn hóa sau đó được lần lượt đưa qua các module 1, 2, 3 để tiến hành các phân tích cần thiết. Ở đầu ra của module 3, các thông tin được hệ thống phát hiện sẽ được đẩy vào cơ sở dữ liệu, làm nền tảng cho việc xây dựng cơ sở tri thức dựa trên công nghệ RDF ở module 4. Trong phạm vi hệ thống của chúng tôi, cơ sở tri thức này sẽ được dùng để cung cấp những dịch vụ truy vấn tìm kiếm đơn giản cho người dùng. Để đơn giản, trong hệ thống của mình, chúng tôi xây dựng cơ sở tri thức từ cơ sở dữ liệu thu được ở module 3 bằng cách chuyển đổi đơn giản cơ sở dữ liệu về một

cấu trúc theo định dạng RDF, sau đó sử dụng cấu trúc này làm cơ sở tri thức của hệ thống. Điểm đặc sắc của hệ thống của chúng tôi nằm ở các module 1, 2 và 3 nhằm trích rút thông tin từ các văn bản đầu vào.

Nhiệm vụ của module 1 là phát hiện ra các thực thể trong các văn bản đã được chuẩn hóa. Các văn bản với các thực thể đã được xác định được đưa vào module 2 để tìm ra các đồng tham chiếu về đại từ trong đó. Cuối cùng, các văn bản được đưa vào module 3 để xác định các quan hệ tiềm tàng. Các quan hệ này chính là các thông tin mà chúng ta muốn trích rút.



Hình 2: Mô tả dữ liệu của hệ thống

3. Module 1 – bài toán phát hiện thực thể (PHTT)

Module 1 thực chất là một chương trình giải quyết bài toán trích rút thực thể có tên (Named Entity Recognition – NER) cho văn bản tiếng Việt. Để giải quyết bài toán này, chúng tôi sử dụng mô hình các trường điều kiện ngẫu nhiên (Conditional Random Fields-CRFs) [9]. Lý do để chúng tôi chọn CRFs vì nó có một số ưu điểm hơn so với các phương pháp khác. Cụ thể, CRFs không tạo ra bất cứ giả định nào về tính độc lập của các quan sát đầu vào như các mô hình Markov ẩn [15], đồng thời giải quyết được vấn đề “hướng nhãn” [9] của các mô hình entropy cực đại [11]. Khi áp dụng CRFs cho bài toán PHTT cho văn bản tiếng Việt, Nguyen C. T. et al [13] đã thu được kết quả đánh giá 85.51% cho F1-score trên tập dữ liệu thực nghiệm của họ.

Để áp dụng CRFs vào module của mình, chúng tôi coi mỗi câu đầu vào như một chuỗi các quan sát (mỗi quan sát là một từ trong câu) đồng thời tiến hành gán nhãn các từ (mỗi từ là một chuỗi các chữ) theo 3 lớp thực thể là con người(PER), địa điểm(LOC) và cơ quan(ORG). Đối với mỗi lớp thực thể trên, các nhãn cho các quan sát có thể có một trong hai dạng *B-C* (bắt đầu của một thực thể lớp *C*) hoặc *I-C* (ở trong của một thực thể lớp *C*). Những quan sát không thuộc vào một trong 3 lớp trên sẽ gán nhãn là “khác”(*O*). Khi đó bài toán PHTT được đưa về bài toán gán một trong $2^*3 + 1 = 7$ nhãn cho mỗi quan sát (từ) trong một câu cho trước. Ví dụ, đối với chuỗi “đồng chí Nông Đức Mạnh”, thông qua vnTagger[10], chúng ta biết được “đồng chí” là một từ, “Nông Đức Mạnh” là một từ ([đồng chí] [Nông Đức Mạnh]). Khi đó cách gán nhãn đúng cho chuỗi gồm 2 quan sát này là: *O B-PER*.

Trong bài báo này, CRFs được coi như là một mô hình đồ thị không định hướng tuyến tính, tức là là một máy hữu hạn trạng thái thỏa mãn đặc điểm Markov thứ tự thứ nhất.

Đặt $\mathbf{o} = (o_1, o_2, \dots, o_T)$ là một chuỗi dữ liệu quan sát nào đó. Đặt S là tập các trạng thái, trong đó mỗi trạng thái được kết hợp với một trong 7 nhãn của ta ở trên. Đặt $s = (s_1, s_2, \dots, s_T)$ là một chuỗi

trạng thái nào đó. Khi đó, CRFs định nghĩa xác suất có điều kiện của một chuỗi trạng thái cho trước một chuỗi quan sát như sau:

$$p_{\lambda}(s|\sigma) = \frac{1}{Z(\sigma)} \exp \left[\sum_k \sum_{t=1}^T \lambda_k f_k(s_{t-1}, s_t, \sigma, t) \right]. \quad (1)$$

Trong đó $Z(\sigma) = \sum_s \exp(\sum_{t=1}^T \sum_k \lambda_k f_k(s'_{t-1}, s'_t, \sigma, t))$ là một số chuẩn hóa, được lấy tổng trên tất cả các chuỗi trạng thái, f_k là hàm đặc điểm trong ngôn ngữ của mô hình entropy cực đại [11] và λ_k là trọng số thu được từ quá trình học được kết hợp với đặc điểm f_k . Trong bài báo của chúng tôi, f_k là các hàm nhị phân kiểm tra các đặc trưng và trạng thái của một dãy quan sát cho trước. Các đặc trưng được sử dụng là các đặc trưng được dùng phổ biến trong các hệ thống trích rút thông tin: đặc trưng chính tả, ngữ pháp, cú pháp, từ điển...

CRFs được huấn luyện bằng cách chọn ra tập trọng số $\lambda = \{\lambda_1, \lambda_2, \dots\}$ để làm cực đại hóa hàm mục tiêu log-likelihood, cho trước tập dữ liệu huấn luyện $D = \{(\sigma^{(k)}, I^{(k)})\} k=1, N$:

$$l = \sum_{j=1}^N \log(p_{\lambda}(I^{(j)}, \sigma^{(j)})) - \sum_k \frac{\lambda_k^2}{2\sigma^2}. \quad (2)$$

Trong đó tổng thứ hai được thêm vào để giảm ảnh hưởng của hiện tượng overfitting gây ra do sự phức tạp của mô hình. Để huấn luyện cho mô hình CRFs trong bài báo của chúng tôi, chúng tôi đã sử dụng phương pháp L-BFGS (quasi-Newton). Việc huấn luyện này được thực hiện trong khối “Huấn luyện CRFs” của module 1.

Để thực hiện quá trình suy diễn (tức là đi gán nhãn cho một chuỗi quan sát cho trước), CRFs đi tìm chuỗi trạng thái với xác suất có điều kiện cho trước một chuỗi quan sát là lớn nhất.

$$s^* = \arg \max_{s^*} p(s|\sigma). \quad (3)$$

Để tìm ra s^* , ta có thể thuật toán qui hoạch động Viterbi (với một số chỉnh sửa phù hợp) như trong mô hình Markov ẩn [15]. Việc suy diễn này được thực hiện trong khối “Phát hiện thực thể” của module 1.

Để giảm bớt công sức, thời gian để xây dựng tập dữ liệu huấn luyện cho CRF, chúng tôi cũng đề xuất một phương pháp học bán giám sát mới cho bài toán PHTT (xem slide trình bày).

4. Module 2 – bài toán phân giải đồng tham chiếu

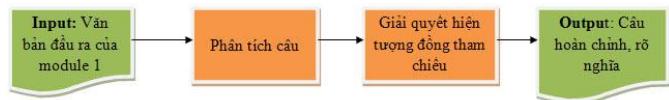
Hiện tượng đồng tham chiếu là hiện tượng sử dụng một từ để thay thế cho một từ, một cụm từ, một phần câu đã được nhắc đến trước đó. Xét một ví dụ đồng tham chiếu:

Tô Ngọc Vân là họa sĩ. Ông ấy vẽ rất đẹp

Trong ví dụ trên, “Ông ấy” chính là đại từ thay thế cho “Tô Ngọc Vân”. Vấn đề đặt ra là làm sao để hệ thống nhận diện được “Ông ấy” chính là từ thay thế cho “Tô Ngọc Vân”.

Có rất nhiều cách tiếp cận khác nhau được đề xuất để giải quyết bài toán đồng tham chiếu cho các ngôn ngữ khác nhau [12]. Tuy nhiên, ở Việt Nam hiện nay, chưa có một hệ thống nào trực tiếp để cập đến việc giải quyết bài toán này. Trong quá trình tìm tòi và nghiên cứu, chúng tôi nhận thấy: giải thuật Hobbs [8] cho kết quả khá tốt, với tỉ lệ thành công 88% cho tiếng Anh. Tuy

nhiên, giải thuật này đòi hỏi phải có bộ phân tích cú pháp, trong khi đó, đối với tiếng Việt chưa có bộ phân tích cú pháp nào thực sự hiệu quả. Hơn nữa, giải thuật này cũng không giải quyết được những đồng tham chiếu đòi hỏi phải dùng ngữ nghĩa mới giải quyết được. Do đó, việc áp dụng giải thuật Hobbs cho tiếng Việt là không khả thi. Chúng tôi cũng đã nghiên cứu các phương pháp học máy, chẳng hạn [14], để giải quyết bài toán nhưng do tập huấn luyện còn hạn chế và tập thuộc tính chưa đầy đủ nên hiệu quả không cao. Để giải quyết bài toán phân giải đồng tham chiếu trong hệ thống của mình, chúng tôi đề xuất một giải thuật mới phù hợp với tiếng Việt hơn. Mô hình của giải thuật là:



Hình 3: Module giải quyết bài toán đồng tham chiếu

- Quá trình phân tích câu. Đây là quá trình nhằm tìm ra hiện tượng đồng tham chiếu trong câu. Đó là sự xuất hiện của các đại từ như: *ông ấy*, *anh ấy*, *cô ấy*... Chúng tôi đã xây dựng một từ điển bao gồm các đại từ trên và các thuộc tính đi kèm để phát hiện ra các đại từ này. Từ điển có dạng sau:

```

<People>
  <Name>
    <w gender="XO" >Anh ấy</w>
    <w gender="XX" >Chị ấy</w>
  </Name>
</People>

```

Hình 4: Từ điển từ đại từ

- Quá trình giải quyết hiện tượng đồng tham chiếu:

Dựa trên ý tưởng của luật Centering [7, 16], mỗi đại từ thay thế (*ông ấy*, *cô ấy*...) sẽ ứng với một tiền ngữ (*Tô Ngọc Vân*...) duy nhất trước nó (chính là danh ngữ được xếp mức ưu tiên cao nhất của câu trước đó). Một cách hình thức, gọi tập các tiền ngữ trong câu thứ $n-1$ là C_f , đại từ cần thay thế trong câu thứ n là $C_b(i)$, thì việc chúng ta phải làm là tìm ra được C_p (tiền ngữ thích hợp nhất) trong C_f tương ứng với $C_b(i)$. C_f được tìm tra bằng cách sử dụng đầu ra của module 1. Để tìm được C_p , chúng tôi xây dựng một tập các ràng buộc với mức độ ưu tiên từ trên xuống dưới. Sau đó, đối với mỗi ràng buộc, những ứng viên nào trong C_f vi phạm ràng buộc này sẽ bị loại (không được xét chọn làm C_p cho $C_b(i)$ nữa). Sau khi kết thúc quá trình kiểm tra với tất cả các ràng buộc, nếu trong C_f có nhiều hơn 1 ứng viên tiền ngữ, thì ta sẽ chọn tiền ngữ nào đứng gần $C_b(i)$ nhất để làm C_p cho $C_b(i)$. Dưới đây là các luật với độ ưu tiên từ trên xuống dưới được chúng tôi sử dụng (dựa trên tập luật của Carbonell, J. G. el al [2]):

- Local Constraints : Tiền ngữ và đại từ đồng tham chiếu phải thống nhất về mặt số lượng (số ít, số nhiều), giống (đực/cái), bản chất (động vật/thực vật)...
- Ở ví dụ trên: (*ông ấy*, *Tô Ngọc Vân*) đều là danh từ chỉ người số ít, có giới tính là nam.
- Case – Role Sentence Constraints : Ngữ nghĩa của tiền tố và đại từ đồng tham chiếu phải thống nhất (một số động từ thường chỉ đi với một số loại đối tượng nhất định, như động từ “ăn” thường đi với các đối tượng chỉ thực ăn,

động từ “nghiên cứu” thường đi với các đối tượng chỉ người hoặc tổ chức...).

Ví dụ : *Nam lấy cái bánh ở trên bàn và ăn nó.*

Ở đây có 2 tiền ngữ có thể thay thế cho đại từ “nó” là “bánh” và “bàn”. Chúng tôi sẽ kiểm tra xem 2 tiền ngữ trên đi với động từ “ăn” thì có hợp lý hay không.

Để kiểm tra ràng buộc này, chúng tôi đã xây dựng bộ từ điển chủ thể-động từ-đối tượng có dạng như sau:

```
<Vt>
  <Headword ObjType="Food" SubType="People">
    <w>ăn</w>
  </Headword>
  <Headword ObjType="Things" SubType="People">
    <w>đi</w>
  </Headword>
</Vt>
```

Hình 5: Từ điển chủ thể-động từ-đối tượng

Dựa vào từ điển, ta có thể thấy “bánh” thuộc lớp “Food” nên từ “nó” là “bánh” chứ không thể là “bàn”.

- ❖ Condition-Constraint : Tiền ngữ và đại từ tham chiếu phải thoả mãn thực tế hành động.

Ví dụ: *Minh cho Tuân một quả táo. Anh ấy ăn nó ngay.* Ở đây, ta có 2 hành động là “cho” và “ăn” và 2 tiền ngữ có thể tham chiếu đến từ “anh ấy” là “Minh” và “Tuân”. Hành động “ăn” sẽ xảy ra khi chủ thể sở hữu quả táo. Vì vậy, chỉ có “Tuân” là phù hợp.

5. Module 3 – bài toán trích rút quan hệ

Module 3 thực chất là một chương trình để giải quyết bài toán trích rút quan hệ (Relation Extraction). Với một số ngôn ngữ trên thế giới, bài toán này đã được giải với mức độ chính xác khoảng 70-80%. Đối với tiếng Việt, trích rút quan hệ vẫn còn là một vấn đề mới mẻ. Phương pháp học máy thành công nhất hiện nay để giải quyết bài toán này, theo hiểu biết của chúng tôi, là phương pháp Support Vector Machines (SVMs) [17]. Năm bắt ý tưởng đó, chúng tôi chọn SVM là phương án để triển khai module trích rút quan hệ cho văn bản tiếng Việt của mình.

Ý tưởng của SVM là tìm một mặt hình học (siêu phẳng) $f(x)$ “tốt nhất” trong không gian n -chiều để phân chia tập dữ liệu huấn luyện (gồm hai loại dữ liệu x_+ và x_-) sao cho tất cả các điểm x_+ thuộc về phía dương của siêu phẳng ($f(x_+) > 0$), các điểm x_- thuộc về phía âm của siêu phẳng ($f(x_-) < 0$). Tuy nhiên, các điểm dữ liệu trong thực tế thường không dễ để phân chia một cách tuyến tính. Vì vậy, các hàm ánh xạ sẽ được sử dụng để ánh xạ dữ liệu từ không gian n -chiều ban đầu sang một không gian m -chiều khác, mà trong không gian m -chiều đó, các dữ liệu có thể phân tách tuyến tính được.

Để huấn luyện SVM trong bài toán của mình (được thực hiện ở khối “Huấn luyện SVMs” của module 3), chúng tôi sử dụng ý tưởng của Giuliano, C. et al [6]. Cụ thể, chúng tôi chuyển các câu dữ liệu huấn luyện (đã được gán nhãn thực thể và quan hệ) sang định dạng mới nhằm thuận tiện hơn cho việc trích rút các đặc trưng như: thể từ loại (danh từ, động từ, tính từ...), nhãn từ loại (con người – *Per*, tổ chức – *Org*, địa điểm - *Loc*)... Với mỗi loại quan hệ cho trước, chẳng hạn là *làm việc ở* (*work_for*), chúng tôi xác định các kiểu thực thể tham gia vào quan hệ đó

(trong trường hợp của *làm việc ở* là con người và tổ chức). Từ đây, những câu trong tập dữ liệu huấn luyện có ít nhất một cặp thực thể loại này sẽ được sử dụng làm dữ liệu huấn luyện cho quan hệ đang xét. Với mỗi câu dữ liệu huấn luyện, chúng tôi xây dựng tập vector ứng với các cặp thực thể tham gia quan hệ trong câu đó (trong trường hợp của ta, nếu câu có a thực thể *Per*, b thực thể *Org* thì chúng ta xây dựng nên $a*b$ vector tương ứng với $a*b$ cặp thực thể *Per-Org*) để làm dữ liệu huấn luyện cho quan hệ đang xét, sử dụng các đặc trưng ở trên. Dựa vào dữ liệu huấn luyện, chúng ta biết được các vector này có tương ứng với quan hệ hay không. Đến đây, chúng ta có một tập các vector huấn luyện cho quan hệ đang xét làm đầu vào cho SVM. Một cặp thực thể quan tâm được gọi tắt là một quan hệ huấn luyện R . Với một quan hệ huấn luyện R , vector $\phi(R)$ tương ứng xác định bởi:

$$\phi(R) = (\phi_{TG}(R), \phi_G(R), \phi_{GS}(R), \phi_I(R), \phi_R(R)). \quad (4)$$

$\phi_{TG}(R)$: đặc trưng cho các văn cảnh ở trước và giữa 2 thực thể trong quan hệ.

$\phi_G(R)$: đặc trưng cho các văn cảnh ở giữa 2 thực thể trong quan hệ.

$\phi_{GS}(R)$: đặc trưng cho các văn cảnh ở giữa và sau 2 thực thể trong quan hệ.

$\phi_I(R)$: đặc trưng cho các văn cảnh của thực thể ở bên trái trong quan hệ.

$\phi_R(R)$: đặc trưng cho các văn cảnh của thực thể ở bên phải trong quan hệ.

Chú ý, ở đây việc ánh xạ từ một không gian ban đầu sang một không gian có thể phân chia tuyến tính đã được thực hiện gộp trong việc tính hàm $\phi(R)$.

Sau khi SVM đã được huấn luyện sử dụng dữ liệu huấn luyện, với một câu đầu vào mới (đã được gán nhãn thực thể), hệ thống cũng tiến hành phân tích đặc trưng, xác định xem câu có chứa cặp thực thể tương ứng với quan hệ quan tâm không, xây dựng một vector cho mỗi cặp thực thể như vậy và sử dụng SVM để xác định xem vector có phải tương ứng với quan hệ đang xét hay không. Các vector được xác định là có tương ứng với quan hệ đang xét chính là đầu ra của module 3. Việc xác định quan hệ cho một câu đầu vào mới được thực hiện ở khối “Phát hiện quan hệ” của module 3.

6. Kết quả thực nghiệm

Thử nghiệm của chúng tôi được tiến hành trên 100 văn bản đầu vào. Mỗi văn bản chứa khoảng 500 chữ. Các văn bản này là các bài viết về các nhà khoa học Việt Nam, được chúng tôi thu thập thủ công trên Internet. Chúng tôi sử dụng 2000 câu tiếng Việt làm dữ liệu huấn luyện cho module 1 – trích rút thực thể và 1500 câu tiếng Việt để huấn luyện cho module 3 - trích rút quan hệ. Để đánh giá hệ thống, chúng tôi sử dụng các đánh giá sau:

- **Precision(P)**: số thực thể (đồng tham chiếu, quan hệ) được gán nhãn đúng chia cho số thực thể (đồng tham chiếu, quan hệ) được gán nhãn
- **Recall(R)**: số thực thể (đồng tham chiếu, quan hệ) được gán nhãn đúng chia cho số thực thể (đồng tham chiếu, quan hệ) có trong văn bản đầu vào
- **F-measure**: $F = 2*P*R/(P+R)$

Chúng tôi đánh giá hệ thống trong 3 công đoạn: công đoạn trích rút thực thể (module 1), công đoạn phân giải đồng tham chiếu (module 2), công đoạn trích rút quan hệ (module 3). Trong đó, module 1 sẽ ảnh hưởng đến hiệu năng của module 2 (vì đầu ra của module 1 là đầu vào của module 2). Cả hai module này sẽ ảnh hưởng đến hiệu năng của module 3. Kết quả về hiệu năng của module 3 sẽ được coi là hiệu năng chung của cả hệ thống. Trong thử nghiệm này, chúng tôi chỉ trích rút các thực thể con người, tổ chức, địa điểm và các quan hệ *sóng_ở*, *làm việc_ở*. Đây cũng chính là các thông tin được đẩy vào cơ sở dữ liệu đầu ra ở module 3 để phục vụ dịch vụ tìm kiếm của hệ thống ở module 4. Các bảng dưới đây cho thấy kết quả thực nghiệm của chúng tôi.

Bảng 1: Kết quả thử nghiệm module 1 + 2

Module 1 + 2	Phát hiện thực thể			Phân giải đồng tham chiếu		
	P	R	F	P	R	F
Con người	93.42	93.03	93.22	90.18	89.54	89.86
Địa điểm	85.46	90.28	87.80	82.06	86.74	84.34
Tổ chức	82.49	83.64	83.06	80.21	83.16	81.66

Bảng 2: Kết quả thử nghiệm module 3

Quan hệ	P	R	F
<i>sóng_ở</i>	86.12	78.95	82.38
<i>làm việc_ở</i>	73.94	79.37	76.56

Kết quả thực nghiệm cho thấy module trích rút thông tin của hệ thống hoạt động tương đối hiệu quả (trích rút được phần lớn các thông tin quan tâm). Module phân giải đồng tham chiếu dù bị ảnh hưởng bởi module 1 cũng cho kết quả tốt. Module trích rút quan hệ do bị ảnh hưởng bởi kết quả của 2 module trước nên hiệu quả trích rút quan hệ bị giảm nhẹ (chúng tôi tiến hành kiểm tra riêng module 3 trên cùng tập dữ liệu đầu vào đã được phát hiện thực thể một cách thủ công thì thu được kết quả $F = 85.28\%$ cho quan hệ *sóng_ở* và $F = 80.95\%$ cho quan hệ *làm việc_ở*). Dù vậy, kết quả toàn cục của hệ thống vẫn ở mức tốt và có thể so sánh được với kết quả của VN-KIM [19]. Tuy nhiên, hướng tiếp cận của chúng tôi là dựa vào học máy nên hệ thống của chúng tôi có tính khả chuyên cao hơn so với VN-KIM [19].

7. Kết luận

Trong bài báo này, chúng tôi đề xuất một hệ thống trích rút thông tin cho các hệ thống xây dựng cơ sở tri thức. Hệ thống gồm 3 module chính: module trích rút thực thể, module phân giải đồng tham chiếu và module trích rút quan hệ. Để triển khai module trích rút thực thể, chúng tôi sử dụng thuật toán Conditional Random Fields [9]. Để triển khai module phân giải đồng tham chiếu, chúng đề xuất một phương pháp phân giải sử dụng ý tưởng của Grosz B. J. et al [7]. Để triển khai module 3, chúng tôi sử dụng thuật toán Support Vector Machines [17]. Các kết quả thực nghiệm cho thấy hệ thống của chúng tôi có hiệu quả trích rút khá tốt trên tập văn bản đầu vào. Hệ thống đã được áp dụng thành công để xây dựng một cơ sở tri thức phục vụ cho việc tìm kiếm thông tin đơn giản về các nhà khoa học Việt Nam. Định hướng phát triển công việc của chúng tôi trong tương lai bao gồm: (i) thử nghiệm hệ thống trên một tập dữ liệu lớn hơn để có được những đánh giá chính xác hơn về hệ thống, (ii) tích hợp

thêm module cho phép tìm kiếm các tài liệu liên quan đến một lĩnh vực cho trước vào hệ thống, (iii) nghiên cứu các thuật toán học bán giám sát và nửa giám sát để tích hợp vào các module học của hệ thống.

8. Lời tri ân

Tập thể tác giả xin bày tỏ lời cảm ơn chân thành đến giáo sư Nguyễn Thanh Thủy, tiến sĩ Lê Thanh Hương, thạc sĩ Sam Rathany, đại học Bách Khoa Hà Nội. Sự đóng góp ý kiến và định hướng phát triển của họ đã giúp ích rất nhiều cho chúng tôi trong quá trình thực hiện bài báo này.

9. TÀI LIỆU THAM KHẢO

- [1] Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., Shadbolt, N. R. 2003. Automatic Ontology-Based Knowledge Extraction from Web Documents, IEEE Intelligent Systems, v.18 n.1, p.14-21.
- [2] Carbonell, J. G. and Brown R. D. 1988. Anaphora Resolution: a Multi-Strategy Approach. In Proceedings of the 12th International Conference on Computational Linguistics COLING'88, Budapest.
- [3] Craven, M., Dipasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K. and Slattery, S. 2000. Learning to Construct Knowledge Bases from the World Wide Web. Artificial Intelligence.
- [4] Cunningham H., et al. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proc. 40th Anniversary Meeting of the Association for Computational Linguistics (ACL 2002), ACL.
- [5] Freitag, D. 1998. Information Extraction from HTML: Application of a General Machine Learning Approach. In: Proceedings of AAAI'98, pp. 517-523.
- [6] Giuliano, C., Lavelli, A. and Romano, L. 2006. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. Proc., 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), Trento, Italy.
- [7] Grosz, B. J., Weinstein, S., Joshi, A. K. 1995. Centering: a Framework For Modeling the Local Coherence of Discourse, Computational Linguistics, v.21 n.2, p.203-225, 1995.
- [8] Hobbs, J. R. 1978, Resolving Pronoun References, Lingua, Vol. 44, pp. 311-338.
- [9] Lafferty, J., McCallum, A. and Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proc. ICML, pages 282-290.
- [10] Le, P. H. vnTagger. <http://www.loria.fr/~lehong/tools/vnTagger.php>
- [11] McCallum, A., Freitag, D., and Pereira F. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. In Proc. ICML 2000, pages 591-598.
- [12] Mitkov, R. 1999. Anaphora Resolution: The State of the Art. Technical Report, University of Wolverhampton, Wolverhampton.
- [13] Nguyen, C. T., Tran, T. O., Phan, X. H. and Ha, Q. T. 2005. Named Entity Recognition in Vietnamese Free-Text and Web Documents Using Conditional Random Fields. The 8th Conference on Some selection problems of Information Technology and Telecommunication, Hai Phong, Vietnam.
- [14] Nøklestad, A. 2009. A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection, PhD thesis, University of Oslo.
- [15] Rabiner. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In Proc. The IEEE, 77(2):257-286.
- [16] Sidner, C. L. 1979. Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse. Ph.D. thesis, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.
- [17] Vapnik, V. 1998. Statistical Learning Theory. John Wiley and Sons, New York.
- [18] Vargas-Vera, M., Motta, E., Domingue, J. and Shum, S. B. and Lanzoni, M. 2001. Knowledge Extraction by Using Ontology-bases Annotation Tool. First International Conference on Knowledge Capture (K-CAP 2001). Workshop on Knowledge Markup and Semantic Annotation, Victoria B.C., Canada.
- [19] VN-KIM: <http://www.dit.hcmut.edu.vn/~tru/VN-KIM/index.htm>

Phát triển nền tảng NS2 nhằm phục vụ mô phỏng các giao thức định tuyến trên mạng cảm biến không dây

Bùi Tiến Quân, Nguyễn Trung Hiếu

Tóm tắt - Các giao thức định tuyến trong mạng cảm biến không dây là một chủ đề nghiên cứu khá mới mẻ và hấp dẫn các nhà khoa học trong thời gian gần đây. Có rất nhiều kết quả, bài báo khoa học được công bố về lĩnh vực này. Hầu hết các bài báo đều đánh giá hiệu năng dựa trên các kết quả mô phỏng. Tuy nhiên, hiện chưa có một công cụ mô phỏng mạnh, chuyên dụng hỗ trợ cho việc mô phỏng các mạng cảm biến không dây. Xuất phát từ thực tế đó, nghiên cứu này sẽ đi sâu tìm hiểu về mạng cảm biến không dây, đặc trưng vật lý, các giao thức truyền thông, đặc biệt là các giao thức định tuyến và xây dựng một công cụ mô phỏng chuyên dụng cho mạng cảm biến không dây dựa trên nền tảng NS2.

Từ khóa - Mạng cảm biến không dây, định tuyến địa lý, NS2.

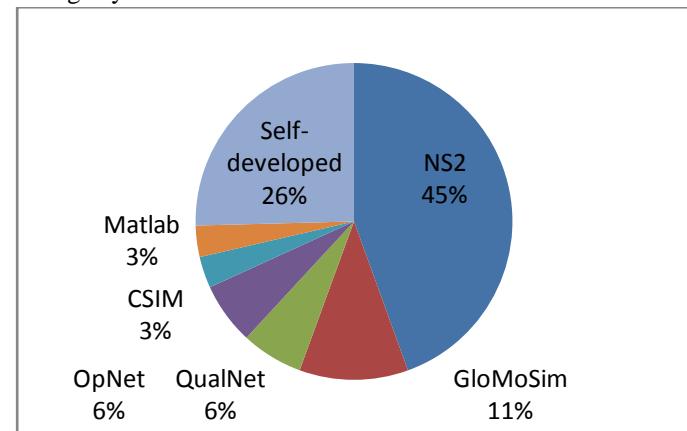
1. GIỚI THIỆU

Mạng cảm biến không dây (MCBKD) là mạng bao gồm nhiều node cảm biến có kích thước nhỏ, có khả năng tính toán và khả năng truyền dẫn không dây. Mạng cảm biến không dây có những ứng dụng quan trọng và ngày càng được sử dụng rộng rãi. Chẳng hạn như: thu thập thông tin tự động, theo dõi kiểm soát trong một môi trường đặc biệt nào đó (khu tr雍t, khu chiến địa, ...), dự báo thiên tai (động đất, sóng thần, ...). Mạng cảm biến không dây có những đặc trưng về năng lượng, khả năng tính toán, điều kiện hoạt động rất khác biệt so với các mạng có dây và không dây thông thường. Làm sao để đáp ứng những đặc trưng đó thực sự là một thử thách với các nhà nghiên cứu. Vì vậy, nghiên cứu về các công nghệ sử dụng riêng trong mạng cảm biến không dây là một chủ đề nghiên cứu nóng rất được quan tâm và phát triển trong thời gian gần đây. Bài toán định tuyến là một vấn đề cơ bản, truyền thống trong mạng máy tính, tuy nhiên định tuyến trong mạng cảm biến không dây lại là một vấn đề mới do đặc thù hanh he về mặt bộ nhớ, tối ưu năng lượng, sự bất ổn của cấu hình mạng (node vao ra). Có rất nhiều nghiên cứu, các thuật toán được công bố trong lĩnh vực này, tiêu biểu như các giao thức SPIN [2], LEACH [2], GPSR [5]

Một trở ngại không nhỏ đối với các nhà nghiên cứu là ... đánh giá chính xác hiệu năng của các nghiên cứu trước khi đem ra ứng dụng thực tế. Hướng tiếp cận phổ biến của các nhà nghiên cứu là các công cụ mô phỏng mạng như NS2, OpNet, QualNet... Tuy nhiên, các công cụ trên hầu hết mới phát triển để thực hiện các mô phỏng mạng có dây và mạng không dây thông thường, chưa có một công cụ chuyên dụng hỗ trợ tốt cho việc mô phỏng các giao thức trong mạng cảm biến không dây. Vì vậy, hầu hết các nghiên cứu về mạng cảm biến không dây đều sử dụng các công cụ tự phát triển (self developed). Bản thân trong quá trình nghiên cứu các giao thức định tuyến trong MCBKD, chúng tôi cũng gặp phải vấn đề này. Xuất phát

từ đó, chúng tôi đưa ra ý tưởng phát triển một công cụ mô phỏng mạnh, chuyên dụng, đáng tin cậy trong mô phỏng các giao thức trên mạng cảm biến không dây để đóng góp cho cộng đồng nghiên cứu mạng nói chung và hỗ trợ công việc nghiên cứu của nhóm nói riêng.

Với mục đích nêu trên, nghiên cứu này chọn nền tảng NS2, công cụ mô phỏng mạng phổ biến nhất hiện nay với hơn 44,44% nhà nghiên cứu sử dụng (hình 1) để phát triển và mở rộng. Mỗi trường mô phỏng NS2 rất lý tưởng để phát triển với mục đích mô phỏng các giao thức mạng cảm biến không dây do bản thân nó đã có sẵn mô hình năng lượng (energy model), các giao thức điều khiển truy nhập như SMAC, 802.15.4[8] và một số giao thức định tuyến có thể kế thừa được như AODV, DSR,... Tuy nhiên kiến trúc của nó khá phức tạp, vì vậy các nghiên cứu trước đây thường chưa đạt được một sự đồng bộ mang tính hệ thống. Đây chính là độ khó của vấn đề phát triển công cụ mạng hỗ trợ mô phỏng mạng cảm biến không dây. Ở nghiên cứu này, chúng tôi sẽ đi sâu nghiên cứu kiến trúc lõi NS2, các đặc trưng của mạng cảm biến không dây và dựa vào đó xây dựng các module lõi (core) mang tính hệ thống và phát triển một số thư viện giao thức cụ thể cho mạng cảm biến không dây.



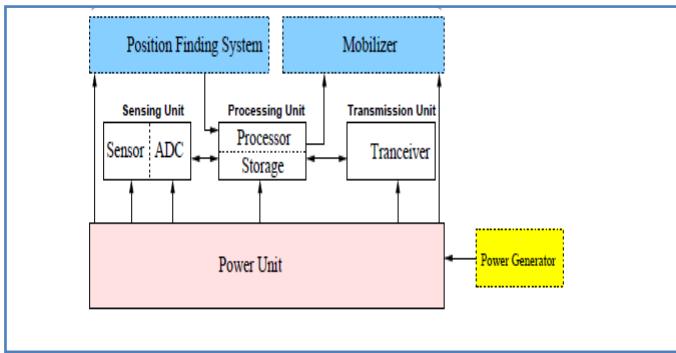
Hình 1: Các công cụ sử dụng mô phỏng mạng được sử dụng trong các bài báo của ACM 2000-2004 (theo thống kê của bài báo [15]).

Phần còn lại của báo cáo này được trình bày như sau: Ở phần 2, chúng tôi sẽ trình bày các kiến thức cơ bản liên quan đến mạng cảm biến không dây và công cụ mô phỏng NS2. Phần tiếp theo sẽ trình bày thiết kế chi tiết của công cụ mô phỏng chuyên dụng cho mạng cảm biến không dây dựa trên nền tảng NS2. Một số ví dụ và trường hợp thử nghiệm (case study) mô phỏng và đánh giá hiệu năng của các thuật toán định tuyến trong mạng cảm biến không dây sẽ được trình bày trong phần 4. Phần 5 sẽ tổng kết lại báo cáo.

2. CÁC VẤN ĐỀ LIÊN QUAN

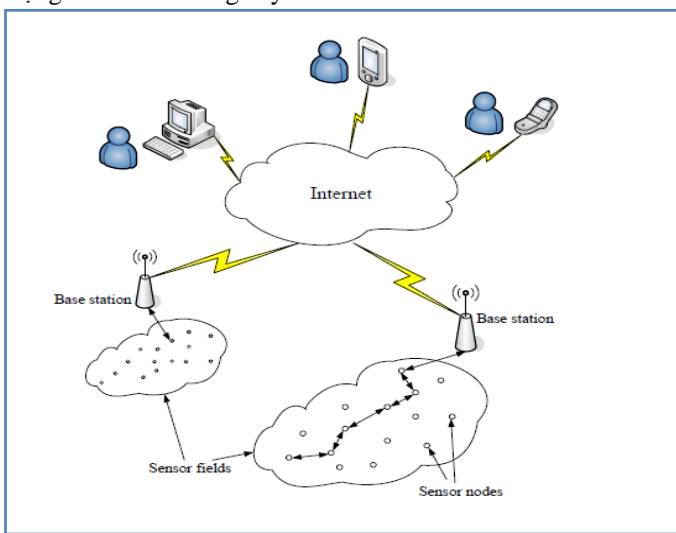
2.1. Đặc trưng của mạng cảm biến không dây

Hình 2 mô tả cấu trúc của một node mạng cảm biến không dây, bao gồm các thành phần: Processing Unit (PU), Sensor Unit, Transmission Unit, Position Finding Unit, Mobilizes Unit, Power Unit [1].



Hình 2: Cấu trúc của node cảm biến

Với sự phát triển vũ bão của công nghệ vi chip, kích thước cảm biến ngày càng bé, kéo theo đó là sự hạn chế về mặt khả năng tính toán của PU, cũng như hạn chế bộ nhớ của sensor. Bộ phận power unit thường là pin, có mức năng lượng hạn chế. Vì vậy làm sao để các node xử lý, tính toán đơn giản, tiết kiệm năng lượng là những thử thách cơ bản khi nghiên cứu về mạng cảm biến không dây.



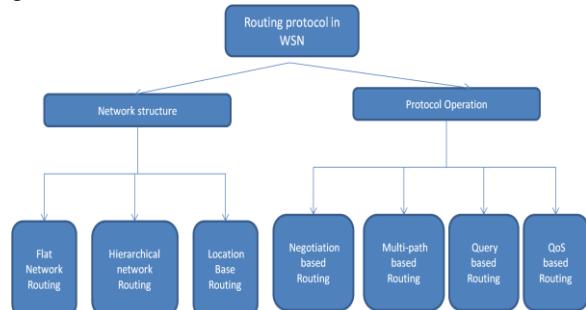
Hình 3: Mô hình chung của mạng cảm biến không dây

Mạng cảm biến không dây thường bao gồm các node cảm biến được đặt trong một vùng cảm biến (sensor field), và một hoặc nhiều trạm trung tâm (base station) được kết nối với mạng internet. Cấu hình của MCBKD thường xuyên thay đổi do khả năng di chuyển và khả năng hỏng hóc của các node mạng (Các node sensor thường được đặt trong tự nhiên nên rủi ro là khá cao.). Một khác, do thường triển khai trong tự nhiên, nên mạng thường hoạt động trong điều kiện địa hình xấu, nhiều vật cản, nhiều nhiễu. Do đó cần đảm bảo tính linh hoạt, khả năng hoạt động tốt khi có nhiều biến đổi là yêu cầu cấp

thiết trong việc thiết kế giao thức cho mạng cảm biến không dây.

2.2. Định tuyến trong mạng cảm biến không dây.

Định tuyến trong mạng cảm biến không dây phải đáp ứng các đặc trưng về bộ nhớ, khả năng tính toán và sự bất ổn của cấu hình mạng. Các giao thức định tuyến trong mạng cảm biến không dây có thể phân loại thành 2 nhóm, nhóm thứ nhất dựa trên cấu trúc mạng và nhóm thứ hai dựa trên các phép toán của các giao thức.



Hình 4: Các giao thức định tuyến trong MCBKD

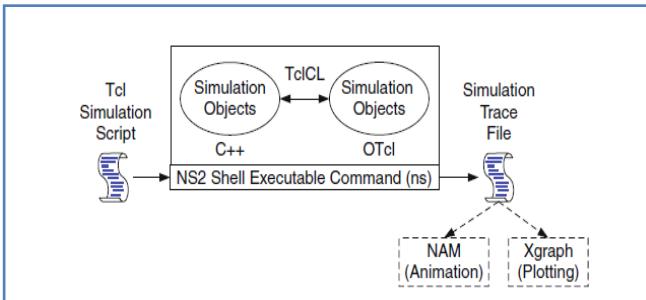
Báo cáo này tập trung vào nhóm thứ nhất và cụ thể đi sâu vào các giao thức định tuyến dựa trên thông tin địa lý. Các giao thức này dựa vào thông tin về vị trí vật lý của các node mạng để định tuyến. Việc lựa chọn node tiếp theo chủ yếu dựa vào ý tưởng của thuật toán tham ăn (Greedy). Ưu điểm nổi bật của các giao thức này là việc xử lý tin ở các node mạng khá đơn giản do các node mạng chỉ cần thông tin về láng giềng của nó, không nhất thiết phải lưu trữ thông tin về toàn mạng. Điều này phù hợp với các ràng buộc về khả năng tính toán và bộ nhớ trong mạng cảm biến. Diễn hình cho các giao thức này là các thuật toán: Flooding, Greedy, GPSR[5], GEAR [10], HAIR.

- Flooding: Nốt mạng sau khi nhận được gói tin cần chuyển tiếp chuyển gói tin cho tất cả các nốt láng giềng xung quanh.
- Greedy: Nốt mạng thực hiện truyền tin tham lam, theo đó gói tin sẽ được truyền tới nốt láng giềng nào gần đích nhất so với các nốt láng giềng còn lại và so với chính nốt đó.
- GPSR (Greedy perimeter stateless routing): Giải thuật phát triển từ Greedy giải quyết vấn đề cực tiểu địa phương (hay nói cách khác là khi việc truyền tin bị tắc).
- GEAR (Geographical and Energy Aware Routing): Mỗi nốt có một giá trị được xác định từ hàm đánh giá chi phí. Qua đó nốt cần truyền tin tìm nốt láng giềng có chi phí nhỏ nhất để gửi tin. Hàm đánh giá chi phí dựa vào 2 yếu tố là (1) khoảng cách tới đích và (2) năng lượng còn lại tại nốt đó.

2.3. Kiến trúc NS2:

NS2(Network Simulation version 2) là công cụ mô phỏng mạng mã nguồn mở được sử dụng phổ biến nhất hiện nay.

NS2 sử dụng đồng thời 2 ngôn ngữ hướng đối tượng là C++ và Otel. Cấu trúc tổng quan của NS2 được mô tả trong hình 5.



Hình 5: Kiến trúc cơ bản NS2

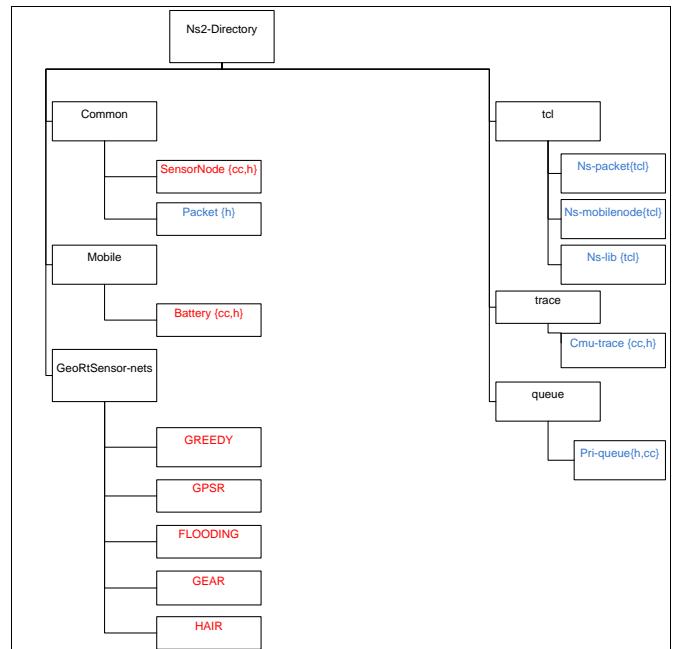
Nền tảng NS2 cung cấp các đối tượng quan trọng phục vụ mô phỏng như: *Node*, *Packet*, *Agent*...

- *Node* (Node, MobileNode, SatNode): là đối tượng thể hiện những thành phần cơ bản của một node mạng. Một node mạng trong mô phỏng là tổng hợp của nhiều thành phần khác nhau dựa trên đối tượng Node.
- *Packet*: là đối tượng thể hiện cho gói tin trao đổi giữa các node.
- *Agent*: là thể hiện của các giao thức. Mỗi giao thức hay ứng dụng trên các node mạng đều được thể hiện dưới dạng một agent. Các phuong thức quan trọng của đối tượng Agent là *recv()*, *send()*, *command()*. *Recv()* sử dụng để quy định cách thức xử lý với một gói tin nhận được của giao thức. *Send()* thể hiện quá trình xử lý tin trước khi gửi đi. *Command()* là một hàm quy định interface của giao thức, cho phép lời gọi từ Otel.

Hiểu rõ các đối tượng [13] và kiến trúc thực thi [14] là tiền đề của việc phát triển NS2. Trong khuôn khổ bài báo chúng tôi không mô tả toàn bộ kiến trúc hệ thống khá phức tạp của NS2, có thể tham khảo tại [13] [14].

3. PHÁT TRIỂN NỀN TẢNG NS2.

Với mục đích xây dựng một công cụ mô phỏng chuyên dụng cho mạng cảm biến không dây, trong nghiên cứu này chúng tôi xây dựng các module mới nhúng trên nền tảng NS2 nhằm cung cấp đối tượng node mạng kiểu cảm biến với đầy đủ các đặc trưng về cảm biến, năng lượng, vị trí, ... Đây là module mang tính hệ thống, đặt nền tảng cho toàn bộ các phát triển của chúng tôi hoặc của các nhóm nghiên cứu khác sau này. Hơn nữa, để hỗ trợ các nhà nghiên cứu trong việc mô phỏng các thuật toán định tuyến, chúng tôi chuẩn bị sẵn các module mô phỏng các giao thức định tuyến tiêu biểu trong mạng cảm biến không dây như FLOODING, GPSR, GEAR... Ngoài ra, để tăng tính tiện ích, để sử dụng của công cụ, chúng tôi nghiên cứu và xây dựng thêm module WTG (*WSN topology generator*), có chức năng tạo kịch bản mô phỏng và module Result Analyzer, có chức năng phân tích kết quả mô phỏng.

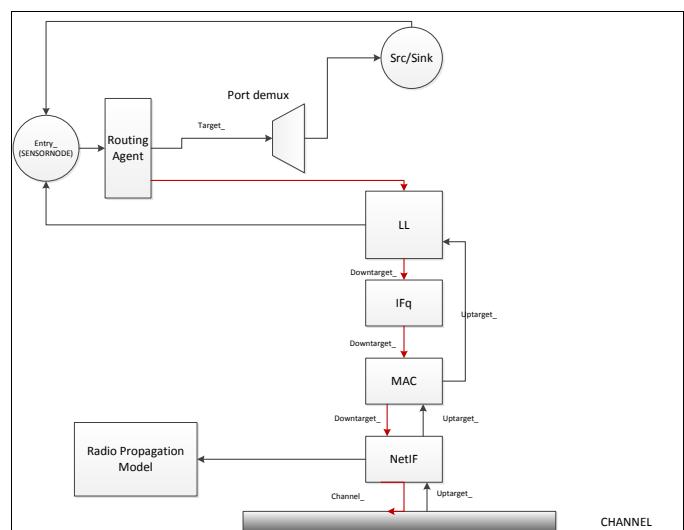


Hình 6: Các thành phần phát triển từ NS2
Màu đỏ: các file thêm mới; màu xanh: các file chỉnh sửa

3.1.SensorNode module:

SensorNode module là module cơ bản nhất, được mở rộng nhằm phục vụ cho việc mô phỏng không chỉ riêng các giao thức định tuyến trong mạng không dây mà toàn bộ các mô phỏng trong mạng cảm biến không dây nói chung. Module này gồm hai đối tượng là SensorNode và Battery(*SensorNode(.cc.h)*, *battery (.cc.h)*).

Đối tượng SensorNode (hình 7), tương tự như đối tượng Node, MobileNode, thể hiện các đặc trưng vật lý như cảm biến, khả năng tính toán, vị trí, địa chỉ basestation, khả năng di chuyển, trạng thái¹ (ON= hoạt động, OFF= dừng hoạt động), ...



Hình 7: Mô hình một node mạng cảm biến

Đối tượng Battery là mở rộng của đối tượng energyModel trong NS2, có nhiệm vụ biểu diễn mô hình năng lượng của node mạng cảm biến. Ngoài các thể hiện sẵn có được kế thừa

từ NS2, Battery thể hiện thêm mức năng lượng sử dụng để cảm biến, tính toán.

3.2. Geographic routing module:

Module này cung cấp thư viện các thuật toán định tuyến dựa trên thông tin địa lý tiêu biểu đã được nghiên cứu và phát triển.

Giao thức	Agent	Thư mục
Flooding	FLOODAgent	flooding
Greedy	GREEDYAgent	greedy
GPSR	GPSRAgent	GPSR
GEAR	GEARAgent	GEAR
HAIR	HAIRAgent	HAIR

Bảng 1: Các thuật toán định tuyến tiêu biểu

Cấu trúc một thư mục *Name* (name dùng chung để chỉ tên các giao thức trên) gồm các file: *nameAgent{.cc,.h}*, *nameNeighbor{.cc,.h}*, *name_packet.h*.

- *name_packet.h*: định nghĩa gói tin sử dụng trong giao thức.
- *nameNeighbor{.cc,.h}*: thể hiện bảng láng giềng, lưu trữ thông tin về các node mạng kế cận mà nó có thể truyền tin được.
- *nameAgent{.cc,.h}* : định nghĩa giao thức, quyết định cách nhận và gửi gói tin . Lớp này kế thừa từ lớp Agent (*/common/Agent*). Trong lớp này, đặc biệt chú ý override các hàm *recv()*, *send()*, *command()*.

3.3. WSN Topology Generator (WTG)

WTG là module được phát triển với mục đích cung cấp giao diện đồ họa trợ giúp việc thiết lập các kịch bản mô phỏng về mạng cảm biến không dây. Module được phát triển dựa trên ngôn ngữ Java, có thể chạy tốt trên các hệ điều hành Linux hay Window. Do tính độc lập của module, chúng tôi đã phát triển nó thành một công cụ riêng. Hiện tại chúng tôi đã hoàn thành phiên bản 1.0 và đang trong quá trình thực hiện phiên bản 1.1.

Với phiên bản 1.0, công cụ có khả năng hỗ trợ xây dựng nhiều kiểu kịch bản bằng các cách rất thân thiện, tiện dụng như:

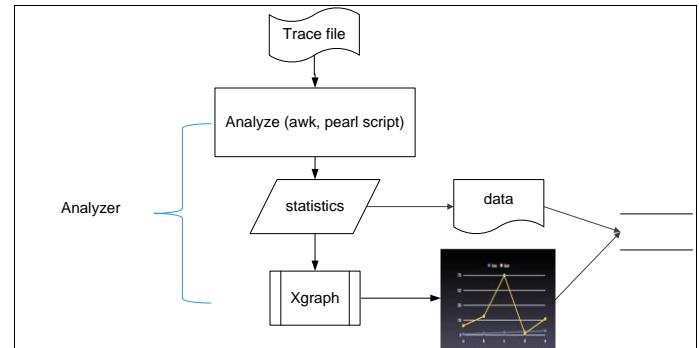
- Tạo và lựa chọn vị trí các node bằng việc kéo thả.
- Cấu hình node theo tùy chọn.
- Thiết lập các giao thức, các tham số.
- Thiết lập các sự kiện mô phỏng (event) như thời gian bắt đầu truyền tin, thời gian kết thúc...
- Cung cấp chế độ tạo kịch bản ngẫu nhiên với số lượng node lớn
- ...

Ngoài ra công cụ còn cung cấp sẵn một số kịch bản và topology có sẵn như topology hình lưới...

3.4 Result Analyzer :

Nếu như WTG là module được phát triển nhằm xây dựng công cụ đồ họa tạo ra kịch bản hay đầu vào của mô trường mô phỏng NS2, thì module result analyzer được phát triển nhằm cung cấp giao diện đồ họa phân tích kết quả mô phỏng. Kết quả mô phỏng của NS2 là một dạng text lưu vết của gói tin, dữ liệu là khá lớn và ở dạng thô. Để thu được thông tin thống kê,

phải sử dụng các công cụ trích rút dữ liệu như awk hay perl. Nhằm hỗ trợ quá trình này, chúng tôi xây dựng công cụ result analyzer, cung cấp khả năng lọc dữ liệu, đưa ra kết quả dưới dạng thống kê hoặc đồ thị.



Hình 8: Mô hình công cụ result analyzer

Hiện tại công cụ có thể cung cấp các chức năng đưa ra các thông số để đánh giá các thuật toán định tuyến như:

- Số lượng gói tin được gửi
- Tỷ lệ gói tin đến đích
- Tiết kiệm năng lượng
- Tối ưu về đường đi
- ...

Công cụ vẫn đang được phát triển và hoàn thiện trong thời gian tới.

4. MÔ PHỎNG, ĐÁNH GIÁ MỘT SỐ GIAO THỨC ĐỊNH TUYẾN TRÊN MẠNG CẢM BIẾN KHÔNG DÂY

Trong phần này chúng tôi sẽ trình bày việc sử dụng công cụ mô phỏng chuyên dụng đã trình bày trong phần trước để mô phỏng, đánh giá và so sánh hiệu năng của một số thuật toán định tuyến trong mạng cảm biến không dây theo một số kịch bản. Cụ thể ở đây sẽ thử nghiệm, đánh giá và so sánh hiệu năng của ba giải thuật: GPSR, AODV [12] và DSR [13]. Việc đánh giá sẽ được thực hiện dựa vào hai tiêu chí: độ dài đường đi của gói tin đến đích và khả năng tiết kiệm năng lượng của giao thức.

4.1. Mô phỏng:

Bà giao thức đều thử nghiệm với các kịch bản gồm 10, 50 và 100 node cảm biến, thời gian mô phỏng là 50s. Các kịch bản đều được sinh tự động bởi công cụ WTG (*WSN topology generator*). Các node mạng sử dụng giao thức MAC 802.11. Mức năng lượng ban đầu: 100, sự tiêu hao năng lượng mặc định ở các trạng thái như trong bảng 2.

Trạng thái	Năng lượng tổn hao (J/s)
Idle	1.0
txPower	1.0
rxPower	2.0
sleepPower	0.01

Bảng 2: Tiêu hao năng lượng/ trạng thái

Kịch bản sử dụng ứng dụng CBR, với các packet_size_=50, interval_time_=2.0.

Để thực hiện mô phỏng với mạng cảm biến không dây, cần cấu hình node mạng:

```
$ns node-config -sensorNode ON \
    -adhocRouting $val(rp) \
    -lltype $val(ll) \
    -macType $val() \
    -ifqType $val(ifq) \
    -ifqLen $val(ifqlen) \
```

Ở đây tùy chọn `-sensorNode ON` để thông báo node mạng sử dụng là node mạng cảm biến. Tùy chọn `-adhocRouting $val(rp)` thiết lập giao thức định tuyến, ở đây để thực hiện mô phỏng 3 giao thức GPSR AODV và DSR nên `$val(rp)` lần lượt là GPSR, AODV và DSR

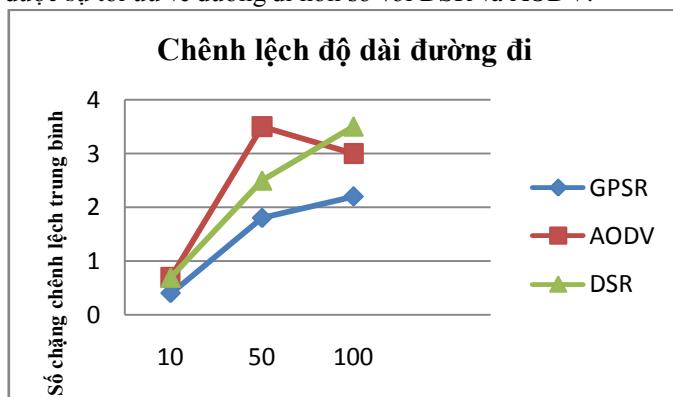
4.2 Đánh giá hiệu năng:

Kết quả mô phỏng được trả về dưới dạng một file text, trace file, và được phân tích dựa vào module Analyzer. Kết quả phân tích tập trung vào hai tiêu chí là: tối ưu đường đi gói tin và khả năng tiết kiệm năng lượng.

Tối ưu về đường đi gói tin

Tối ưu đường đi gói tin được đánh giá bằng độ dài đường đi tới đích của gói tin so với đường đi tối ưu hay chính là chênh lệch số chặng (gói tin đến đích) và số chặng tối ưu có thể có. $Path\ length = \text{số chặng} - \text{số chặng tối ưu}$

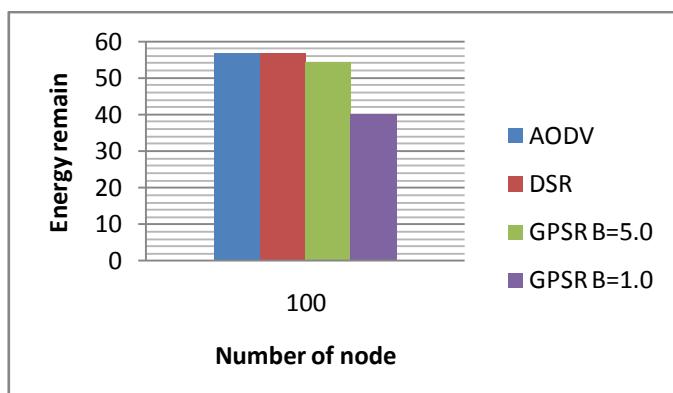
Từ kết quả mô phỏng (hình 8) có thể thấy GPSR gần đạt được sự tối ưu về đường đi hơn so với DSR và AODV.



Hình 9: Đồ thị độ tối ưu về đường đi

Tiết kiệm năng lượng:

Tiêu chí tiết kiệm năng lượng được tính theo tổng mức năng lượng còn lại của các node trên tổng mức năng lượng ban đầu.



Hình 10: Khả năng tiết kiệm năng lượng của các giao thức

Ở đây B là thời gian định kỳ gửi các gói tin beacon trong giao thức GPSR. Đồ thị về mức năng lượng ở kích thước 100 node mạng cho thấy rằng trong trường hợp về năng lượng GPSR không tốt bằng AODV và DSR.

5. KẾT LUẬN

Phát triển nền tảng NS2 nhằm phục vụ cho các mô phỏng giao thức định tuyến trong mạng không dây là kết quả của nghiên cứu này. Với hướng mở rộng hơn với các bài toán mạng cảm biến không dây khác như ứng dụng hay các giao thức điều khiển truy nhập (MAC), nhóm sẽ tiếp tục bổ sung thư viện nhằm hoàn thiện hơn nữa. Đồng thời, nghiên cứu sẽ phát triển để tiếp cận với bài toán định tuyến cảm biến không dây trong điều kiện địa hình xấu, bài toán cá nhân hóa mà nhóm đang nghiên cứu.

6.LỜI CẢM ƠN

Chúng tôi xin chân thành cảm ơn TS Nguyễn Khanh Văn và Ths Nguyễn Phi Lê (giảng viên viện CNTT&TT trường ĐH Bách Khoa) đã giúp đỡ, hướng dẫn chúng tôi rất nhiều trong quá trình thực hiện nghiên cứu này!

TÀI LIỆU THAM KHẢO

- [1] Ian F. Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, Erdal Cayirci “A survey on Sensor Network” IEEE Communication Magazine August 2002
- [2] Jamal N.Al-Karaki, Ahmed E.Kamal “Routing techniques in wireless sensor network: a survey”
- [3] C.Shen, C.Srisathapornphat , C. Jaikakeo “Sensor Information Networking Architecture and Application” IEEE Pers Commun. August 2001 , pp.52-59
- [4] K.Sohrabi et al “A protocol for self-organization of WSN” IEEE Pers Commun. Oct 2000 , pp.16-19
- [5] B.Karp , H.T Kung “GPSR: Greedy perimeter stateless routing for wireless sensor network” in the Proceeding of 6th Annual ACM/IEEE International Conference on Mobile Computing and Networking (Mobile ‘00) Boston,MA,August 2000.
- [6] T.He et al., “SPEED: A stateless protocol for real time communication in sensor network” the Proceeding International Conference on Distributed Computing System, May 2003.
- [7] N. Bulusu, J Heideman, D. Estrin; “GPS-less low cost outdoor localization for very small devices”, Technical report 00-079 Computer Science department, USC Apr 2000
- [8] Sunil K., Vineet S., Jing Deng ; “Medium Access Control protocol in ad-hoc wireless sensor network : A survey”
- [9] Q.Fang,J.Gao, L.J Guibas “Locating and bypassing routing hole in sensor network” In Infocom 2004, vol 4, March 2004
- [10] Y.Yu, R.Govindan, D.Estrin “Geographical and Energy Aware Routing : A recursive data dissemination protocol for WSN”
- [11] Charles E. Perkins, Elizabeth M.Royer “Ad-hoc On-Demand Distance Vector Routing”
- [12] David B. Johnson, David A. Maltz “Dynamic Source Routing in Ad hoc Wireless Networks”
- [13] NS Manual
- [14] Introduction to Network Simulation 2
- [15] S.Kurkowski, T.Camp and M. Colagrossi. “MANET simulation studies: The current state and new simulation tool” Technical report, Department of Math and Computer Sciences,Colorado School of Mines, MCS-05-02, February 2005.

Xây dựng thư viện khung song song dữ liệu cho hệ thống nhiều bộ xử lý đồ họa

Nguyễn Minh Tháp, Ngô Huy Hoàng

Tóm tắt—Bộ xử lý đồ họa (GPU) có tốc độ tính toán và băng thông bộ nhớ cao hơn của bộ xử lý trung tâm (CPU) nhiều lần. Kiến trúc của GPU phù hợp với lớp bài toán song song dữ liệu. Tuy nhiên mô hình lập trình cho GPU là CUDA vẫn gây khó khăn cho người sử dụng. Bài báo trình bày công việc xây dựng một bộ thư viện khung song song dữ liệu cho GPU có giao diện đơn giản nhưng tận dụng được hiệu năng cao của GPU. Thư viện này có hỗ trợ cả hệ thống một và nhiều GPU. Thư viện bao gồm các khung song song dữ liệu – các hàm tính toán song song trên các phần tử của dữ liệu, gọi tắt là các khung. Các khung này là nền tảng xây dựng của nhiều ứng dụng cấp cao hơn. Các thử nghiệm trên các khung và các ứng dụng tính tích vô hướng, độ lệch chuẩn, hệ số tương quan Pearson sẽ minh chứng tính tiện dụng và hiệu năng cao của thư viện.

Từ khóa—GPU Computing, Algorithmic Skeletons, Data Parallelism, CUDA.

1. GIỚI THIỆU

Ngày nay, bộ xử lý đồ họa (GPU) có tốc độ tính toán và băng thông bộ nhớ cao hơn của bộ xử lý trung tâm (CPU) nhiều lần [1]. Nguyên nhân là do GPU được thiết kế cho công việc tính toán nhiều và có tính song song cao trong đồ họa. Do đó, trong GPU có nhiều thành phần dành cho xử lý dữ liệu là các bộ ALU hơn là bộ nhớ đệm và thành phần điều khiển.

Với kiến trúc như vậy, GPU phù hợp với lớp bài toán tính toán song song dữ liệu – tính toán giống nhau được thực hiện song song trên nhiều phần tử của dữ liệu. Nhiều ứng dụng xử lý trên dữ liệu lớn có thể áp dụng mô hình tính toán song song dữ liệu để tăng tốc. Ví dụ một ứng dụng là bài toán tô trát ảnh 3D: các điểm ảnh có thể được xử lý song song với nhau.

Tuy nhiên mô hình lập trình cho GPU là CUDA vẫn gây khó khăn cho người sử dụng. Người sử dụng CUDA phải phân chia bài toán cần giải quyết thành các bài toán con có thể được giải quyết song song và xem xét các vấn đề truyền thông, đồng bộ giữa các bài toán con này. Sau đó, người dùng còn phải ánh xạ chúng vào mô hình lập trình CUDA. Ngoài ra, lập trình trên CUDA để đạt được hiệu năng cao là không đơn giản, đòi hỏi có

Công trình này được thực hiện dưới sự bảo trợ của Trung tâm Tính toán hiệu năng cao, trường Đại học Bách khoa Hà Nội.

Nguyễn Minh Tháp, sinh viên lớp Kỹ sư tài năng-Công nghệ thông tin, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (diện thoại: 0169-6312-091, e-mail: towernguyenminh@gmail.com).

Ngô Huy Hoàng, sinh viên lớp Tin Pháp, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (diện thoại: 098-7862-360, e-mail: hoang.ngo.h@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

kiến thức sâu về kiến trúc cả phần cứng và phần mềm.

Để giải quyết khó khăn cho người dùng, chúng tôi xây dựng bộ thư viện trên ngôn ngữ C có giao diện đơn giản nhưng tận dụng được hiệu năng cao của GPU. Bộ thư viện bao gồm những hàm tính toán song song dữ liệu cơ bản, có ứng dụng cao gọi là các *khung* (skeletons) song song dữ liệu, gọi tắt là khung.

Thư viện bao gồm các khung *dịch* (shift), *ánh xạ* (map), *thu gọn* (reduce), *quét* (scan). Các khung áp dụng một phép tính toán song song lên các phần tử của một cấu trúc dữ liệu. Ví dụ, khung ánh xạ áp dụng hàm bình phương lên các phần tử của một mảng thu được một mảng có các phần tử là bình phương của các phần tử của mảng ban đầu.

Thư viện được lập trình để chạy trên cả một và nhiều GPU nhằm tận dụng tối đa tài nguyên của hệ thống. Việc sử dụng nhiều GPU hứa hẹn sẽ làm tăng tốc độ và khả năng tính toán.

Các thử nghiệm được thực hiện là phép tích vô hướng, độ lệch chuẩn và hệ số tương quan Pearson. Đây là các ví dụ điển hình áp dụng các khung trong thư viện. Kết quả đạt được là tốc độ tính toán sử dụng thư viện gấp hàng chục hàng trăm lần tốc độ tính toán trên CPU. Ngoài ra, tốc độ sử dụng thư viện cũng không thua kém nhiều việc lập trình trực tiếp cho GPU – công việc vốn tốn kém thời gian và công sức.

Các phần còn lại của bài báo được tổ chức như sau: phần 2 trình bày về khung song song dữ liệu, phần 3 trình bày thực thi trên CUDA, phần 4 thử nghiệm và đánh giá, cuối cùng phần kết luận nằm ở phần 5.

2. KHUNG SONG SONG DỮ LIỆU

Các khung song song dữ liệu có trong thư viện bao gồm *dịch* (shift), *ánh xạ* (map), *thu gọn* (reduce) và *quét* (scan). Các khung áp dụng một phép tính toán song song lên một cấu trúc dữ liệu. Có nhiều cấu trúc dữ liệu như véc tơ, ma trận, cây,... Các khung trong thư viện hiện tại được thực hiện trên cấu trúc dữ liệu véc tơ. Một véc tơ có n phần tử được sắp xếp theo thứ tự liên tiếp và đánh chỉ số từ 0 đến $n-1$. Kí hiệu một véc tơ n phần tử như sau:

$$[x_0, x_1, \dots, x_{n-1}]$$

Các phần sau sẽ trình bày các khung song song trên dữ liệu véc-tơ này.

2.1. Khung dịch (shift)

Khung dịch bao gồm dịch trái và dịch phải, kết hợp lại ta có định nghĩa khung dịch k vị trí với phần tử chèn id (thường là 0) trên một véc tơ n phần tử sinh ra một véc tơ có cùng số phần tử. Nếu $k \geq 0$ là phép dịch trái, các phần tử có chỉ số nhỏ hơn k bằng id , các phần tử khác bằng phần tử có chỉ số nhỏ hơn k trong véc

tờ ban đầu:

$$\begin{aligned} & \text{shift}(k, id, [x_0, x_1, \dots, x_{n-1}]) \\ &= [id, \dots, id, x_0, \dots, x_{n-1-k}], k \geq 0 \end{aligned}$$

Nếu $k < 0$ là phép dịch phải, các phần tử có chỉ số lớn hơn $n-1-k$ bằng id , các phần tử khác bằng phần tử có chỉ số lớn hơn $|k|$ trong vec tờ ban đầu:

$$\begin{aligned} & \text{shift}(k, id, [x_0, x_1, \dots, x_{n-1}]) \\ &= [x_k, \dots, x_{n-1}, id, \dots, id], k < 0 \end{aligned}$$

Khung dịch có tính song song dữ liệu. Thật vậy, nếu ta sử dụng n bộ xử lí mỗi bộ thực hiện dịch một phần tử thì chúng có thể thực hiện song song với nhau:

```
for i=0 to n-1 do in parallel
  if (i-p >= 0) and (i-p < n) then
    x[i]=x[i-p];
  else
    x[i]=id;
  end if
end for
```

Trong trường hợp này, độ phức tạp của khung dịch lần lượt là $O(1)$.

2.2. Khung ánh xạ (map)

Khung ánh xạ lên một vec tờ áp dụng một hàm số một biến f lên các phần tử của một vec tờ:

$$map(f, [x_0, x_1, \dots, x_{n-1}]) = [f(x_0), f(x_1), \dots, f(x_{n-1})]$$

Khung ánh xạ lên hai vec tờ áp dụng một hàm số hai biến f lên các cặp phần tử tương ứng của hai vec tờ:

$$\begin{aligned} & map2(f, [x_0, x_1, \dots, x_{n-1}], [y_0, y_1, \dots, y_n]) \\ &= [f(x_0, y_0), f(x_1, y_1), \dots, f(x_{n-1}, y_{n-1})] \end{aligned}$$

Khung ánh xạ lên một vec tờ và một hằng số áp dụng một hàm số hai biến f lên các phần tử của vec tờ và hằng số đó:

$$\begin{aligned} & map2b1(f, [x_0, x_1, \dots, x_{n-1}], a) \\ &= [f(x_0, a), f(x_1, a), \dots, f(x_{n-1}, a)] \end{aligned}$$

Sử dụng n bộ xử lí mỗi bộ thực hiện hàm số trên một phần tử:

```
for i=0 to n-1 do in parallel
  f(x[i]);
  //hoặc f(x[i], y[i]);
  //hoặc f(x[i], a);
end for
```

Nếu thời gian tính hàm f là $O(1)$ và số lượng bộ xử lí là $O(n)$ thì độ phức tạp của các khung ánh xạ là $O(1)$.

2.3. Khung thu gọn (reduce)

Khác với các khung dịch và ánh xạ, khung thu gọn có đầu ra là một phần tử duy nhất. Khung thu gọn áp dụng một phép toán hai ngôi \oplus lên các phần tử của vec tờ:

$$reduce(\oplus, [x_0, x_1, \dots, x_{n-1}]) = x_0 \oplus x_1 \oplus \dots \oplus x_{n-1}$$

Để có thể song song hóa được thì phép toán \oplus phải thỏa mãn tính chất kết hợp. Để thực thi cho GPU nó còn phải thỏa mãn thêm tính chất giao hoán. Rất may các phép toán thường sử dụng thường thỏa mãn các tính chất này như phép cộng, phép nhân, phép và, phép hoặc, phép lấy max và phép lấy min.

Để đơn giản ta minh họa trong trường hợp $n = 2^k$, dùng n bộ xử lí thực hiện khung thu gọn theo mô hình cây cân bằng:

```
p=n/2;
while p>0 do
  for i=0 to p-1 do in parallel
    x[i]=x[2i]⊕x[2i+1];
  end for
end while
```

Độ phức tạp khung thu gọn là $O(log n)$.

Khung thu gọn có thể kết hợp với khung ánh xạ để tạo thành khung phức hợp. Ví dụ kết hợp khung thu gọn và khung ánh xạ lên hai vec tờ:

$$\begin{aligned} & reduce \circ map2(\oplus, f, [x_0, x_1, \dots, x_{n-1}], [y_0, y_1, \dots, y_n]) \\ &= reduce(\oplus, map2(f, [x_0, x_1, \dots, x_{n-1}], [y_0, y_1, \dots, y_n])) \\ &= f(x_0, y_0) \oplus f(x_1, y_1) \oplus \dots \oplus f(x_{n-1}, y_{n-1}) \end{aligned}$$

Sử dụng các khung phức hợp thay thế các khung cơ bản sẽ đạt được hiệu năng cao hơn [2] dẫn đến việc phải thực thi các khung này.

2.4. Khung quét (scan)

Khung quét còn có tên gọi khác là *tổng trước* (prefix-sum). Khung này áp dụng một phép toán hai ngôi \oplus lên các phần tử của vec tờ thu được đầu ra là một vec tờ:

$$\begin{aligned} & scan(\oplus, [x_0, x_1, \dots, x_{n-1}]) = \\ & [x_0, x_0 \oplus x_1, \dots, x_0 \oplus x_1 \oplus \dots \oplus x_{n-1}] \end{aligned}$$

Điều kiện đối với phép toán \oplus của khung quét cũng giống như phép toán của khung thu gọn.

Để đơn giản ta minh họa trong trường hợp $n = 2^k$, dùng n bộ xử lí thực hiện khung quét theo kĩ thuật con trỏ nhảy:

```
for i=0 to k-1 do
  for j=2^{i-1} to n-1 do in parallel
    x[j]=x[j]⊕x[j-2^{i-1}];
  end for
end for
```

Độ phức tạp khung thu gọn là $O(log n)$.

3. THỰC THI THƯ VIỆN TRÊN CUDA

Phần này sẽ trình bày công việc thực thi thư viện khung song song dữ liệu trên CUDA. Thư viện bao gồm các phần: cấu trúc vec tờ và các hàm làm việc với vec tờ như hàm tạo, hàm hủy, các khung là các hàm tính toán trên vec tờ, cách định nghĩa hàm và phép toán dùng làm đầu vào cho các khung. Nhưng trước hết là phần giới thiệu về mô hình lập trình song song CUDA.

3.1. Mô hình lập trình song song CUDA

Như đã giới thiệu ở phần trên, CUDA là một ngôn ngữ lập trình cho GPU và là một mở rộng của C [1]. Chính vì vậy thư viện được lập trình trên C. Một số khái niệm cơ bản trong CUDA:

Kernel là đoạn chương trình được biên dịch để thực hiện song song cho GPU.

Kernel được thực hiện trên GPU bởi một *grid* bao gồm nhiều *thread* - đơn vị thực hiện xử lí nhỏ nhất. Các thread được nhóm

vào các *block*, một grid bao gồm một hoặc nhiều block. Các thread trong cùng block có thể được đồng bộ và cùng truy nhập bộ nhớ chia sẻ. Các thread của các block khác không thể truy nhập bộ nhớ chia sẻ của block này cũng như đồng bộ với các thread cho đến khi kernel kết thúc thực hiện.

Bộ nhớ toàn cục là bộ nhớ của GPU mà tất cả các thread có thể truy nhập được. Muốn sử dụng bộ nhớ này phải sử dụng các lệnh sao chép dữ liệu của CUDA.

Mỗi block được thực thi bởi một *đa xử lý theo luồng*, viết tắt *SM* (Stream Multi-processor). SM không thực hiện song song toàn bộ các thread trong block mà theo từng warp – một tập các thread có id liên tiếp nhau. Số lượng thread trong warp có thể thay đổi khi kiến trúc thay đổi, hiện nay là 32.

3.2. Cấu trúc véc tơ

Cấu trúc véc tơ phục vụ việc tính toán trên GPU, được định nghĩa là một C struct tên *Vector* gồm hai trường: trường *dim* chứa số phần tử của véc tơ, trường *data* là con trỏ trả về đoạn bộ nhớ toàn cục trên GPU lưu trữ các phần tử của véc tơ. Trong phiên bản cho nhiều GPU, con trỏ *data* được thay thế bởi mảng con trỏ *data* với mỗi phần tử của mảng là một con trỏ trả về đoạn bộ nhớ toàn cục trên mỗi GPU lưu trữ một phần các phần tử của véc tơ.

Việc lưu trữ số lượng số phần tử trên mỗi GPU là không cần thiết. Bởi vì cách lưu trữ đảm bảo việc tính số lượng này là đơn giản: số lượng phần tử trên GPU thứ i ($0 \leq i < \text{số_GPU}$) hoặc là $\text{dim}/\text{số_GPU} + 1$ nếu $i < \text{dim \% số_GPU}$ hoặc là $\text{dim}/\text{số_GPU}$ nếu ngược lại.

Các hàm làm việc với véc tơ:

Hàm tạo có hai hàm tạo

```
makeVoidVector(Vector *Vtr,
               int inDim)
```

khởi tạo véc tơ Vtr có inDim phần tử và

```
makeVector(Vector *Vtr,
           Datatype *inData,
           int inDim)
```

khởi tạo véc tơ Vtr có inDim phần tử sao chép từ con trỏ inData.

Hàm hủy deleteVector (Vector Vtr) giải phóng bộ nhớ của véc tơ.

Trước khi sử dụng các khung để tính toán, cần sao chép dữ liệu lên véc tơ và sau khi tính toán xong, lại sao chép lại dữ liệu (đã thay đổi) từ véc tơ. *Các hàm sao chép dữ liệu* lên và từ véc tơ

```
setVector(Vector *Vtr,
          Datatype *inData)
```

và getVector (Datatype *outData,
 Vector *Vtr)

3.3. Các khung song song dữ liệu

Khung dịch shift (Vector *outVtr,
 Vector *inVtr,
 int k,
 Datatype id)

Khung ánh xạ map (Vector *outVtr,
 Vector *inVtr,
 Function f)

```
map2 (Vector *outVtr,
       Vector *inVtr1,
       Vector *inVtr2,
       Function f)
map2b1 (Vector *outVtr,
        Vector *inVtr,
        Datatype *inScalar,
        Function f)
```

Khung thu gọn reduce (Datatype *outScalar,
 Vector *inVtr,
 Operator op)

Khung phức hợp

```
mapreduce (Datatype *outScalar,
           Vector *inVtr,
           Function f,
           Operator op)
```

```
map2reduce (Datatype *outScalar,
            Vector *inVtr1,
            Vector *inVtr2,
            Function f,
            Operator op)
```

```
map2b1reduce (Datatype *outScalar,
               Vector *inVtr,
               Datatype *inScalar
               Function f,
               Operator op).
```

Khung quét scan (Vector *outVector,
 Vector *inVtr,
 Operator op)

Trong đó, các tham số vào, ra bao gồm cả các hàm (Function), phép toán (Operator) đã được mô tả ở phần trước.

Việc thực thi các khung này đòi hỏi phải cải biến thuật toán song song lí thuyết của chúng phù hợp với kiến trúc GPU và mô hình lập trình GPU. Cụ thể:

Đối với các khung dịch, ánh xạ, các phần tử được chia đều cho từng thread thực hiện song song. Tuy nhiên việc phân chia đảm bảo các thread trong cùng warp thao tác trên các phần tử ở ô nhớ kế tiếp vì các thao tác này có thể được phần cứng gộp lại: 32 phép đọc, ghi trong 1 warp có thể được gộp lại thành 2 phép.

Đối với khung thu gọn, ban đầu ta cũng phân chia như trên vì số phần tử lớn hơn số thread. Các thread trong cùng block thực hiện thu gọn theo thuật toán song song đã nêu ở phần trên. Cuối cùng do không có cơ chế đồng bộ các block, cần một kernel để thu gọn các kết quả của các block.

Khung quét được thực hiện với ý tưởng cần phân chia dữ liệu của véc tơ thành nhiều đoạn, thực hiện quét song song trên các đoạn, cuối cùng trên mỗi đoạn cần cập nhật thêm một giá trị bằng thu gọn của tất cả các phần tử thuộc các đoạn trước. Khung quét được thực hiện theo thứ tự: đầu tiên các warp quét trên các phần tử của nó, một warp tiến hành quét các kết quả cuối của các warp trong block và cập nhật các kết quả của block. Cuối cùng để cập nhật các kết quả trên các block, cần lặp lại việc quét trên các kết quả cuối của các block.

3.4. Định nghĩa hàm và phép toán

Để định nghĩa hàm và phép toán, kĩ thuật con trỏ hàm (hoặc đối tượng hàm trong C++) thường được sử dụng [3], ở đây là con trỏ hàm cho GPU. Tuy nhiên, lập trình CUDA là phụ thuộc kiến trúc của GPU. Kiến trúc của GPU được sử dụng trong cài đặt và thử nghiệm thư viện có số phiên bản là 1.3 không hỗ trợ con trỏ hàm cho GPU [1]. Vì vậy để định nghĩa hàm và phép toán sử dụng trong các khung chỉ còn cách dùng các macro tựa hàm. Ví dụ, định nghĩa hàm bình phương có thể dùng macro

```
#define sqr(x) ((x)*(x))
```

Một ví dụ khác, để định nghĩa phép toán cộng có thể dùng

```
#define plus(x,y) ((x)+(y))
```

3.5. Chọn cấu hình block, thread tốt nhất

Việc chọn cấu hình số lượng thread trong một block và số lượng block trong grid của kernel có ảnh hưởng tới hiệu năng đạt được. Các cấu hình khác nhau cho các hiệu năng rất khác nhau, hiệu năng cấu hình tốt đem lại gấp nhiều lần hiệu năng của cấu hình kém.

Trong thư viện có tích hợp khả năng tự chọn cấu hình tốt hoàn toàn trong suốt đối với người sử dụng. Do đó người dùng không cần quan tâm tới đặc điểm kĩ thuật này của lập trình CUDA.

3.6. Thực thi cho nhiều GPU

Trước đây CUDA (từ phiên bản 3 trở về trước) chỉ cho phép mỗi luồng của CPU chỉ quản lý được một GPU tại một thời điểm, nếu chuyển sang quản lý GPU khác thì toàn bộ dữ liệu trên GPU cũ sẽ mất. Các luồng khác nhau không thể chia sẻ dữ liệu dù trên cùng một GPU. Tuy nhiên, phiên bản CUDA 4.0 đã khắc phục nhược điểm này. Thư viện hiện tại sử dụng phiên bản CUDA này.

Khung dịch không được thực thi cho nhiều GPU vì nó dẫn đến việc GPU này tham chiếu đến dữ liệu trên GPU khác làm phức tạp quá trình quản lý dữ liệu, hiệu năng đạt được không cao lên so với chạy trên một GPU.

Khung ánh xạ được thực hiện hoàn toàn song song trên các GPU vì quá trình xử lí trên các GPU là độc lập.

Khung thu gọn sau khi thực hiện thu được kết quả từng phần trên các GPU cần một bước cuối cùng thu gọn các kết quả này.

Khung quét thực hiện trên từng GPU, trong đó phần tử cuối của kết quả trên từng GPU là thu gọn của phần dữ liệu trên GPU đó. Để đi tới kết quả cuối cùng cần cập nhật kết quả trên từng GPU thêm một giá trị bằng thu gọn của các kết quả trên các GPU có chỉ số thấp hơn nó (phép cập nhật chính là khung ánh xạ map2b1) mà các thu gọn này là thu gọn của các phần tử cuối đã đề cập ở trên.

4. THỬ NGHIỆM VÀ ĐÁNH GIÁ

Các thử nghiệm bao gồm tính tích vô hướng, độ lệch chuẩn, tham số tương quan Pearson. Thử nghiệm được thực hiện trên máy pnode3 của cụm bkluster, cấu hình máy: hai bộ xử lí đồ họa kép GTX 295 có 4 GPU kiến trúc 1.3, bộ xử lí trung tâm Intel Xeon 3040 1.86 GHz với 2 GB RAM.

4.1 Tích vô hướng

Tích vô hướng của hai vec tơ là tổng của các tích các cặp tọa độ tương ứng của hai vec tơ:

$$dot([x_0, x_1, \dots, x_{n-1}], [y_0, y_1, \dots, y_{n-1}])$$

$$= x_0 * y_0 + x_1 * y_1 + \dots + x_{n-1} * y_{n-1}$$

Tính theo khung là đầu tiên áp dụng khung map2 với hàm tích, rồi áp dụng tiếp khung reduce với phép cộng:

$$reduce(+, map2(*, [x_0, x_1, \dots, x_{n-1}], [y_0, y_1, \dots, y_{n-1}]))$$

Hoặc chỉ cần áp dụng một khung phức hợp map2reduce:

$$map2reduce(+, *, [x_0, x_1, \dots, x_{n-1}], [y_0, y_1, \dots, y_{n-1}])$$

Bảng 1. Kết quả thử nghiệm tích vô hướng

log ₂ n	Tuần tự	Hai khung 1 GPU	Một khung 1 GPU	Một khung 4 GPU
20	3.5254	0.1986	0.1066	0.0859
21	6.9870	0.3672	0.1836	0.1061
22	13.9735	0.6947	0.3380	0.1461
23	27.9252	1.3831	0.6468	0.2230
24	56.1517	2.9475	1.2907	0.3779

Đơn vị thời gian: ms.

Từ kết quả thử nghiệm trong **Bảng 1** ta thấy: trên 1 GPU, sử dụng một khung phức hợp cải thiện tốc độ so với nhiều khung cơ bản 1,9-2,3 lần, sử dụng khung phức hợp cho 1 và 4 GPU cải thiện tốc độ so với lập trình tuần tự trên CPU lần lượt là 33-44 và 41-149 lần và chạy trên 4 GPU có thể tăng tốc so với 1 GPU gấp 3,4 lần với n=2²⁴.

4.2 Độ lệch chuẩn

Độ lệch chuẩn của một tập mẫu được tính bằng căn bậc hai trung bình bình phương hiệu các phần tử cho trung bình của chúng. Như vậy để tính độ lệch chuẩn trước hết cần tính trung bình.

$$\bar{x} = \frac{x_0 + x_1 + \dots + x_{n-1}}{n}$$

$$dev([x_0, x_1, \dots, x_n]) = \sqrt{\frac{\sum_{i=0}^{n-1} (x_i - \bar{x})^2}{n}}$$

Tính theo khung:

$$\bar{x} = \frac{reduce(+,[x_0, x_1, \dots, x_{n-1}])}{n}$$

$$dev = \sqrt{\frac{map2b1(f,[x_0, x_1, \dots, x_{n-1}], \bar{x})}{n}}, \text{trong đó}$$

$$f(x, a) = (x - a)^2$$

Từ kết quả thử nghiệm trong **Bảng 2** ta thấy: trên 1 GPU, sử dụng cấu hình block, thread tốt sẽ tăng tốc so với cấu hình kém là 6,4-10,1 lần, sử dụng 1 và 4 GPU tăng tốc so với lập trình tuần tự trên CPU lần lượt là 26-44 và 23-127 lần và chạy trên 4 GPU có

thể chậm hơn 1 GPU 1,1 lần duy nhất với $n=2^{20}$ nhưng nhanh hơn 2,9 lần với $n=2^{24}$.

Bảng 2. Kết quả thử nghiệm độ lệch chuẩn

log ₂ n	Tuần tự	1 GPU c/hình kém	1 GPU c/hình tốt	4 GPU c/hình tốt
20	3.5842	0.8831	0.1388	0.1582
21	7.1638	1.7199	0.2203	0.1749
22	14.3563	3.3931	0.3814	0.2135
23	28.6714	6.7392	0.6961	0.2964
24	58.3738	13.4313	1.3233	0.4599

Đơn vị thời gian: ms.

4.3 Hệ số tương quan Pearson

Hệ số tương quan Pearson là phép đo độ phụ thuộc tuyến tính của hai véc tơ, kí hiệu r. Cách tính:

$$r = \frac{n \sum_{i=0}^{n-1} x_i y_i - \sum_{i=0}^{n-1} x_i \sum_{i=0}^{n-1} y_i}{\sqrt{(n \sum_{i=0}^{n-1} x_i^2 - (\sum_{i=0}^{n-1} x_i)^2)(n \sum_{i=0}^{n-1} y_i^2 - (\sum_{i=0}^{n-1} y_i)^2)}}$$

sum1 = reduce(+,[x₀, x₁, ..., x_{n-1}])

sum2 = reduce(+,[y₀, y₁, ..., y_{n-1}])

sumpr =

map2reduce(+,mul,[x₀, x₁, ..., x_{n-1}],[y₀, y₁, ..., y_{n-1}])

sumsql = mapreduce(+,sqr,[x₀, x₁, ..., x_{n-1}])

sumsq2 = mapreduce(+,sqr,[y₀, y₁, ..., y_{n-1}])

trong đó: mul(x, y) = x * y, sqr(x) = x²

$$r = \frac{n * sumpr - sum1 * sum2}{\sqrt{(n * sumsql - sum1^2)(n * sumsq2 - sum2^2)}}$$

Bảng 3. Kết quả thử nghiệm hệ số tương quan Pearson

log ₂ n	Tuần tự	1 GPU thư viện	1 GPU tối ưu	4 GPU thư viện
20	2.9958	0.4073	0.3771	0.3903
21	6.1352	0.6148	0.4901	0.4549
22	11.9716	1.0897	0.6510	0.5671
23	24.2893	2.0286	0.9745	0.8072
24	48.6886	3.9145	1.6190	1.2814

Đơn vị thời gian: ms.

Từ kết quả thử nghiệm trong **Bảng 3** ta thấy: tuy sử dụng thư viện chậm hơn so với việc lập trình lại toàn bộ bài toán cho GPU khoảng 1.1-2.4 lần nhưng bù lại, sử dụng thư viện là đơn giản hơn rất nhiều, không đòi hỏi kiến thức về lập trình song song và kiến trúc GPU, sử dụng thư viện trên 1 GPU và 4 GPU tăng tốc so với lập trình tuần tự trên CPU lần lượt là 7,4-12,4 lần, sử dụng 4 GPU tăng tốc so với 1 GPU là 1-3 lần.

5. KẾT LUẬN

Bài báo đã trình bày một thư viện khung song song dữ liệu được lập trình chạy trên hệ thống nhiều bộ xử lý đồ họa. Thư viện có giao diện đơn giản, trong suốt với người dùng, không đòi hỏi người dùng có kiến thức về kiến trúc GPU và mô hình lập trình CUDA.

Một chức năng trong suốt quan trọng là việc tự động lựa chọn cấu hình block, thread tốt nhất. Việc lựa chọn này giúp tăng đáng kể hiệu năng của thư viện.

Do mô hình lập trình CUDA không hỗ trợ con trỏ hàm trên kiến trúc GPU 1.3, thư viện đã cung cấp một phương thức khai báo hàm sử dụng cho các khung song song dữ liệu đó là phương thức sử dụng C macro tự hàm. Phương pháp này có thể thay thế

Thư viện hỗ trợ cả hệ thống một và nhiều GPU. Kết quả thử nghiệm cho thấy khi chạy trên hệ thống nhiều GPU, hiệu năng thu được cao hơn với số lượng phần tử đủ lớn. Hiệu năng đạt được ở cả hai trường hợp đều cao hơn lập trình tuần tự nhiều lần. Hiệu năng này tiệm cận với hiệu năng đạt được khi lập trình lại bài toán cho GPU, công việc vốn tốn thời gian, công sức hơn nhiều.

LỜI TRI ÂN

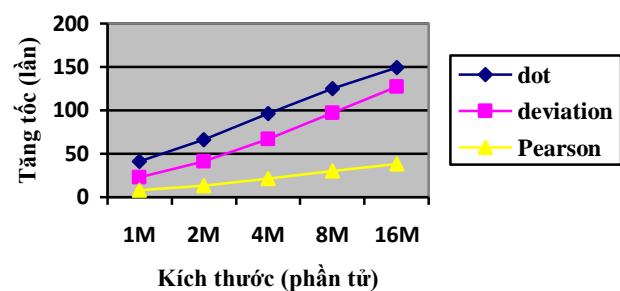
Em xin chân thành cảm ơn sự hướng dẫn tận tình của thày Nguyễn Hữu Đức, anh Lê Đức Tùng.

Em cũng xin cảm ơn sự giúp đỡ tạo điều kiện của thày Nguyễn Thanh Thủy, giám đốc trung tâm tính toán hiệu năng cao và các anh, chị trong trung tâm.

TÀI LIỆU THAM KHẢO

- [1] NVIDIA CUDA. “NVIDIA CUDA Programming Guide version 3.2”.
- [2] K. Matsuzaki, K. Emoto, H. Iwasaki and Z. Hu. “A Library of Constructive Skeletons for Sequential Style of Parallel Programming.” In InfoScale ’06: Proceedings of the 1st international conference on Scalable information systems, page 13, New York, NY, USA, 2006. ACM.
- [3] SkeTo project. Homepage: <http://www.ipl.t.u-tokyo.ac.jp/sketo/>

Hình 1. Hiệu năng nhiều bộ xử lí



Hệ thống giám sát năng lượng tòa nhà sử dụng công-tơ điện tử và hệ thống truyền tin trên đường điện lưới

Nguyễn Trọng Nhật Quang

Tóm tắt – Bài báo trình bày một giải pháp thiết kế hệ thống giám sát điện năng sử dụng cho các tòa nhà với việc xây dựng và phát triển thiết bị đo chính xác, dựa trên nền tảng truyền thông là mạng lưới điện cung cấp cho chính các thiết bị sử dụng điện trong tòa nhà. Các thông tin thực tại ở điểm đo đặc cho các phòng, ban trong tòa nhà có thể được gửi về trung tâm xử lý, và lưu trữ trong cơ sở dữ liệu. Giải pháp hệ thống đề xuất có thể giúp cho các cơ quan, trường học, các tòa nhà công ty kiểm soát được lượng điện năng tiêu thụ, hạn chế tối đa tình trạng sử dụng lãng phí điện năng.

Từ khóa – Power Line, Energy Meter, Building Energy Metering System.

1. GIỚI THIỆU

Vấn đề giám sát điện năng tiêu thụ ở các tòa nhà như các giảng đường, cơ quan, trường học... đang được thế giới và Việt Nam đầu tư nghiên cứu. Nhu cầu cấp thiết hiện nay của ngành năng lượng không chỉ riêng ở nước ta là cần phải tiết kiệm và nâng cao khả năng sử dụng hiệu quả nguồn điện năng. Mặc dù các hộ sản xuất, kinh doanh, hộ gia đình đã áp dụng nhiều biện pháp: bố trí lại thiết bị điện dùng trong các giờ, sản xuất giờ thấp điểm, đề ra các nội quy tiết kiệm điện, sử dụng các bóng đèn hiệu suất cao... nhưng thực trạng ở nước ta vẫn còn hiện tượng sử dụng điện khá lãng phí, nhất là ở khu vực công cộng, trụ sở cơ quan, chiếu sáng quảng cáo, nhiều phòng làm việc buông rèm bật điện, không tắt thiết bị điện khi ra về...

Để có thể giám thiều được tối đa tình trạng sử dụng điện lãng phí, cần có các giải pháp giúp điều chỉnh linh động, kịp thời các hiện tượng sử dụng điện lãng phí. Một trong các giải pháp được nhiều nhà khoa học, nhà chính xác quan tâm chú ý là các hệ thống giám sát năng lượng. Để đảm bảo hệ thống ấy có thể tích hợp và phổ cập rộng rãi, thì các hệ thống đó cần phải có chi phí hợp lý, tính tin cậy cao, nhỏ gọn, dễ lắp đặt, thay thế phát triển... Đó cũng là các mục tiêu cho hệ thống được đề xuất trong bài báo này.

Một trong các thành phần không thể thiếu của các hệ thống giám sát năng lượng là các thiết bị đo đếm năng lượng hay các công-tơ điện (Energy Meter). Các công-tơ điện truyền thông sử dụng ở các hộ dân chỉ phục vụ mục đích ghi đọc từ phía công ty điện lực. Ý tưởng cơ bản cho hệ thống là cần thu thập được các thông tin thời gian thực về giá trị tích lũy của công-tơ điện như công suất toàn phần, công suất tiêu thụ, hệ số $\cos\Phi$... và lưu trữ các thông tin ấy ở trung tâm. Để thực hiện được ý tưởng này, bài báo đề xuất phát triển một sản phẩm công-tơ điện điện tử. Công-tơ điện điện tử có giao diện cơ bản vẫn hỗ trợ khả năng đọc/ghi số công-tơ, đặc biệt có thêm khả năng giao tiếp với hệ thống.

Trong một môi trường áp dụng cụ thể như cơ quan, giảng đường, trường học... các điểm đo đặc năng lượng thường không đặt tập trung mà phân tán, thông tin cần được thu thập về cũng rất ít khi đặt trực tiếp tại các điểm đo đặc mà cần có cơ chế thu thập dữ liệu vào một trung tâm xử lý (máy tính). Để thực hiện được điều này, các điểm đo đặc cần phải được kết nối một hạ tầng truyền thông. Theo quan điểm về chi phí, khả năng tích hợp, mở rộng, tính tin cậy, sẵn dùng, công nghệ truyền thông được lựa chọn cho hệ thống đề xuất là sử dụng công nghệ truyền thông trên đường truyền tải điện (Power Line Communication System), với ưu điểm nổi bật nhất là đường truyền tải điện – cũng là kênh truyền thông vật lý được đưa sẵn tới tất cả các điểm cần đo đếm năng lượng. Đồng thời, với hạ tầng truyền thông như vậy cho phép lắp đặt, mở rộng các module điều khiển, đóng ngắt...

2. NỀN TẢNG LÝ THUYẾT CỦA BÀI BÁO

Khái niệm về các hệ thống giám sát năng lượng cho các tòa nhà được các nhà nghiên cứu đề xuất từ đầu những năm 1970, nhưng chỉ thực sự được nhắc đến trên thị trường khi xuất hiện các thiết bị điện tử mật độ tích hợp cao với khả năng đo đặc phức tạp các đại lượng về năng lượng, các cảm biến ánh sáng, độ ẩm, nhiệt độ... Trước khi có sự ra đời của các thiết bị điều chế-giải điều chế (modem) các hệ thống như vậy là các hệ cơ điện tập trung và dữ liệu truyền trực tiếp tới trung tâm xử lý (máy tính). Tiền bộ trong ngành xử lý tín hiệu số, công nghệ

Công trình này được thực hiện dưới sự bảo trợ của công ty CP-CNTT&TT Bách Khoa – BK-ict

Nguyễn Trọng Nhật Quang, sinh viên lớp Kỹ sư tài năng Công nghệ thông tin, khóa 51, Viện Công nghệ thông tin và truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 097-610-7538, email: nhatquangnt.88@gmail.com)

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

vi điện tử cho ra đời nhiều IC modem đã nới rộng khoảng cách, giảm thiểu kích thước và tính linh hoạt cho các hệ thống như vậy. Nhờ đó, các hệ thống giám sát năng lượng được tích hợp vào các hệ thống quản lý lớn hơn, gồm nhiều thành phần tạo thành các hệ thống quản lý tòa nhà phức tạp. Bên cạnh đó, thị trường cũng có xu hướng tới áp dụng các hệ thống nhỏ hơn vào các tòa nhà quy mô vừa phải nhằm đáp ứng nhu cầu quản lý, giám sát ngày càng cao của con người. Hệ thống giám sát năng lượng cho các tòa nhà bao gồm 3 thành phần chính:

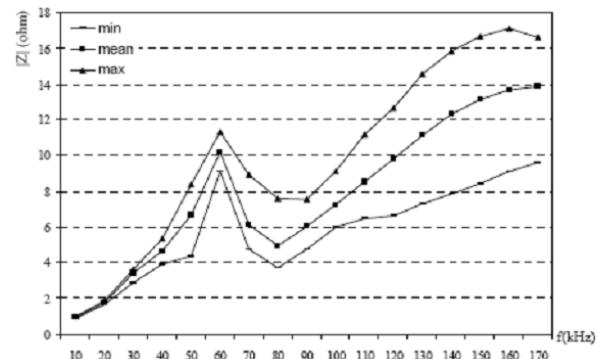
- Các module đo đếm năng lượng:** gồm có khối nguồn cung cấp, bộ phận cảm biến đo đặc, bộ phận điều khiển (có thể có) và giao diện truyền thông để truyền dữ liệu về trung tâm. Trong rất nhiều trường hợp, quá trình truyền dữ liệu không nhất thiết là chỉ có một phía làm chủ quá trình truyền thông. Bên cạnh việc đo đếm năng lượng, các module này có thể tích hợp thêm các cảm biến đo lưu lượng khí đốt, lưu lượng nước... nhằm tận dụng giao diện truyền thông
- Hệ tầng truyền thông:** được sử dụng để truyền dữ liệu và thông tin điều khiển giữa các module đo đếm và trung tâm xử lý. Công nghệ truyền thông thường được sử dụng hiện nay là: sử dụng đường truyền thuê bao PSTN, sử dụng đường truyền tái điện PLC, sử dụng công nghệ không dây. Các công nghệ sẽ quy định các thành phần của hệ thống
- Trung tâm thu thập và xử lý dữ liệu:** bao gồm modem, thiết bị lưu trữ, xử lý, ở mức đơn giản là các máy tính PC. Ở các hệ thống lớn hơn có thể bao gồm nhiều cáp, các thiết bị thu thập cầm tay (Hand Held Device), các Server lưu trữ, hệ thống mạng liên kết ở trung tâm xử lý...

3. TRUYỀN TIN TRÊN ĐƯỜNG ĐIỆN LUỐI

Truyền tin trên đường truyền tái điện là phương pháp truyền dữ liệu sử dụng đường dây dẫn tín hiệu là mạng điện cao thế (35kV hoặc hơn), mạng điện trung thế (10kV) hoặc mạng điện hạ thế (380/220V) cung cấp cho các hộ gia đình, cơ quan... Tín hiệu điều chế vào mạng lưới điện có thể là tương tự hoặc từ tín hiệu số, thông qua các IC điều chế thành tương tự. Mạng lưới điện có một ưu điểm rất lớn nếu tận dụng đó là tính phủ rộng khắp, mạng tối từng phòng, từng thiết bị sử dụng, đồng thời có thể sử dụng luôn nguồn năng lượng điện cung cấp để hoạt động thiết bị, không cần phải hệ thống nguồn cấp riêng. Tuy vậy, vấn đề khó khăn chính gặp phải do lưới điện không được thiết kế cho việc truyền tin tốc độ cao, mạng truyền tái được thiết kế chính cho tín hiệu ở tần số 50Hz (60Hz) và tối đa là 400Hz. Do đó, tín hiệu ở tần số cao bị suy hao lớn, khó truyền đi xa. Đường điện lưới bị can nhiễu từ rất nhiều thiết bị trên đường truyền, đồng thời các quy định về dài tần số ở một số quốc gia cũng hạn chế ở một dải hẹp làm cho thiết kế hệ tầng mạng truyền thông càng khó khăn.

Đường điện từ trạm biến thế (hạ thế) tới các hộ gia đình, các tòa nhà, các phòng phân tán ở trên một diện tích rộng lớn, biên độ hiệu dụng của tín hiệu điện khác nhau, cáp dẫn điện

khác nhau, ... dẫn đến đặc tính trở kháng của đường truyền thông khác nhau, lại không cố định do sự bất thường của tài tham gia vào mạng điện, các nhiễu xuyên âm trên đường dây... Tất cả các yếu tố đó quyết định công nghệ truyền thông cần phải sử dụng.



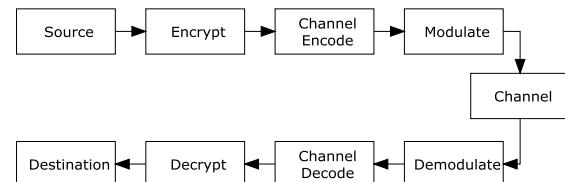
Hình 1. Đặc tính tái – tần số của mạng điện gia đình [3]

Theo chuẩn CELENEC EN50065 quy định về dài tần truyền thông băng hẹp trên đường tái điện giới hạn trong miền tần số từ 3kHz tới 148.5kHz. Mỹ yêu cầu sử dụng dài tần số từ 45kHz tới 450kHz.

4. MODEM TRUYỀN TIN PLC

Bài báo đề xuất thiết kế modem truyền tin trên đường điện lưới nhằm phục vụ cho hạ tầng truyền tin của hệ thống giám sát năng lượng. Sử dụng đường điện lưới làm kênh kết nối, hệ thống có thể lắp đặt nhiều modem truyền tin. Các modem này có thể thực hiện các nhiệm vụ:

- Tích hợp làm giao diện truyền thông cho các cảm biến, thiết bị đo đếm
- Tích hợp các giao diện điều khiển nhằm thực hiện các tác vụ điều khiển như đóng/cắt mạch điện...
- Thực hiện nhiệm vụ của hạ tầng truyền thông: đóng vai trò như các bộ repeater mở rộng khoảng cách truyền thông của hệ thống, khắc phục vấn đề suy hao tín hiệu trên đường truyền.



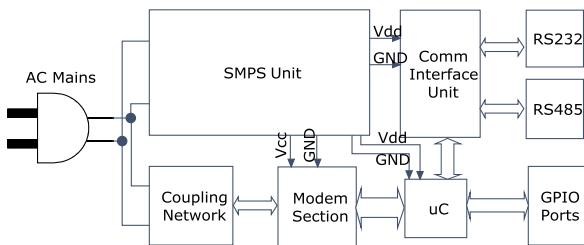
Hình 2. Mô hình truyền thông

Trong một hệ thống truyền thông trên đường tái điện, ở phía phát, dữ liệu số truyền đi được mã hóa theo yêu cầu của hệ thống, rồi đưa vào bộ mã hóa kênh với mục đích làm tăng khả năng chịu lỗi khi truyền tin bằng cách sử dụng các bit dữ thừa (redundancy), sau khi điều chế thành tín hiệu tương tự được khuếch đại và ghép (coupling) vào mạng điện thông qua một mạch ghép tín hiệu. Ở bên thu, tín hiệu nhận được (thông qua mạng ghép lưới điện) được giải điều chế, giải mã hóa kênh, giải mã mật và đưa trở lại cho bên nhận.

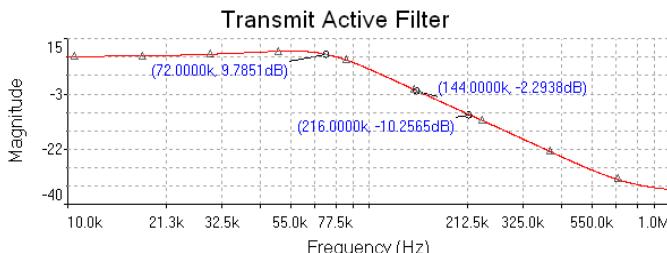
Modem sử dụng cho hệ thống đè xuất sử dụng IC điều chế FSK ST7540 của hãng STmicroelectronics, cho phép tốc độ điều chế tối đa 4800bps phù hợp với yêu cầu của hệ thống.



Hình 3. Modem PLC FSK 72kHz



Hình 4. Sơ đồ khái niệm các thành phần phần cứng



Hình 5. Đáp ứng tần số biên độ của mạch lọc tầng phát

Các tính năng chính

- Kỹ thuật điều chế sử dụng BFSK với tần số trung tâm F0 = 72kHz
- Hỗ trợ hai giao diện truyền tin RS232 và RS485 cho phép:
 - o Đóng vai trò như modem truyền tin với máy tính qua giao tiếp RS232
 - o Đóng vai trò như bộ chuyển đổi giao thức (transceiver) cho các thiết bị có giao diện RS485
- Modem có cơ chế cho phép thiết lập bằng phần mềm vai trò truyền thông (là Master hay Slave)
- Modem không cần sử dụng nguồn điện phụ trợ, hoạt động nhờ lấy nguồn từ lưới điện với hiệu suất cao (~70%), gọn nhẹ
- Hỗ trợ các đầu ra điều khiển (GPIO Ports) cho phép tích hợp các bộ phận:
 - o Điều khiển Rơ le đóng ngắt

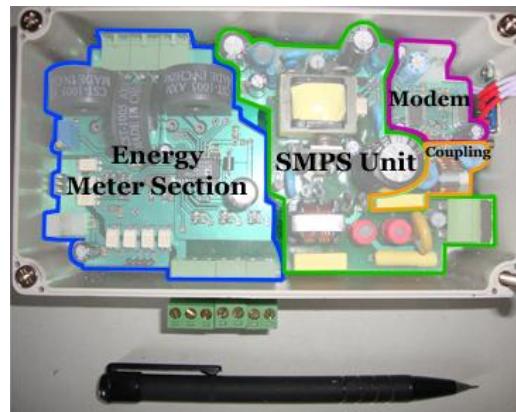
- o Điều khiển công suất (triac, thrysistor)...

Điều kiện hoạt động

- Điện áp hoạt động: 110V – 220V, 50Hz
- Nhiệt độ: 10° - 50°C
- Độ ẩm: < 95%

5. CÔNG TƠ NĂNG LƯỢNG ĐIỆN TỬ

Công tơ có nhiệm vụ đo đếm năng lượng tiêu thụ của mạng điện cần đo đặc. Công tơ điện tử cho phép hiển thị thông tin trực quan trên màn hình điện tử, đồng thời cho phép thu thập thông tin cần thiết.



Hình 6. Module đo đếm năng lượng

Bài báo đề xuất thiết kế của module đo năng lượng dựa trên công nghệ xử lý tín hiệu số (DSP – Digital Signal Processing) tích hợp trên IC xử lý tín hiệu ADE7758 của hãng Analog Devices. Quá trình đo đặc được thực hiện từ việc lấy mẫu tín hiệu trên các kênh tín hiệu dòng điện (qua bộ chuyển đổi về tín hiệu điện áp) và kênh tín hiệu điện áp, sau đó thực hiện quá trình biến đổi sang miền số và thực hiện các quá trình lọc tín hiệu, tích phân, lấy giá trị tuyệt đối... bên trong IC xử lý. Giá trị thu thập được có thể hiển thị lên trên màn hình LCD một cách trực quan cho người sử dụng. IC cho phép thực hiện đo đếm các đại lượng: công suất tiêu thụ (W), công suất phản kháng (VAR), công suất toàn phần (VA), năng lượng tiêu thụ, năng lượng phản kháng, năng lượng toàn phần, các giá trị hiệu dụng: điện áp, dòng điện, hệ số công suất cosφ, tần số lưới điện, các hiện tượng tiêu thụ: quá áp, quá dòng, sụt áp, sụt dòng...

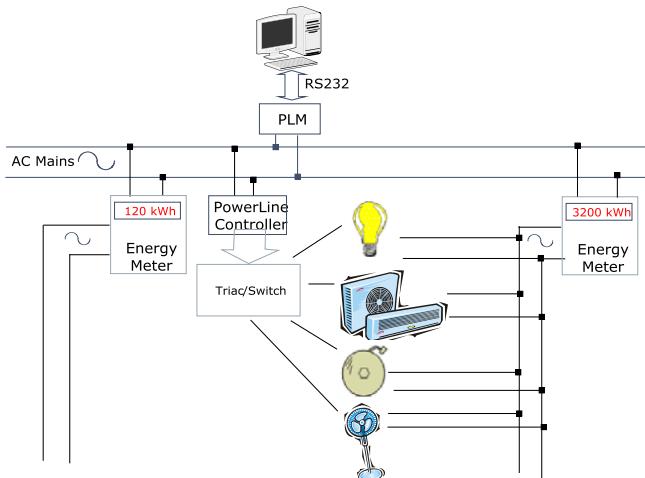
Module đo đếm năng lượng bao gồm các thành phần:

- Bộ phận đo đếm năng lượng bao gồm các mạch lọc tần số, mạch tạo tín hiệu điện áp tham chiếu, mạch chuyển đổi tín hiệu trên kênh điện áp, dòng điện, mạch giao tiếp số với hệ vi xử lý
- Khối nguồn SMPS theo kiến trúc flyback cung cấp các nguồn điện cách ly nhằm đảm bảo các yêu cầu về: nhiễu, an toàn, hiệu suất sử dụng đảm bảo cho tín hiệu đo chính xác, và hao phí điện năng nhỏ nhất

- Khối modem PLC giao tiếp cho phép thu thập giá trị đo đếm về trung tâm thông qua lưới điện
- Khối xử lý, điều khiển gồm có các thành phần: Vi điều khiển, bộ phận đồng hồ thời gian thực (RTC), bộ phân điều khiển giao tiếp với IC đo đếm năng lượng, bộ phận giao tiếp modem PLC, bộ phận giao tiếp qua mạng RS232, RS485 với mục đích tích hợp và mở rộng hệ thống.

6. HỆ THỐNG GIÁM SÁT NĂNG LƯỢNG TÒA NHÀ

Hệ thống đề xuất được áp dụng cho các mô hình vừa và nhỏ, trong đó, một máy tính được kết nối với modem truyền tín hiệu đóng vai trò làm trung tâm điều khiển, thu thập dữ liệu thông qua giao diện truyền thông RS232. Các module đo đếm năng lượng được đặt tại đầu vào các điểm đo đặc: như các phòng họp, các ban. Các module truyền tin có thể được tích hợp với các board điều khiển đóng cắt (relay array) thực hiện đóng/ngắt kết nối với các loại thiết bị điện: chiếu sáng, điều hòa, hệ thống quạt làm mát... Đồng thời có thể tích hợp thêm cơ chế cảnh báo thông qua hệ thống còi, chuông...



Hình 7. Sơ đồ ứng dụng hệ thống

Nguyên lý hoạt động của hệ thống

Tất cả các module thành phần đều được tích hợp modem truyền tin trên đường điện lưới. Dữ liệu thu thập được hoặc cần truyền đi được mã hóa thích hợp và điều chế vào sóng mang theo kỹ thuật điều chế di tần (Frequency Shift Keying). Tần số trung tâm được lựa chọn (75kHz) nằm trong dải tần số ít bị ảnh hưởng nhất bởi các thiết bị điện tử, được dịch tần 1.2 kHz (ở tốc độ truyền tin 2400bps) tương ứng với bit 0 và bit 1. Tín hiệu tương tự sau khi được khuếch đại được ghép vào lưới điện.

Các modem truyền tin hoạt động ở chế độ bán-song công (half-duplex) với giao thức master/slave (chủ - tớ). Để giảm thiểu tối đa yêu cầu xử lý cho phía các modem, toàn bộ quá trình điều khiển luồng dữ liệu được thực hiện phía master – là máy tính đóng vai trò trung tâm dữ liệu. Các thiết bị còn lại đóng vai trò slave – đáp ứng lại các yêu cầu về dữ liệu và điều

khiển từ phía master. Tốc độ truyền tin của kênh truyền có thể thiết lập ở nhiều tốc độ (baudrates) khác nhau: 600bps, 1200bps, 2400bps và 4800bps trong đó, tốc độ mặc định được lựa chọn là 2400bps. Do hàm lượng và tần suất thông tin trao đổi trên kênh truyền thông không nhiều và không cao, nên tốc độ không cần cao, nhưng cần phải đảm bảo được vấn đề tin cậy. Đây là một yêu cầu rất khắt khe đối với việc thiết kế truyền thông của hệ thống, bởi đặc tính tải luôn biến đổi, can nhiễu đường truyền tải lớn nên cần phải áp dụng các biện pháp tổng thể để đảm bảo tính tin cậy trong truyền thông.

Ngoài các biện pháp về kỹ thuật lọc nhiễu, kỹ thuật khuếch đại tín hiệu, kỹ thuật phối ghép lưới điện, thực hiện dưới phần cứng của thiết bị, giao thức truyền tin của hệ thống cũng đóng vai trò quyết định tới hiệu quả truyền thông. Giao thức được thiết kế trên hai tầng chính: tầng liên kết dữ liệu (cụ thể là tầng con truy nhập kênh – MAC với kỹ thuật CSMA) và tầng vật lý với các kỹ thuật:

- Mã kiểm tra CRC-16 (Cyclic Redundancy Code) kiểm tra tính toàn vẹn của các khung tin gửi đi
- Kỹ thuật mã hóa đơn giản theo giải thuật TEA (Tiny Encryption Algorithm) của tác giả David Wheeler và Roger Needham
- Mã sửa lỗi hướng tới FEC (Forward Error Correction) nhằm nâng cao khả năng phát hiện lỗi và thực hiện sửa lỗi.
- Kỹ thuật trộn bit (Interleaving) kết hợp với kỹ thuật FEC cho phép nâng cao khả năng chống nhiễu.

Quá trình truyền thông giữa thiết bị chủ với thiết bị tớ được thực hiện khi có yêu cầu dữ liệu từ phía thiết bị chủ, gửi gói tin yêu cầu tới địa chỉ của thiết bị tớ. Do mô hình của mạng các điểm đo đặc ít có thay đổi nên địa chỉ của các module đo đếm được gán tĩnh. Để đảm bảo vấn đề về giảm thiểu điện năng tiêu thụ của toàn bộ hệ thống, các module đo đếm được thiết kế hoạt động chủ yếu ở chế độ tiết kiệm năng lượng (power saving modes), giảm thiểu tối đa lưu lượng truyền thông mạng. Định kỳ hoặc theo yêu cầu, thiết bị chủ mới tiến hành thực hiện quá trình thu thập dữ liệu.

7. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Với khả năng thu thập thông tin một cách trực tiếp, tin cậy, hệ thống đề xuất sẽ đóng vai trò quan trọng trong việc giảm thiểu tối đa tình trạng sử dụng điện năng lãng phí đang còn tồn tại ở nhiều tòa nhà, công sở, trường học, bệnh viện... trong cả nước. Ưu điểm nổi bật nhất của hệ thống là với chi phí nhỏ, cài đặt đơn giản, khả năng tích hợp và phát triển tốt, phù hợp với mô hình và điều kiện thực trạng ở nước ta.

Để đảm bảo các yêu cầu cao hơn về khả năng tin cậy, tính dễ dùng, và các yêu cầu mở rộng hơn về tính năng của hệ thống, tác giả đang tiếp tục nghiên cứu:

- Thực hiện các tầng cao hơn trong gói giao thức truyền thông của hệ thống qua việc tham khảo và nghiên cứu

- phát triển hệ giao thức COSEM và DLMS. Hướng tới giao thức đa-thiết bị chủ (MultiMaster)
- Cài đặt các dịch vụ hỗ trợ tương ứng trên phần mềm điều khiển trung tâm...

8. LỜI TRI ÂN

Tác giả xin gửi lời cảm ơn sâu sắc tới Thầy giáo hướng dẫn ThS. Bùi Quốc Anh đã cung cấp nhiều tài liệu và lời khuyên quý giá giúp tác giả hoàn thành bài báo này. Tác giả cũng xin bày tỏ lòng biết ơn tới tập thể lãnh đạo và anh chị em trong công ty CPCNTT-TT Bách Khoa BK-ict đã tài trợ hết sức to lớn về mặt vật chất, kỹ thuật và công nghệ cho tác giả. Trong suốt thời gian thực hiện nghiên cứu, không thể không kể tới sự giúp đỡ tinh thần quý giá từ phía gia đình, thầy cô và bạn bè.

9. TÀI LIỆU THAM KHẢO

- [1]. Bernard Sklar, *Digital Communications – Fundamentals and Applications*, Prentice Hall, Second Edition
- [2]. William Stallings, *Data and Computer Communications*, Prentice Hall of India, Fifth Edition, 1999
- [3]. E. Mainardi, S. Banzi, M. Bonfe, S. Beghelli, *A Low-cost Home Automation System based on Power-Line Communication Links*, 22nd ISARC 2005 Ferrara (Italy)
- [4]. OPEN meter, D2.1 Part 2, Version: 2.3, *Description of current State-of-the-art of technology and protocols – Description of State-of-the-art PLC-based access technology*, May 2009.
- [5]. IEC 62056-53, International Standard, Electricity metering – *Data Exchange for meter reading, tariff and load control – Part 53 COSEM application layer*
- [6]. IEC 62056-21 International Standard, Electricity metering – *Data Exchange for meter reading, tariff and load control – Part 21 Direct local data exchange, First Edition*, 2002-05
- [7]. IEC 62056-42 International Standard, Electricity metering – *Data Exchange for meter reading, tariff and load control – Part 42 Physical Layer Services and Procedures using Connection-Oriented Asynchronous Data Exchange*
- [8]. IEC 62056-62 International Standard, Electricity metering – *Data Exchange for meter reading, tariff and load control – Part 62 Interface Classes*
- [9]. Andrew S. Tanenbaum, *Computer Networks*, Prentice Hall, Fourth Edition
- [10]. Xavier Carcelle, *Power Line Communications in Practice*, Artech House, 2006
- [11]. Echelon Corporation, *LonTalk Protocol Specification*, Version 3.0,
- [12]. Kevin Ackerman, David Dodds, Carl McCrosky, *Protocol to Avoid Noise in Power Line Networks*, IEEE 2005
- [13]. Frederick Emmons Terman, *Radio Engineers' Handbook*, First Edition, McGRAW-HILL Book, New York 1943
- [14]. Alan V. Oppenheim, Ronald W. Schafer, John R. Buck, *Discrete-Time Signal Processing*, Prentice Hall, Second Edition
- [15]. Austin Harney, *Smart Metering Technology Promotes Energy Efficiency for a Greener World*, Analog Dialogue 43-01, January 2009
- [16]. David J. Wheeler, Roger W. Needham, *TEA – a Tiny Encryption Algorithm*, Computer Laboratory, Cambridge University England.

Nâng cao chất lượng tín hiệu tiếng nói

Nguyễn Đức Hải

I. Tóm tắt

Nhiều nền được thêm vào tiếng nói có thể làm suy giảm hiệu quả của các hệ xử lý âm thanh số được sử dụng cho các ứng dụng như nhận dạng tiếng nói, nén tiếng nói... hệ thống âm thanh số sẽ được sử dụng trong rất nhiều môi trường, và hiệu quả cần đạt được ở mức độ gần với tiếng nói tự nhiên. Để đảm bảo cho khả năng thực thi, tác động của nhiều nền có thể được giảm bằng cách sử dụng micro lọc âm (noise - cancelling micro), sự thay đổi bên trong của các giải thuật xử lý âm thanh để rõ ràng của tín hiệu bị nhiễu hay quá trình giảm nhiễu tiền xử lý.

Trong bài viết này trình bày hai bộ lọc giảm nhiễu tự nhiên đó là Wiener và Spectral Subtraction. Sự so sánh kết quả của hai bộ lọc.

II. Lý thuyết

A. Bộ lọc Spectral Subtraction

1. Mô hình nhiễu cộng

Trong một cửa sổ, nhiễu $n(k)$ được thêm vào tín hiệu tiếng nói $s(k)$ được tín hiệu ký hiệu là $x(k)$. Ta có:

$$x(k) = s(k) + n(k). \quad (1)$$

bên đồi Fourier cho ta:

$$X(e^{j\omega}) = S(e^{j\omega}) + N(e^{j\omega}) \quad (2)$$

trong đó:

$$\begin{aligned} x(k) &\leftrightarrow X(e^{j\omega}) \\ X(e^{j\omega}) &= \sum_{k=0}^{L-1} x(k) e^{-j\omega k} \quad (3) \\ x(k) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega k} d\omega. \end{aligned}$$

2. Ước lượng trừ phô

Bộ lọc trừ phô $H(e^{j\omega})$ được tính toán bằng cách thay thế phô nhiễu $N(e^{j\omega})$ với phô cái mà có thể dễ dàng đạt được. Độ lớn của $|N(e^{j\omega})|$ của $N(e^{j\omega})$ được thay thế bằng bình quân giá trị $\mu(e^{j\omega})$ được lấy trong vùng không có tiếng nói, và pha $\theta_N(e^{j\omega})$ của nhiễu

được thay thế bằng pha $\theta_x(e^{j\omega})$ của $X(e^{j\omega})$. Kết quả trừ này cho kết quả ước lượng trừ phô $\hat{S}(e^{j\omega})$

$$\hat{S}(e^{j\omega}) = [|X(e^{j\omega})| - \mu(e^{j\omega})] e^{j\theta_x(e^{j\omega})}. \quad (4)$$

$$\text{hay } \hat{S}(e^{j\omega}) = H(e^{j\omega}) X(e^{j\omega})$$

$$\text{với } H(e^{j\omega}) = 1 - \frac{|N(e^{j\omega})|}{|X(e^{j\omega})|}$$

$$\mu(e^{j\omega}) = E\{|N(e^{j\omega})|\}$$

3. Sai số phô

Sai số phô được định nghĩa:

$$\varepsilon(e^{j\omega}) = \hat{S}(e^{j\omega}) - S(e^{j\omega}) = N(e^{j\omega}) - \mu(e^{j\omega}) e^{j\theta_x(e^{j\omega})}.$$

hay còn được gọi là độ lệch đường bao phô. Một số bước đơn giản để giảm việc không chính xác của sai số bao gồm các bước sau: 1. Trung bình biên độ 2. chỉnh nửa sóng 3. giảm nhiễu thừa 4. giảm tín hiệu cộng trong miền không có tiếng nói

4. Trung bình biên độ

Khi mà sai số phô bằng với độ lệch giữa phô nhiễu N và trung bình của chính nó μ , thì trung bình cục bộ của biên độ có thể được sử dụng cho việc giảm sai số. Thay thế $|X(e^{j\omega})|$ với $\overline{|X(e^{j\omega})|}$ trong đó

$$\overline{|X(e^{j\omega})|} = \frac{1}{M} \sum_{i=0}^{M-1} |X_i(e^{j\omega})|$$

$X_i(e^{j\omega})$, với i th là chỉ số cửa sổ thời gian biến đổi của $x(k)$, do đó cho ta

$$\hat{S}_A(e^{j\omega}) = [\overline{|X(e^{j\omega})|} - \mu(e^{j\omega})] e^{j\theta_x(e^{j\omega})} \quad (6)$$

lý do của việc lấy trung bình này là sai số phô trở thành xấp xỉ:

$$\varepsilon(e^{j\omega}) = \hat{S}_A(e^{j\omega}) - \hat{S}(e^{j\omega}) \cong \overline{|N|} - \mu \quad (7)$$

trong đó

$$\overline{|N(e^{j\omega})|} = \frac{1}{M} \sum_{i=0}^{M-1} |N_i(e^{j\omega})|.$$

theo đó, trung bình mẫu của $|N(e^{j\omega})|$ dần tới $\mu(e^{j\omega})$ nếu lấy trung bình trên một đoạn đủ lớn.

Vấn đề rõ ràng với chỉnh sửa này là tiếng nói là tĩnh, và theo đó chỉ có trung bình thời gian ngắn là được phép. Kết quả DRT cho ta thấy việc lấy trung bình nhiều hơn 3 cửa sổ chồng sẽ làm giảm hiệu quả. Nhược điểm của việc lấy trung bình là lấy tín hiệu âm trong thời gian ngắn.

5. Chỉnh nửa sóng

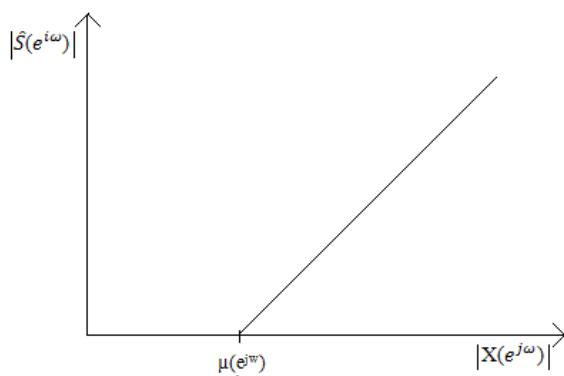
Đối với mỗi tần số ω nơi mà biên độ phổ tín hiệu nhiễu $|X(e^{j\omega})|$ nhỏ hơn trung bình biên độ phổ nhiễu $\mu(e^{j\omega})$, thì đầu ra được xét là 0. Việc điều chỉnh này có thể được thực hiện một cách khá đơn giản bằng chỉnh nửa sóng $H(e^{j\omega})$

Sự ước lượng sau đó trở thành

$$\hat{S}(e^{j\omega}) = H_R(e^{j\omega}) X(e^{j\omega})$$

trong đó

$$H_R(e^{j\omega}) = \frac{H(e^{j\omega}) + |H(e^{j\omega})|}{2}$$



Hình 1. Mối liên hệ giữa $X(e^{j\omega})$ và $\hat{S}(e^{j\omega})$

Ưu điểm của chỉnh nửa sóng là nền nhiễu có thể giảm bằng $\mu(e^{j\omega})$. Mặc dù vậy, các tiếng vang nhiễu có giá trị thấp có thể bị loại trừ. Sự bất tiện của phương pháp này được thể hiện rõ trong trường hợp khi mà tổng tiếng nói nhiễu tại tần số ω nhỏ hơn $\mu(e^{j\omega})$. Sau đó tiếng thông tin tiếng nói tại tần số tương ứng sẽ bị xóa, đây là điều dễ thấy.

6. Giảm nhiễu thừa

Sau khi chỉnh nửa sóng, nhiễu cộng được lấy trên μ còn lại. Trong phần không có tiếng nói thì sai số

$N_R = N - \mu e^{j\theta_n}$, nó sẽ được gọi là nhiễu thừa. Nhiều thừa này sẽ có biên độ nằm giữa 0 và một giá trị cực đại đạt được trong vùng không có tiếng nói. Biên đổi ngược lại trong miền thời gian, nhiễu thừa sẽ có âm giống như tổng của các tiếng vang tạo ra với tần số cơ bản ngẫu nhiên, nó được bật và tắt tại một tốc độ khoảng 20 ms. Trong miền có tiếng nói, nhiễu thừa vẫn có thể được nhận ra.

Khả năng nghe thầm của nhiễu thừa có thể được giảm bằng các tính năng của việc lấy từng frame 1 cách ngẫu nhiên. Cụ thể, tại 1 vùng tần số, khi mà tiếng nói thừa sẽ dao động ngẫu nhiên tại độ lớn tại mỗi cửa sổ phân tích, nó có thể được ngăn chặn bằng việc thay thế bởi các giá trị hiện tại của chính nó với giá trị nhỏ nhất được chọn từ các cửa sổ liền kề. Việc lấy giá trị nhỏ nhất chỉ được sử dụng khi mà biên độ của $\hat{S}(e^{j\omega})$ nhỏ hơn giá trị lớn nhất của nhiễu thừa, và nó có giá trị khác nhau trên mỗi cửa sổ, rất có khả năng rằng đây là phổ nhiễu, theo đó có thể giảm bằng cách thay bằng giá trị nhỏ nhất. Thứ hai, nếu biên độ của $\hat{S}(e^{j\omega})$ nhỏ hơn giá trị lớn nhất nhưng lại gần với giá trị không đổi, đây là một khả năng cao rằng phổ tại tần số là tiếng nói năng lượng thấp, theo đó, việc lấy giá trị nhỏ nhất sẽ được giữa lại thông tin, và thứ 3, nếu $\hat{S}(e^{j\omega})$ lớn hơn giá trị lớn nhất, đây là phần tiếng nói, ta chỉ cần bỏ qua. Số lượng của giảm nhiễu sử dụng kỹ thuật thay thế được xem xét đương lượng tới điều đạt được bằng trung bình với 3 cửa sổ. Tuy vậy, với cách tiếp cận này, các vùng tần số năng lượng lớn sẽ không được lấy trung bình với vùng khác. Nhược điểm của kỹ thuật này là việc lưu trữ nhiều được yêu cầu cho việc lưu trữ nhiều thừa lớn nhất và giá trị biên độ cho 3 cửa sổ liền kề.

Kỹ thuật giảm nhiễu thừa này được thực hiện như sau:

$$|\hat{S}_i(e^{j\omega})| = |\hat{S}_i(e^{j\omega})| \text{ nếu}$$

$$|\hat{S}_i(e^{j\omega})| \geq \max |N_R(e^{j\omega})|$$

$$|\hat{S}_i(e^{j\omega})| = \left\{ \min |\hat{S}_j(e^{j\omega})| \mid j = i-1, i, i+1 \right\} \text{ nếu}$$

$$|\hat{S}_i(e^{j\omega})| < \max |N_R(e^{j\omega})| \quad (7)$$

trong đó $N_R(e^{j\omega}) = N - \mu e^{j\theta_n}$

và $\max |N_R(e^{j\omega})|$ = giá trị lớn nhất của kết quả giảm nhiễu vùng không có tiếng nói

7. Giảm tín hiệu cộng trong vùng không có tiếng nói

Trong vùng không có tiếng nói, $\hat{S}(e^{j\omega})$ sẽ bao gồm nhiều thửa nó lưu lại sau khi chỉnh sửa sóng và sự lựa chọn giá trị nhỏ nhất. Thông thường giá trị trung bình tì số năng lượng được giảm nhỏ nhất bằng 12dB. Điều này gợi ý kết quả cho việc tìm vùng không có tiếng nói cho bởi:

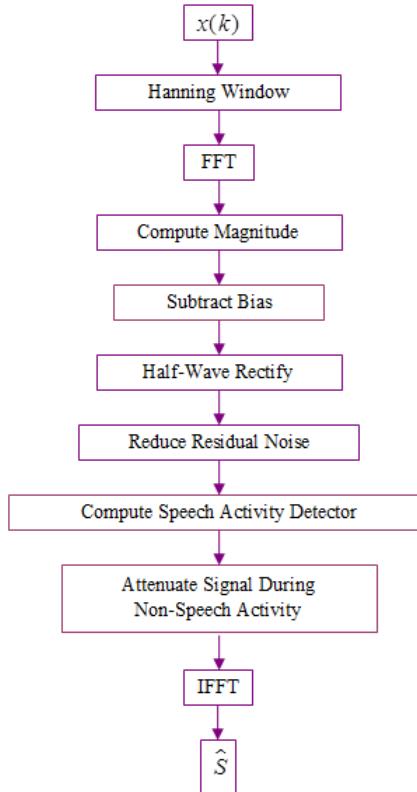
$$T = 20 \log_{10} \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{\hat{S}(e^{j\omega})}{\mu(e^{j\omega})} \right| d\omega \right]$$

nếu T nhỏ hơn -12dB, thì của số được phân loại là vùng không có tiếng nói. Mặt khác, trung bình tốt nhất của khối lượng giảm nhiễu được tìm thấy khoảng -30dB. Theo đó phô đầu ra ước lượng bao gồm việc giảm đầu ra trong vùng không có tiếng nói cho bởi:

$$\hat{S}(e^{j\omega}) = \begin{cases} \hat{S}(e^{j\omega}) & T \geq -12dB \\ cX(e^{j\omega}) & T \leq -12dB \end{cases}$$

trong đó $20 \log_{10} c = -30dB$

Mô hình lọc Spectral Subtraction[5]



B. Bộ lọc Wiener

Theo mô hình tiếng nói và nhiễu là nhiễu cộng

$$x(k) = s(k) + n(k)$$

việc ước lượng tín hiệu ban đầu về mặt lý thuyết là xác định được đáp ứng xung

$$\hat{s}(k) = \sum_{l=-\infty}^{\infty} h(l)x(k-l) \quad (2.1)$$

chú ý rằng tín hiệu tiếng nói ban đầu và tín hiệu nhiễu là độc lập, nên đáp ứng tần số của bộ lọc Wiener cho bởi:

$$G(\Omega) = \text{IDFT}\{h(l)\} = \frac{Pss(\Omega)}{Pss(\Omega) + Pnn(\Omega)} \quad (2.2)$$

Trong đó $Paa(\Omega)$ ký hiệu cho mật độ phô của tín hiệu và DTFT là biến đổi Fourier rời rạc trong miền thời gian. Theo đó trong trường hợp tín hiệu tĩnh, phô của tín hiệu ra nâng cao được tính toán theo công thức:

$$\hat{S}(\Omega) = \frac{Pss(\Omega)}{Pss(\Omega) + Pnn(\Omega)} Y(\Omega) = \frac{Pss(\Omega)}{Pxx(\Omega)} Y(\Omega) = G(\Omega)Y(\Omega) \quad (2.3)$$

$G(\Omega)$ thường được gọi là chức năng tăng phô. Với bộ lọc Wiener thì chức năng này phụ thuộc vào nhiễu đầu vào $x(k)$ hay đối với biến đổi Fourier $X(\Omega)$ và chỉ khi tín hiệu tiếng nói bị rối loạn là tĩnh. Tuy vậy, việc thực hiện của bộ lọc Wiener không được đầy đủ

như với bộ lọc này có đáp ứng xung vô hạn và đáp ứng tần số liên tục.

Với việc thực thi kết hợp hệ thống tổng hợp và phân tích phô trên, phương thức tăng được ước lượng tại tần số trung tâm của vùng phô. Hơn nữa, như tín hiệu tiếng nói và nhiễu là không cố định, việc xấp xỉ ngắn hạn cho phô năng lượng cân được sử dụng. Tuy nhiên, cho cách tiếp cận xử lý theo đoạn được phác thảo ở trên, chúng ta mong rằng sẽ thu được kết quả mong muốn. Tương tự như bộ lọc Wiener trong 3.6, đầu ra của bộ lọc cho từng đoạn tín hiệu tại thời điểm k , $\hat{S}(k) = (X_0(k), \dots, X_{M-1}(k))^T$, khi đó đầu ra được tính toán bởi phép nhân chập

$$\hat{S}(k) = G(k) \otimes X(k), \text{ trong đó } G(k) = (G_0(k), \dots, G_{M-1}(k))^T$$

với giá trị nhỏ nhất của ước lượng $E\{(\hat{S}\mu(k) - S\mu(k))^2\}$ ta có:

$$G\mu(k) = \frac{E\{|S\mu(k)|\}^2}{E\{|S\mu(k)|\}^2 + E\{|N\mu(k)|\}^2} = \frac{\eta_\mu(k)}{1 + \eta_\mu(k)} \quad (2.4)$$

trong đó $\eta_\mu(k) = \frac{E\{|S\mu(k)|\}^2}{E\{|N\mu(k)|\}^2}$ là tỷ số tín hiệu trên nhiễu trước (prior signal to noise ratio-SNR). $E\{|S_\mu(k)|^2\} = \sigma_{s,\mu}^2(k)$ và $E\{|N_\mu(k)|^2\} = \sigma_{n,\mu}^2(k)$ là năng lượng của tín hiệu tiếng nói bị rối loạn và tín hiệu nhiễu trong miền tần số μ tương ứng.

Tuy nhiên, công thức (2.4) chỉ dựa trên hệ tuyến tính, giải pháp không tuyến tính thường yêu cầu các kiến thức của hàm mật độ xác suất (probability density function-pdf) của hệ số phô tiếng nói và nhiễu. Ta tiếp cận theo hướng quyết định (*decision-directed*) sau đó quay lại ước lượng của đoạn trước và phối hợp với ước lượng hiện thời của SNR.

$$\gamma_\mu(k) - 1 = \frac{|Y_\mu(k)|^2}{E\{|N_\mu(k)|^2\}} - 1 = \frac{R_\mu^2(k)}{\sigma_{n,\mu}^2(k)} - 1 \quad (2.5)$$

Như vậy ước lượng priori SNR đạt được là:

$$\widehat{\eta}_\mu(k) = \alpha_\eta \frac{|S_\mu(k-r)|^2}{E\{|N_\mu(k)|^2\}} + (1 - \alpha_\eta) \max(0, \gamma_\mu(k) - 1) \quad (2.6)$$

trong đó kết hợp với điều kiện không âm và α_η là biến tron. $\gamma_\mu(k)$ là tỷ số SNR sau. Trong các điều kiện SNR thấp, việc ước lượng này bị sai lệch rõ ràng. Sự sai lệch này có thể được giảm nếu áp dụng phép toán lấy max được áp dụng cho tổng của 2 phần:

$$\widehat{\eta}_\mu(k) = \max(0, \alpha_\eta \frac{|S_\mu(k-r)|^2}{E\{|N_\mu(k)|^2\}} + (1 - \alpha_\eta)(\gamma_\mu(k) - 1)) \quad (2.7)$$

III. Kết quả

Ở phần này thực hiện việc lấy dữ liệu từ định dạng file wave 1 kênh 16 bit, tần số lấy mẫu 16kHz. Việc thu âm ở trong phòng Lab

Trong phần đánh giá kết quả, chúng ta sẽ chú ý đến 2 tham số đánh giá. Thứ nhất đó là tỷ số tín hiệu trên nhiễu, điều này đã nhắc đến trong công thức (2.4). Thứ hai là trung bình khoảng cách các đường bao phô. Chúng ta sẽ lần lượt chú ý tới 3 khoảng cách, tín hiệu ban đầu/sau lọc, tín hiệu trước và sau lọc, tín hiệu ban đầu và trước lọc.

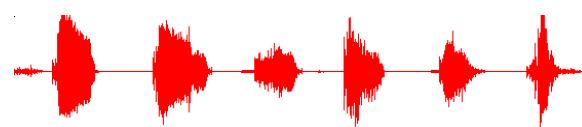
Trong bài thực hành, lấy tín hiệu trong thu trong văn phòng được coi là tín hiệu sạch. Sau đó được làm nhiễu, với tỷ số công suất nhiễu được chọn. Và cùng đánh giá tín hiệu sau lọc. Dưới đây là các kết quả đạt được.



Hình 2. Tín hiệu ban đầu



Hình 3. Nghiên SNR=10 dB



Hình 4. Tín hiệu sau lọc Spectral Subtraction



Hình 5. Tín hiệu sau lọc Wiener

Khoảng cách các đường bao phô

	SNR	Bộ Lọc	Tỷ số tín hiệu SNR sau lọc	Ban đầu / sau lọc	Trước / sau lọc	Ban đầu / trước lọc
10	Bộ lọc trừ phô	63.922	2.7384	7.6237	5.6903	
10	Bộ lọc Wiener	59.9019	4.2883	9.4565	5.6903	
20	Bộ lọc trừ phô	64.068	2.6726	7.5989	5.6903	
20	Bộ lọc Wiener	59.9019	4.2883	9.4565	5.6903	
*						

Tỷ số tín hiệu/nhiễu 65.2521

Hình 6. Kết quả đánh giá

Rõ ràng ta thấy, bộ lọc trừ phô cho kết quả khá tốt về cả mặt tín hiệu trong miền thời gian và trong miền phô. Kết quả nghe cũng khẳng định rõ điều này. Tín hiệu sau lọc nghe gần bằng với kết quả ban đầu, tuy nhiên ở các âm có công suất thấp, sau lọc sẽ mất một lượng thông tin đáng kể. Và khi nghe sẽ bị nhở đi

IV. Tài liệu tham khảo

1. “Speech enhancement” của J. Benesty, S.Makino và J.Chen
2. Noise Reduction in Speech Applications - CRC Press
7. Giải thuật tạo phân bố chuẩn
<http://download.oracle.com/javase/1.4.2/docs/api/java/util/Random.html#nextGaussian%28%29>
8. tables-of-integrals-series-and-products-7ed

- 3.
- Digital_Processing_of_Speech_Signals_L._Rabiner,_R._Schafer_(1978)_WW
4. “Speech Enhancement Using Adaptive Filters And Independent Component Analysis” của Tomasz Rutkowski, Andrzej Cichocki và Allan Kardec Barros
5. Suppression of Acoustic Noise in Speech Using Spectral Subtraction, Steven F.Boll
6. Phân bố chuẩn
http://vi.wikipedia.org/wiki/Ph%C3%A2n_b%C3%B9_chu%C1%BA%BA%A9n

Xây dựng hệ thống bán hàng tương tác dựa trên nền tảng mạng cảm biến không dây

Phạm Đức Anh, Trương Quốc Tú

Tóm tắt— Xã hội ngày càng phát triển, con người ngày càng quan tâm đến hình thức bên ngoài. Trang phục quần áo, giày dép, túi sách phản náo nói lên phong cách, thái độ, địa vị của mỗi cá nhân trong xã hội. Với mong muốn mang lại cho người sử dụng khả năng lựa chọn sản phẩm trực quan, nhanh và phù hợp nhất với cơ thể, đề tài nghiên cứu xây dựng và phát triển hệ thống bán hàng tương tác trên nền tảng mạng cảm biến không dây. Với mục tiêu chính là cho phép khách hàng có thể theo dõi trực tiếp hình ảnh của bản thân được hiển thị qua mô hình 3D và dễ dàng thay đổi trang phục thích hợp, đề tài đã sâu nghiên cứu và áp dụng các công nghệ tiên tiến hiện nay bao gồm công nghệ hỗ trợ xử lý mô hình 3D XNA do Microsoft đưa ra, công nghệ mạng cảm biến không dây miwi hỗ trợ tay cầm điều khiển lựa chọn sản phẩm và công nghệ DLNA(Chuẩn kết nối các thiết bị số) hiển thị hình ảnh trên các màn hình LCD.

Từ khóa—Bán hàng tương tác, mạng cảm biến không dây miwi, xây dựng model 3D, chuẩn DLNA.

1. GIỚI THIỆU

Hệ thống bán hàng tương tác được xây dựng nhằm mang lại cho người sử dụng sự tiện dụng, chuyên nghiệp trong việc lựa chọn trang phục. Với yêu cầu của một cửa hàng thời trang, đề tài đã nghiên cứu và đưa ra một giải pháp tổng thể bao gồm: xây dựng phần mềm quản lý bán, thiết kế tay điều khiển và các nút trung tâm trên nền tảng mạng cảm biến không dây sử dụng chip MRF24J40 của Microchip hỗ trợ cho việc chọn sản phẩm hiển thị qua mô hình 3D trên màn hình LCD đặt tại các vị trí khác nhau trong cửa hàng.

2. MẠNG CẢM BIẾN KHÔNG DÂY MIWI

Giao thức MiWi được xây dựng dựa trên tầng MAC và PHY của chuẩn giao thức IEEE 802.15.4 và được sử dụng cho phát triển một mạng không dây đơn giản trên băng tần 2.4Ghz với tốc độ truyền thông thấp vào khoảng 250 kbps. Đặc điểm chính của chuẩn này là tính mềm dẻo, tiêu hao ít năng lượng, chi phí nhỏ, tốc độ truyền dữ liệu thấp trong khoảng không gian nhỏ, thuận tiện khi áp dụng trong khu vực nhà riêng, văn phòng.

Công trình này được thực hiện dưới sự hướng dẫn của TS. Nguyễn Hồng Quang - Giảng viên Viện Công nghệ thông tin và Truyền thông- Trường Đại học Bách Khoa Hà Nội và đồng hướng dẫn của ThS. Phạm Văn Thuận - Giảng viên Viện Công nghệ thông tin và Truyền thông- Trường Đại học Bách Khoa Hà Nội.

Phạm Đức Anh, sinh viên lớp Kỹ thuật máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0975790500, e-mail: phamducanh.bk@gmail.com).

Trương Quốc Tú, sinh viên lớp Kỹ thuật máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0984227223, e-mail:tutq88@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

2.1. Thành phần của mạng:

- PAN coordinator (FFD): là thiết bị trung tâm có nhiệm vụ khởi tạo mạng, cấp phát địa chỉ mạng và tổ chức bảng kết nối (binding table).

- Coordinator (FFD): sử dụng để mở rộng dải phủ sóng của mạng. Nó cho phép nhiều nodes hơn tham gia vào mạng. Đồng thời có thể sử dụng FFD cho các chức năng giám sát và điều khiển.

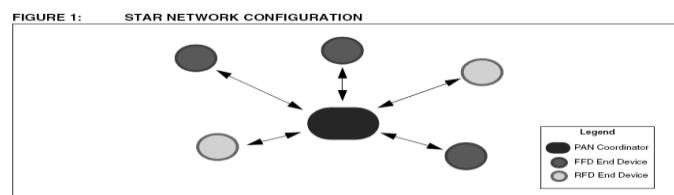
- End Device (RFD or FFD): sử dụng cho các chức năng giám sát và điều khiển.

2.2. Kiến trúc mạng:

Trong các loại thiết bị mạng MiWi thì thiết bị có vai trò quan trọng nhất là PAN coordinator. PAN coordinator là thiết bị sẽ khởi tạo mạng, chọn kênh và PAN ID của mạng(PAN ID-Số hiệu định danh cho 1 mạng). Tất cả các thiết bị tham gia vào PAN phải nghe theo chỉ thị của PAN coordinator.

2.2.1. Cấu hình mạng Star

Một cấu hình mạng Star Network gồm một node PAN coordinator và một hoặc vài thiết bị kiểu end devices khác. Trong mạng Star, tất cả các thiết bị end devices chỉ được phép giao tiếp với PAN coordinator. Nếu một thiết bị end device này cần truyền dữ liệu với một thiết bị end device khác, nó sẽ truyền dữ liệu của nó tới PAN coordinator, sau đó PAN coordinator sẽ

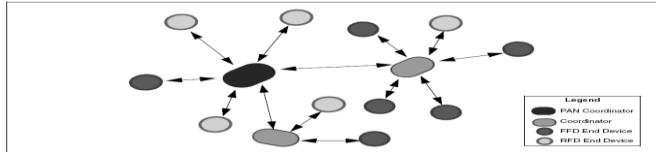


truyền dữ liệu đó tới thiết bị nhận.

2.2.2. Cấu hình mạng Cluster-Tree

Một mạng Cluster Tree gồm có một PAN coordinator, các thiết bị coordinator và các thiết bị end device. Trong cấu trúc này, PAN coordinator đóng vai trò là nút gốc của mạng, các coordinator khác là các thành phần trung gian kết nối tới nút gốc và các thiết bị end-device tương ứng. Các gói tin trong mạng Cluster-Tree được truyền theo các đường có sẵn ban đầu trong kiến trúc mạng. Khi các gói tin được định tuyến qua nhiều hơn một nút để tới thiết bị nhận, mạng Cluster Tree có thể được xem như mạng multi-hops.

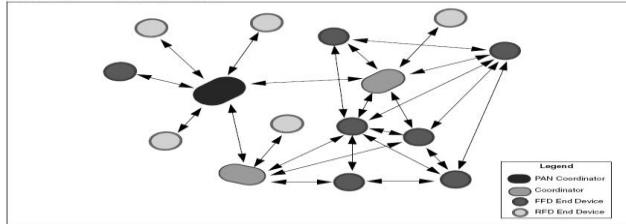
FIGURE 2: CLUSTER TREE TOPOLOGY



2.2.3. Cấu hình mạng Mesh

Mạng mesh có kiến trúc tương tự như mạng Cluster Tree, tuy nhiên mạng mesh có đặc điểm nổi trội hơn là cho phép các FFD có thể định tuyến các gói tin trực tiếp tới các FFD khác thay vì phải phải theo cấu trúc cây. Kiến trúc này làm tăng độ tin cậy của việc truyền dữ liệu trong mạng. Giống như mạng Cluster Tree, mạng Mesh cũng là multi-hops.

FIGURE 3: MESH NETWORK



2. XÂY DỰNG MÔ HÌNH 3D VỚI 3DSOM VÀ HIỂN THỊ MÔ HÌNH VỚI XNA

2.1. Xây dựng mô hình 3D với 3DSOM

Hiện nay, có khá nhiều phần mềm hỗ trợ xây dựng mô hình 3D cho nhân vật có thể kể tới như 3DSmax, Maya, 3DSom... Tuy nhiên với yêu cầu thực tế của việc tạo ra mô hình 3D của khách hàng một cách nhanh chóng, chính xác phục vụ cho mục tiêu thay đổi trang phục quần áo, đề tài đã sử dụng phần mềm 3DSom để tạo ra mô hình của khách hàng.

2.1.1. Yêu cầu

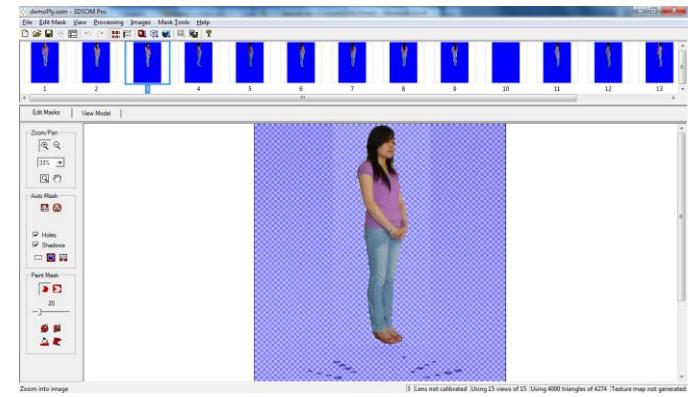
- Dữ liệu đầu vào được yêu cầu là hình ảnh của khách hàng thực hiện tại 15 vị trí khác nhau.

- Các hình ảnh đầu vào cần có kích thước phù hợp, kích thước thông thường là 1440 x 2180 px.

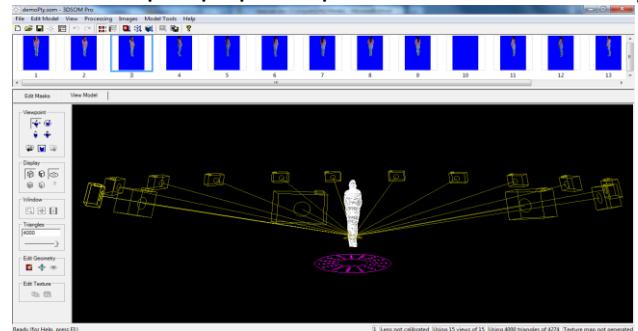
2.1.2. Các bước thực hiện

Việc xây dựng một mô hình 3D khách hàng gồm có 4 bước chính:

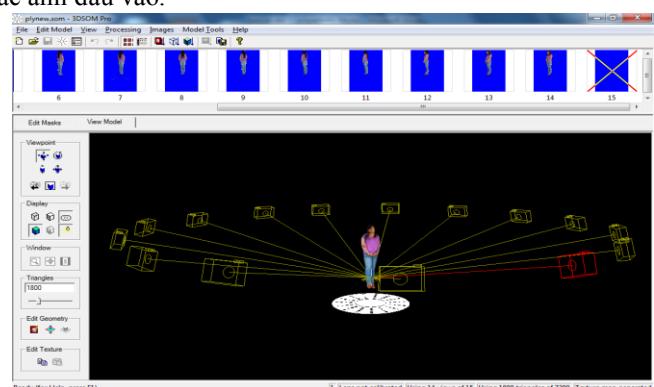
- **Bước 1:** đưa 15 ảnh đầu vào, tiến hành lọc các ảnh chỉ để lại hình ảnh của khách hàng, loại bỏ khung cảnh xung quanh.



- **Bước 2:** thực hiện tạo bề mặt mô hình 3D của khách hàng.



- **Bước 3:** thực hiện thêm trang phục lên mô hình, dựa trên các ảnh đầu vào.



2.2. Hiển thị mô hình 3D với XNA

XNA được phát triển bởi Microsoft bắt đầu năm 2004 Microsoft. XNA không chỉ là một framework như DirectX, nó còn chứa nhiều công cụ và thậm chí là một IDE tùy biến bắt nguồn từ Visual Studio hỗ trợ việc lập trình dễ dàng hơn. Ưu điểm nổi bật của XNA là thực hiện nhanh, quản lý tài nguyên hiệu quả và đơn giản, hỗ trợ nhiều ngôn ngữ. Mô hình ứng dụng XNA Framework chia làm 3 thành phần cốt lõi :

- Engine đồ họa XNA trong Microsoft.Xna.Framework.dll
- Mô hình ứng dụng game XNA trong Microsoft.Xna.Framework.Game.dll
- Kênh nội dung XNA trong Microsoft.Xna.Framework.Content.Pipeline.dlls

2.2.1. Engine đồ họa XNA

Sau khi hình thành trong một khoảng thời gian khá dài,

engine đồ họa XNA đã được phát triển khá phong phú, hỗ trợ cho người lập trình nhiều công cụ hiệu quả, giúp việc lập trình trở nên dễ dàng, nhanh chóng bởi nó bao gồm đầy đủ các phương thức quan trọng trong vấn đề xử lý 3D như: camera, light, model... Đồng thời hỗ trợ tốt cho việc xử lý các animate Model, pointLight và directionLight.

2.2.2. Mô hình ứng dụng XNA

Một mô hình ứng dụng XNA có rất nhiều thành phần, trong đó 3 phương thức quan trọng nhất bao gồm :

Initizline() : nạp tất cả các nội dung của ứng dụng, thiết lập các khung cảnh ban đầu và các khởi tạo.

Update (GameTime Time) : update được gọi trước mỗi khung hình được vẽ để cập nhật thời gian, nhập dữ liệu, âm thanh và các thành phần khác trong ứng dụng mà không hiển thị trên màn hình.

Draw (GameTime Time): draw được gọi mỗi khung hình để vẽ mọi thứ lên màn hình.

2.2.3. Kênh nội dung XNA

Kênh nội dung XNA được sử dụng để đưa biên dịch và nạp các tài nguyên của ứng dụng như texture, mô hình 3D, shader và các file âm thanh... Nhờ có kênh nội dung này, người lập trình giảm thiểu được khá nhiều công sức trong việc viết mã tùy chỉnh các thành phần đồ họa, mô hình 3D. Kênh nội dung không chỉ được chứa trong 1 file dll, có 5 file khác nhau bao gồm:

- Microsoft.Xna.Framework.Content.Pipeline.dlls: chứa các chức năng cơ bản.
- Microsoft.Xna.Framework.Content.Pipeline.EffectImporter : sử dụng để biên dịch và nhập các shader.
- Microsoft.Xna.Framework.Content.Pipeline.FBXImporter: là file lớn nhất chưa nhiều mã dùng để xử lý file mô hình .fbx và hỗ trợ nhiều tính năng như tạo da, làm xương.
- Ms.Xna.Framework.Content.Pipeline.TextureImporter : sử dụng để tùy chọn các file texture trong ứng dụng. Những file này có thể là file dds và đã được định dạng kiểu DirectX (Định dạng tốt nhất cho texture và hỗ trợ nén phân cứng).
- Microsoft.Xna.Framework.Content.Pipeline.Ximporter: cho phép xử lý file mô hình 3D .x.

3. CÔNG NGHỆ DLNA

DLNA (Digital Living Network Alliance) là chuẩn kết nối các thiết bị số do Sony đưa ra cùng với sự tham gia của hơn 250 công ty thành viên khác. Theo dự tính đến năm 2014, hơn 1 tỷ thiết bị sẽ được kích hoạt công nghệ này. Mục đích chung của việc sử dụng công nghệ DLNA nhằm làm cho người tiêu dùng dễ dàng hơn trong việc sử dụng thiết bị, dễ dàng hơn trong việc chia sẻ các bức ảnh số, nhạc và video. Cho tới tháng Năm 2010, hơn 8000 loại thiết bị đã đạt chứng chỉ "DLNA Certified".

Các thiết bị hỗ trợ công nghệ DLNA được phân chia thành 3 nhóm chính, gồm có:

- Thiết bị mạng:
 - + Digital Media Server (DMS)

- + Digital Media Player (DMP)
- + Digital Media Renderer (DMR)
- + Digital Media Controller (DMC)
- + Digital Media Printer (DMPr)

- Thiết bị cầm tay:

- + Mobile Digital Media Server (M-DMS)
- + Mobile Digital Media Player (M-DMP)
- + Mobile Digital Media Uploader (M-DMU)
- + Mobile Digital Media Downloader (M-DMD)
- + Mobile Digital Media Controller (M-DMC)

- Thiết bị hạ tầng:

- + Mobile Network Connectivity Function (M-NCF)
- + Media Interoperability Unit (MIU)

Đặc điểm nổi bật của công nghệ DLNA là cho phép các thiết bị điện tử chia sẻ dữ liệu với nhau một cách thông minh. Nhờ việc sử dụng công nghệ này, người sử dụng có thể truy cập các ứng dụng trên nhiều thiết bị chỉ cần một máy chủ duy nhất. Với ưu điểm này, DLNA hứa hẹn sẽ là công nghệ chủ yếu trong tương lai.

4. XÂY DỰNG NỀN TẢNG PHẦN CỨNG, FIRMWARE VÀ PHẦN MỀM QUẢN LÝ BÁN HÀNG

4.1. Tổng quan hệ thống

Yêu cầu đặt ra cho hệ thống bán hàng tương tác dựa trên nền tảng mạng cảm biến không dây có yêu cầu đưa ra đối với việc xây dựng phần cứng là thiết kế các bộ tay cầm điều khiển kết nối không dây với nút điều khiển trung tâm qua các nút trung gian. Tay cầm điều khiển cho phép người sử dụng lựa chọn mã sản phẩm qua các nút bấm và truyền dữ liệu tới nút trung gian (coordinator) của nó, và các nút trung gian truyền dữ liệu tương ứng đến nút trung tâm (PAN coordinator) được kết nối với máy tính qua cổng USB.

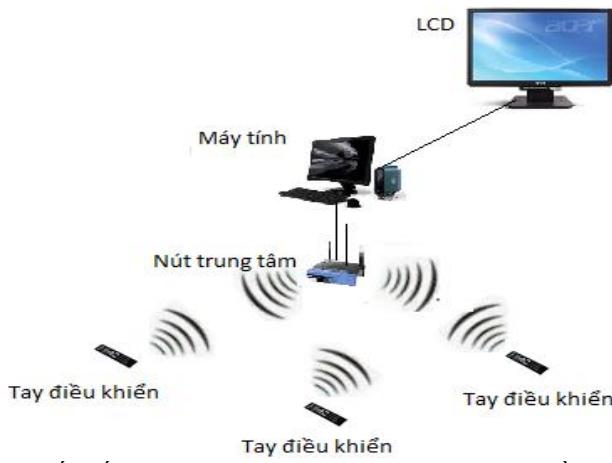
Sơ đồ hệ thống:



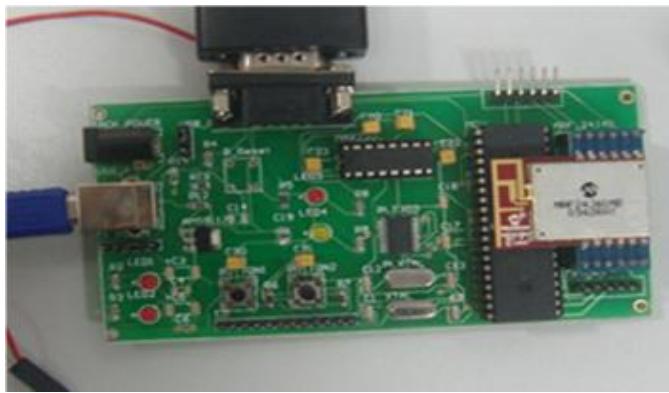
4.2. Xây dựng nền tảng phần cứng và firmware

Hệ thống phần cứng được xây dựng cho ứng dụng hiện tại bao gồm : nút trung tâm (PAN coordinator) và tay cầm điều khiển (end devices). Nút trung tâm được kết nối với máy tính qua cổng USB, gửi dữ liệu lên máy tính sau đó hiển thị lên màn hình LCD hình ảnh tương ứng.

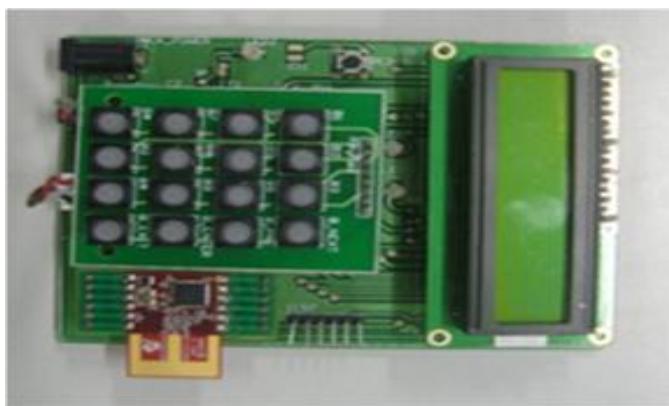
Sơ đồ hệ thống:



- Thiết kế nút trung tâm (PAN coordinator) bao gồm các modul chính sau: modul nguồn chuyển đổi từ nguồn 5V về 3.3V qua AMS1117, modul phát sóng miwi dùng chip MRF_24J40, modul kết nối với máy tính qua cổng USB sử dụng PL2303 và modul kết nối máy tính qua cổng COM sử dụng Max232.



- Thiết kế các tay điều khiển bao gồm các modul chính sau: modul nguồn chuyển đổi từ nguồn 5V về 3.3V qua AMS1117, modul phát sóng miwi dùng chip MRF_24J40, các phím bấm gắn ngoài qua header 8 chân.



4.3. Phần mềm quản lý bán hàng

4.3.1. Các giao dịch thực hiện

Giao dịch 1: Bán hàng

Khách hàng lựa chọn sản phẩm trong cửa hàng, nếu không có nhân viên tại quầy nhân viên sẽ thông báo cho cửa hàng trưởng, cửa hàng trưởng yêu cầu nhân viên tìm hàng trong kho. Cuối mỗi ca, nhân viên kiểm tra hàng và báo kết quả cho cửa hàng trưởng.

Giao dịch 2: Xuất kho

Khi có yêu cầu xuất kho, nhân viên kho sẽ kiểm tra xem còn hàng yêu cầu hay không. Nếu có hàng, nhân viên kho xuất hàng yêu cầu sau đó kê khai vào phiếu xuất kho rồi cuối mỗi ca nộp lại cho cửa hàng trưởng để cửa hàng trưởng xác nhận và nộp cho công ty.

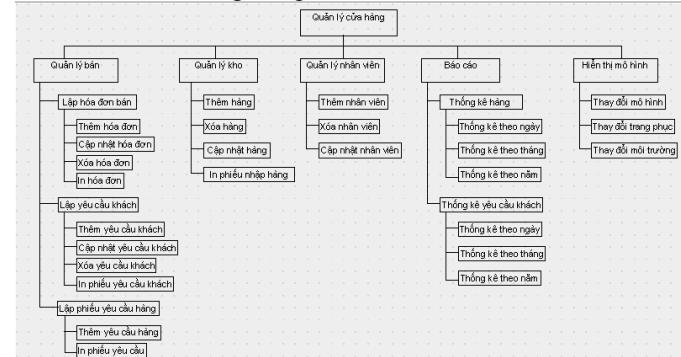
Giao dịch 3: Nhập kho

Khi có hàng mới về, cửa hàng trưởng sẽ đối chiếu số lượng hàng thực tế với chứng từ kèm theo. Sau đó thủ kho sẽ nhập hàng theo số lượng đã kiểm kê vào sổ kho và nhận hàng. Nếu có sai sót, hàng được giữ lại chờ xử lý.

Giao dịch 4: Thanh toán

Khi khách đã đồng ý mua bất kỳ sản phẩm nào có tại cửa hàng hay lấy từ kho, cửa hàng trưởng sẽ ghi lại mã hàng, số lượng, và giá tiền khách phải trả vào bảng kê hàng hóa và nhận thanh toán từ khách hàng. Cuối ngày, cửa hàng trưởng sẽ nộp lại bản kê khai bán hàng và phiếu xuất kho ngày hôm đó cho công ty.

4.3.2. Các chức năng của phần mềm



Các chức năng chính của phần mềm là:

- Quản lý bán
- Quản lý kho
- Quản lý nhân viên
- Báo cáo
- Hiển thị mô hình 3D(Thay đổi trang phục)

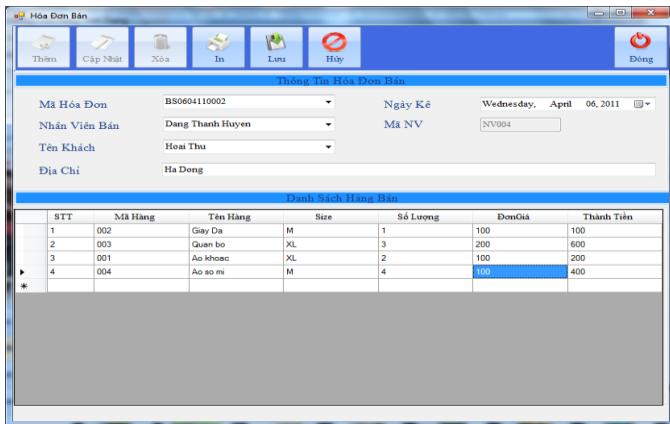
Giao diện của phần mềm:



- Giao diện form hiển thị mô hình 3D :



- Giao diện form hóa đơn bán:



5. KẾT LUẬN

Sau thời gian thực hiện đề tài này, em tự đánh giá mình đạt được các kết quả như:

- Tìm hiểu về mạng cảm biến không dây.
- Tìm hiểu công nghệ XNA xử lý các mô hình 3D.
- Tạo và điều khiển mô hình 3D của khách hàng.
- Thiết kế và xây dựng hệ thống node mạng và tay điều khiển tương ứng phục vụ cho hệ thống bán hàng tương tác.
- Xây dựng phần mềm quản lý cửa hàng thời trang.

6. LỜI TRI ÂN

Em xin chân thành cảm ơn sự giúp đỡ của thầy Nguyễn

Hồng Quang và sự hướng dẫn, chỉ bảo tận tình của thầy giáo Phạm Văn Thuận. Trong quá trình thực hiện đề tài, thầy đã luôn giúp chúng em có những định hướng tốt nhất, đồng thời giúp đỡ tháo gỡ những vấn đề nảy sinh. Em cũng xin cảm ơn các thầy cô trong Viện công nghệ thông tin và truyền thông đã truyền đạt cho em những kiến thức nền tảng cho em có thể xây dựng, phát triển đề tài khoa học này và các đề tài sau. Cuối cùng, xin cảm ơn các bạn trong lớp kỹ thuật máy tính đã đóng góp nhiều ý kiến xây dựng giúp đề tài hoàn thành tốt nhất. Em xin chân thành cảm ơn!

7. TÀI LIỆU THAM KHẢO

- [1] XNA Game Studio 3.0
- [2] Professional XNA Game Programming: For Xbox 360 and Windows – Benjamin Nitschke
- [3] Microchip MiWi™ P2P Wireless Protocol.
- [4] MiWi™ Wireless Networking Protocol Stack.
- [5] Microchip Wireless (MiWi™) Application Programming Interface – MiApp.
- [6] Microchip Wireless (MiWi™) Application Programming Interface – MiMac.
- [7] Website: <http://www.microchip.com>.
- [8] Website: <http://msdn.com>
- [9] Website: <http://wikipedia.org>

Nghiên cứu và xây dựng mô hình mạng của Network-on-Chip

Tạ Thị Hà Thu

Tóm tắt— Network On Chip (NoC) là một kiến trúc thiết kế chip hoàn toàn mới, thay vì thiết kế chip trên mô hình bus (đơn tầng hay phân tầng) truyền thống, NoC mở ra một hướng đi thiết kế chip trên mô hình mạng. NoC được nghiên cứu trên thế giới từ những năm 2005, tại phòng lab của các trường đại học, phòng R&D của các công ty chuyên sản xuất chip danh tiếng như Xilinx, Altera.

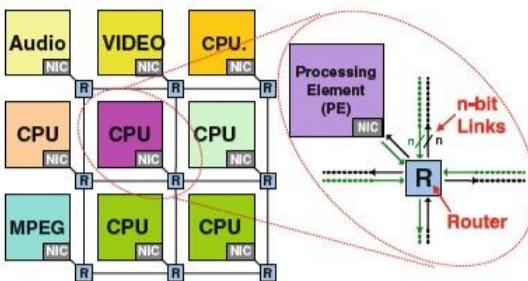
Hiện tại chưa có sản phẩm chính thức nào được đưa ra thị trường trên mô hình mới này, tất cả đều nằm ở dạng nghiên cứu và thử nghiệm.

Với mong muốn tìm hiểu một kiến trúc mới cũng như từng bước xây dựng hoàn thiện một sản phẩm trên kiến trúc này sinh viên thực hiện nghiên cứu kiến trúc NoC đồng thời thiết kế xây dựng một mạng Mesh trên cơ chế truyền tin cậy sử dụng ngôn ngữ VHDL trên công cụ Quartus 7.2 và kit DE2.

Từ khóa — Network-on-Chip, mạng mesh, ack, định tuyến, flit, lane

1. GIỚI THIỆU

Với kiến trúc bus tại một thời điểm chỉ có một master được chiếm dụng đường truyền. Hiện tại, SoC tích hợp khoảng 5 bộ xử lý và không nhiều hơn 10 bus master. Đường bus tỏa ra không hiệu quả khi xét đến năng lượng bởi vì việc truyền dữ liệu được thực hiện broadcast – dữ liệu đến được nhận với sự tiêu hao lớn nhất. Và Network-on-chip(NoC) ra đời với ý tưởng thay thế kiến trúc bus cũ, áp dụng mô hình mạng đưa vào chip. Các modul RAM, CPU, thiết bị ngoại vi sẽ được kết nối với nhau theo mô hình này với mong muốn sử dụng băng thông hiệu quả hơn, giảm được tốn hao năng lượng - điều này vô cùng quan trọng với thiết bị di động.



Công trình này được thực hiện dưới sự hướng dẫn của TS. Nguyễn Kim Khánh- Giảng viên Viện Công nghệ thông tin và Truyền thông- Trường Đại học Bách Khoa Hà Nội

Tạ Thị Hà Thu, sinh viên lớp Kỹ thuật máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0984685994, e-mail: hathu.bk1@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

2. TIẾP CẬN TỔNG QUAN

Nói đến mô hình mạng chúng ta cần quan tâm topology, định tuyến, điều khiển luồng, định dạng packet, chất lượng dịch vụ là tin cậy hay không tin cậy ...

Một số Topology như : mesh(mạng lưới), sao, ...

Và tùy vào topology mà người thiết kế đưa ra thuật toán định tuyến hay cơ chế điều khiển luồng phù hợp.

Ở tầng mạng NoC vẫn sử dụng định tuyến như mạng internet thông thường, ở tầng điều khiển luồng thì packet này được chia thành đơn vị nhỏ hơn được gọi là flit. Và như thế, các flit được luân chuyển qua router. Việc chuyển flit có thể được truyền theo cơ chế bắt tay, credit-base ...

Hình 1- chính là một mô hình mạng mesh 3×3 gồm có 9 router, các modul của chip truyền thông với router bới network-interface-card (NIC)

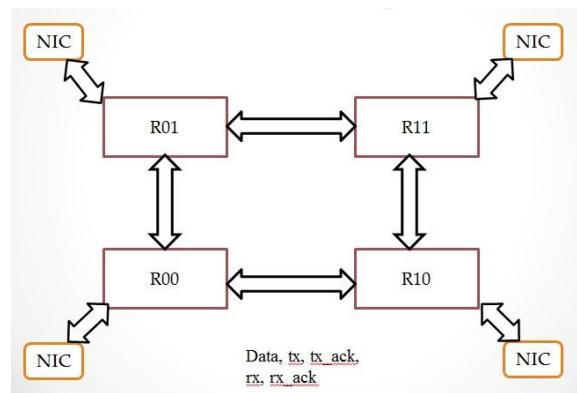
Như vậy, NoC có hai vấn đề lớn :

- Thiết kế NIC để nó chuyển đổi dữ liệu giữa modul đã tồn tại ở mô hình bus với router trong NoC.
- Thiết kế mô hình mạng của NoC: các router truyền thông với nhau như thế nào – Và đây chính là nội dung đề tài nghiên cứu.

3. THIẾT KẾ MÔ HÌNH MẠNG CỦA NETWORK ON CHIP

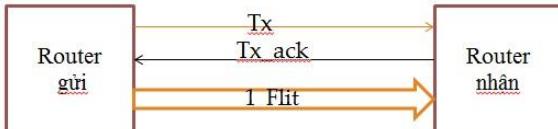
Mạng thiết kế với mô hình 4 router kết nối theo topology mesh- 2×2 ; định tuyến XY và cơ chế truyền tin cậy sử dụng ack. Mô hình này được xây dựng trên ngôn ngữ miêu tả phần cứng VHDL sử dụng phần mềm Quantus 7.2 và kit DE2

3.1 Mô hình mạng mesh 2×2



Các router truyền thông với nhau và với NIC của riêng chúng. Coi như các gói tin được đưa vào và đưa ra khỏi mạng từ các NIC này. Router thực hiện định tuyến và chuyển từng packet đến đúng địa chỉ và không mất dữ liệu.

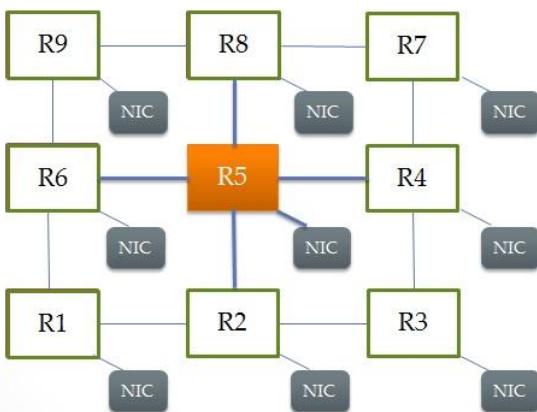
Với mô hình ack trên internet thông thường thì ack sẽ được gửi đi mỗi khi truyền một packet. Với NoC, packet được chia nhỏ hơn thành các flit để đảm bảo về bộ đệm. Đồng nghĩa với việc ack sẽ được gửi mỗi khi có flit cần truyền. Một cách chung nhất, các router sẽ trao đổi tín hiệu tx, rx (yêu cầu truyền flit); ack_tx, ack_rx (bộ đệm còn trống, đồng ý nhận flit) ; data cần truyền đi.



Hình 3. Cơ chế truyền tin cậy sử dụng ack

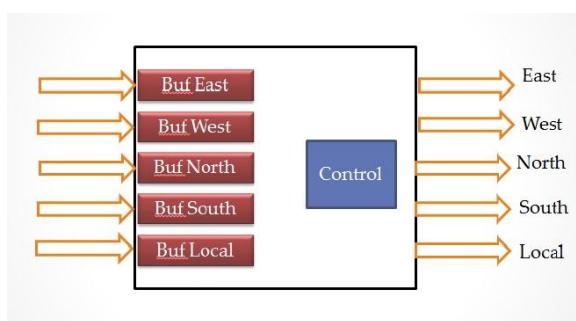
Vì cơ chế hoạt động của các router này đều giống nhau nên việc thiết kế một router với đầy đủ chức năng là trọng tâm của hệ thống.

3.2. Kiến trúc của Router



Hình 4 – mạng mesh tổng quát, Router có tối đa 5 cổng truyền nhận

Trong mesh 2×2 này thì mỗi Router chỉ có 3 cổng truyền nhận, nhưng mạng mesh tổng quát có kích thước $n \times n$ ($n > 2$) thì một router sẽ có tối đa 5 cổng truyền nhận



Hình 5. Kiến trúc Router

Router sẽ có một bộ đệm tại mỗi một cổng nhận dữ liệu, sau khi được định tuyến và ack, dữ liệu sẽ đi từ bộ đệm này ra một trong 5 cổng – được gọi là cổng ra. Cách thức thiết kế bộ đệm này được gọi là bộ đệm một phía cổng In. Ngoài ra thì cũng có loại bộ đệm 2 phía In-Out (2 cổng ra, vào đều có bộ đệm), hoặc bộ đệm một phía cổng Out.

Cách đánh địa chỉ bộ đệm theo East, West, North, South và Local theo đúng mô hình mạng mesh đưa ra.

Ví dụ với hình số 4: R5 liên kết với R8 theo hướng bắc - North ; với R4 theo hướng đông - East; R2 theo hướng nam -South, R6 theo hướng tây - West và với card NIC là Local. Dữ liệu mà các router, hay card NIC gửi đến router 5 này sẽ được xếp vào các bộ đệm có địa chỉ tương ứng. Tương tự như vậy với việc gửi dữ liệu đi.

Bộ control sẽ xác định cổng cần ra cho mỗi flit cũng như thời điểm flit được chuyển ra ngoài, vì tại một thời điểm chỉ có duy nhất một flit được chiếm đường truyền tại một cổng ra.

Trước khi đi vào chi tiết kiến trúc từng component của Router, mình đưa ra cách thức tổ chức một packet

3.2.1. Tổ chức packet

Packet là đơn vị dữ liệu được thực hiện ở tầng mạng, căn cứ vào header của packet mà Router xác định được địa chỉ cần chuyển packet đến. Xét đến tầng datalink thì packet lại được chia nhỏ thành các flit, lần lượt từng flit này sẽ được truyền đi. 1 packet = Header flit + counter flit + các data flit

- Header flit chứa địa chỉ nguồn và địa chỉ đích – phần này để định tuyến
- Counter flit lưu kích thước số flit data của packet – phần này để giải phóng việc một packet chiếm giữ đường truyền.



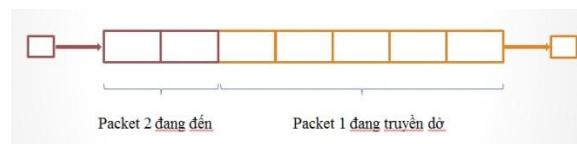
Hình 6. Tổ chức packet

3.2.2. Kiến trúc bộ đệm

Router có 5 bộ đệm tại 5 cổng vào và hoạt động của các tập bộ đệm này là như nhau.

Thay vì sử dụng bộ đệm theo cơ chế tuần tự, bộ đệm được thiết kế với cơ chế song song

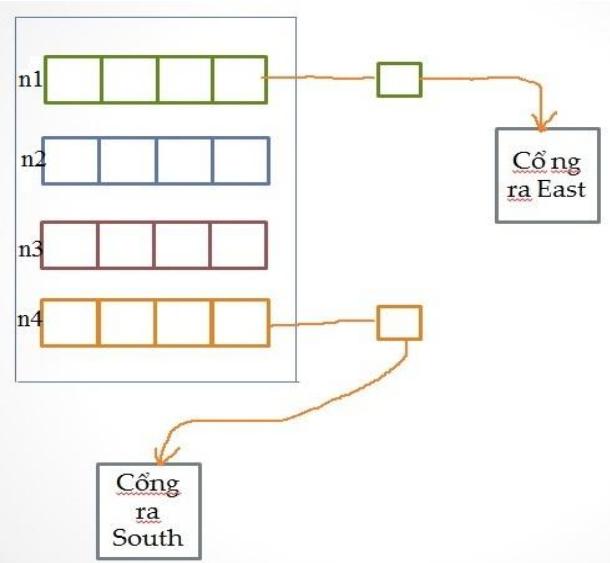
- Cơ chế tuần tự



Hình 7. Bộ đệm truyền tuần tự

Các packet được đẩy ra theo đúng thứ tự khi nó đến, khi một packet truyền xong, thì packet tiếp theo mới được chiếm đường truyền. Điều này cũng đồng nghĩa với việc nếu như gói packet 1 bị nghẽn thì packet 2 cũng bị nghẽn

- Cơ chế song song



Hình 8. Bộ đệm song song

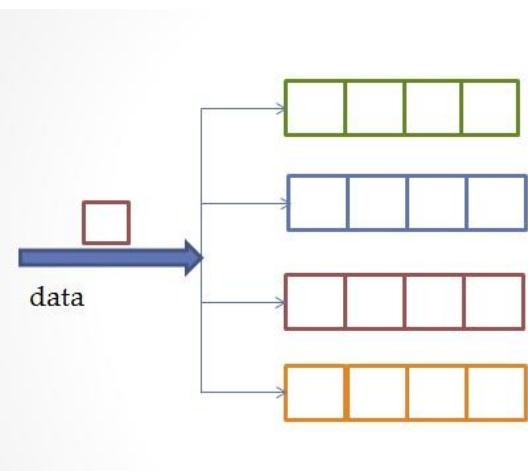
Thay vì sử dụng duy nhất một bộ đệm như ở dạng tuần tự, bộ đệm song song này có n bộ đệm con ($n \geq 2$) – được gọi là lane, mỗi lane lưu các packet khác nhau.

VD ở hình 7: lane n1 và n4 lưu hai packet khác nhau, và chúng cũng định tuyến ra hai cổng khác nhau. Nếu như packet ở lane 1 bị nghẽn thì packet ở lane 4 vẫn được chuyển ra.

Về hiệu quả truyền gói tin thì bộ đệm song song tỏ ra hiệu quả hơn nhưng về mặt điều khiển thì bộ đệm song song hoạt động phức tạp hơn bộ đệm tuần tự nhiều.

3.2.2.1. Chức năng nhận dữ liệu

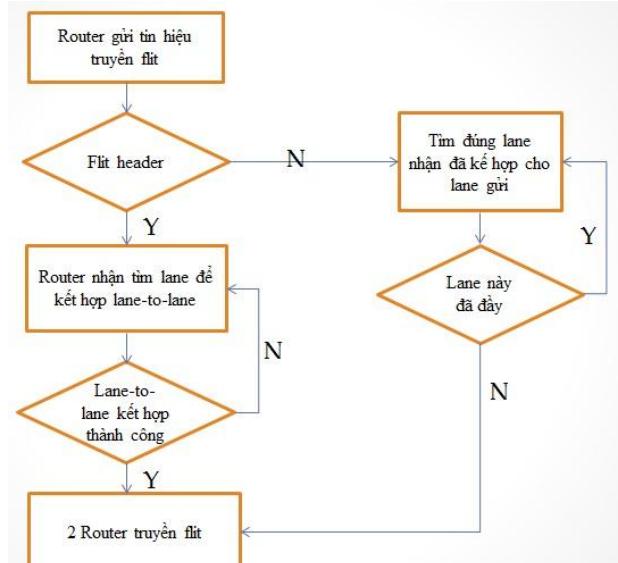
3.2.2.1.1. Giới thiệu chung



Hình 9. Nhận dữ liệu của bộ đệm

Khi có một flit đến bộ đệm, thì flit này sẽ được chuyển vào lane nào?

Vì đơn vị truyền giữa các Router là flit còn đơn vị định tuyến là packet mà một packet lại có nhiều flit nên để hạn chế việc mỗi khi cần truyền một flit lại phải định tuyến lại các flit của cùng một packet chỉ được xếp vào cùng một lane ở bên nhận cũng như bên gửi. Thuật ngữ lane-to-lane thể hiện sự kết hợp lane ở bên gửi với lane ở bên nhận.



Hình 10. Trình tự thực hiện của bộ đệm nhận khi có flit

Trên đây là sơ đồ chung các bước bộ đệm nhận thực hiện khi có một flit truyền đến nó. Nếu flit truyền đến là flit header thì bộ đệm nhận sẽ phải chọn một lane phù hợp (lane chưa kết hợp với lane nào ở bên gửi và lane này chưa đầy). Nếu sự kết hợp là thành công thì lane này sẽ là “độc quyền” của packet chứa flit vừa đến. Và sự kết hợp này chỉ được giải phóng khi bộ đệm gửi dụng cờ endConnect.

Một giả thiết đặt ra là flit đến nhưng bộ đệm bên nhận không thể đáp ứng như vậy sẽ dẫn đến mất dữ liệu. Và với cơ chế truyền tin cậy sẽ khắc phục được nhược điểm này với việc sử dụng ack trước khi truyền một flit, còn phần nhận flit vẫn tuân theo sơ đồ chung Hình 9

3.2.2.1.2. Tín hiệu rx và ack_rx

Rx : là tín hiệu yêu cầu nhận flit gửi đến bộ đệm.

Ack_rx là tín hiệu phản hồi từ bộ đệm, nếu ack_rx = 1 khi bộ đệm chưa đầy. Và bên gửi có thể truyền flit

Vì bộ đệm nhận có 4 lane, nếu tại một thời điểm chỉ có một tín hiệu tx gửi đến thì sẽ không phát huy được hết tiện ích của bộ đệm song song. Bởi lẽ nếu tín hiệu rx đang gửi đến yêu cầu một lane đã kết hợp (đang truyền dở packet) mà lane đó đã đầy thì rx này bị hủy bỏ. Nên thay vì để duy nhất một tín hiệu rx, rx sẽ là một vector có 4 phần tử, điều này cũng có nghĩa rằng, một cổng ra sẽ chỉ cho phép 4 tín hiệu rx từ các lane ra ngoài.

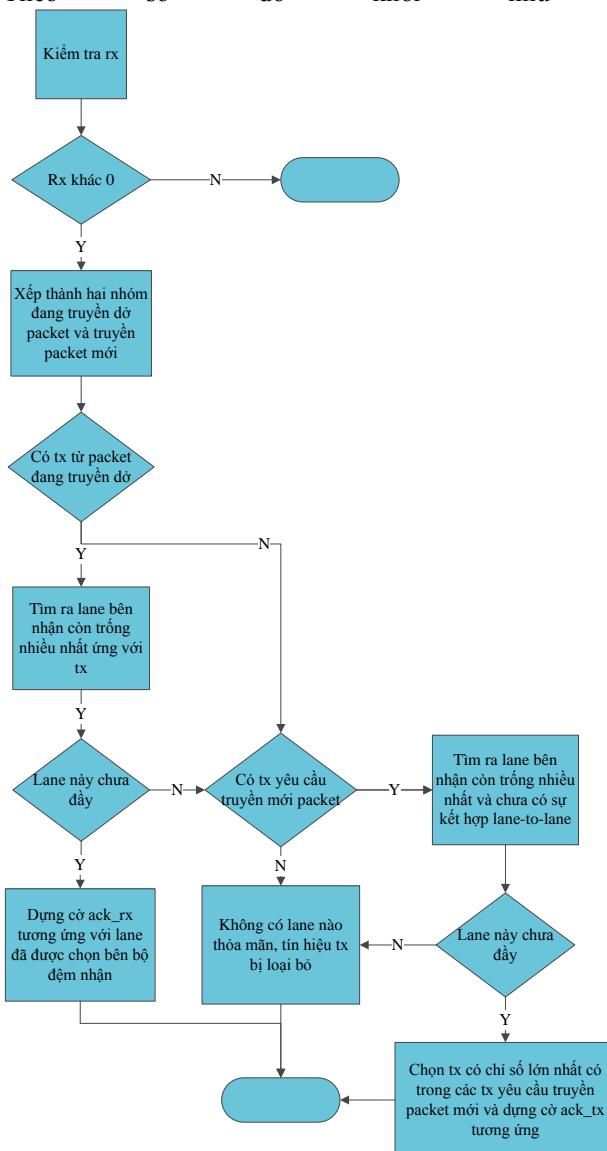
Khi vector rx đến, sẽ có những phần tử của tx = 1 tức là có nhiều hơn 1 tín hiệu yêu cầu nhận flit từ các lane khác nhau. Mà bộ đệm nhận chỉ đáp ứng được duy nhất một rx vậy chế độ phân xử nào giúp bộ đệm chọn được.

Mình đã đưa ra 2 cơ chế phân xử như sau

- Khi tx gửi đến, sẽ có tx từ các packet đang truyền dở và packet mới. Bước phân xử một ưu tiên các lane đang truyền dở packet
- Qua bước phân xử một sẽ chọn ra tập các tx cùng cấp: cùng là truyền packet dở, hoặc cùng là truyền packet mới. Nếu là packet đang truyền dở: chọn lane bên nhận còn trống nhiều nhất và tìm được tx tương ứng. Nếu là packet mới lane nhận sẽ là lane còn trống

nhiều nhất còn lane gửi là lane có chỉ số vector cao nhất có thể.

Theo sơ đồ khái niệm sau



Hình 11.Biểu đồ lựa chọn rx của modul nhận dữ liệu

Kết thúc hai quá trình phân xử, nếu bộ đếm thỏa mãn thì sẽ có duy nhất một phần tử trong vector ack_tx bằng 1 và ngược lại tx bị từ chối thì ack_tx = 0.

3.2.2.2. Chức năng gửi dữ liệu

Bộ đếm ngoài chức năng nhận dữ liệu từ cổng ra của Router lân cận nó còn có chức năng gửi dữ liệu đồng thời quản lý việc giải phóng sự kết hợp lane-to-lane.

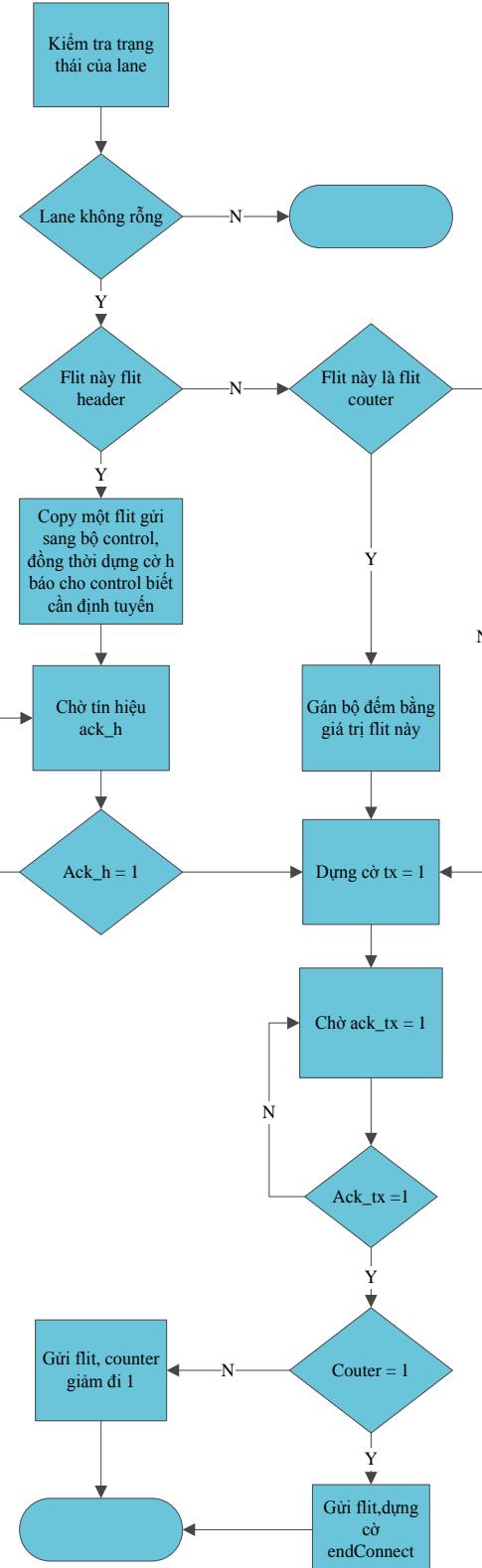
Nếu như modul nhận tại một thời điểm chỉ nhận được duy nhất một flit thì modul gửi có khả năng gửi nhiều flit, mỗi lane có thể gửi được 1 flit tại một thời điểm. Các lane trong bộ đếm hoạt động song song và độc lập nhau.

Xét hoạt động gửi flit của một lane

Khi lane có flit, nó sẽ kiểm tra đó có phải là flit header hay không, nếu là flit header, nó gửi flit này sang bộ control để thực hiện định tuyến, khi việc định tuyến xong, nó dụng cờ rx và

chờ ack_rx từ Router nhận. Khi có ack_rx từ bên gửi, nó đẩy data ra ngoài.

Trong quá trình truyền dữ liệu, bộ đếm gửi sẽ giám sát quá trình truyền bằng cách sử dụng counter flit = flit thứ hai của packet và dựng cờ endConnect khi toàn bộ packet đã được truyền xong. Bên nhận nhận được cờ này và sẽ tự hủy kết nối lane – to – lane



Hình 12.Biểu đồ gửi dữ liệu

3.2.3. Kiến trúc bộ control

Bộ control có hai chức năng là định tuyến và cấp đường truyền cho các gói tin. Với chức năng nhận dữ liệu thì bộ đệm không cần có sự can thiệp của bộ control mà tự nó tính toán và điều khiển được. Nhưng ở chức năng gửi, khi có flit header bộ đệm sẽ yêu cầu bộ control định tuyến, và khi gửi tín hiệu tx, bản thân bộ đệm không quan tâm tín hiệu tx này được gửi đến cổng nào và ack_tx được nhận từ cổng nào, việc này sẽ do bộ control điều phối.

3.2.3.1. *Chức năng định tuyến*

Để gửi được gói tin một cách chính xác, bộ control sẽ phải thực hiện định tuyến. Vì đề tài này tập trung đến việc xây dựng mô hình mạng và cơ chế điều khiển luồng nên thuật toán định tuyến được áp dụng khá đơn giản – định tuyến XY

- Định tuyến XY là dạng định tuyến mà đường truyền được xác định trước khi gói tin được truyền đi, nói một cách khác thì đường đi của packet chỉ phụ thuộc vào đích và nguồn.

Flit header lưu địa chỉ của Router đích và nguồn và dưa vào thông tin này, bộ control điều khiển hướng ra của gói tin. Vì tại mỗi thời điểm chỉ có 3 tín hiệu tx được gửi ra ở một cổng, tức là sẽ có nhiều nhất 3 lane sẵn sàng gửi flit ra cổng này. Nên tại một thời điểm bộ control cũng chỉ cấp quyền hoạt động cho 3 lane trên một cổng (trong khi số lane nhiều nhất muôn đầy gói tin ra một cổng là $5 \times 4 = 20$ lanes). Các lane chưa được định tuyến sẽ nằm ở trạng thái chờ cho đến khi cờ endConnect được dựng lên.

Và tại một thời điểm thì control cũng chỉ định tuyển được cho duy nhất một lane nên nó sẽ phải đưa ra thứ tự định tuyển

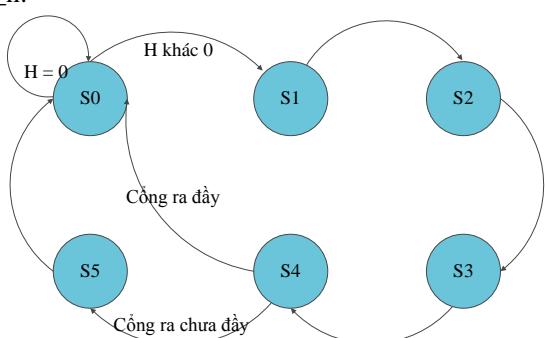
Sẽ có hai trang thái là Pre – Next

Cả bộ đệm và lane được điều xếp theo cơ chế vòng tròn
EAST → WEST → NORTH → SOUTH → LOCAL

Lane 0 \rightarrow lane 1 \rightarrow lane 2 \rightarrow lane 3

Vd: nếu lane hiện tại đang định tuyến là lane 1 của bộ đệm West thì lane được ưu tiên tiếp theo sẽ là lane 2 của bộ đệm NORTH

Kết thúc việc định tuyến, bộ control sẽ có danh sách các lane đã được định tuyến, địa chỉ cổng packet của lane này sẽ được chuyển đi. Khi cần định tuyến thì lane gửi tín hiệu h của nó, sau khi định tuyến xong thì bộ control sẽ gửi trả lại bộ đệm tín hiệu ack h .



Hình 13. Biểu đồ trang thái định tuyến gói tin

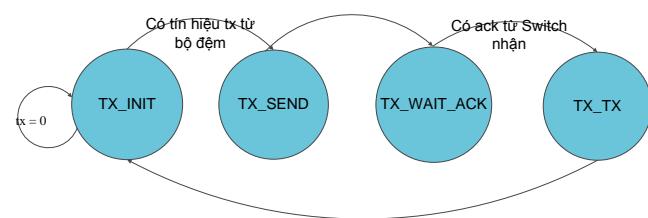
Sự thay đổi trạng thái gắn liền với sự kiện sùn dương của đồng hồ

- ✓ S0: Trạng thái khởi tạo
 - ✓ S1: Trạng thái lựa chọn cổng định tuyến tiếp theo
 - ✓ S2: Trạng thái lựa chọn lane định tuyến tiếp theo
 - ✓ S3: Chuyển tham số cổng và lane
 - ✓ S4 : Định tuyến cho lane và kiểm tra tình trạng của cổng ra
 - ✓ S5: dựng cờ ack_h gửi lại cho bộ đệm

3.2.3.2. Cáp đường truyền cho các gói tin

Cáp đường truyền cho các gói tin bao gồm các chức năng: sáp các tín hiệu tx công ra để gửi đến Router nhận; nhận tín hiệu ack_tx từ các cổng và trả về đúng lane đã gửi tx. Bộ đệm phải đảm bảo không có nhiều hơn hoặc bằng hai lane định tuyến chung một cổng ra mà nhận ack_tx cùng một thời điểm.

Mỗi một công ra có một thanh ghi 4 phần tử tương ứng với các tín hiệu tx mà các lane gửi đến, nên bộ control phải biết rằng từng phần tử của thanh ghi này đang tương ứng với lane nào để nó chuyen tín hiệu ack tx về cho các lane chính xác



Hình 14. Biểu đồ trạng thái cấp đường truyền cho gói tin
Sự thay đổi trạng thái gắn liền với sự kiện sườn dương của
đồng hồ

- ✓ TX_INIT : trạng thái khởi tạo
 - ✓ TX_SEND: Trạng thái này sắp các tx vào các lane của cổng ra rồi gửi đi
 - ✓ TX_WAIT_ACK: Trạng thái chờ tín hiệu ack từ Switch nhận
 - ✓ TX TX gửi tín hiệu ack tx về đúng lane đã yêu cầu tx

4. KẾT LUẬN

Sau quá trình thực hiện đề tài này mình đã thực hiện được

- Tìm hiểu kiến trúc Network-on-chip
 - Thiết kế mô hình mạng mesh 2×2 theo cơ chế truyền tin cậy sử dụng ack
 - Xây dựng và kiểm thử trên phần mềm Quantus 7.2 và kit DE2.

5. LỜI TRI ÂN

Em xin chân thành cảm ơn sự giúp đỡ của thầy Nguyễn Kim Khánh. Trong quá trình thực hiện đề tài thầy đã luôn tận tình hướng dẫn, gợi ý cho em nhiều ý tưởng để đưa ra được thiết kế. Em xin chân thành cảm ơn các thầy cô trong Viện công nghệ thông tin và truyền thông đã truyền cho em những kiến thức nền tảng giúp em hoàn thành nghiên cứu. Xin gửi lời cảm ơn tới giáo sư Zhonghai Lu đại học Stockholm đã gửi cho em những tài liệu rất quý về NoC. Cuối cùng em xin gửi lời cảm ơn tập thể lớp Kỹ Thuật Máy Tính đã nhiệt tình hỗ trợ em trong

quá trình thực hiện đề tài.

6. TÀI LIỆU THAM KHẢO

- [1] Zhonghai Lu, “Design and Analysis of On-Chip Communication for Network-on-Chip platforms”
- [2] Tobias Bjerregaard and Shankar mahadevan, “A survey of Research and practices of Network-on-Chip”
- [3] Axel Jantsch, Royal Institute of technology,stockholm “NOC architecture”
- [4] Rostislav Dobkin, “Creditbase Communication in NoCs”.

Nghiên cứu mạng cảm biến không dây và ứng dụng trong hệ thống xếp chỗ tự động

Trần Duy Phương

Tóm tắt - Ngày nay,các mạng không dây WiFi, WiMax, 3G đã dần trở nên phổ biến và không còn xa lạ với mọi người. Công nghệ không dây đang hướng tới việc kết nối các thiết bị gia dụng cũng như kết nối các bộ phận chức năng trong ngôi nhà để có thể điều chỉnh và kiểm soát từ xa các hệ thống gaz, điện nước, ánh sáng ...Với mong muốn xây dựng một hệ thống truyền tin tin cậy, đề tài này đã đi sâu nghiên cứu về giao thức MiWi(một phiên bản thu gọn Zigbee của Microchip) và xây dựng một mạng cảm biến không dây sử dụng chip MRF24J40 của microchip để truyền thông tin trong hệ thống xếp chỗ tự động.

Từ khóa - Mạng cảm biến không dây, MiWi, MRF 24J40, Wireless sensor network.

1. GIỚI THIỆU

Ứng dụng xếp chỗ tự động QMS(Queue Management System) đang được dần trở nên phổ biến ở các nơi hành chính như ngân hàng , bệnh viện,hải quan...Nhờ có các hệ thống phục vụ này mà những nơi hành chính đó trở nên văn minh và lịch sự hơn, tiện lợi và nhanh chóng hơn. Hiện nay các hệ thống như vậy hiện đang được triển khai trên nền tảng mạng RS485 phải đi dây và nhiều khi là bất tiện ở những nơi đã được xây dựng từ trước, gây mất mỹ quan.Vì thế đề tài này nhằm ứng dụng mạng cảm biến không dây vào việc triển khai hệ thống xếp chỗ tự động nhằm mục đích khắc phục điều đó.

2. MẠNG MIWI/IEEE 802.15.4

Giao thức MiWi được xây dựng dựa trên tầng MAC và PHY của chuẩn giao thức IEEE 802.15.4 và được sử dụng cho phát triển một mạng không dây đơn giản trên băng tần 2.4Ghz với tốc độ truyền thông thấp vào khoảng 250 kbps. Giao thức cung cấp các đặc tính như tìm, tạo và kết nối tới một mạng, khám phá các nodes trong một mạng và định tuyến chúng. Đặc điểm chính của chuẩn này là tính mềm dẻo, tiêu hao ít năng lượng, chi phí nhỏ, tốc độ truyền dữ liệu thấp trong khoảng không gian nhỏ, thuận tiện khi áp dụng trong khu vực nhà riêng, văn phòng.

Công trình này được thực hiện dưới sự hướng dẫn của Ths.Phạm Văn Thuận-giảng viên Viện Công Nghệ Thông Tin và Truyền Thông, trường Đại Học Bách Khoa Hà Nội.

Trần Duy Phương, sinh viên lớp Kỹ thuật máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: (+84)0944-525-386, e-mail: tranduyphuong1988@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

2.1.Thành phần của mạng:

-PAN coordinator(FFD): là một thiết bị mạng có nhiệm vụ khởi tạo mạng,cấp phát địa chỉ mạng và tổ chức bảng kết nối(binding table)

-Coordinator(FFD):sử dụng để mở rộng dải phủ sóng của mạng.Nó cho phép nhiều nodes hơn tham gia vào mạng.Cũng có thể dùng cho các chức năng giám sát và điều khiển

-End Device(RFD or FFD): sử dụng cho các chức năng giám sát và điều khiển.

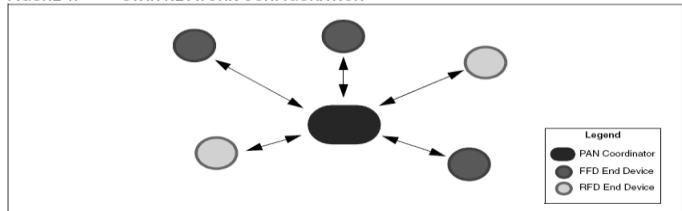
2.2.Kiến trúc mạng:

Trong các loại thiết bị mạng MiWi thì thiết bị có vai trò quan trọng nhất là PAN coordinator. PAN coordinator là thiết bị sẽ khởi tạo mạng, chọn kênh và PAN ID của mạng.

2.2.1. Cấu hình Star Network:

Một cấu hình mạng Star Network gồm một node PAN coordinator và một hoặc vài thiết bị kiểu end devices khác.Trong mạng Star Net,tất cả các thiết bị end devices chỉ được phép giao tiếp với PAN coordinator.

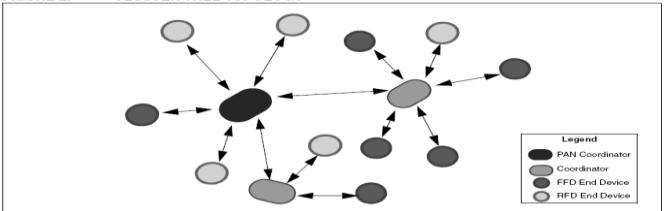
FIGURE 1: STAR NETWORK CONFIGURATION



2.2.2. Cấu hình Cluster-Tree Network:

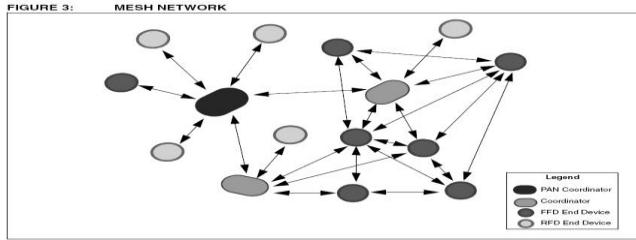
Trong một Cluster Tree Net vẫn sẽ có một PAN coordinator,tuy nhiên các thiết bị coordinators khác được cho phép tham gia vào mạng.Việc này tạo nên cấu trúc tree-like-structure.Ở đây,PAN coordinator là gốc của cây(the root of the tree),các coordinators sẽ là các nhánh của cây(the branches of the tree) và các end devices như là các lá của cây(the leaves of the tree).

FIGURE 2: CLUSTER TREE TOPOLOGY



2.2.3. Cấu hình Mesh Network:

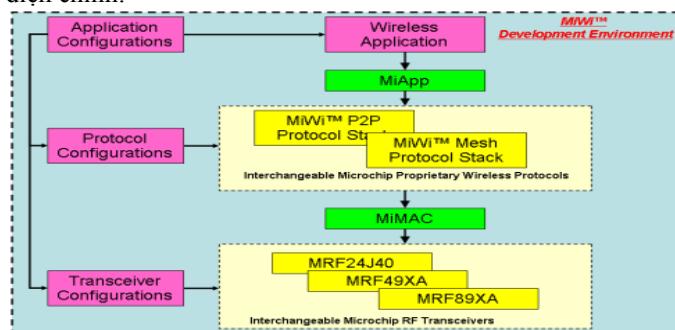
Một Mesh Network tương tự như một cấu hình Cluster Tree Network, ngoại trừ rằng FFDs có thể định tuyến các messages trực tiếp tới các FFDs khác thay vì phải theo cấu trúc cây.Các Messages tới RFDs vẫn phải đi qua node cha của RFDs (the RFDs'parent node).



2.3.Giới thiệu về giao thức MiWi:

Môi trường phát triển MiWi là giải pháp không dây của tập đoàn Microchip. Giải pháp không dây này giúp cho khách hàng dễ dàng phát triển các ứng dụng mạng không dây và giảm thời gian thương mại hóa sản phẩm.

Gói giải pháp này hỗ trợ các giao thức như MiWi Mesh và MiWi P2P. Môi trường phát triển MiWi bao gồm 2 tầng giao diện chính:

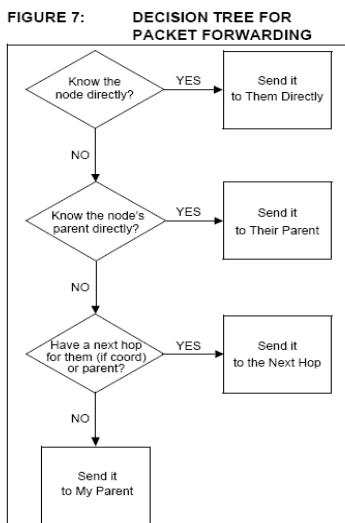


-Giao diện MiApp cung cấp các hàm cho phép phát triển trên tầng ứng dụng.

-Giao diện MiMac cung cấp các hàm cho phép giao tiếp với các modun thu phát MRF ở phía dưới qua giao tiếp SPI.

Với những người phát triển ứng dụng thì chỉ cần quan tâm sử dụng các dịch vụ hỗ trợ trên giao diện MiApp là có thể triển khai được một ứng dụng sử dụng mạng cảm biến không dây MiWi một cách dễ dàng.

2.3.1. Định tuyến trong giao thức MiWi: Định tuyến trong mạng MiWi trở nên dễ hơn khi chúng ta biết các neighboring coordinator. Gửi một packet tới một node khác sẽ tuân theo sơ đồ logic sau:



Nguyên tắc định tuyến trong mạng MiWi là dựa trên short address. Khi một gói tin được gửi từ node này tới node khác thì bao giờ gói tin cũng được gửi tới node cha của node gửi đầu

tiên. Sau đó node cha này sẽ căn cứ vào địa chỉ nhận mà sẽ quyết định forward gói tin đi đâu. Như vậy thuật toán định tuyến trong MiWi tuân theo mô hình cây.

2.3.2. Khám phá một node trong mạng:

Khi 2 nodes giao tiếp trên mạng giao thức MiWi chúng sử dụng Short Address của chúng. Nếu một topology mạng từng thay đổi, sẽ rất hữu ích để tìm kiếm lại thiết bị. Bởi vì địa chỉ Short Address được gán cho thiết bị bởi cha của nó nên địa chỉ Short Address của thiết bị có thể thay đổi. Nếu rơi vào trường hợp này thì sau đó địa chỉ Short Address của thiết bị phải được khám phá trước khi việc truyền thông được thiết lập lại. Không giống như Short Address, EUI của thiết bị không bao giờ thay đổi và là duy nhất. Nếu một thiết bị biết EUI của thiết bị khác, nó sẽ có thể phân biệt được thiết bị này so với các thiết bị khác. Tìm kiếm mạng nhờ vào EUI trở nên quan trọng trong việc thiết lập lại sự truyền thông với các nodes đã thay đổi. Giao thức MiWi cung cấp đặc tính tìm kiếm mạng dựa vào EUI.

Có hai loại gói tin Stack được định nghĩa để hỗ trợ tìm kiếm dựa trên EUI trên mạng là:

EUI_ADDRESS_SEARCH_REQUEST và EUI_ADDRESS_SEARCH_RESPONSE

+**Bước 1:** Yêu cầu tìm kiếm được unicast tới coordinator đầu tiên.(sequence 1)

+**Bước 2:** Broadcast yêu cầu đó giữa các coordinator trong mạng với EUI của thiết bị cần xác định.

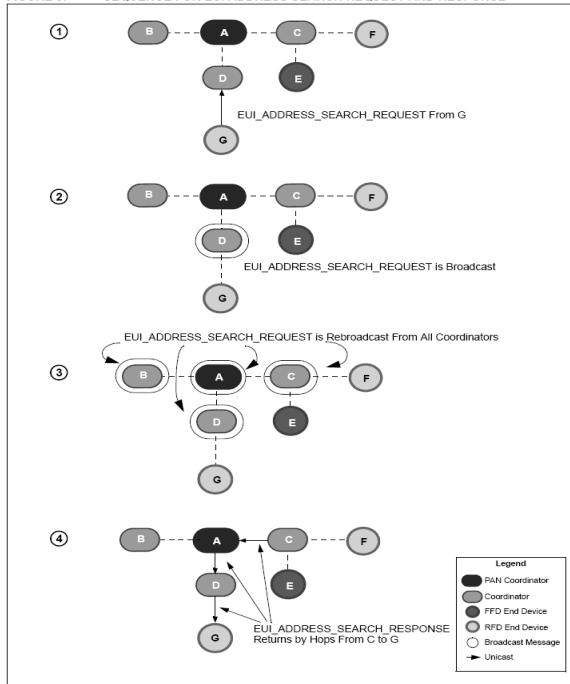
+**Bước 3:** Rebroadcast giữa các coordinator cho tới khi bộ đếm Hops(hops counter) triệt tiêu.

+**Bước 4:** Nếu một trong các coordinator hiện tại chứa node con cần tìm, nó sẽ gửi về gói tin

EUI_ADDRESS_SEARCH_RESPONSE cùng với EUI của thiết bị và địa chỉ ShortAddress của nó.

Chú ý: **EUI_ADDRESS_SEARCH_RESPONSE** được gửi theo cơ chế unicast node by node(từ coordinator chứa node con đó) Xem hình vẽ:

FIGURE 8: SEQUENCE FOR EUI ADDRESS SEARCH REQUEST AND RESPONSE



2.3.3. Gói tin trong giao thức MiWi:

Giao thức MiWi sử dụng khuôn dạng gói tin tầng MAC IEEE 802.15.4 cho tất cả các gói tin của nó. Trên tầng này sẽ đặt thêm phần MiWi protocol Header mà chưa thông tin cần cho định tuyến và xử lý gói tin. Khuôn dạng Header:

+**Hops:** Chỉ ra số hops mà gói tin được cho phép truyền lại(00h nghĩa là không truyền lại gói tin-1byte)

+**Frame Control:** Trường này là một bản đồ bit(bit map) để định nghĩa các đặc tính của gói tin.(1 byte)

+**Dest PANID:** đây là PAN ID của node đích cuối cùng(final destination node-2bytes in MiWi)

+**Dest Short Address:** Địa chỉ Short Addr của đích(2 bytes)

+**Source PANID:** PAN ID của node mà gửi gói tin(2 bytes)

+**Source Short Address:** Địa chỉ short address của node gửi gói tin(2 bytes)

+**Sequence Number:** Số hiệu có thể được sử dụng để theo dõi trạng thái của gói tin khi chúng di chuyển qua mạng(1byte)

+**Report Type:** Nhóm các thông điệp trong gói tin này. Các gói tin tạo từ Stack đều cho Report Type là 00h. Các reports người dùng định nghĩa từ 01h→FFh.(1byte)

+**Report ID:** Loại thông điệp chứa trong gói tin này(1 byte)

FIGURE 6: MIWI™ PROTOCOL PACKET HEADER FORMAT

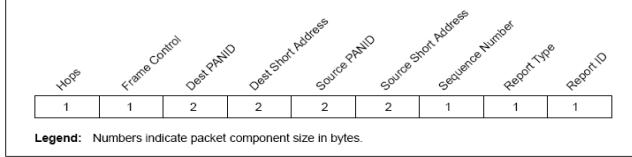


TABLE 3: FRAME CONTROL BIT FIELD

bit 7-3	0	0	0	0	0	x	1	0
bit 7	r	r	r	r	r	ACKREQ	INTRCLST	ENCRYPT
								bit 0

bit 7-3 **Reserved:** Maintain as '0' in this implementation
 bit 2 **ACKREQ:** Acknowledge Request bit
 When set, the source device requests an upper layer Acknowledgement of receipt from the destination device.
 bit 1 **INTRCLST:** Intra Cluster bit
 Reserved in this implementation, maintain as '1'.
 bit 0 **ENCRYPT:** Encrypt bit
 When set, data packet is encrypted at the application level.

Note: Abbreviated bit names are for convenience of display only; they are not an official part of IEEE 802.15.4™.

2.3.4. Các thao tác cơ bản với giao thức MiWi(MiWi Stack):

Các hoạt động cơ bản của giao thức mạng không dây MiWi được thực thi thông qua một tập các lời gọi hàm cơ bản

2.3.4.1. Hình thành một mạng

Một thiết bị có khả năng trở thành PAN coordinator có thể quyết định rằng không có một mạng thích hợp nào có sẵn và tự tạo mạng cho chính nó.

2.3.4.2. Tham gia vào mạng(joining a network):

Khi tìm kiếm mạng hoàn thành,một thiết bị phải xem xét kĩ lưỡng danh sách kết quả mạng và quyết định mạng nào sẽ được chấp nhận kết nối đến.Nếu một mạng thích hợp được khám phá, một thiết bị có thể cố gắng tham gia vào mạng đó.

2.3.4.3. Gửi dữ liệu:

MiWi stack cung cấp cho ta hai bộ đệm phục vụ cho việc truyền và nhận dữ liệu.Trước tiên ta sẽ nói về cách truyền dữ liệu:

- Đầu tiên ta phải xóa bộ đệm truyền để ghi dữ liệu.Ta sử dụng hàm sau:

```
MiApp_FlushTx();
#define MiApp_FlushTx() {TxData = PAYLOAD_START;}
```

Đây chỉ là một macro cho phép reset bộ đệm truyền bằng cách reset con đếm TxData.

- Sau đó ta ghi lần lượt các byte dữ liệu muốn truyền vào bộ đệm bằng cách sử dụng hàm:

```
MiApp_WriteData(a);
#define MiApp_WriteData(a) TxBuffer[TxData++] = a
```

Đây cũng chỉ là một macro cho phép gán dữ liệu vào bộ đệm đồng thời tăng con đếm TxData.

- Tiếp theo tùy vào loại hình truyền broad cast hay Unicast mà ta sẽ sử dụng các hàm sau:

```
BOOL MiApp_BroadcastPacket(BOOL SecEn);
```

Hàm này sẽ gửi gói tin nằm trong TxBuffer tới tất cả các node trong dải sóng radio của nó.Tức là mọi node trong dải sóng radio của nó đều có thể nhận được gói tin này. Tham số SecEn chỉ ra rằng gói tin có mã hóa khi SecEn=True và không được mã hóa khi SecEn=False. Hàm trả về True để chỉ ra rằng việc broadcast gói tin là thành công và là False nếu thất bại.

```
BOOL MiApp_UncastConnection(BYTE ConnectionIndex, BOOL SecEn);
```

Hàm này sẽ gửi gói tin nằm trong TxBuffer tới chỉ node có thứ tự là ConnectionIndex nằm trong bảng ConnectionTable(hay NetWorkTable) của node gửi.

```
BOOL MiApp_UncastAddress(BYTE *DestinationAddress, BOOL PermanentAddr, BOOL SecEn);
```

Hàm này sẽ gửi gói tin nằm trong TxBuffer tới thiết bị có địa chỉ nằm trong DestinationAddress.

Tham số PermanentAddr để chỉ ra rằng địa chỉ thiết bị đích là permanent address (hay long address) hoặc alternative network address (hay short address) .

2.3.4.4. Nhận dữ liệu:

MiWi Stack cung cấp cho ta một hàm kiểm tra sự xuất hiện của các gói tin bằng cách dùng hàm:

```
BOOL MiApp_MessageAvailable(void);
```

Hàm trả về true nếu có một gói tin được nhận và ngược lại trả về false nếu không có gói tin nào được nhận cả. Hàm này còn có một nhiệm vụ quan trọng khác đó là duy trì sự hoạt động của Stack.

Khi một gói tin được nhận thì toàn bộ thông tin của gói tin được lưu trong một biến toàn cục là RxMessage có kiểu là cấu trúc RECEIVED_MESSAGE.Chú ý là gói tin nhận về ở đây nằm trên tầng ứng dụng.

Sau khi nhận về gói tin và xử lý nó ta phải gọi hàm:

```
void MiApp_DiscardMessage(void);
```

Hàm này có nhiệm vụ là loại bỏ gói tin hiện tại trong bộ đệm nhận, mục đích của nó là giải phóng tài nguyên cho hệ thống để sẵn sàng nhận gói tin tiếp theo.

Ví dụ:

```
if( TRUE == MiApp_MessageAvailable() )
{
    // handle the received message in global variable RxMessage
    .....
    // discard the received message after processing
    MiApp_DiscardMessage();
}
```

3. XÂY DỰNG NỀN TẢNG PHẦN CỨNG, FIRMWARE VÀ PHẦN MỀM HOÀN CHỈNH GIÁM SÁT VÀ QUẢN LÝ TRÊN PC

3.1. Vấn đề bài toán đặt ra:

Hệ thống xếp hàng tự động là một hệ thống cho phép phục vụ khách hàng một cách tự động theo nguyên tắc ai đến trước được phục vụ trước

Hệ thống có n quầy dịch vụ. Mỗi dịch vụ sẽ được thực hiện xếp hàng tự động.

Bước 1: Khi mỗi khách hàng đến và muốn thực hiện một dịch vụ nào đó ở một quầy dịch vụ, khách hàng sẽ được cấp một phiếu trên đó có ghi số thứ tự được thực hiện với một dịch vụ ở một quầy tương ứng. (Sẽ có gắng thiết kế hệ thống cấp phiếu tự động khi khách hàng nhập thông tin yêu cầu vào máy cấp phiếu, tạm thời sẽ có người thực hiện và viết phiếu)

Bước 2: Khách hàng sẽ ngồi chờ ở khu vực chờ và đợi đến lượt mình. Khi đến lượt sẽ có thông báo trên bảng LED (bảng chính) (có thể tích hợp hệ thống phát tiếng nói thông báo cho khách hàng biết)

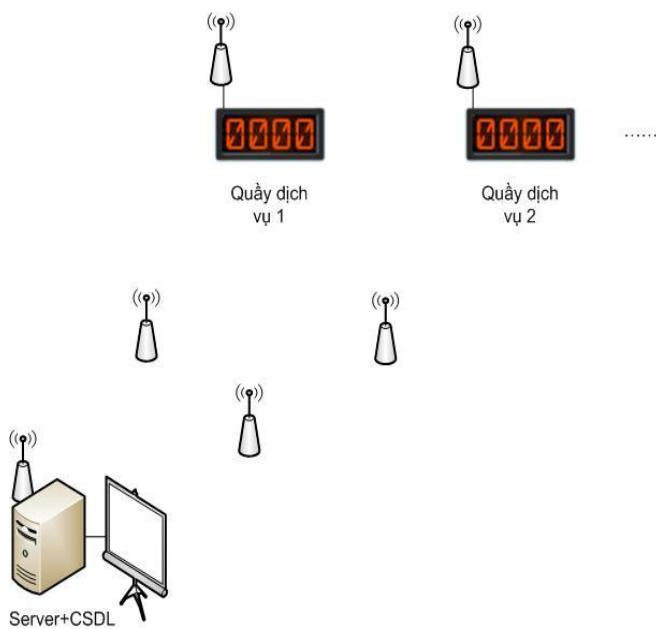
Bước 3: Khách hàng đến quầy giao dịch. Trên bảng của quầy sẽ hiển thị sẽ hiển thị số thứ tự của khách hàng đang được phục vụ.

Sau bước 3, Nếu khách hàng muốn thực hiện tiếp dịch vụ thì có thể quay lại bước 1 hoặc bước 2 nếu đã lấy phiếu.

Chú ý: sau 3 phút khách hàng không đến quầy dịch vụ thì số thứ tự tương ứng của khách hàng sẽ bị bỏ qua.

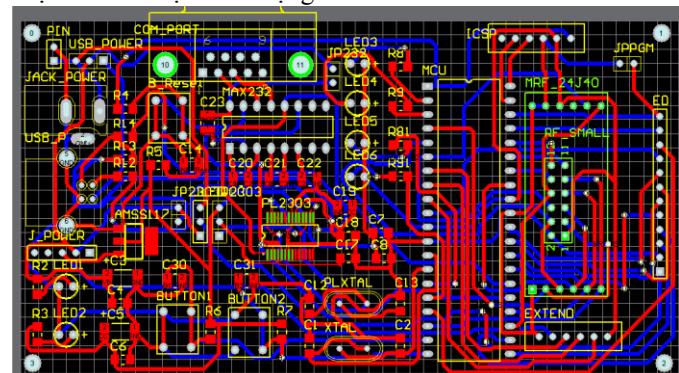
3.2. Phân tích thiết kế hệ thống về phần cứng:

Dưới đây là hình ảnh minh họa tổng quan của toàn bộ hệ thống xếp chỗ tự động. Hệ thống gồm các quầy phục vụ, máy tính điều khiển trung tâm kết nối cơ sở dữ liệu và các node mạng phục vụ truyền thông không dây.



Ta sẽ có các node mạng tương ứng với các quầy dịch vụ gọi là ClientNode và node mạng ứng với module điều khiển chính là ServerNode, ngoài ra còn có các node trung gian và một node mạng PAN coordinator đóng vai trò là bộ điều phối giúp thiết lập mạng MiWi để các node khác có thể kết nối tới

Mạch in của một node mạng:

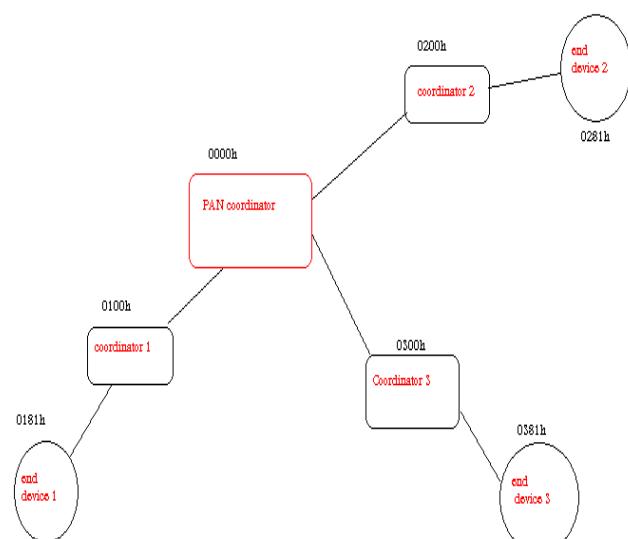


Mạch thực tế:



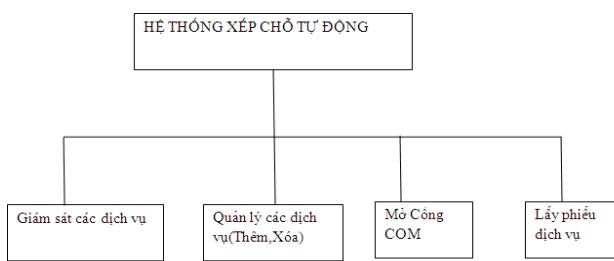
3.3. Thiết kế firmware với kịch bản:

ServerNode đóng vai trò là end device 1, ClientNode1 là end device 2, ClientNode2 là end device 3. Khi ClientNode 1 có yêu cầu, nó sẽ gửi gói tin yêu cầu tới ServerNode qua coordinator2 rồi qua coordinator 1 tới ServerNode. Tương tự khi ClientNode 2 có yêu cầu, nó sẽ gửi gói tin yêu cầu tới ServerNode qua coordinator3 rồi qua coordinator 1 tới ServerNode. Theo lý thuyết, khoảng cách giữa các node mạng từ 5-50m. Trong điều kiện phòng thí nghiệm ta đặt các node tương đối cách nhau khoảng 5 m để test thử hệ thống. Khoảng cách khi test rơi vào khoảng 5-10m thì hệ thống chạy ổn định.



3.4.Phần mềm điều khiển và giám sát trên PC:

-Biểu đồ phân cấp chức năng:



-Cơ sở dữ liệu quản lý dịch vụ:

Column Name	Data Type
ServiceName	nvarchar(50)
MacAddress	nvarchar(50)
MaxNum	int
CurrentNum	int

+**ServiceName:** là tên của dịch vụ được sử dụng.

+**MacAddress:** là địa chỉ LongAddress của quầy dịch vụ ClientNode được quản lý.

+**MaxNum:** là số thứ tự mà khách hàng cuối cùng sử dụng dịch vụ này vừa lấy.

+**CurrentNum:** là số thứ tự đang được phục vụ trên quầy dịch vụ này.

-Form chính:



-Form giám sát:



3.5.Kiểm nghiệm hệ thống:

Quá trình test thử hệ thống được tiến hành trong thời gian từ 13-04-2011 đến 26-04-2011.Qúa thử nghiệm hệ thống chạy khá ổn định trong điều kiện phòng thí nghiệm, thời gian đáp ứng của hệ thống khi có một yêu cầu từ các quầy dịch vụ là từ 2 đến 3 giây. Tuy nhiên khi đưa hệ thống ra môi trường bên ngoài phòng thí nghiệm thì đôi khi quá trình thiết lập mạng của các node trong mạng diễn ra khá lâu. Một nguyên nhân có thể

là do hiện tượng nhiễu. Nhìn chung hệ thống vẫn cần được cải tiến trước khi có thể đưa ra làm một sản phẩm thương mại.

4. KẾT LUẬN

Sau thời gian thực hiện đề tài này, em tự đánh giá mình đạt được các kết quả như:

- Tìm hiểu về mạng cảm biến không dây ,điều khiển không dây.
- Tìm hiểu về giao thức MiWi/IEEE802.15.4 và xây dựng thành công hàm lấy về địa chỉ Short Address khi biết địa chỉ Long Address từ mô tả phương pháp của Microchip.
- Thiết kế và xây dựng hệ thống các nodes mạng phục vụ cho hệ thống xếp chỗ tự động.

-Xây dựng phần mềm điều khiển và giám sát hệ thống xếp chỗ tự động trên PC.

-Định hướng sẽ xây dựng một hệ thống nền tảng (FrameWork) giúp phát triển ứng dụng trên hạ tầng mạng cảm biến không dây dễ dàng hơn. Đó là các hàm API cho phép người dùng phát triển ứng dụng dựa trên hạ tầng mạng cảm biến không dây trong suốt với nền tảng phần cứng tương tự như lập trình socket trên giao thức TCP/IP.

5. LỜI TRI ÂN

Em xin chân thành cảm ơn sự giúp đỡ tận tình của thầy Phạm Văn Thuận. Trong quá trình làm đề tài, thầy luôn tận tình chỉ bảo cho em những kinh nghiệm quý báu để em có thể hoàn thành đề tài này một cách tốt nhất. Em cũng xin cảm ơn các thầy cô trong Viện công nghệ thông tin và truyền thông đã truyền đạt cho em những kiến thức quý báu là nền tảng cho em có thể nghiên cứu đề tài khoa học này và các đề tài về sau. Cuối cùng em xin cảm ơn các bạn trong phòng CSLAB đã đóng góp cho em những ý kiến hay để em có thể hoàn thành đề tài này.

6. TÀI LIỆU THAM KHẢO

- [1] Microchip MiWi™ P2P Wireless Protocol.
- [2] MiWi™ Wireless Networking Protocol Stack.
- [3] Microchip Wireless (MiWi™) Application Programming Interface – MiApp.
- [4] Microchip Wireless (MiWi™) Application Programming Interface – MiMac.
- [5] Website: <http://www.microchip.com>.

Giải pháp camera giám sát giao thông trên nền tảng mạng 3G

Trịnh Thị Mây

Tóm tắt - Hiện nay tắc đường đã và đang là một hiện trạng đau đầu tại các thành phố lớn (Tp Hồ Chí Minh, Hà Nội..), và nhờ sự trợ giúp của kênh VOV giao thông thì tình trạng đó đã giảm đi một phần. VOV giao thông sử dụng hệ thống camera giao thông lắp tại điểm ùn tắc, nhưng ta biết chi phí cho hệ thống đó còn khá cao (chi phí tại mỗi điểm trên dưới 200 – 300 triệu với mạng Wireless LAN riêng

http://cafeviet.com.vn/index.php?option=com_atomentry&task=view&id=1228). Với chi phí cao như vậy nên số lượng camera được lắp tương đối hạng chế (hiện tại 36 chiếc tại TP Hồ Chí Minh, 100 chiếc tại Hà Nội) và hệ thống cũng cần sự hỗ trợ của nhiều đội VOV giao thông lưu động. Trước tình hình như vậy, tôi đã xây dựng một giải pháp camera giám sát giao thông trên nền tảng mạng 3G với sự kết hợp của Kit FriendlyArm Mini2440, module camera, module USB 3G để xây dựng lên camera client tại từng điểm ùn tắc

Từ khóa - FriendlyArm, linux, USB 3G driver, video streaming

1. GIỚI THIỆU

Hiện nay, trong giải pháp video streaming, đầu thu video thường là các IP camera với giá hơn 5 triệu VND trở lên, hoặc phải đi kèm với phần cứng kèm kẹp như máy tính để truyền tải dữ liệu về trung tâm qua mạng ADSL, hoặc kết hợp với Wifi rồi qua ADSL,... Vì vậy đề án nghiên cứu này đưa ra giải pháp video streaming gọn nhẹ và giảm bớt chi phí của một client camera và góp phần phổ cập ứng dụng video streaming cho mọi nơi với nhiều tiện ích của ứng dụng (giám sát, e – learning, quan sát tình trạng giao thông...).

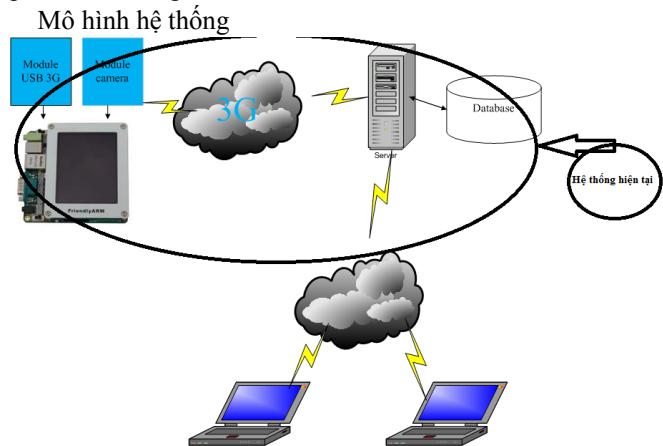
Giải pháp này tận dụng những đặc điểm tối ưu của kit FriendlyArm: có module camera, module USB, màn hình hiển thị hình ảnh thu được, cổng COM giúp ta giao tiếp với kit, có thể chạy hệ điều hành Linux, giá hiện tại của một kit là 3,3 triệu VND nếu ta chỉ cần một số module cần thiết như trên thì có thể giảm bớt chi phí của kit hơn nữa. Ở server, ta tận dụng cấu hình máy cao để hiển thị nhiều client camera đồng thời thao tác với các video được truyền về (xem offline, xem online, chụp hình

Trịnh Thị Mây, sinh viên lớp Kỹ thuật máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0976660727, e-mail: maytt.bk@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

ảnh thu về được từ client và lưu trữ lại). Và quan trọng nhất là giải pháp được xây dựng trên nền tảng mạng 3G nên rất di động, dễ lắp đặt, và có thể mang thiết bị tới bất cứ đâu để thực hiện truyền tải hình ảnh giao thông trực tiếp. Vì ta cũng biết mạng 3G mới bắt đầu phát triển ở thị trường Việt Nam, và khi nó xuất hiện đã góp phần làm tăng tính di động cho các ứng dụng mạng ở khắp mọi nơi.

Trong bài nghiên cứu này, tôi sẽ trình bày các nội dung sau: lý do chọn các thành phần trong giải pháp, xây dựng camera client, xây dựng camera server, kết quả đạt được hiện tại, một số hướng phát triển tương lai.



2. MỘT SỐ VẤN ĐỀ VÀ PHƯƠNG ÁN GIẢI QUYẾT

- Mạng truyền Video streaming có thể là : LAN, Wireless, 3G. Camera client trong vấn đề giám sát giao thông cần tính di động cao nhưng với mạng LAN và Wireless thì ta cần xây dựng mạng dây hoặc Wireless LAN phức tạp và chi phí cao. Chính vì vậy tôi đã chọn phương án là mạng 3G để tăng tính di động và không bị ảnh hưởng bởi hạ tầng mạng đi kèm và cũng giảm chi phí.

- Truyền Video streaming thì có hai xu hướng là: truyền theo luồng video, hoặc truyền theo từng ảnh về. Với việc truyền theo luồng video thì ta có một số công cụ hỗ trợ như vlc thi cần cấu hình thiết bị lớn, thêm vào đó một số thiết bị di động như iPhone, BlackBerry cũng không hỗ trợ việc truyền theo luồng video. Vì những lý do như vậy thì trong giải pháp được đưa ra trong nghiên cứu sử dụng phương pháp truyền theo từng ảnh về server để tiết kiệm bộ nhớ sử dụng.

3. XÂY DỰNG CAMERA CLIENT

3.1. Lý do chọn thiết bị cho giải pháp

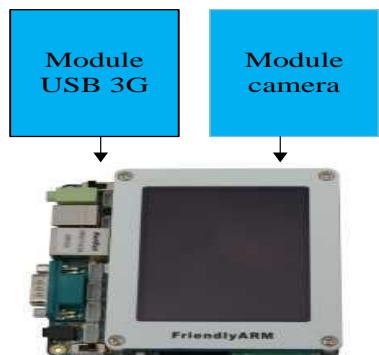


Kit FriendlyArrn cung cấp cho ta một số module, và đặc điểm cần thiết cho giải pháp:

- Module USB để ta tích hợp module camera, module USB 3G.
- Chạy hệ điều hành Linux từ version 2.6, với hệ điều hành này ta có thể dễ dàng phát triển với mã nguồn mở và xây dựng ứng dụng thu thập video streaming từ camera và truyền về server. Và với Linux version 2.6 có hỗ trợ driver của USB 3G, driver USB camera.
- Kit hỗ trợ nạp thêm một số thư viện hỗ trợ việc lập trình ứng dụng client camera: qt, opencv...
- Hỗ trợ chế độ khởi động NOR với giao diện nạp hệ điều hành một cách dễ dàng và thuận tiện.

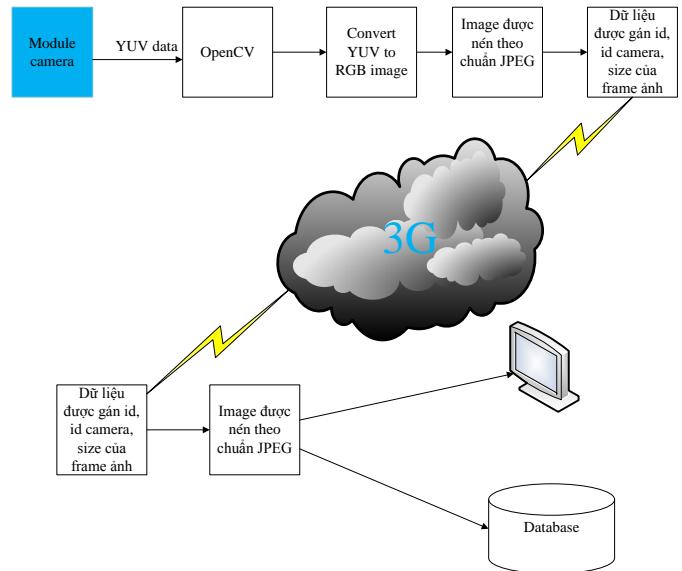
Với những tiện ích trên phù hợp với những yêu cầu đặt ra với giải pháp tôi đưa ra trong bài nghiên cứu, tôi đã quyết định sử dụng kit FriendlyArm. Và như vậy, sau khi rút gọn thiết bị với cấu hình tối thiểu như trên, một sản phẩm chuyên dụng dành cho camera giao thông có thể được sản xuất với giá thành nhỏ hơn 3,3 triệu VNĐ (giá hiện tại của kit FriendlyArm tại pnLab.vn).

3.2. Xây dựng camera client.



Camera client cần xây dựng 2 module:

- Module USB 3G.
- Module camera.



3.2.1 Mô đun USB 3G

Để xây dựng module USB 3G ta cần thực hiện các công việc:

- Kit nhận USB 3G với chức năng là modem.
- Thực hiện kết nối đến nhà cung cấp dịch vụ 3G để có thể vào mạng.

Đầu tiên để hệ điều hành nạp trên kit có nền tảng driver cho các tính năng trên thì ta cần phải dịch lại nhân Linux có tích hợp driver cho USB 3G bằng các khai báo VendorID và ProductID của USB 3G đó vào file option.c (nếu chưa có) của thuộc tính usbserial. Đặc điểm trên phục vụ cho việc chuyển mode storage sang serial cho USB 3G giúp cho Kit nhận USB 3G là usb modem (cổng ttyUSB) chứ không phải là usb storage sau này. Để phục vụ cho việc thiết lập kết nối với nhà cung cấp dịch vụ mạng 3G thì ta cần có driver cho giao thức PPP (point to point). Vì vậy ta phải chọn thuộc tính ppp trong driver network của nhân Linux. Sau khi chọn các thuộc tính driver thì ta sẽ dịch lại nhân và thu được một nhân hệ điều hành nạp cho kit.

Sau khi nạp hệ điều hành cho kit, để kit nhận USB 3G với chức năng là usb modem thì ta cần công cụ usb_modeswitch. Đây là một công cụ chuyển mode cho thiết bị đa chức năng cổng USB. Công cụ này thực hiện chức năng bằng cách lấy các tham số quan trọng từ một file cấu hình (trong trường hợp này là file cấu hình của từng USB 3G tương ứng) và thực hiện các công việc khởi tạo và trao đổi nhờ sự hỗ trợ của thư viện libusb. Mỗi một USB 3G thì có một file cấu hình tương ứng về vendorid, productid, message content tương ứng với USB 3G đó. Khi sử dụng usb_modeswitch load file cấu hình của USB 3G thì cổng usb sẽ được chuyển sang mode serial.

Cuối cùng ta thiết lập kết nối đến nhà cung cấp dịch vụ 3G. Để thiết lập, ta cũng cần có một công cụ hỗ trợ là pppd domain. PPP là một giao thức giúp truyền các gói được gửi qua mạng và đóng gói các gói tin đó để có thể được truyền qua mode serial dễ dàng. Khi ta có tài khoản internet với một ISP, thì ISP đó sử dụng PPP để giao tiếp với các tài khoản dialup. Vì vậy ta sử dụng công cụ pppd để có thể gửi các lệnh AT để thiết lập kết nối:

```

TIMEOUT 15
ABORT "DELAYED"
ABORT "BUSY"
ABORT "ERROR"
ABORT "NO DIALTONE"
ABORT "NO CARRIER"
TIMEOUT 15
"" ATZ
OK ATQ0V1E1S0=0&C1&D2+FCLASS=0
OK AT+CGDCONT=1,"IP","v-internet"
OK AT$QCPDPP=1,1,""
OK ATDT*99#
CONNECT

```

Ví dụ trên tương ứng với mạng Viettel, đối với mạng Mobiphone, Vinaphone thì ta sẽ thay v-internet thành m-wap, m3 – world. Sau khi nhận các lệnh AT thì ISP sẽ gửi thông tin Local IP, remote IP, DNS cho Kit trong mạng 3G. Sau đó ta cần cấu hình bảng định tuyến của Kit. Và từ đây ta có kết nối thành công với mạng.



3.2.1. Mô đun Camera

Như giải pháp đã trình bày ở trên, camera ta sẽ lấy theo từng ảnh để truyền về server để tiết kiệm bộ nhớ và phù hợp cho việc phát triển trên một số thiết bị di động sau này. Để lấy dữ liệu camera từ công video (cổng vào của dữ liệu camera), tôi sử dụng thư viện openCV để hỗ trợ việc lấy dữ liệu ảnh với các hàm : `cvCreateCameraCapture(int id)` – hàm tạo camera, `cvQueryFrame(CvCapture camera)` – hàm lấy các frame ảnh từ camera đã được tạo.

Ta biết mạng 3G có chất lượng không ổn định, tốc độ thấp và chi phí cao nên kích thước của một ảnh truyền đi và số lượng ảnh được truyền trên một giây cũng là vấn đề. Và để phù hợp với chi phí, tình trạng mạng và mục đích của giải pháp (giám sát giao thông tại các thời điểm tắc đường thì ta sẽ lấy dữ liệu về 8 tiếng trong một ngày) tôi đã chọn thông số sau trong giải pháp: thông số dưới đây là tính theo giá mạng bình thường của Mobiphone, mỗi nhà cung cấp dịch vụ có giá riêng và gói cước riêng do đó chi phí có thể giảm đi so với tính toán dưới đây.

Kích thước ảnh (200 x 200)	Ảnh / 1 s	Số tiền / 1 tháng (8h/24h) (VNĐ/tháng)
7kB (JPEG)	5	$\frac{7 * 5 * 3600 * 8}{1024} * 65 = 1,919,531$

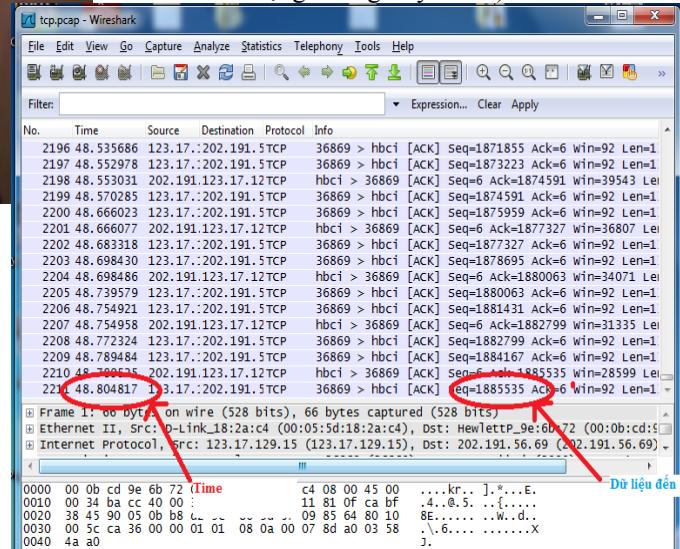
4. XÂY DỰNG CAMERA SERVER.

Yêu cầu của camera server là:

- Quản lý được nhiều client đến.
- Có thể chọn client để hiển thị.
- Cho phép xem online hoặc offline.
- Cho phép chụp ảnh từ các camera client.

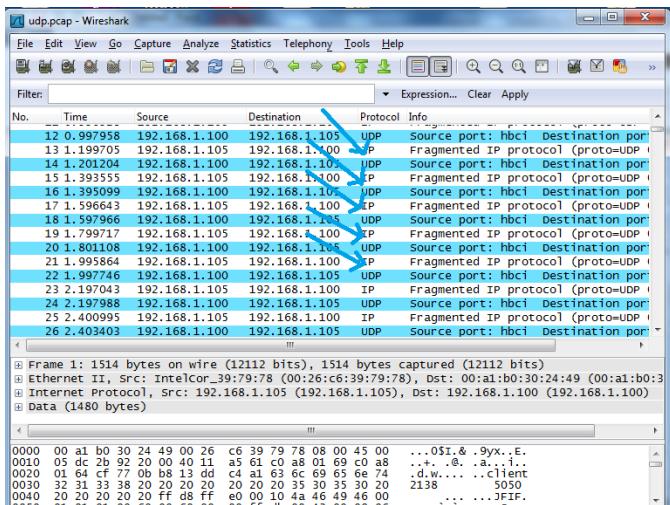
Ở đây vấn đề truyền nhận video streaming thì ta có hai phương án sử dụng: dùng giao thức UDP và TCP. Trong mô hình của ta là client gửi video về cho server, server phải quản lý được các luồng client riêng rẽ nên đối với TCP là kết nối có thiết lập thì ta có thể dễ dàng quản lý nhiều client riêng rẽ, còn với UDP thì là giao thức không có thiết lập kênh kết nối nên việc quản lý nhiều client ở phía server là không được linh động. Hiện nay một số service tổ chức theo mô hình như trong giải pháp nêu ra và chúng cũng sử dụng giao thức TCP để truyền nhận video giữa client và server.

Về vấn đề tốc độ thì khi test theo TCP và UDP thì tôi cũng thấy chất lượng không chênh lệch quá nhiều (với video 5 hình/s thì khi truyền theo TCP thì ta được ~4,5 ảnh trong khi truyền UDP là 5 ảnh với tình trạng đường truyền tốt.)



Với trường hợp TCP:

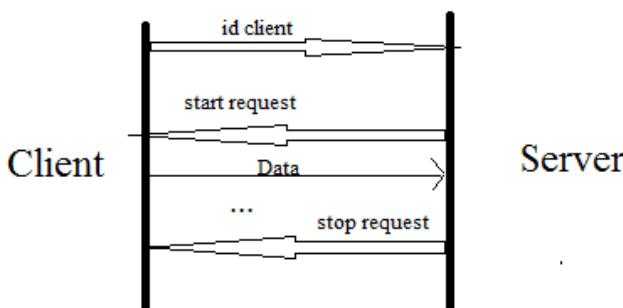
$$\text{số ảnh} = \frac{2211 + 1368}{2 * 7000 + 49} = 4.49$$



Khi sử dụng giao thức UDP thì ta nhận được 5 gói.

Vì vậy ta dùng giao thức TCP để truyền video trong giải pháp nêu trong bài nghiên cứu là có thể phù hợp hơn UDP.

Từ những khảo sát trên, tôi đã chọn sử dụng giao thức TCP để truyền video streaming trong giải pháp này. Truyền giữa client và server theo một số quy tắc sau:



Một số trường hợp client tự động ngắt kết nối thì Server xác định được kết nối được ngắt từ client.

Để lưu trữ video truyền về thì tôi dùng cơ sở dữ liệu Microsoft SQL server 2005 để lưu trữ theo: một client có nhiều video, video tại các thời điểm khác nhau. Video được lưu với thời điểm bắt đầu lưu của từng thời điểm do vậy ta có thể dễ dàng tìm được video tại thời điểm mà ta muốn xem.

Hỗ trợ cho việc xây dựng giao diện đẹp mắt tôi có sử dụng thư viện DotNetBar.

5. KẾT QUẢ ĐẠT ĐƯỢC VÀ ĐÁNH GIÁ

Kết quả:

- Camera client: đã thu nhận được video từ camera và truyền về server qua mạng 3G theo từng ảnh. Với kích thước chương trình chạy là 600kB. Client với phần cứng: CPU 400MHz, RAM 64MB SDRAM thì khi chạy chương trình tiêu tốn 40% bộ nhớ và 70% CPU. Khi chạy hai chương trình một lúc trên cùng kit thì CPU và ngăn nhớ được phân đều và vẫn thực hiện được các chức năng như bình thường.
- Camera server: đã quản lý được nhiều client (thực hiện với 4 client 1 lúc), có thể xem offline, online, chụp ảnh của từng client.

Hình ảnh phía client và server:



Dánh giá: Đã thực hiện được các yêu cầu đề ra với giải pháp, khi truyền trong tình trạng mạng kém thì xảy ra hiện tượng video truyền về bị trễ không những vẫn giữ được sự đảm bảo của thông tin truyền về.

6. MỘT SỐ HƯỚNG PHÁT TRIỂN

Với những đặc điểm hữu ích của Kit FriendlyArm và thư viện openCV (hỗ trợ việc xử lý ảnh) tôi xin đưa ra một số hướng phát triển thêm với nghiên cứu này:

- Thu viện xử lý ảnh openCV và module cmos camera (thu được ảnh với chất lượng cao) ta có thể phát triển ứng dụng nhận dạng đối với ảnh thu được từ cmos camera.
- Với module âm thanh của kit FriendlyArm thì ta có thể phát triển thêm truyền âm thanh về server với ứng dụng video conference thông qua mạng 3G. Với tiện ích kit nhỏ gọn thì làm cho ứng dụng này được phổ biến rộng rãi mọi nơi mọi lúc.
- Phát triển hệ thống trên nền tảng web – based để mọi người có thể truy cập mọi lúc mọi nơi thông qua website được quản lý bởi server.

7. LỜI TRI ÂN

Để thực hiện được bài báo này, tôi xin cảm ơn TS Nguyễn Kim Khánh, thầy Nguyễn Đức Tiến, thầy Trần Tuấn Vinh đã giúp đỡ em trong quá trình thực hiện nghiên cứu.

8. TÀI LIỆU THAM KHẢO

- [1] Prentice Hall Open Source Software Development Series,
"C gui programming with qt4 2nd edition"
- [2] Micro2440 Chinese User Manual - 20100609
- [3] Jonathan Corbet, Alessandro Rubini, and Greg Kroah-Hartman "Linux device driver third edition"
- [4] Gary Bradski and Adrian Kaehler, "Learning OpenCV"
- [5] <http://ilive.vn>
- [6] http://www.draisberghof.de/usb_modeswitch

Hệ thống định vị - hỗ trợ quản lý học sinh tiểu học trên nền tảng GPS-GSM/GPRS

Đinh Thanh Tùng, Đặng Thanh Huyền

Tóm tắt—Hệ thống định vị - hỗ trợ quản lý học sinh tiểu học trên nền tảng GPS-GSM/GPRS là **hệ thống tổng thể bao gồm phần cứng và phần mềm nhằm xác định vị trí, hỗ trợ quản lý các đối tượng di động là học sinh tiểu học.** Phần cứng là thiết bị thu nhận tín hiệu GPS, xử lý, và truyền dữ liệu thông qua GSM/GPRS lên server. Phần mềm có vai trò giám sát, quản lý và xem lại lịch sử di động của đối tượng. Bên cạnh các kỹ thuật xác định vị trí dựa vào tín hiệu GPS, nhóm đã triển khai nghiên cứu và áp dụng thành công kỹ thuật định vị trong trường hợp mất tín hiệu GPS dựa vào vị trí các trạm BTS để tăng tính chính xác và ổn định của hệ thống (kỹ thuật này được áp dụng khi đối tượng di chuyển vào các khu vực nhà cao tầng, di chuyển trong nhà hoặc các khu vực tín hiệu GPS bị nhiễu).

Từ khóa— Định vị GPS, Google map, Cell ID.

1. GIỚI THIỆU

Hệ thống định vị GPS (Global Positioning System) là hệ thống xác định vị trí dựa trên vị trí các vệ tinh nhân tạo do bộ quốc phòng Mỹ quản lý và phát triển. Ngày nay GPS có rất nhiều ứng dụng trong cuộc sống ở nhiều lĩnh vực khác nhau : hỗ trợ quản lý phương tiện, con người, trong trắc địa, khảo sát môi trường, xây dựng các hệ thống dẫn đường, không người lái, các thiết bị trinh sát trong quân sự quốc phòng. Năm 2010, Trung Quốc đã triển khai phát 20.000 bộ thiết bị định vị cho học sinh tiểu học và trung học. Việt Nam cũng đã có quyết định các phương tiện vận tải công cộng và đường dài bắt buộc có trang bị thiết bị định vị từ tháng 7/2011. Với mong muốn áp dụng công nghệ GPS vào thực tế ở Việt Nam, nhóm nghiên cứu đã chọn đề tài xây dựng và triển khai hệ thống định vị - hỗ trợ quản lý học sinh tiểu học trên nền tảng GPS – GSM/GPRS và các công nghệ định vị không phụ thuộc vào GPS.

2. ĐẶT VẤN ĐỀ

Hiện nay, giải pháp định vị và hỗ trợ giám sát quản lý được thế giới ứng dụng rộng rãi. Ở Việt Nam một số hệ thống định vị bắt

Đinh Thanh Tùng, sinh viên lớp Kỹ thuật máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0977 376 197, e-mail: tungdt.bk@gmail.com).

Đặng Thanh Huyền, sinh viên lớp Kỹ thuật máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (e-mail: dangthanhhuyen88@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

đầu được triển khai ở các doanh nghiệp vận tải như Bình Anh, VietMap.... Tuy nhiên mục tiêu hướng tới cũng những hệ thống này chủ yếu là người dùng khối cơ quan doanh nghiệp hơn là khối người dùng cá nhân, do đó khi áp dụng giải pháp này vào cho hệ thống định vị hỗ trợ quản lý học sinh tiểu học đã gặp khá nhiều khó khăn :

- Giá thành thiết bị còn khá cao, độ ổn định của thiết bị chưa đảm bảo.
- Phụ thuộc vào bản đồ số chung của hệ thống, không hỗ trợ người dùng cá nhân truy cập vào hệ thống.
- Hiệu năng của thiết bị chưa được đảm bảo, chưa khắc phục được trường hợp đối tượng di chuyển vào khu vực nhà cao tầng, trong nhà, khu vực mất sóng GPS.

Ngoài ra người dùng có thể sử dụng một giải pháp khác là sử dụng các thiết bị định vị cá nhân cầm tay của các hãng như Garmin, Trimble.... Việc sử dụng các thiết bị định vị cá nhân này đã khắc phục được một số hạn chế của giải pháp trên :

- Thiết bị hoạt động khá ổn định, độ chính xác tương đối cao.
- Tích hợp sẵn hệ thống bản đồ số trên thiết bị, không phụ thuộc vào bản đồ số chung của hệ thống.

Tuy nhiên giải pháp trên vẫn còn một số hạn chế :

- Giá thành sản phẩm còn cao do sản phẩm nhập từ nước ngoài nguyên chiếc.
- Việc sử dụng còn phức tạp, không phù hợp với đối tượng học sinh.
- Không hỗ trợ chức năng giám sát, cảnh báo.
- Chưa khắc phục được trường hợp đối tượng di chuyển vào khu vực nhà cao tầng, trong nhà, khu vực mất sóng GPS.

Qua quá trình tìm hiểu và và đánh giá thực tế, nhóm nghiên cứu đã xây dựng giải pháp và triển khai hệ thống khắc phục được các hạn chế trên

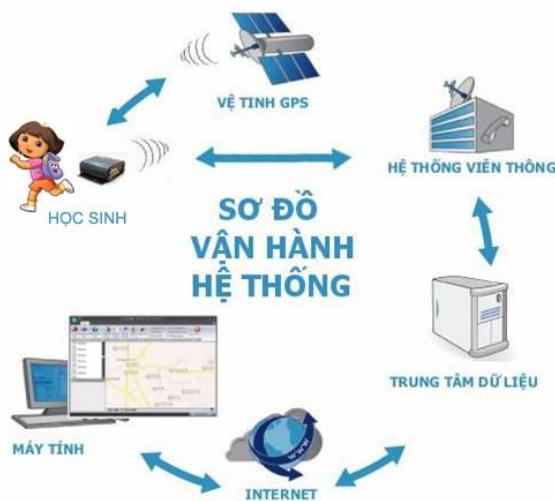
- Xây dựng module phần cứng có chức năng thu nhận tín hiệu GPS và truyền nhận dữ liệu qua GPRS giá thành chấp nhận được với độ ổn định và chính xác tương đối cao.
- Tích hợp tối đa các giải pháp thu GPS để đa dạng hóa với thực tế Việt Nam : hỗ trợ thu GPS qua các giao tiếp Bluetooth với dòng Trimble, qua giao tiếp USB với dòng Garmin và lấy dữ liệu GPS trực tiếp từ các IC chuyên dụng.
- Xây dựng hệ thống giám sát và hỗ trợ quản lý. Hệ thống hỗ trợ cảnh báo và giám sát qua tin nhắn, giúp

đơn giản hóa tối đa cho người dùng. Khi học sinh di chuyển vào các khu vực trong vùng đặt giới hạn, tin nhắn cảnh báo sẽ tự động được gửi vào các số điện thoại được cài đặt sẵn.

- Tích hợp xử lý trường hợp mất sóng GPS. Khi đối tượng di chuyển vào khu vực nhà cao tầng, hay di chuyển trong nhà, sóng GPS yếu, hệ thống sẽ tự động chuyển sang chế độ định vị dựa trên Cell-ID của các trạm BTS. Hệ thống sẽ khoanh vùng và xác định bán kính sai số tại điểm mất sóng GPS.

3. MÔ HÌNH VÀ GIẢI PHÁP THỰC HIỆN

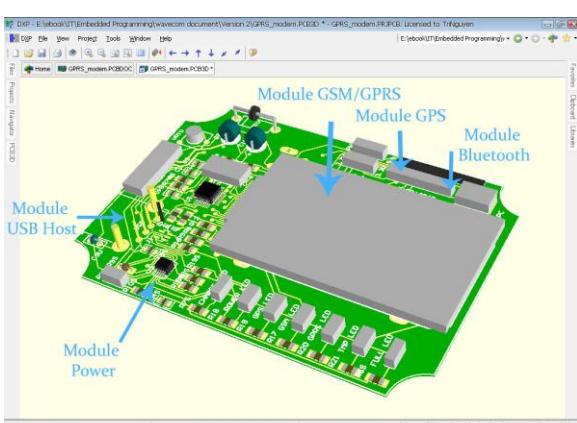
3.1 Mô hình tổng quan của hệ thống



Hình 1. Mô hình tổng quan hệ thống

Hệ thống bao gồm module phần cứng là hộp đen được đựng trong balo của học sinh. Module này có chức năng thu tín hiệu GPS từ vệ tinh và truyền dữ liệu qua GPRS về server. Trên Server có phần mềm xử lý dữ liệu và đưa vào cơ sở dữ liệu. Sau cùng, người dùng có thể sử dụng phần mềm trên desktop hoặc qua web browser để thiết lập, giám sát và quản lý đối tượng học sinh.

3.2 Thiết kế phần cứng



Hình 2. Thiết kế phần cứng

Thiết bị phần cứng bao gồm các module chính như : module

truyền dữ liệu GSM/GPRS - truyền nhận dữ liệu không dây qua GPRS về Server, module truyền nhận dữ liệu qua Bluetooth – hỗ trợ giao tiếp với các thiết bị - hệ nhúng khác, module USB Host Chuẩn giao tiếp này giúp thiết bị nhúng detect và truyền nhận với hầu hết các thiết bị USB), module GPS – thu tín hiệu GPS...

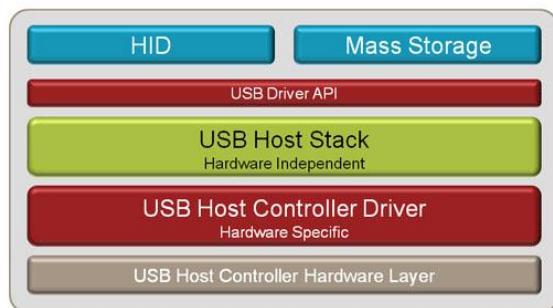
3.2.1 Module truyền dữ liệu GSM/GPRS

Qua khảo sát về yêu cầu tính năng và độ nhỏ gọn của kích thước của sản phẩm, nhóm nghiên cứu đã chọn giải pháp truyền nhận GSM/GPRS là module Q24 của Wavecom hỗ trợ dải tần EGSM/GPRS 850/900/1800/1900 MHz. Kích thước 58.3*32.2*3.9mm. Trọng lượng 12g. Lập trình trên nền tảng hệ điều hành OpenAT OS. Q24 hỗ trợ các giao tiếp : SPI, Keyboard, UART, SIM, GPIO, ADC, DAC...

3.2.2 Module Bluetooth

Sử dụng IC của Wavecom. Module Bluetooth có khả năng tìm kiếm và nhận dạng hầu hết các thiết bị Bluetooth. Chuẩn giao tiếp truyền dữ liệu qua giao tiếp Bluetooth ngày càng được phổ biến và có nhiều thiết bị hỗ trợ. Việc tích hợp module Bluetooth sẽ giúp cho thiết bị kết nối và truyền dữ liệu với các thiết bị khác linh hoạt hơn.

3.2.3 Module USB Host



Hình 3. Kiến trúc USB Host

Sử dụng IC Max3421 của Maxim, module USB Host được thiết kế và lập trình trên hệ điều hành OpenAT với chức năng nhận dạng các thiết bị USB và truyền nhận dữ liệu GPS từ các thiết bị USB như Garmin, eXplorit,... Chuẩn giao tiếp USB Host là chuẩn giao tiếp tương đối khó và chưa được thực hiện nhiều ở Việt Nam. Khác với việc lập trình cho hệ nhúng thành một thiết bị USB (USB device), người lập trình phải tự viết các driver riêng biệt để nhận biết và thiết lập giao thức truyền dữ liệu cho từng thiết bị.

3.2.4 Module thu tín hiệu GPS

Module GPS sử dụng dữ liệu GPS trực tiếp từ module UB93 của Holux. Định dạng dữ liệu xuất ra tuân theo chuẩn NMEA với các bản tin GPGGA, GPRMC,... cho biết thời gian, tọa độ và vận tốc của thiết bị. Hiện nay có khá nhiều IC có chức năng thu GPS, việc nhóm nghiên cứu lựa chọn IC UB93 đảm bảo về hiệu năng và giá thành của sản phẩm.

3.3 Thiết kế phần mềm

Với mục đích xây dựng một hệ thống tổng thể nhằm xác định vị trí, từ đó hỗ trợ vào việc quản lý, giám sát đối tượng trẻ nhỏ

(mà cụ thể là học sinh tiểu học), hệ thống cần đưa ra được một giải pháp phần mềm có các đặc điểm:

- Tính thời gian thực
- Tính chính xác
- Tính trực quan

Với tính thời gian thực, hệ thống cần xây dựng một chương trình có đặc điểm: chạy ổn định 24/24h, luôn luôn lắng nghe và tiếp nhận tất cả các kết nối gửi đến công kết nối của chương trình. Sau khi chấp nhận kết nối đến từ các client, server và client có thể gửi – nhận dữ liệu cho nhau qua socket.

Với tính chính xác, hệ thống xây dựng giao thức chung giữa client và server nhằm đảm bảo server của hệ thống chỉ xử lý dữ liệu từ các client của hệ thống. Giao thức chung giữa client và server là định dạng về gói tin được truyền nhận giữa client và server, bao gồm có quy định về phần mở đầu và kết thúc gói tin, các trường và độ dài các trường trong gói tin.

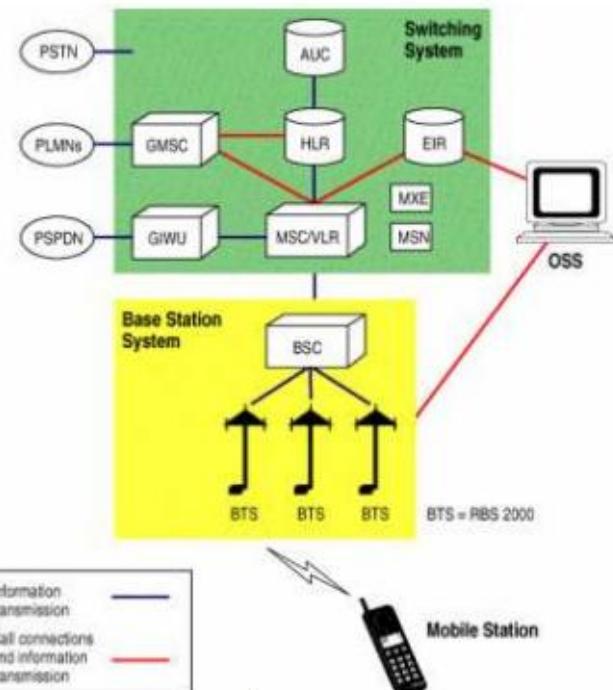
Với tính trực quan, hệ thống muốn đưa đến người dùng hình ảnh cụ thể về vị trí hiện tại cũng như vết lịch sử của người dùng (chính là các client). Để làm được, giải pháp nhóm đưa ra là xây dựng một module phần mềm dạng Web-based, trên đó tích hợp bản đồ GoogleMap. Bản đồ GoogleMap (<http://maps.google.com>) được cung cấp miễn phí cho mọi người dùng, thêm vào đó bằng cách sử dụng các hàm API GoogleMap cung cấp, sử dụng cơ sở dữ liệu từ phía các client gửi về, phần mềm có thể hiển thị được vị trí hiện tại của người dùng, có thể giúp người dùng xem lại lịch sử đường đi trong quá khứ giúp cho việc giám sát và quản lý các đối tượng di động.

Điểm khác biệt trong hệ thống này so với các hệ thống đã tồn tại là việc định vị trong trường hợp GPS. Như đã biết thì hệ thống định vị toàn cầu GPS hoạt động rất tốt trong môi trường không bị che chắn. Tuy nhiên, hệ thống này tỏ ra làm việc kém hiệu quả và đôi khi là không làm việc trong các môi trường bị che khuất như tòa nhà cao tầng, trường học, bệnh viện, rừng rậm, hầm ngầm... Trong trường hợp này, thay vì sử dụng tín hiệu GPS, hệ thống đã sử dụng thu nhận và xử lý thông tin về các trạm thu phát sóng (các trạm BTS) nhằm định vị đối tượng. Đặc biệt là trong điều kiện hiện tại, khi mà số lượng các nhà cao tầng ngày càng tăng, đồng thời với số lượng thuê bao di động cũng ngày càng tăng, làm cho số lượng trạm BTS tăng nhanh, giải pháp định vị bằng GSM ngày càng khả thi và có tác dụng thiết thực. Đây chính là điểm khác biệt của hệ thống so với các hệ thống đã tồn tại.

3.3 Giải pháp định vị bằng GSM

3.3.1 Tổng quan về mạng GSM và phương pháp định vị trong mạng GSM

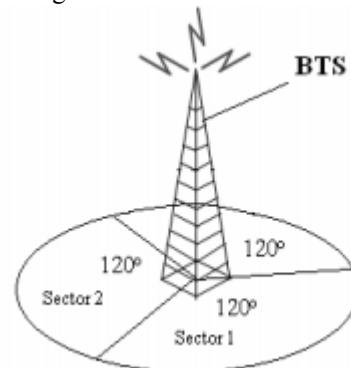
GSM – Global System for Mobile communications – là một trong những công nghệ về mạng di động phổ biến nhất trên toàn thế giới. Cho đến nay, ở Việt Nam có khoảng hơn 150 triệu thuê bao di động.



Hình 4. Cấu trúc mạng GSM

Một mạng GSM bao gồm 3 thành phần chính:

- **Hệ thống chuyển mạch (MSS – Mobile Switching System):** thực hiện chuyển mạch cuộc gọi giữa những người sử dụng điện thoại di động, và giữa di động với thuê bao mạng cố định. Quản lý các thông tin về thuê bao, xác thực, bảo mật...
- **Trạm thu phát gốc (BSS – Base Station System):** xử lý công việc liên quan đến truyền phát sóng radio. Gồm có các BSC, BTS. BTS – Base Transceiver Station – Trạm truyền phát sóng radio phủ sóng một vùng hình tròn (hoặc lục giác đều) với bán kính phủ sóng khoảng 300 – 400m trong khu vực thành phố và vài km trong khu vực nông thôn.



Hình 5. Trạm BTS

Vùng phủ sóng của BTS lại được chia thành 3 sector đều nhau một góc là 120°. Vùng phủ sóng của mạng là kết hợp vùng phủ sóng của nhiều BTS. Vài BTS lại được quản lý bởi một BSC – Base Station Controller – trạm gốc điều khiển.

- **Trạm di động (MS – Mobile Station):** Gồm một thiết

bị di động và một module xác nhận thuê bao (SIM – Subscriber Identity Module)

Để quản lý các thuê bao di động, mạng GSM sử dụng hai đơn vị cơ sở dữ liệu: HLR và VLR. HLR – Home Location Register – bộ ghi địa chỉ thường trú; đơn vị cơ sở dữ liệu này chứa một phần thông tin được cập nhật thường xuyên về vị trí hiện thời của MS (MS hiện đang có mặt tại vùng phục vụ của MSC nào) cho phép các cuộc gọi tới MS được kết nối tới MSC mà tại đó MS đang bị gọi hiện diện; ngoài ra HLR còn chứa các thông tin về thuê bao như các dịch vụ phụ và các thông số nhận thức liên quan tới quá trình nhận thực thuê bao. VLR – Visitor Location Register – bộ ghi địa chỉ tạm trú có chức năng theo dõi mọi MS đang có trong vùng MSC của nó kể cả các MS là thuê bao của các công ty điện thoại di động khác song hoạt động ngoài vùng HLR của chúng. Việc quản lý di động của các MS trong mạng được thực hiện thông qua quá trình cập nhật vị trí (Location Updating – LU) của MS với sự tham gia của các đơn vị cơ sở dữ liệu là HLR và VLR. MS thường xuyên thông báo cho Mạng di động số điện thoại công cộng (PLMN) về vị trí của mình bằng cách thường xuyên cập nhật vị trí thông qua MSC/VLR để đổi mới nội dung của HLR. Để hỗ trợ quá trình này, các PLMN được chia thành các LA (Location Area), mỗi LA này bao gồm một số cell và được đặc trưng bằng mã LA duy nhất trên toàn thế giới LAI (Location Area Identity). Một giá trị LAI có 3 thành phần:

- MCC – Mobile Country Code – số thập phân có 3 chữ số.
- MNC – Mobile Network Code – số thập phân có 2 hoặc 3 chữ số.
- LAC – Location Area Code

Hai giá trị MCC và MNC được dùng để xác định một mạng PLMN trong một nước. Giá trị LAC được dùng để xác định 1 trong 65536 vùng trong một mạng PLMN.

Số này được phát quảng bá thường xuyên tới mọi MS. Các MS có thể di chuyển tự do trong LA mà không cần cập nhật vị trí. Chỉ khi nào MS nhận thấy có sự thay đổi về LAC nó mới phát ra một yêu cầu báo cập nhật vị trí. Vậy là bằng cách truy nhập vào cơ sở dữ liệu VLR của nhà mạng có thể biết được vị trí của thuê bao đang thuộc LA nào, mở ra hướng có thể định vị thuê bao trong mạng GSM.

Có rất nhiều kỹ thuật xác định vị trí thuê bao trong bao GSM. Trên cơ sở module phần cứng đã xây dựng cùng với điều kiện kỹ thuật và cơ sở hạ tầng hiện nay ở nước ta, kỹ thuật dễ dàng triển khai và ít tốn kém nhất là kỹ thuật Cell-ID.

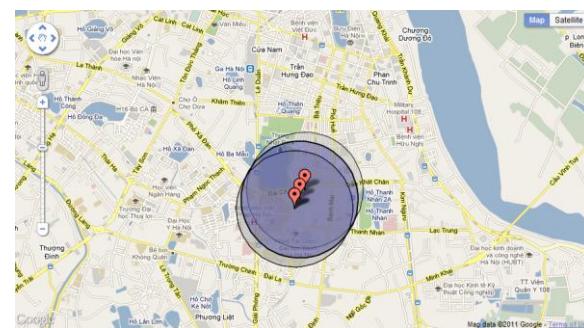
3.3.2 Kỹ thuật Cell-ID định vị trong mạng GSM

Cell-ID còn được gọi là CGI (Cell Global Identity) được sử dụng trong mạng GSM, GPRS và WCDMA. Đây là cách xác định vị trí thuê bao đơn giản nhất. Phương pháp này yêu cầu xác định vị trí của BTS mà MS đang trực thuộc. Tuy nhiên vì MS có thể ở bất kỳ vị trí nào trong Cell nên độ chính xác của phương pháp này phụ thuộc vào kích cỡ Cell. Nếu Cell ở vùng đô thị, mật độ đông thì kích cỡ bé nên độ chính xác cao hơn. Nếu MS ở vùng nông thôn, mật độ BTS thấp thì độ chính xác có thể lên đến vài

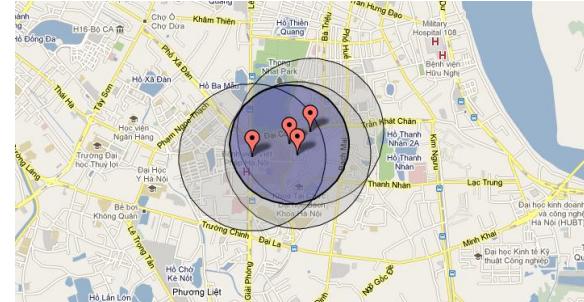
chục km.

3.3.3 Giải quyết bài toán

Trong kỹ thuật định vị Cell-ID, mỗi trạm BTS được xác định bởi 4 giá trị: MCC, MNC, LAC, CID. Trong trường hợp mất tín hiệu GPS, bằng cách cầu hình module phần cứng ta có thể thu được các giá trị trên của nhiều nhất 7 trạm BTS mà MS nằm trong vùng phủ sóng của các BTS đó. Dựa trên các giá trị này, ta gửi 1 request đến địa chỉ của Google GearAPI, sẽ thu được vị trí – tọa độ (kinh độ, vĩ độ) và bán kính phủ sóng của trạm BTS. Bài toán định vị từ đây được đưa về bài toán: Tìm đường tròn nhỏ nhất ngoại tiếp phần giao nhau của (nhiều nhất) ba đường tròn (lựa chọn trong số tối đa 7 Cell-ID thu nhận được). Xác định đường đường tròn này đồng nghĩa ta đã xác định được khu vực của đối tượng.



Hình 6. Trường hợp có thông tin về 2 trạm BTS



Hình 7. Trường hợp có thông tin về 3 trạm BTS

4. KẾT QUẢ ĐẠT ĐƯỢC

4.1 Kết quả

Trong quá trình nghiên cứu và triển khai đề tài, nhóm nghiên cứu đã xây dựng và thử nghiệm thành công hệ thống có chức năng :

- Thu nhận tín hiệu GPS
- Truyền dữ liệu lên Server qua GPRS
- Hỗ trợ giao tiếp với các thiết bị thu GPS thông dụng trên thị trường : Garmin, Trimble...
- Hỗ trợ lấy vị trí của học sinh qua tin nhắn của bố mẹ
- Tự động gửi tin nhắn cảnh báo về số điện thoại của bố mẹ trong trường hợp đi vào các vùng đặt cảnh báo.
- Hỗ trợ lấy thông tin trạm BTS để xác định vị trí tương đối trong trường hợp mất tín hiệu GPS



Hình 8. Sản phẩm phần cứng



Hình 9. Quan sát đối tượng ở thời điểm hiện tại



Hình 10. Đưa ra vị trí tương đối của đối tượng trong trường hợp mất GPS

4.2 Đánh giá

Qua quá trình triển khai và thử nghiệm hệ thống đã đạt được một số kết quả sau :

- Module phần cứng chạy khá ổn định, có khả năng tự reset và thiết lập lại kết nối trong trường hợp kết nối tới Server bị lỗi.
- Hệ thống phần mềm chạy trên Server ổn định, xử lý được nhiều Client kết nối.
- Chức năng định vị, tự động nhắn tin cảnh báo chạy ổn định.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong khi thực hiện đề tài này, nhóm đã xây dựng và triển khai hệ thống định vị - hỗ trợ quản lý học sinh tiêu học trên máy chủ <http://202.191.56.69> và đánh giá hệ thống khá ổn định. Nhóm sẽ tiếp tục nghiên cứu và phát triển các tính năng của hệ thống như : hỗ trợ quay số và gọi trực tiếp trong trường hợp khẩn cấp, Mở rộng và nâng cấp phần mềm quản lý đáp ứng được số lượng lớn người dùng.

6. LỜI TRI ÂN

Nhóm tác giả xin được gửi lời cảm ơn sâu sắc đến thầy Ths. Trần Tuấn Vinh – Giảng viên bộ môn Kỹ Thuật Máy Tính, viện

Công Nghệ Thông Tin và Truyền Thông, đại học Bách Khoa Hà Nội và thầy Ts. Nguyễn Kim Khánh – viện phó viện Công Nghệ Thông Tin và Truyền Thông, trường đại học Bách Khoa Hà Nội cùng các thầy giáo phòng Thí nghiệm Hệ Thống Máy Tính, viện Công Nghệ Thông Tin, trường đại học Bách Khoa Hà Nội đã hết lòng giúp đỡ, hướng dẫn và chỉ dạy tận tình trong quá trình thực hiện công trình.

7. TÀI LIỆU THAM KHẢO

- [1] Stephen Hinch, Stephen W. Hinch, "Outdoor Navigation With GPS",
- [2] Lawrence Letham, "GPS Made Easy : Using Global Positioning Systems," .
- [3] Bernhard Hofmann-Wellenhof, Elmar Wasle, Herbert Lichtenegger. "GNSS: Global Navigation Satellite Systems"

Hệ thống định vị qua bước chân người trong môi trường không có GPS với chi phí thấp.

Nguyễn Đình Thuận

Tóm tắt - Hệ thống định vị toàn cầu GPS hoạt động rất tốt trong môi trường không bị che chắn. Tuy nhiên hệ thống này tố ra làm việc kém hiệu quả và đôi khi là không làm việc trong các môi trường bị che khuất như tòa nhà cao tầng, trường học, bệnh viện, rừng rậm, hầm ngầm... Trong trường hợp xảy ra sự cố, thảm họa thì việc theo dõi các lực lượng cứu hỏa, cứu hộ... khi họ vào trong khu vực này là rất cần thiết. Xuất phát từ ý tưởng đó, bài báo đã trình bày phương pháp sử dụng sử dụng các thiết bị có chi phí thấp vào việc phát hiện vị trí của con người trong môi trường bị che chắn này.

Từ khóa—Pedestrian navigation, inertial navigation system, IMU, MARG, foot step navigation, PDR.

1. GIỚI THIỆU

Để đảm bảo sự linh hoạt cần thiết cho lực lượng cứu hộ khi thực hiện nhiệm vụ di chuyển trong các hệ thống tòa nhà cao tầng, trường học, bệnh viện thường ít khi sử dụng các phương tiện hỗ trợ, do đó 1 hệ thống định vị hỗ trợ họ phải có đặc điểm là thật nhỏ, gọn nhẹ và tương đối chính xác là điều bắt buộc.

Hệ thống định vị quán tính (INS) sử dụng cảm biến gia tốc(accelerometer), vận tốc góc (gyroscope) và từ tính (magnetometer) đã ra đời từ lâu nhưng nhược điểm của chúng là kích thước và sai số lớn. Gần đây nhờ sự tiến bộ của ngành vi điện tử nên kích thước các cảm biến này là tương đối nhỏ tuy nhiên sai số của chúng cũng vẫn còn rất lớn.

Sử dụng các sensor này cùng với module truyền nhận dữ liệu từ xa như Bluetooth hoặc RF sẽ giúp kích thước của hệ thống nhỏ gọn và giá thành chấp nhận được. Kết hợp với việc sử dụng các thuật toán lọc nhiễu hợp lý hệ thống sẽ cho kết quả tương đối chính xác.

2. MÔ HÌNH TOÁN HỌC CỦA HỆ THỐNG

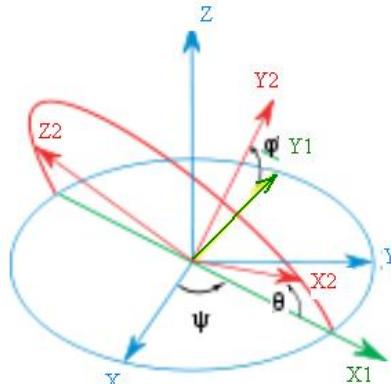
Để có thể xác định được vị trí 1 vật ta cần chọn 1 hệ trục tọa độ chung. Ở đây ta gọi là hệ tọa độ định vị(hệ này có tâm tại 1 điểm chọn trước trên mặt đất, trục Z hướng từ tâm trái đất lên, trục X hướng theo từ trường trái đất, trục Y xác định theo quy tắc bàn tay phải). Giá trị trả về từ các sensor là giá trị ứng với hệ trục tọa

Công trình này được thực hiện dưới sự hướng dẫn của Th.S Trần Tuấn Vinh-Bộ môn Kỹ Thuật Máy Tính – Viện Công Nghệ Thông Tin và Truyền Thông

Nguyễn Đình Thuận, sinh viên lớp Kỹ thuật máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 303-497-3650, e-mail: dinthuanbk88@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

độ vật thể. Mô hình chuyển đổi từ hệ trục tọa độ định vị(XYZ) sang hệ trục tọa độ vật thể (xyz) như sau



Hình 1. Góc Euler

Mô hình sử dụng 3 góc Euler

ϕ : Góc quay quanh trục X_2

θ : Góc quay quanh trục Y_1

ψ : Góc quay quanh trục Z

Thứ tự quay như sau:

$$XYZ \xrightarrow{C_\psi} X_1Y_1Z \xrightarrow{C_\theta} X_2Y_1Z_1 \xrightarrow{C_\phi} X_2Y_2Z_2 \equiv xyz$$

Do đó ma trận chuyển đổi từ hệ tọa độ định vị sang hệ tọa độ vật thể sẽ là:

$$C_n^b = C_\phi C_\theta C_\psi (1)$$

hay

$$C_n^b = \begin{pmatrix} c(\theta)c(\psi) & c(\theta)s(\psi) & -s(\psi) \\ s(\phi)s(\theta)c(\psi) - c(\phi)s(\psi) & s(\phi)s(\theta)s(\psi) + c(\phi)c(\psi) & s(\phi)c(\theta) \\ c(\phi)s(\theta)c(\psi) + s(\phi)s(\psi) & c(\phi)s(\theta)s(\psi) - s(\phi)c(\psi) & c(\phi)c(\theta) \end{pmatrix} (2)$$

Đạo hàm 3 góc Euler được tính dựa trên vận tốc góc theo công thức:

$$d \begin{bmatrix} \phi \\ \theta \\ \psi \end{bmatrix} = \begin{bmatrix} \omega_x + (\omega_y s\phi + \omega_z c\phi) \tan \theta \\ \omega_y c\phi - \omega_z s\phi \\ (\omega_y s\phi + \omega_z c\phi) / c\theta \end{bmatrix} dt (3)$$

c ở đây là cos, s là sin

Từ đây ta có thể tính được góc Euler tại thời điểm bất kỳ

$$\begin{bmatrix} \phi \\ \theta \\ \psi \end{bmatrix}_k = \begin{bmatrix} \phi \\ \theta \\ \psi \end{bmatrix}_{k-1} + d \begin{bmatrix} \phi \\ \theta \\ \psi \end{bmatrix}$$

Một phương pháp khác để biểu diễn ma trận xoay đó là sử dụng quaternion. Quaternion được đề xuất bởi Hamilton(1843)⁽¹⁾ và

ngày nay nó được sử dụng rộng rãi thay cho việc sử dụng ma trận chuyển đổi (9 thành phần) thì quaternion chỉ gồm có 4 thành phần.

Mô hình sử dụng Quaternion

$${}^A_B q = [q_1 \quad q_2 \quad q_3 \quad q_4] \quad (4)$$

Nghịch đảo của 1 quaternion

$${}^B_A q = [q_1 \quad -q_2 \quad -q_3 \quad -q_4] \quad (5)$$

Công thức nhân 2 quaternion

$$\begin{aligned} a \otimes b &= [a_1 \quad a_2 \quad a_3 \quad a_4] \otimes [b_1 \quad b_2 \quad b_3 \quad b_4] \\ &= \begin{bmatrix} a_1 b_1 - a_2 b_2 - a_3 b_3 - a_4 b_4 \\ a_1 b_2 + a_2 b_1 + a_3 b_4 - a_4 b_3 \\ a_1 b_3 - a_2 b_4 + a_3 b_1 + a_4 b_2 \\ a_1 b_4 + a_2 b_3 - a_3 b_2 + a_4 b_1 \end{bmatrix} \quad (6) \end{aligned}$$

Khi đó việc chuyển đổi hệ trục A sang hệ trục B, quaternion sẽ được tính toán lại như sau:

$${}^B X = {}^A_B q \otimes {}^A X \otimes {}^B_A q \quad (7)$$

Và đây là công thức tính đạo hàm của quaternion

$${}^B \dot{q} = \frac{1}{2} {}^B q \otimes {}^B \omega \quad (8)$$

Trong đó $\omega = [0 \quad \omega_x \quad \omega_y \quad \omega_z]$

Từ đó ta có thể tính được quaternion tại 1 thời điểm bất kỳ.

3. CÁC VẤN ĐỀ KHÓ KHĂN KHI THỰC HIỆN HỆ THỐNG.

3.1. Tính toán giá trị thực sự của gia tốc, vận tốc góc và từ tính

Đặc điểm của các sensor sử dụng trong định vị quán tính là dữ liệu được trả về dưới dạng số thông qua chuyển đổi ADC hoặc giao tiếp I2C. Tuy nhiên có 1 nhược điểm là tất cả các sensor này đều bị ảnh hưởng rất mạnh bởi môi trường bên ngoài như nhiệt độ, từ tính, kể cả chuyển động quay của trái đất.

Mà theo công thức tính khoảng cách

$$x = x_0 + \frac{at^2}{2}$$

nếu gia tốc chỉ cần lệch 1 giá trị nhỏ thì sau 1 khoảng thời gian đủ lớn thì vị trí lệch đi bao nhiêu? Do đó việc xác định độ lệch và giá trị chính xác của các sensor là việc rất quan trọng khi thực hiện hệ thống này.

3.2. Xác định hướng

Hướng ở đây tức là ta nói đến việc tính toán góc quay trên mặt phẳng OXY trong hệ tọa độ định vị góc ψ . Góc này có thể tính chỉ dựa vào accelerometer sensor đo gia tốc và magnetometer (sensor định hướng dựa vào từ trường trái đất-la bàn) hoặc accelerometer và gyroscope (sensor đo vận tốc góc). Nhưng nhược điểm của nó là ở magnetometer nhiều tương đối lớn, còn gyroscope thì xảy ra hiện tượng trôi sau 1 khoảng thời gian nhất định.

3.3. Phát hiện bước chân người.

Có nhiều phương pháp khác nhau giúp xác định bước chân người, nhưng khó khăn lớn nhất trong việc xác định bước chân người là khi người 1 người di chuyển với vận tốc lớn. Khi đó các thông số ở pha đứng yên và di chuyển có sự phân biệt không rõ ràng dẫn đến rất khó xác định được đó có phải là 1 bước chân hay không.

3.4. Tính toán khoảng cách bước chân người.

Chuyển động của bàn chân người là loại chuyển động phức tạp và nó lại không là cố định với từng người. Nên nhìn chung việc áp dụng công thức $x = x_0 + \frac{at^2}{2}$ là không chính xác.

Việc tìm ra công thức khác để ước lượng khoảng cách giữa các bước chân người là điều cần thiết.

3.5. Sai lệch so với thực tế

Xác định thiểu bước, thừa bước, góc hướng dù chỉ sai lệch rất nhỏ cũng sẽ gây nên sai lệch lớn trên quỹ đạo đường đi. Do đó khi đã có 1 bản đồ số ở 1 khu vực nhất định, thì ta cần xây dựng 1 thuật toán để hiệu chỉnh vị trí dựa vào thông tin từ bản đồ.

4. THỰC HIỆN HỆ THỐNG

4.1. Tính toán giá trị thực sự của các sensor

Do không có các thiết bị chuyên dụng để có thể xác định thật chính xác giá trị thực sự của các sensor nhưng ta có thể xác định 1 cách tương đối chính xác như sau

Đối với accelerometer:

Giả định rằng giá trị thu được và giá trị thực tế là tỉ lệ tuyến tính(nhìn chung giả thuyết này là chấp nhận được trong môi trường có nhiệt độ thay đổi không nhiều) tức là:

$$Y_{real} = Gain \times X_{measured} + Bias$$

Y_{real} : Giá trị thực tế của thông tin(gia tốc(m/s²), vận tốc góc(rad/s)...)

$X_{measured}$: Giá trị số trả về từ sensor

Gain: Hệ số chuyển đổi

Bias: Độ sai lệch so với trạng thái đứng yên.

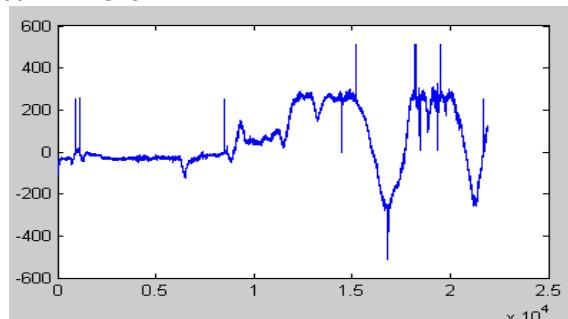
Với accelerometer ta xoay chậm thiết bị để tạo thành 1 mặt cầu. Do ta xoay chậm nên việc xoay gây ra gia tốc là rất nhỏ so với gia tốc trọng trường của trái đất. Tuy nhiên do các sensor rất hay xảy ra nhiễu ngẫu nhiên(hình 1). Ta cần loại bỏ nhiễu này mà không cần thiết bảo toàn số lượng mẫu do đó ta có thể áp dụng 1 bộ lọc đơn giản để loại bỏ nhiễu này. Không nên áp dụng bộ lọc trung vị vì ta cần đảm bảo độ trung thực của dữ liệu:

Xi sẽ bị loại bỏ nếu nó thỏa mãn:

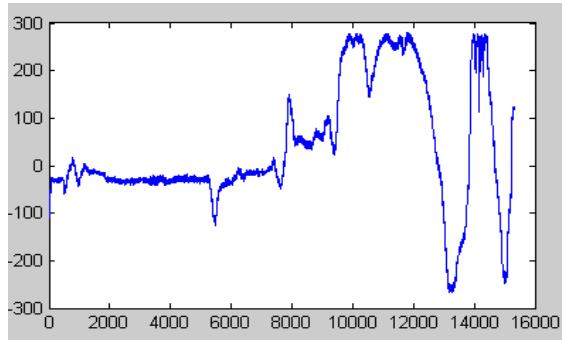
$$\left\{ \begin{array}{l} \left| \sum_{i=w}^{i-1} X_k - X_i \right| > thresh \quad (9) \\ \left| \sum_{i=1}^{i+w} X_k - X_i \right| > thresh \end{array} \right.$$

Kích thước cửa sổ w và ngưỡng (thresh) ta có thể chọn ở mức cảm thấy chấp nhận được

Dưới đây là bảng so sánh giá trị trước và sau khi lọc của cảm biến gia tốc ADXL345



Hình 1. Giá trị trước khi lọc của gia tốc theo trục x



Hình 2. Giá trị sau khi lọc của gia tốc theo trục x(w=10,thresh=60)

Khi đó $\max(a_x) = +1g$

$\min(a_x) = -1g$

$$\text{hay } \begin{cases} +1g = \text{Gain} \times \max(a_x) + \text{Bias} \\ -1g = \text{Gain} \times \min(a_x) + \text{Bias} \end{cases}$$

Từ đây ta xác định được Gain và Bias

Với magnetometer:

Theo công thức (11) để xác định hướng bằng việc sử dụng magnetometer ta không cần quan tâm đến giá trị cụ thể của nó. Do đó việc xác định bias và Gain là hoàn toàn tương tự như accelerometer tức là khi ta xoay thiết bị tạo thành 1 măt cầu thì $\text{Gain}_m = \frac{\max(m) - \min(m)}{2}$ và $\text{Bias}_m = \frac{\max(m) + \min(m)}{2}$

Đối với Gyroscope:

Với cảm biến vận tốc góc thì khi ta để ở trạng thái đứng yên thì giá trị mà ta thu được chính là độ lệch. Và giá trị Gain sẽ được tính

$$\text{Gain} = \frac{V_{ref}}{V_{sensitivity} \times \text{ADC}} \quad (10)$$

V_{ref} là điện áp tham chiếu cho ADC

$V_{sensitivity}$ là điện áp tương ứng với $1^\circ / s$

ADC là độ phân ly của bộ chuyển đổi ADC có thể là $2^8, 2^{10}, 2^{12}$ tùy từng loại.

4.2. Xác định góc hướng

a. Sử dụng cảm biến từ tính (magnetometer) và cảm biến gia tốc (accelerometer)

Magnetometer cảm biến hướng của từ trường trái đất (đường sức điện trường tại điểm đo) và accelerometer cảm biến gia tốc trọng trường của trái đất và gia tốc của thiết bị (trong trường hợp gia tốc gây ra là nhỏ thì sensor này chỉ còn cảm nhận gia tốc trọng trường). Do đó có thể sử dụng 2 sensor này để xác định hướng của thiết bị tại 1 vị trí bất kỳ.

Trong trường hợp gia tốc gây ra trên thiết bị là nhỏ, ta có:

$$C_n^b \times \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix}$$

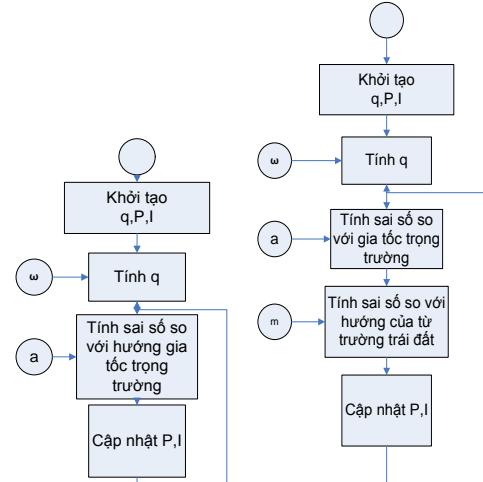
từ đây ta sẽ tính được ϕ, θ . Mặt khác ta cũng có

$$C_\phi C_\theta \times \begin{bmatrix} m_x \\ m_y \\ m_z \end{bmatrix} = \begin{bmatrix} X_h \\ Y_h \\ Z_h \end{bmatrix}$$

Nên ta suy ra

$$\psi = a \tan 2\left(\frac{Y_h}{X_h}\right) \quad (11)$$

b. Sử dụng accelerometer và gyroscope kết hợp thuật toán PI(hình 3)



Hình 3. Accelerometer và Gyroscope

Hình 4. Sử dụng thêm magnetometer

Giá trị của q được tính theo công thức:

$$\dot{q} = \frac{1}{2} q \otimes \omega + P + I \quad (12)$$

Và sai số được tính theo công thức:

$$e = e_g = q \otimes g \otimes q' \times a \quad \text{với } g = [0 \ 0 \ 0 \ 1]$$

Cập nhật P và I

$$P = K_p \cdot e$$

$$I = I + K_I \cdot e \quad (13)$$

c. Sử dụng accelerometer, gyroscope, magnetometer và thuật toán PI

$$e = e_g + e_m \quad (14) \quad \text{với } e_m = q \otimes m_n \otimes q' \times m \quad (15)$$

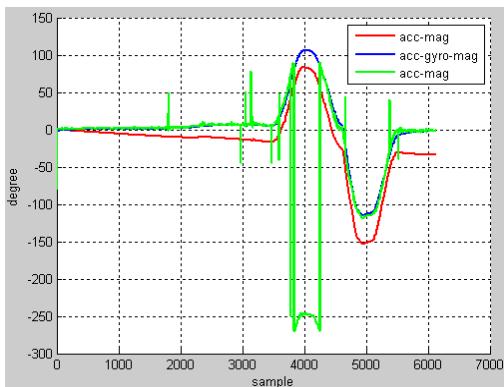
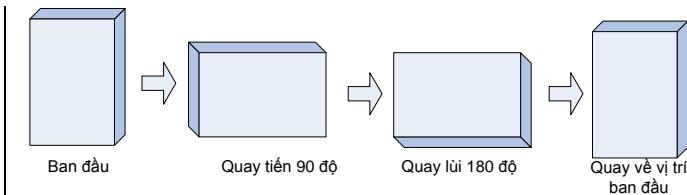
Thực tế đã chứng tỏ rằng chỉ cần sử dụng 2 trục tọa độ là có thể biểu diễn được hướng của từ trường trái đất. Trong hệ định vị hướng của từ trường trái đất được lựa chọn như sau:

$m_n = [0 \ b_x \ 0 \ b_z]$. Trong đó nếu để magnetometer

song song với bề mặt trái đất thì $b_x = \sqrt{m_x^2 + m_y^2}$ và

$$b_z = m_z$$

d. Kết quả so sánh của 3 phương pháp tính khác nhau
Ta xoay thiết bị như sau



Hình 5. Bản so sánh các kết quả thực hiện tính toán góc hướng với 3 phương pháp khác nhau.

Có thể thấy phương pháp sử dụng kết hợp accelerometer và magnetometer (màu xanh lá cây) cho kết quả bị nhiễu tương đối lớn, còn khi kết hợp accelerometer và gyroscope (màu đỏ) thì kết quả bị lệch khi quay thiết bị về vị trí ban đầu. Còn phương pháp sử dụng kết hợp cả accelerometer, gyroscope và magnetometer (màu xanh dương) và thuật toán PI là tốt hơn cả, do đó đây cũng là phương pháp được chọn để xác định hướng trong hệ thống này.

4.3. Phát hiện bước chân người

Các khảo sát thực tế cho thấy khi con người bước đi thì ở vận tốc góc và gia tốc dài ở chân biến đổi khá mạnh, nên ta có thể dựa vào các thông tin đó để quyết định đó có phải là khi chúng ta bước hay không. Để giúp loại bỏ thông tin gia tốc cũng như vận tốc góc lưu lại khi ta bước nhanh, ta dùng phương sai của gia tốc. Thuật toán xác định như sau

$$\text{-Bước 1. Tính } \mathbf{a} = \sqrt{a_x^2 + a_y^2 + a_z^2}$$

$$\text{-Bước 2. Tính } Ea_i = \frac{1}{2w+1} \sum_{k=i-w}^{i+w} (X_k - \bar{X})^2 \text{ Trên từng cùa số.}$$

*-Bước 3. Chọn ngưỡng (K) của pha dừng yên và pha di chuyển
-Bước 4. Quyết định là bước chân người thỏa mãn điều kiện*

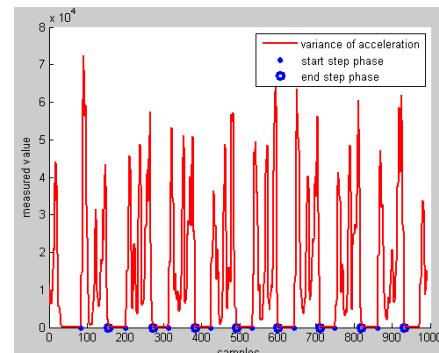
$$\begin{cases} -Ea_i > 0 \\ -\sum_{k=i-w}^w X_k = 0 \\ -Step_k - Step_{k-1} > 2w \end{cases}$$

Kết quả đạt được

Lần	Kiểu bước	Giá trị thực tế(bước)	Giá trị tính toán (bước)
1	Chậm	93	92

	Bình thường	82	82
	Nhanh	71	69
2	Chậm	92	92
	Bình thường	82	82
	Nhanh	72	72

Bảng 1. So sánh độ chính xác của giải thuật phát hiện bước chân người.



Hình 6. Biểu đồ thể hiện pha ban đầu và pha kết thúc đối với từng bước chân người.

4.4 Tính toán khoảng cách bước chân người khi bước

Bằng giá trị tính toán thực tế và của nhiều tác giả đã đề xuất thì khoảng cách bước chân người khi đi bộ là tỉ lệ với tần suất bước và phương sai của gia tốc

Ta dùng công thức $D = \alpha F + \beta E + \gamma$ (16)

Trong đó

D: Khoảng cách bước chân người

F: Tần suất chân người bước

E: Phương sai của gia tốc đo được

Ta cần tìm α, β, γ sao cho $\sum_{i=1}^n (D_i - (\alpha F_i + \beta E_i + \gamma))^2$ là nhỏ nhất hay

$$\frac{dD}{d\alpha} = \frac{dD}{d\beta} = \frac{dD}{d\gamma} = 0 \quad (17).$$

Giải hệ này cho ta α, β, γ .

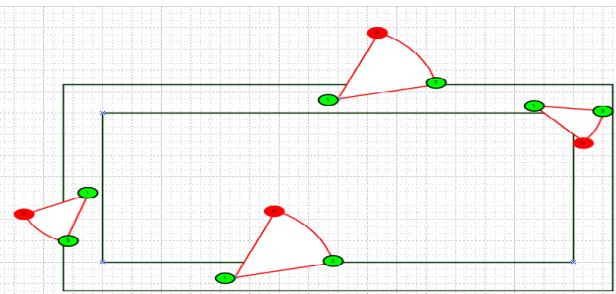
Để có thể xác định chính xác các tham số thì tốt nhất ta nên chọn 3 bộ dữ liệu (di chậm, bình thường và nhanh) của từng người

4.5 Giải thuật phù hợp dữ liệu với bản đồ số

Khi ta di chuyển trong 1 khu vực đã biết trước (chẳng hạn trong 1 tòa nhà) rõ ràng chỉ sau 1 bước chân mà vị trí của chúng ta đã thay đổi từ phòng này sang phòng này sang phòng khác, hoặc từ lối đi vào phòng nào đó (trừ trường hợp ta đang ở cửa phòng) là điều không thể, điều đó có nghĩa việc xác định vị trí của chúng ta đã bị sai lệch \rightarrow hiệu chỉnh lại vị trí người ở những điểm vi phạm là điều cần thiết để ta có thể theo vết người đó 1 cách chính xác.

Một giải thuật hiệu chỉnh vị trí đơn giản có thể thực hiện như sau:

Khi vị trí của 1 người bị vi phạm (màu đỏ) thì ta sẽ đưa vị trí đó về quỹ đạo đúng dựa vào đặc điểm về hướng đi của người đó và bảo toàn khoảng cách bước chân (màu xanh)



Hình 7. Hiệu chỉnh vị trí 1 cách đơn giản

Giải thuật

Bước 1. Cập nhật $\begin{cases} A \leftarrow (X_i, Y_i) \\ B \leftarrow (X_i + D \cos(\psi), Y_i + D \sin(\psi)) \\ Poly(i) = Poly(i-1) \end{cases}$

Bước 2. Tính toán kiểu giao

$$C \leftarrow AB \cap Poly(i)$$

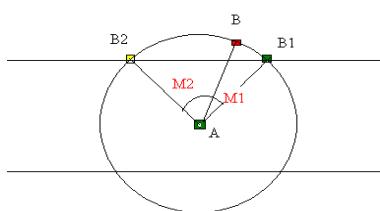
Nếu C là tường thì cập nhật lại B theo công thức sau

Tìm $B1, B2$ là giao của đường tròn (A, AB) với $Poly(i)$ khi đó điểm được chọn sẽ là điểm có góc nhỏ hơn điểm còn lại (điểm $B1$) hay $\text{Cos}(M1) > \text{cos}(M2)$

$$B \leftarrow B_1$$

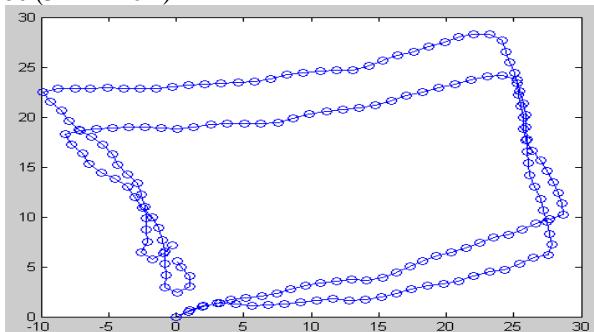
Nếu C là điểm giao thì ngoài việc cập nhật lại điểm B ta cần phải cập nhật lại $Poly$ để nó thể hiện được đường biên mới trong các bước tính toán sau

$$\begin{cases} B \leftarrow B_1 \\ Poly \leftarrow \{Map\} \end{cases}$$

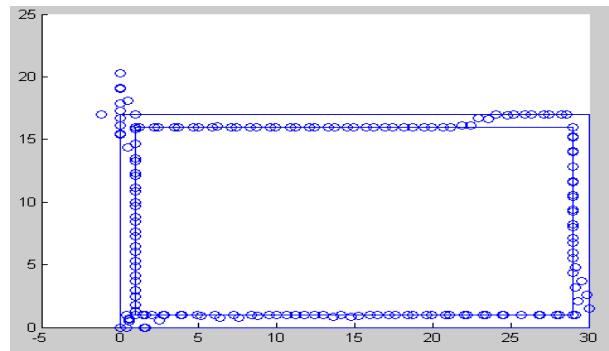


5. KẾT QUẢ THỬ NGHIỆM

Tiến hành thử nghiệm, thiết bị được gắn vào chân. Sau đó người đi bộ tiến hành đi vòng quanh tầng 6 thư viện điện tử, có kích thước (32m x 16m)



Hình 8. Dữ liệu tính toán, không dùng kết hợp bản đồ.



Hình 9. Dùng kết hợp với bản đồ số (giải thuật tích hợp đơn giản)

Dựa vào kết quả đạt được có thể thấy rằng khi được kết hợp với bản đồ số, thì hệ thống sẽ cho kết quả đáng tin cậy hơn nhiều tuy nhiên vẫn có những vị trí mà giải thuật này chưa thể hiệu chỉnh.

Định hướng tiếp theo của đề tài sẽ hướng đến việc xây dựng giải thuật tích hợp bản đồ tối ưu hơn để có thể theo vết đối tượng chính xác hơn.

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã đưa ra phương pháp để giúp tính toán hiệu chỉnh dữ liệu từ các sensor cảm biến giá thành thấp có được kết quả chính xác ở mức chấp nhận được. Kết quả thực nghiệm cũng cho thấy việc ứng dụng hệ thống này là hoàn toàn khả thi và có thể triển khai trên diện rộng. Đồng thời bài báo cũng mở ra hướng nghiên cứu mới đó là việc ứng dụng các loại cảm biến này trong việc phát hiện các chuyển động của cơ thể người, đặc biệt giúp ích trong việc chăm sóc người già, trẻ em.

Hướng phát triển tiếp theo của nghiên cứu này sẽ được tập trung vào việc định vị chính xác vị trí bằng việc kết hợp với bản đồ số và sử dụng lọc Particle. Hơn nữa việc chuyển sang định vị 3D là điều thực sự cần thiết trong hệ thống nhà cao tầng, trường học bệnh viện để trước khi nó có thể được ứng dụng thực tế.

7. LỜI TRI ÂN

Để thực hiện thành công đề tài này tôi xin gửi lời cảm ơn chân thành nhất với thầy giáo Th.S Trần Tuấn Vinh, người đã luôn theo sát, định hướng và tận tình hướng dẫn tôi trong quá trình nghiên cứu, thực hiện cũng như viết bài báo này.

8. TÀI LIỆU THAM KHẢO

- [1] O.H. Madgwick “Quaternion” 2010.
- [2] Esmat Beckir “Introduction to modern navigation”
- [3] William Premerlani and Paul Bizard “Direction Cosine Matrix IMU”
- [4] Madgwick “An efficient orientation filter for inertial and inertial/magnetic sensor arrays” April 30, 2010.
- [5] Robert Grover Brown, “Introduction to Random Signals and Applied Kalman Filtering”
- [6] William Premerlani and Paul Bizard “Direction Cosine Matrix IMU”
- [7] Ross Grote Stirling “Development of a Pedestrian Navigation System Using Shoe Mounted Sensors”, 2004
- [8] Fredrik Brannstromm “Positioning techniques alternative to GPS”

Phát triển hệ thống dẫn đường bằng giọng nói và giám sát từ xa bằng camera sử dụng công nghệ 3G trên nền tảng kit Friendlyarm

Nguyễn Thành Luân

Tóm tắt - Ngày nay, định vị GPS được phát triển và ứng dụng rộng rãi trong quản lý và giám sát phương tiện. Tuy nhiên, các hệ thống giám sát phương tiện hiện tại chưa có chức năng dẫn đường. Mặt khác, các hệ thống dẫn đường hiện tại cũng chưa có tính năng giám sát phương tiện đặc biệt là giám sát từ xa bằng camera. Bài báo này đưa ra giải pháp triển khai, xây dựng thiết bị cho phép giám sát phương tiện, dẫn đường bằng tiếng nói và hình ảnh, đồng thời quản lý phương tiện từ xa bằng Camera trên nền tảng mạng 3G. Giải pháp được xây dựng trên nền tảng thiết bị của FriendLy-ARM sử dụng hệ điều hành Linux thay thế cho lập trình firmware thông thường trên các thiết bị nhúng. Điểm cốt lõi của giải pháp là xây dựng driver USB 3G cho thiết bị; triển khai công nghệ Google Map lên thiết bị nhúng; nghiên cứu và triển khai các ứng dụng mã nguồn mở trên nền tảng ARM-Linux; xử lý, tổng hợp các file wav đã được ghi âm, kết hợp dữ liệu thu từ thiết bị GPS và googleAPI để hỗ trợ và dẫn đường cho chủ phương tiện; truyền dữ liệu camera từ thiết bị, phục vụ xem online và offline.

Từ khóa - GPS, Google Map, NMEA-1083, Mjpg-Streamer, OpenVPN.

1. GIỚI THIỆU

Công nghệ GPS(Global Positioning System) phát triển rộng rãi trên thế giới phục vụ giám sát và dẫn đường cho các phương tiện. Ở Việt Nam đã có một số công ty triển khai các hệ thống giám sát phương tiện như Bình Anh, VietMap... Tuy nhiên các hệ thống này chưa có chức năng dẫn đường. Ngược lại, các hệ thống dẫn đường hiện tại như TOMTOM, My Dean (Hàn Quốc) chưa có chức năng giám sát phương tiện đặc biệt giám sát từ xa bằng Camera.

Tôi đã lựa chọn đề tài với mong muốn đưa ra giải pháp tổng thể đảm bảo đầy đủ chức năng giám sát phương tiện, dẫn đường, và giám sát từ xa bằng Camera, cụ thể như sau:

- Dẫn đường: Triển khai công nghệ GoogleAPI lên thiết bị nhúng, dẫn đường bằng hình ảnh và âm thanh.
- Giám sát phương tiện: Thu thập dữ liệu từ module GPS và truyền dữ liệu về Server theo định dạng NMEA 1083.
- Giám sát từ xa bằng hình ảnh: Truyền dữ liệu Camera từ phương tiện theo chuẩn Mpeg.

Nguyễn Thành Luân, sinh viên lớp Kỹ thuật máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 01663859916, e-mail: thanhluanbk88@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường
Đại học Bách Khoa Hà Nội.

2. VẤN ĐỀ VÀ GIẢI PHÁP.

2.1. Giải pháp cho thiết bị trên xe.

Điều kiện cần để đảm bảo được 3 tính năng đặt ra cho bài toán là thiết bị phải có đầy đủ các module sau: module LCD (hiển thị bản đồ), module Speaker (dẫn đường bằng âm thanh), module UART (nhận dữ liệu GPS) và module USB (Nhận dữ liệu Camera). Ngoài ra, thiết bị phải đủ nhỏ gọn để có thể đặt trên xe.

Từ những yêu cầu đặt ra, tôi đã lựa chọn thiết bị FriendlyARM – Micro2440, với những thông số kỹ thuật như sau:

- **CPU:** 400 MHz Samsung S3C2440A ARM920T (max freq. 533 MHz).
- **RAM:** 64 MB SDRAM, 32 bit Bus.
- **Flash:** 256 NAND Flash and 2 MB NOR Flash with BIOS.
- **EEPROM:** 1024 Byte (I2C).
- **Ext. Memory:** SD-Card socket.
- **Serial Ports:** 1x DB9 connector (RS232), total: 3x serial port connectors.
- **USB:** 1x USB-A Host 1.1, 1x USB-B Device 1.1.
- **Audio Output:** 3.5 mm stereo jack.
- **Audio Input:** Connector + Condenser microphone.
- **Ethernet:** RJ-45 10/100M (DM9000)
- **RTC:** Real Time Clock with battery
- **Beeper:** PWM buzzer
- **Camera:** 20 pin Camera interface (2.0 mm)
- **LCD Interface:**
 - + TFT Displays.
- **OS Support:**
 - + Windows CE 5 and 6.
 - + Linux 2.6.
 - + Android.

Window CE là hệ điều hành mã nguồn đóng, rất khó có thể can thiệp tới mức driver. Linux và Android đều là hệ điều hành mã nguồn mở, có tính khả thi cao trong việc xây dựng driver cho các thiết bị, tôi đã lựa chọn hệ điều hành Linux là giải pháp của mình.

Ngoài ra thiết bị cần sử dụng thêm module GPS thu nhận tín hiệu GPS, module USB 3G để truyền dữ liệu GPS, Video về server và hiển thị bản đồ online.

Các module cần thiết cho 1 thiết bị Client đặt trên xe được biểu diễn như hình vẽ sau :



Hình 1: Thiết bị di động và các modules.

2.2. Giải pháp dẫn đường và giám sát phương tiện.

Có nhiều giải pháp để xác định vị trí đối tượng như định vị quán tính, định vị trên nền tảng mạng GSM. Tuy nhiên, hiện nay, GPS vẫn là phổ biến nhất.

Trong giải pháp của mình, tôi đã lựa chọn module GPS UB93 với độ nhạy thu rất cao lên tới -159 dB; thời gian fix dữ liệu nhanh; trả về dữ liệu theo định dạng NMEA 1083 qua cổng UART.

2.2.1) Dẫn đường

Để dẫn đường cho phương tiện có thể dẫn đường bằng hình ảnh bằng cách sử dụng bản đồ google map miễn phí, cùng với các hàm google API cho phép hiển thị vị trí hiện tại của phương tiện ở trung tâm của bản đồ và vẽ đường đi khi lựa chọn đường đi từ điểm đầu đến điểm cuối. Tuy nhiên, nếu chỉ dẫn đường bằng hình ảnh thì nhiều lúc sẽ khó khăn cho chủ phương tiện. Vì vậy, để chủ phương tiện chủ động hơn, tôi đã nghiên cứu giải pháp dẫn đường bằng âm thanh. Thiết bị sẽ gửi yêu cầu thông tin đường đi tới Google Maps Server. Dữ liệu trả về từ google dưới dạng XML, sau khi bóc tách dữ liệu, chương trình sẽ tổng hợp tiếng nói để dẫn đường cho người lái xe. Thông tin sẽ liên tục được cập nhật để đảm bảo tính chính xác cho chủ phương tiện.

Có nhiều phương pháp để tổng hợp tiếng nói. Hệ thống dẫn đường bằng tiếng nói với lượng từ vựng ít nên tôi đã lựa chọn phương pháp tổng hợp tiếng nói ở mức độ đơn giản là ghép nối các file wav đã được ghi âm sẵn từ trước.

Cấu trúc 1 file wav như sau:

The Canonical WAVE file format

endian	File offset (bytes)	field name	Field Size (bytes)
big	0	ChunkID	4
little	4	ChunkSize	4
big	8	Format	4
big	12	Subchunk1ID	4
little	16	Subchunk1Size	4
little	20	AudioFormat	2
little	22	NumChannels	2
little	24	SampleRate	4
little	26	ByteRate	4
little	28	BlockAlign	2
little	32	BitsPerSample	2
big	36	Subchunk2ID	4
little	40	Subchunk2Size	4
little	44	data	Subchunk2Size

Hình 2: Cấu trúc file WAV

Giả sử cần tổng hợp 2 file wav (file1, file2), file tổng hợp có cấu trúc như sau:

+ Subchunk2Size = Subchunk2Size (file1) + Subchunk2Size (file2).

+ ChunkSize = ChunkSize (file1) + ChunkSize(file2).

+ Data = Data (file1) nối tiếp Data (file2).

+ Các trường còn lại giữ nguyên của file1.

2.2.2) Giám sát phương tiện.

Dữ liệu GPS được truyền về Server (202.191.56.69) theo định dạng dữ liệu NMEA 1083. Server sẽ lưu trữ, xử lý dữ liệu, và phần mềm trên máy tính giúp hiển thị vị trí hiện tại của phương tiện, xem lại hành trình và lập báo cáo về đường đi của phương tiện.

2.3. Giải pháp truyền dữ liệu Video.

Có 2 giải pháp để truyền dữ liệu Video, bao gồm:

- Embedded Client: Thiết bị đóng vai trò client, truyền dữ liệu socket về server.
- Embedded Server: Thiết bị đóng vai trò là server, trả về dữ liệu Camera khi hệ thống quản lý yêu cầu.

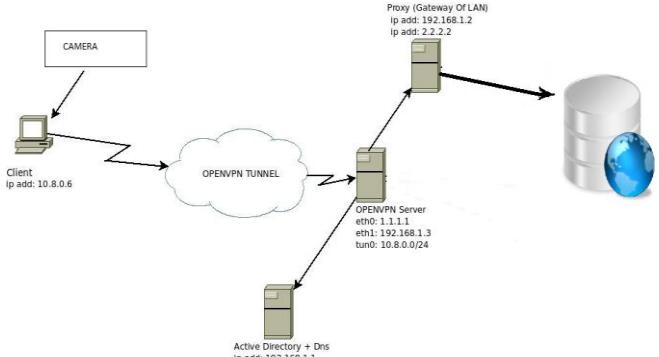
Tận dụng điểm mạnh của thiết bị đã hỗ trợ sẵn web-embedded server built-in, tôi sử dụng giải pháp Embedded Server để triển khai việc truyền dữ liệu Video qua mạng 3G. Để triển khai được phương án này trên thiết bị cần có:

2.3.1. Mjpg-Streamer

Mjpg-Streamer là 1 phần mềm mã nguồn mở cho phép nhận dữ liệu từ Camera, lưu trữ dưới dạng Web server và truyền dữ liệu về Client thông qua giao thức UDP.

2.3.2. OpenVPN

Để KIT đóng vai trò 1 web server thì KIT phải có 1 IP tĩnh. Hiện tại, các thiết bị USB 3G chưa hỗ trợ IP tĩnh, giải pháp của tôi đưa ra là dùng OpenVPN để tạo mạng LAN ảo giữa thiết bị và máy tính kết nối internet.



Hình 3: Mô hình Open-VPN.

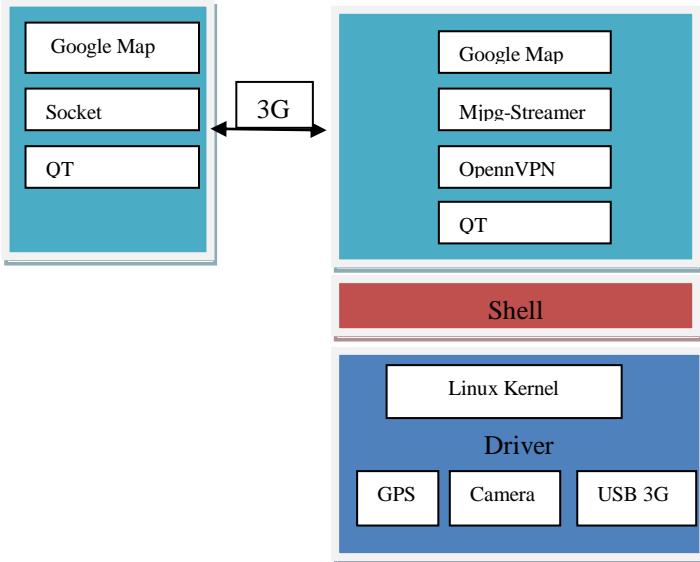
OpenVPN là phần mềm mã nguồn mở chạy trên cả hệ điều hành Linux, Windows.

- OpenVPN Server chạy trên 1 server có IP tĩnh, tạo ra các mã xác thực cho các OpenVPN client. Sau khi các client request đến với đúng mã xác thực được tạo ra, OpenVPN server sẽ tạo ra 1 địa chỉ IP ảo cho các client và tạo thành 1 mạng LAN ảo.
- Các package phụ thuộc của OpenVPN bao gồm: OpenSSL, LZO. OpenSSL là package dùng để xác thực

khi OpenVPN-Client request đến OpenVPN Server. LZO là package dùng để nén dữ liệu truyền qua đường truyền OpenVPN.

2.4. Mô hình tổng quan của hệ thống.

Tùy các giải pháp trên, mô hình tổng quan của hệ thống được thể hiện trong hình vẽ sau:



Hình 5: Mô hình tổng quan hệ thống.

3. THỰC HIỆN.

3.1. Biên dịch Linux Kernel cho thiết bị.

+ Download Linux Kernel, trong bài báo này tôi sử dụng Linux Kernel 2.6.32.2.

+ Chuyển tới thư mục mã nguồn của Linux Kernel.

```
$ cd Linux2.6.32.2
```

+ Lựa chọn các thông số cấu hình cho Linux Kernel về phần cứng và drivers cho phù hợp với thiết bị.

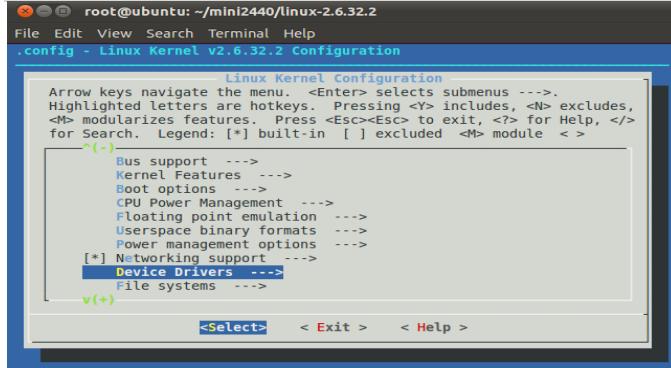
```
$ make menuconfig
```

+ Biên dịch ảnh của Linux Kernel. File ảnh kết quả nằm tại thư mục arch/boot.

```
$ make
```

+ Biên dịch các module cho Linux OS.

```
$ make modules
```



Hình 6: Biên dịch Linux Kernel.

3.2. Xây dựng driver USB 3G.

Mỗi thiết bị USB có ProductID và VendorID riêng. Khi cắm

thiết bị USB 3G hệ điều hành Linux sẽ nhận thiết bị dưới dạng USB-Mass Storage.

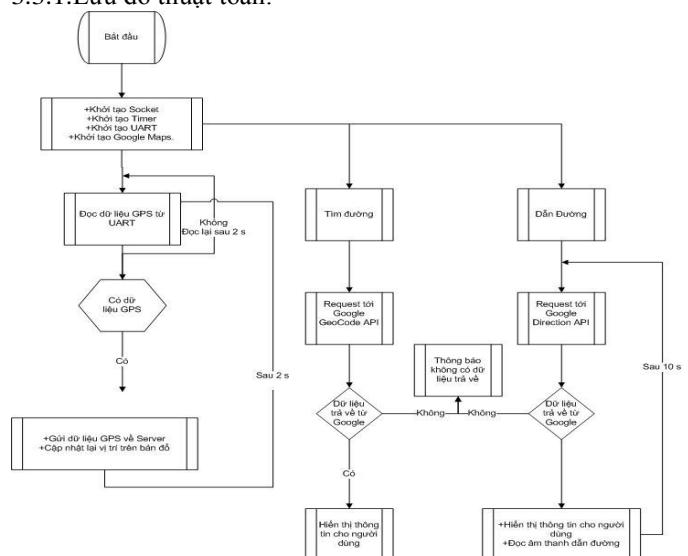
Để thiết bị hoạt động với vai trò của Modem, phải chuyển thiết bị sang chế độ Serial, sau đó gửi các lệnh AT để cấu hình mạng cho thiết bị, cụ thể công việc bao gồm:

- Tái biên dịch Linux kernel, thêm ProductID, VendorID của thiết bị vào driver Serial; Đồng thời, lựa chọn driver PPP (Point to Point Protocol) trong kernel. PPP sẽ tạo ra interface cho card mạng.
- Dịch phần mềm mã nguồn mở USB-ModeSwitch và chạy trên thiết bị, phần mềm sẽ chuyển thiết bị sang chế độ hoạt động Serial.
- Dịch phần mềm mã nguồn mở pppd (Point to point protocol Daemon), phần mềm sẽ gửi các lệnh AT command cấu hình cho Modem.
- Sửa bảng định tuyến để thiết bị có thể vào được mạng.

Điểm khó khăn của giai đoạn này là biên dịch phần mềm mã nguồn mở USB-ModeSwitch cho thiết bị ARM-Linux, cụ thể là distribution Qtopia. Khác biệt của Distribution Linux này so với distribution Ubuntu hay Fedora chạy trên Desktop hay trên thiết bị nhúng khác là không có các package hay thư viện có sẵn giúp ta cài đặt từ dòng lệnh hay dịch thông qua các source code của các package thông thường; ngoài ra, cấu trúc thư mục root file của distribution Qtopia cũng khác so với cấu trúc thư mục so với các distribution Linux thông thường. Vì vậy, để biên dịch USB-ModeSwitch, trước hết phải tìm tất cả các package và các thư viện mà nó sử dụng, sau đó download source code của các thư viện đó về, biên dịch bằng toolchain cho ARM-Linux trên hệ điều hành Ubuntu(Fedora) và chuyển tất cả các thư viện xuống KIT với cấu trúc thư mục tương tự như đã biên dịch trên hệ điều hành Ubuntu.

3.3. Lập trình ứng dụng dẫn đường QT-Embedded trên thiết bị nhúng.

3.3.1. Lưu đồ thuật toán.



Hình 7: Lưu đồ thuật toán

3.3.2. Nhận dữ liệu GPS.

Bản tin NMEA trả về gồm có các bản tin: AAM, ALM, APA, GGA, GSA, RMC....Trong đó tôi chỉ sử dụng bản tin

\$GPRMC vì bản tin này đầy đủ thông tin về thời gian, vận tốc, và vị trí.

3.3.3. Request tới google để lấy thông tin địa điểm.

Gửi Request tới địa chỉ:

<http://maps.googleapis.com/maps/api/geocode/output?parameters>

Trong đó:

Output: là định dạng dữ liệu trả về (Json hoặc Xml).

Parameters: là tham số request (địa chỉ hoặc vị trí theo latitude, longitude).

3.3.4. Request tới google để lấy thông tin đường đi.

Gửi Request tới địa chỉ:

<https://maps.googleapis.com/maps/api/directions/output?parameters>

Trong đó:

Output: là định dạng dữ liệu trả về (Json hoặc Xml).

Parameters:

+ origin (*required*) — địa chỉ hoặc vị trí (latitude, longitude) điểm đầu.

+ destination (*required*) — địa chỉ hoặc vị trí (latitude, longitude) điểm cuối.

3.4. Biên dịch Open-VPN

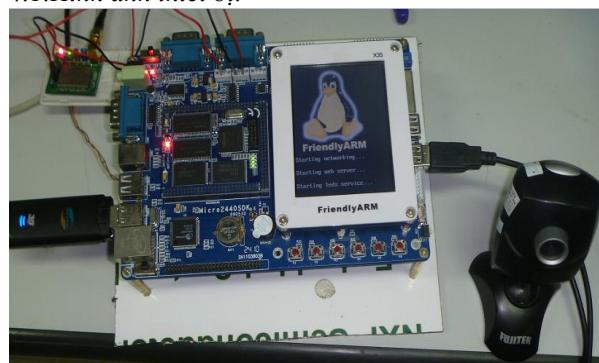
Tương tự như USB-ModeSwitch, OpenVPN là 1 ứng dụng mã nguồn mở, trong đó, ngoài việc thêm các thư viện, cần phải có thêm driver TUN/TAP để tạo ra card mạng ảo cho OpenVPN. Cụ thể thêm vào Linux kernel.

```
... Network device support
< > Dummy net driver support
< > Bonding driver support
< > M-C-VLAN support (EXPERIMENTAL)
< > SCL (serial line load balancing) support
[+] Universal TUN/TAP device driver support
< > Virtual ethernet pair device
< > HY Device support and infrastructure --->
[*] Ethernet (10 or 100Mbit) --->
[+] Ethernet (1000 Mbit) --->
[+] Ethernet (10000 Mbit) --->
Wireless LAN --->
*** Enable WiMAX (Networking options) to see the WiMAX drivers ***
[!] Network Adapters --->
[+] Wan interfaces support --->
< > PPP (point-to-point protocol) support
< > SLIP (serial line) support
< > Network console logging support (EXPERIMENTAL)
```

Hình 9: Thêm Driver TUN/TAP.

4. KẾT QUẢ ĐẠT ĐƯỢC.

4.1. Hình ảnh thiết bị.



Hình 7: Thiết bị.

4.2. Kết quả kết nối internet bằng USB 3G:

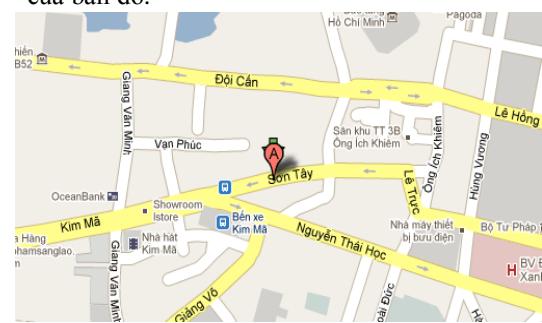


Hình 10: Kết quả kết nối Internet.

4.3. Kết quả phân dàn đường

a. Dẫn đường bằng hình ảnh trên KIT.

- Người lái xe tại vị trí luôn thấy mình tại vị trí trung tâm của bản đồ.



Hình 11: Kết quả hiển thị vị trí tại trung tâm của bản đồ.

b. Vẽ đường đi trên KIT.



Hình 12: Kết quả dẫn đường bằng hình ảnh

b. Dẫn đường bằng âm thanh.

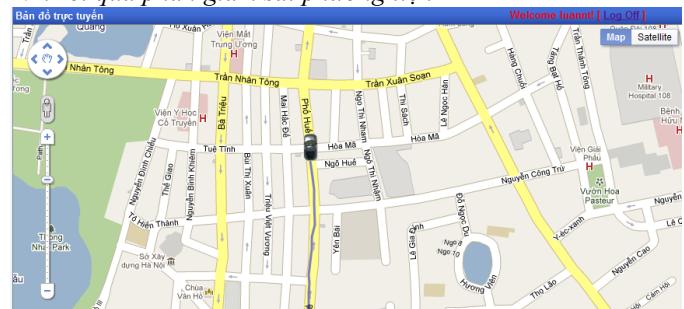
Tổng hợp tiếng nói đơn giản với lượng từ vựng ít, ví dụ:

+Đi thẳng ba trăm mét.

+Hai mươi mét rẽ trái.

+Mười mét nữa rẽ phải.

4.4. Kết quả phân giám sát phương tiện



Hình 13: Xem lại lịch trình giám sát phương tiện

4.5.Kết quả truyền dữ liệu Video.

Hệ thống đã được kiểm nghiệm truyền Video qua mạng 3G và sử dụng OpenVPN chạy thử trong thời gian 3 tiếng. Kết quả chạy ổn định. Video được lưu trữ vào cơ sở dữ liệu và xem lại log.

4.6.Hạn chế và hướng phát triển.

4.6.1.Hạn chế.

-Chương trình mới chạy thử nghiệm trên 1 thiết bị, chưa được chạy thử và kiểm nghiệm trên nhiều thiết bị.

-Màn hình hiện tại với kích thước 320x240 còn hơi bé để người lái xe có thể quan sát được vị trí và đường đi.

4.6.2.Hướng phát triển.

-Triển khai hệ thống trên màn hình kích thước lớn hơn (800x600).

-Triển khai hệ thống với nhiều thiết bị. Phát triển hệ thống thành sản phẩm thị trường.

5. LỜI TRI ÂN.

Tôi xin chân thành cảm ơn thầy giáo Th.S Trần Tuấn Vinh cùng các thầy cô giáo trong phòng thí nghiệm hệ thống máy tính – viện công nghệ thông tin và truyền thông đã tận tình giúp đỡ tôi trong quá trình thực hiện đề tài này.

6. TÀI LIỆU THAM KHẢO.

- [1] Loc Truong and Brijesh Singh, "Building a Small Embedded Linux Kernel Example" May-2008.
- [2] Jonathan Corbet, "Linux device driver third edition", January 27, 2005.
- [3] Jasmin Blanchette,Mark Summerfield, and I. N. Sneddon, "C++ GUI Programming with Qt 4".
- [4] NMEA Data Format, wikipedia.
- [5] Google Maps JavaScript API V2 Reference.
- [6] Waveform Audio File Format, wikipedia.
- [7] <http://openvpn.net>
- [8] MINI2440 Development Board User Manual & Schematic.

Xây dựng nền tảng phát triển ứng dụng quảng cáo dựa trên công nghệ Led 3d

Nguyễn Thị Phương Ly, Mai Xuân Chiến

Tóm tắt -Bạn đang băn khoăn, phân vân trong việc tìm kiếm một món quà đặc đáo nhưng không kém phần ý nghĩa để tặng cho chàng/nàng hoặc cho bè bạn và người thân của mình trong dịp lễ Valentine/sinh nhật/kỉ niệm ngày quen của 2 người, v.v..

Khối Led 3D của chúng tôi sẽ giúp bạn truyền tải đi những nội dung, thông điệp hoặc những hình ảnh đầy ý nghĩa đến cho người thân hoặc bè bạn của mình bằng việc hiển thị cũng như thay đổi hình ảnh, nội dung theo dạng lập thể 3 chiều cùng với các hiệu ứng hoạt họa thật bắt mắt và sinh động.

Từ khóa—công nghệ, hiệu ứng, Led 3d, quảng cáo

1. GIỚI THIỆU

LED (viết tắt của Light Emitting Diode, có nghĩa là diốt phát quang) là các diốt có khả năng phát ra ánh sáng hay tia hồng ngoại, từ ngoại. Hiện nay, công nghệ LED ở Việt Nam còn khá mới mẻ, các hiệu ứng và ứng dụng LED chủ yếu được tạo nên từ các phần mềm Nhập khẩu và mới chỉ dừng lại ở việc nghiên cứu vào các ứng dụng được xây dựng trên nền tảng 2D mà chưa được áp dụng và phát triển nhiều ở công nghệ 3D. Với mong muốn tận dụng ưu thế của Led, đề tài được phát triển để áp dụng nhiều vào công nghệ 3D, không chỉ dừng lại ở hình lập phương mà còn có thể phát triển là hình trụ, hình cầu...

Mục đích ứng dụng của đề tài là thể hiện được tính *Mỹ thuật, Sáng tạo, Ứng dụng và Kinh tế*.

Đề phát triển đề tài từ nền tảng đèn LED, chúng tôi đã đưa ra mô hình kiến trúc tổng thể bao gồm cả Phần cứng và Phần mềm, cụ thể như sau :

- Xây dựng phần mềm mô phỏng điều khiển và tạo hiệu ứng trên Led 3d sử dụng thư viện DirectX 9.0.
- Xây dựng nền tảng lập trình cho người phát triển tạo các hiệu ứng và thử nghiệm trên máy tính.
- Phần mềm phát triển trên máy tính có thể kết nối với khôi LED 3D và gửi câu lệnh hiệu ứng trực tiếp xuống khôi LED 3D để có thể thực hiện các hiệu ứng mà người sử dụng mong muốn.
- Xây dựng khôi LED 3D (16 triệu màu) và mạch điều khiển LED sử dụng chip chuyên dụng của hãng NXP (NXP được sáng lập bởi PHILIPS, hãng chiếu sáng hàng đầu thế giới).

Công trình này được thực hiện dưới sự hướng dẫn của ThS. Phạm Văn Thuận- Giảng viên Viện Công nghệ thông tin và Truyền thông- Trường Đại học Bách Khoa Hà Nội

Nguyễn Thị Phương Ly, sinh viên lớp Kỹ thuật máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (Điện thoại: 0979371373, e-mail: phuongly0902@gmail.com).

Mai Xuân Chiến, sinh viên lớp Kỹ thuật máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (Điện thoại: 01255764647, e-mail: maixuanhien89@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

2. CÔNG NGHỆ 3D SỬ DỤNG DIRECTX 9.0

2.1. Xây dựng phần mềm mô phỏng điều khiển và tạo hiệu ứng trên Led 3d sử dụng thư viện DirectX 9.0

2.1.1. Tìm hiểu môi trường lập trình DirectX

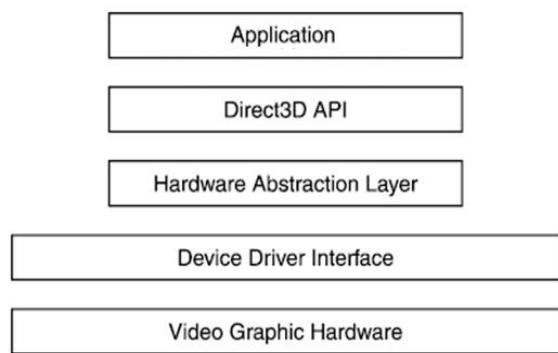
DirectX là một tập hợp thư viện các hàm API được Microsoft thiết kế, cung cấp một giao diện lập trình cấp thấp để liên kết tới các phần cứng của PC chạy trên hệ điều hành Windows, mỗi một đối tượng API của DirectX cung cấp một khả năng truy cập khác nhau tới từng loại phần cứng của hệ thống, nó bao gồm hệ thống đồ họa, âm thanh và kết nối mạng.

DirectX được chia làm 3 lớp:

- Lớp nền tảng
- Lớp phương tiện
- Lớp thành phần

Nền tảng của DirectX bao gồm 2 lớp: lớp hàm API và lớp (giao diện) thiết bị (HAL – Hardware Abstraction Layer). Các hàm trên lớp API sẽ kết nối tới phần cứng thông qua lớp HAL.

Lớp HAL cung cấp một giao diện đã được chuẩn hoá cho DirectX, ngoài ra nó có khả năng “nói chuyện” trực tiếp với phần cứng thông qua các xác lập đối với các loại thiết bị điều khiển của chính nó. Và bởi vì HAL cần phải biết cách hoạt động của phần cứng cũng như của các thiết bị điều khiển nên HAL được viết bởi chính các nhà phát triển phần cứng. Ta không bao giờ phải kết nối tới HAL một cách trực tiếp trong quá trình viết ứng dụng của mình. Thay vào đó chúng ta sẽ sử dụng kết nối gián tiếp thông qua các hàm mà DirectX cung cấp.



Hình 4-1 Kiến trúc DirectX 3D

Bằng việc sử dụng thư viện có sẵn của Microsoft DirectX. Direct3D, chúng tôi đã xây dựng phần mềm mô phỏng dễ dàng và thuận tiện giúp người sử dụng có thể tùy ứng điều khiển, tạo hiệu ứng trên Led 3D một cách linh hoạt cũng như trực quan nhất.

2.1.2. Khái niệm cơ bản về 3D

- Hệ trục tọa độ XYZ :

Trục tọa độ Oxyz gồm một gốc tọa độ (0,0,0) và 3 trục tọa độ (x, y, z) đối một vuông góc với nhau. Trong lập trình 3D sử dụng hệ trục tọa độ bàn tay phải, nghĩa là trục y hướng lên trên trục x hướng từ trái qua phải và z có hướng tiến ra ngoài màn hình theo chiều dương của mỗi trục. Điều đó có nghĩa để tiến vào không gian 3D xuyên qua màn hình, chúng ta phải có tọa độ z là âm (-) và giảm dần.

- Ma trận :

Ma trận (Matrix) là căn bản cho mọi biến đổi trong không gian 3D, chúng ta sử dụng nó để biến đổi object (di chuyển, xoay, thay đổi tỉ lệ) và sử dụng effect cũng như lập trình camera và tạo hình chiếu phối cảnh.

- Vertex :

Vertex là một điểm được định vị (không có thể hiện nào mà cho chúng ta có thể nhìn thấy) trên không gian 3D, thường chúng ta không draw vertex mà chỉ khai báo tọa độ cho nó, sau đó nối các điểm vertex lại với nhau và Draw các đường thẳng hay hình tam giác được tạo ra từ các vertex.

- Hình tam giác, đường thẳng và điểm là 3 đối tượng hình học quan trọng nhất vì từ chúng bạn có thể tạo ra bất cứ hình học không gian nào (hình cube tạo từ 12 tam giác chẳng hạn).
- Vertex có thể chứa màu và texture, khi được tạo luôn có tọa độ, do đó ta lưu nó ở dạng Vector3 có sẵn của DirectX.

- Vertex Buffer(VB):

Vertex Buffer là các điểm lưu giữ các vị trí của vertex, chính xác là 1 indices lưu giá trị của một hay nhiều vertex. Vì quá trình xử lý vertex buộc chúng ta phải trích xuất tài nguyên từ RAM sang Card graphic nên sử dụng indices sẽ ngăn ngừa đc tình trạng trên nên máy chạy nhanh hơn. Tất nhiên Indices là không bắt buộc. Ngoài ra khi có nhiều vertex trùng nhau thì sử dụng indices sẽ làm gọn chương trình.

Tập hợp các indices ta gọi là Vertex Buffer, đc xem như cái nâng đỡ cho hệ thống vertex của chúng ta.

Vertex Buffer là 1 buffer ở GPU hoặc AGP. Direct 3D sử dụng Vertex Buffer để lưu danh sách các đỉnh. Vertex Buffer giống như một dải các bytes, mỗi mảng các đỉnh được cấp phát tại mỗi khoảng (offset) nhất định trên dải đó. Khi muốn vẽ, Direct3D xác định danh sách các đỉnh bằng cách truy nhập Vertex Buffer theo địa chỉ của Vertex Buffer và offset chứa danh sách các đỉnh. Tùy vào định dạng vertex (vertex format) mà kích thước của Vertex Buffer được cấp phát khác nhau. Vertex format trong Direct3D rất mềm dẻo. Ta có thể lựa chọn vertex format tùy vào ứng dụng của mình.

Sau khi cấp phát Vertex Buffer, tiếp theo là ghi danh sách các đỉnh trên vùng nhớ đó.

- Basic Model

Chúng ta ko thể dùng vertex để draw mọi thứ trong đồ họa, có những thứ rất phức tạp mà phải cần đến các công cụ phần mềm 3D (3DSM, Maya...) để tạo ra các mô hình 3D (model). DirectX

sẽ sử dụng các model này (như hình người, cây cối...) làm sprite hay dùng chúng để tạo ra không gian trong game (skyBox).

Chúng ta cũng có thể load ra các hiệu ứng đã có sẵn cho khối Cube 3d của chúng ta.

Model mà DirectX có thể sử dụng là .FBX và .X. Ngoài ra DirectX có thể sử dụng nhiều loại model khác nhau nếu chúng ta tạo thêm các bộ xử lý cho Content của DirectX.

-Basic Effect:

Đây là hiệu ứng cơ bản mà DirectX đã cài sẵn cho chúng ta sử dụng, nó có nhiệm vụ chính là Draw một model, biến đổi model thông qua WorldMatrix của model đó, thêm một số hiệu ứng về ánh sáng (không có đồ bóng mà nó chỉ hỗ trợ 3 hướng chiếu sáng tùy chọn) và sương mù (fog)..

Cách sử dụng khá đơn giản bạn chỉ cần thiết lập vài tham số của Basic Effect là dùng được.

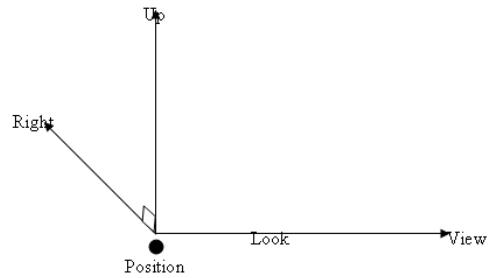
- Camera:

Camera thực sự là góc nhìn trong 3D. Gồm có năm thành phần sau:

- Vị trí đặt camera (Position)
- Mục tiêu cần nhìn vào (View)
- Hướng nhìn của camera (Look)
- Hướng phía trên của camera (Up)
- Vector pháp tuyến (Right) hướng ra từ mặt phẳng tạo bởi hai vectơ (Up, Look).

Để tạo một camera ta cần Matrix sử dụng method CreateLookAt, method cần các tham số để hoạt động:

- Position: vị trí đặt camera
- Target: điểm mà camera luôn hướng đến
- Vector3Up: một vector luôn dc camera hiểu là vector đó hướng lên trên, thường chúng ta chọn Vector3.Up => trục Y hướng lên trên, nếu bạn chọn Vector3.Right (1,0,0) thì trục X sẽ hướng lên trên theo quan sát của camera mà ta lập trình.



3.CÔNG NGHỆ LED

3.1. Tao khối LED 3D

Các Led được ghép với nhau tạo thành khối 5x5x5. Như vậy ta sẽ điều khiển với 5 mặt và 25 cột. Các anode được nối với nhau thành 5 mặt các cathode được nối với nhau thành 25 cột. Đây là một phần rất quan trọng trong đề tài, nó quyết định đèn thẩm mỹ và độ chắc chắn của khối LED.

3.2. Tim hiểu IC điều khiển mạch.

Hiện nay trên thế giới có rất nhiều công ty sản xuất về các IC chuyên dụng về điều khiển LED như Philips, ATG Electronics... Sau khi tìm hiểu và được sự gợi ý của Thầy hướng

dẫn, tôi đã chọn sử dụng IC PCA9635 của NXP. Đây là một IC chuyên dụng rất hay về điều khiển LED, với hiệu suất điều khiển cao, tiêu tốn ít năng lượng, điều khiển hiệu ứng tốt. Ở đây chúng ta sử dụng LED 3 màu nên mỗi LED sẽ có 3 chân điều khiển màu được nối vào IC, do vậy để điều khiển 125 con LED ta cần 5 IC PCA làm slave để điều khiển, sử dụng giao thức I2C.

Chip sử dụng làm master thì tôi đã sử dụng dòng ATMEGA vì đây là dòng chip khá phổ biến và được nhiều người sử dụng.

4. GIẢI PHÁP THỰC HIỆN VÀ KHÓ KHĂN

Như chúng ta đã biết, việc tạo hiệu ứng Led dù trên công nghệ 2D hay 3D thì việc lập trình cho phần cứng (Firmware) cũng gặp phải khó khăn, từ thuật toán đến thử nghiệm. Do đó, phần mềm này được xây dựng nhằm khắc phục khó khăn đó cho người phát triển trong công nghệ quảng cáo, đặc biệt là công nghệ 3D có thể tạo ra các hiệu ứng như mong muốn.

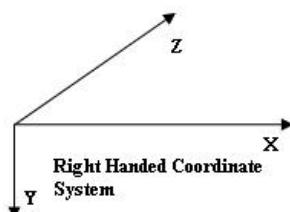
Các chức năng chính của phần mềm:

- Hiển thị khối LED 3D 5x5x5
- Có thể điều khiển sáng từng LED như mong muốn.
- Đưa ra bộ thư viện các hiệu ứng có sẵn.
- Có thể tạo ra hiệu ứng đơn giản tùy ý người sử dụng và dễ dàng điều khiển quay.
- Chọn màu như ý cho LED.
- Chọn hiệu ứng hiển thị trong bộ hiệu ứng đã tạo.
- Kết nối cổng COM, hiển thị phần cứng.

4.1. Xây dựng nền tảng lập trình cho người phát triển tạo các hiệu ứng và thử nghiệm trên máy tính

Nếu muốn tạo các khối 3D cơ bản như hình lập phương, hình đa giác, hay hình cầu, thư viện của chúng tôi có thể đáp ứng mà không cần sử dụng tới thư viện có sẵn của Microsoft. Từ thư viện này, người lập trình có thể phát triển và mở rộng theo ý muốn.

Thư viện sử dụng hệ trục tọa độ bàn tay phải và có cùng hệ tọa độ OXY như GDI+:



Các lớp cơ bản trong thư viện:

1. Camera:

Thiết lập các góc nhìn, khoảng cách, góc quay quanh theo trục X, trục Y, trục Z... Thiết lập hệ tọa độ và hình chiếu... Khi màn hình máy tính chỉ hiện thị được hình ảnh 2 chiều thì chúng phải làm dẹt đối tượng 3d trên hệ quy chiếu 2 chiều. Lớp Camera được tạo để biến đổi điểm 3D thành 2D. Lớp Camera.cs có 3 thuộc tính quan trọng :

- Location
- FocalDistance
- Quaternion

Khi camera di chuyển thì Location của nó thay đổi. Và khi camera quay thì Quaternion của nó thay đổi.

2. Cuboid:

Thiết lập các điểm 3D, các cạnh để tạo nên hình khối phù hợp. Có thể áp hình ảnh hoặc vật liệu lên từng bề mặt của hình khối theo ý.

3. Quaternion:

Trong toán học, quaternion là 1 hệ thống không có tính giao hoán mà mở rộng ra cả với số phức. Chúng cung cấp cho người dùng 1 hệ thống các kí hiệu tiện lợi trong việc biểu diễn và quay đổi tượng trong không gian 3 chiều. Quaternion có 4 chiều, 1 chiều thực w và 3 chiều ảo $\mathbf{i}x + \mathbf{j}y + \mathbf{k}z$ mô tả 1 trực quay và 1 góc quay.

Công thức định nghĩa Quaternion như sau:

$$\mathbf{q} = w + xi + yj + zk = w + (x, y, z) = \cos(a/2) + \mathbf{u}\sin(a/2)$$

Ở đây \mathbf{u} là đơn vị vector, và a là góc quay quanh \mathbf{u} .

- \mathbf{v} là vector cơ bản trong không gian 3 chiều, được coi như 1 quaternion với tổng tọa độ thực w bằng 0.

Công thức tính như sau: \mathbf{qv}^{-1} .

4. Shape3d:

Là lớp cơ sở để lưu trữ các điểm 3d cho việc vẽ hình khối và cung cấp các tiện ích để vẽ các khối 3d.

5. FreeImageTransformation:

Bạn có thể dùng lớp này để thay đổi hình ảnh mà bạn thích cho mặt trước của hình 3D.

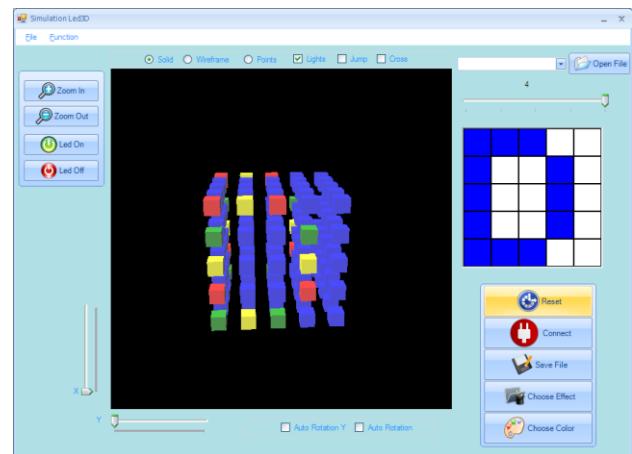
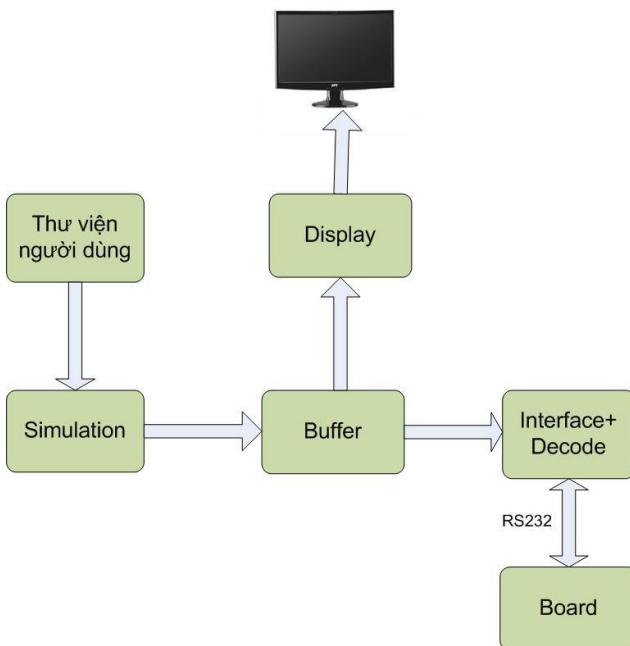
4.2. Phần mềm phát triển trên máy tính có thể kết nối với khối LED 3D và gửi câu hình hiệu ứng trực tiếp xuống khối LED 3D đó để có thể thực hiện các hiệu ứng này theo ý muốn của người sử dụng.

- Có thể bạn sẽ băn khoăn liệu rằng sản phẩm của chúng tôi có thể đưa ra sản phẩm như ý không? Thì mô hình LED 3D chúng tôi tạo ra sẽ là kết quả để khẳng định điều đó.

- Phần mềm mô phỏng cung cấp 1 bộ thư viện các hiệu ứng có sẵn, thích hợp và thuận tiện cho người lập trình phần cứng có thể dựa vào đó tìm ra phương pháp giải quyết phù hợp cho mình.

- Cung cấp đủ các chức năng để gửi hiệu ứng xuống khối LED 3D qua cổng COM sao cho hiệu ứng là đúng nhất và trực quan nhất.

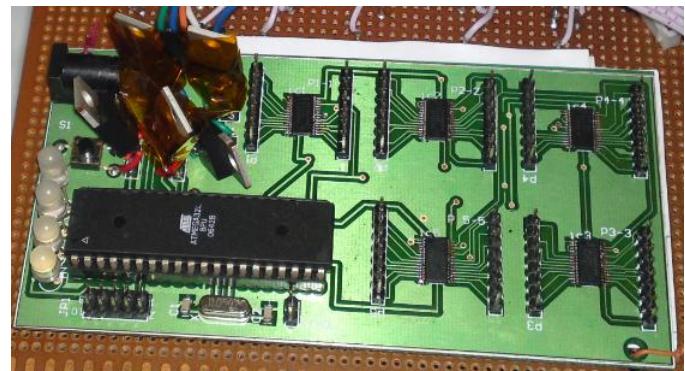
Mô hình phát triển phần mềm:



Bước 3:

Cho phép người sử dụng kết nối trực tiếp để truyền hiệu ứng xuống khói LED 3D qua RS232.

4.3 Thiết kế mạch phần cứng điều khiển



Kết nối truyền nhận UART với máy tính, phân tích dữ liệu và đưa ra hiển thị LED

Dữ liệu truyền xuống sẽ được lưu và phân tích thành phần trong đó, đưa ra tọa độ, màu sắc của từng LED. Sau đó hiển thị trên khói LED 3D.

4.4 Khó khăn phát sinh trong quá trình thực hiện đề tài

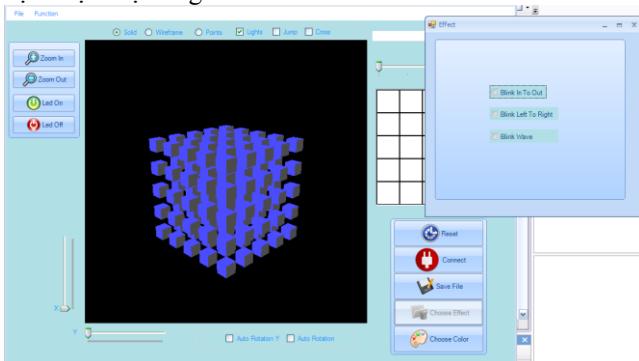
- Như chúng ta đã biết, phần cứng là một vấn đề hết sức quan trọng trong đề tài, vì đây là một sản phẩm chỉ có thể thực hiện trên mạch thật, không có giải pháp mô phỏng. Do vậy phải làm một mạch đơn giản để có thể điều khiển các tính năng của PCA9635 trên LED đơn màu. Khi điều khiển được rồi thì ta sẽ dễ dàng điều khiển được nhiều LED hơn. Nhưng với 125 con LED, để điều khiển 5 mặt, mỗi mặt ta phải điều khiển 25 con LED như vậy dòng của mạch LED rất lớn, từ 2->4 Ampe. Do đó việc dùng transistor thường thì không thể cấp đủ áp điều khiển LED. Ta phải chuyển sang dùng FET(p). Ở đây tôi dùng IRF9540 để có thể cấp đủ áp cho cả 25 LED sáng đồng thời.

- Việc chọn LED làm sao để độ sáng như nhau là rất khó, vì vậy để giải quyết vấn đề này là phải mắc những điện trở khác nhau trên từng chân của LED, nhưng do đây là vấn đề khá phức tạp vì mỗi LED 3 màu phải có 3 điện trở đi kèm, như thế sẽ phải dùng rất nhiều điện trở, hơn nữa vì dòng quá lớn nên áp roi trên điện trở là rất lớn nên trở sẽ bị hỏng, mà mạch lại không thể

Các bước thực hiện:

Bước 1:

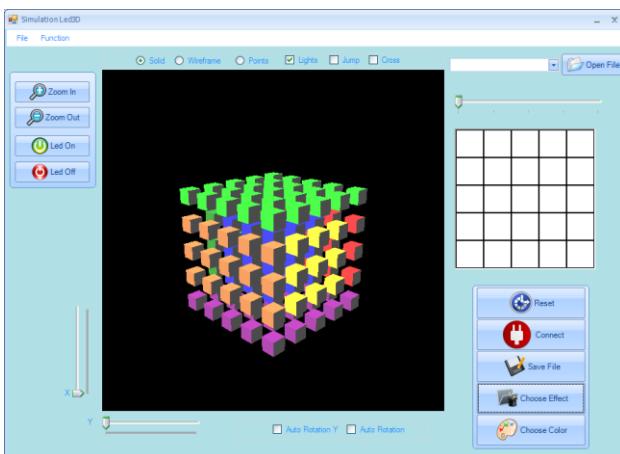
Lựa chọn hiệu ứng sẵn có từ form Effects.



Bước 2:

Hiển thị hiệu ứng lên mô hình 3D

- Hiệu ứng có sẵn :



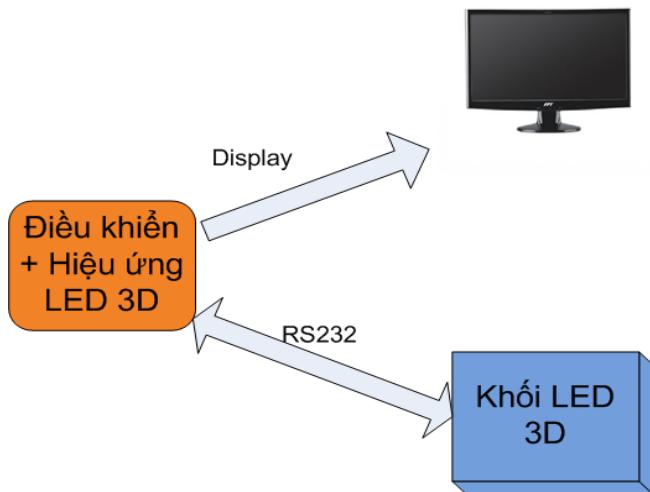
- Tạo hiệu ứng tùy biến theo người dùng

sáng.(LED 3 màu chỉ màu đỏ là sáng ở nguồn 2,2V là đủ,còn 2 màu còn lại phải cấp nguồn 3,3 V).

- Quá trình hàn LED thành khối 5x5x5 cũng là cả một vấn đề, nó đòi hỏi hàn phải chắc chắn và phải đều thì hiệu ứng không bị méo.Có rất nhiều giải pháp như:hàn khung rồi hàn LED lên,gắn từng chân LED với nhau...

- Quá trình kết nối điều khiển: Đây là phần rất quan trọng khi hoàn thành xong phần cứng. Phải phân tích dữ liệu nhận về qua cổng COM, tham chiếu đến từng tọa độ, sau đó mới tham chiếu đến màu ở từng chân LED.

Mô hình phát triển tổng quan:

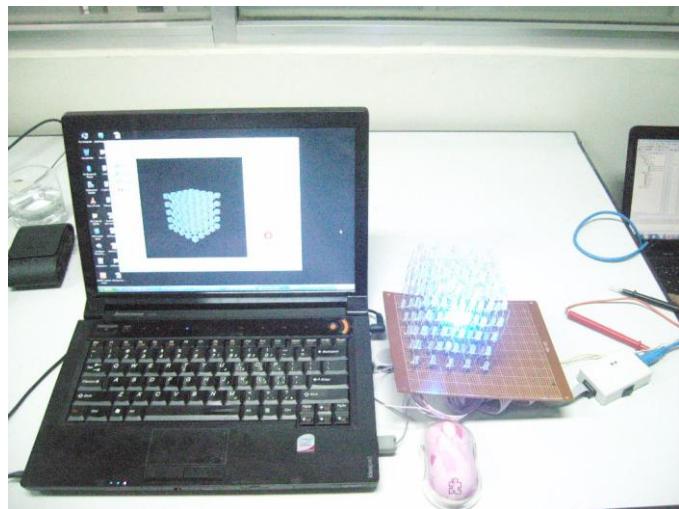


5.KẾT QUẢ THỰC HIỆN ĐỀ TÀI

Sau khi thực hiện đề tài này, chúng tôi tự đánh giá mình đạt được kết quả như sau:

- Tìm hiểu môi trường lập trình DirectX 9.0
- Tìm hiểu cách xây dựng vật thể trong công nghệ 3D
- Xây dựng được nền tảng phát triển và ứng dụng công nghệ 3D thuận lợi cho người phát triển có thể tạo ra các hình khối 3D theo mong muốn.
- Xây dựng được phần mềm mô phỏng tạo hiệu ứng và điều khiển từng đèn LED trên khối LED 3D.
- Kết nối thành công với phần cứng thông qua cổng COM.
- Decode và truyền hiệu ứng thành công qua cổng COM,
- Xây dựng được khối LED 3D Cube phần cứng.
- Tìm hiểu giao thức I2C để đưa vào kết nối giữa ATMEGA với PCA 9635.
- Tìm hiểu IC chuyên dụng để điều khiển LED PCA 9635.
- Xây dựng và lập trình hiệu ứng khối LED 3D.

Hình ảnh mô phỏng sau khi đã kết nối



6. LỜI TRI ÂN

Chúng tôi xin gửi lời cảm ơn đến thầy Phạm Văn Thuận, thầy Trần Tuấn Vinh đã tận tình giúp đỡ, trực tiếp chỉ bảo, hướng dẫn và định hướng cho chúng tôi trong suốt quá trình làm đề tài khoa học này. Cuối cùng chúng tôi xin cảm ơn các bạn trong lớp Kỹ thuật máy tính đã nhiệt tình đưa ra các góp ý xây dựng đề tài để đề tài hoàn thiện hơn. Nhờ đó mà bài báo này mới được thực hiện. Chúng tôi xin chân thành cảm ơn.

7. TÀI LIỆU THAM KHẢO

- [1] Tom Miller,"Managed DirectX 9 Kick Start Graphics and Game Programming"
- [2] Alexandre Santos Lobao and Ellen Hatton,"Game Programming with DirectX 9.0"
- [3] Wendy Jones," Beginning DirectX 9.0"
- [4] <http://www.riemers.net>
- [5] <http://www.gamespp.com>
- [6] Datasheet PCA9635

Ứng dụng xác thực khuôn mặt trong kiểm tra hộ chiếu

Nguyễn Viết Thành Trung

Tóm tắt - Báo cáo này trình bày về một hệ thống bán tự động, cho phép trợ giúp nhân viên hải quan trong việc kiểm tra ảnh hộ chiếu. Hệ thống này thẩm định tính hợp lệ của ảnh hộ chiếu thông qua xác thực khuôn mặt, dựa trên phương pháp trích chọn đặc trưng khuôn mặt sử dụng giải thuật phân tích thành phần cơ bản (Principal Component Analysis – PCA). Hệ thống đã được tiến hành cài đặt và thử nghiệm, bước đầu đã cho những kết quả nhất định.

Từ khóa – Face verification, OpenCV, Passport image, PCA

1. Giới thiệu

Từ những năm 60 của thế kỷ 20, các nghiên cứu về nhận dạng, xác thực khuôn mặt đã được thực hiện. Trong số các hướng tiếp cận để giải quyết bài toán này, hướng tiếp cận dựa trên diện mạo, sử dụng các mô hình của thống kê xác xuất cho kết quả khá quan sát. Đã có nhiều ứng dụng của xác thực khuôn mặt được triển khai trong thực tế, phổ biến nhất là chức năng đăng nhập máy tính cá nhân sử dụng khuôn mặt. Trong báo cáo này, em xin trình bày về một ứng dụng của xác thực khuôn mặt đang được nghiên cứu và phát triển ở nhiều quốc gia trên Thế giới: ứng dụng xác thực khuôn mặt vào kiểm tra ảnh trong hộ chiếu. Kiểm tra ảnh trong hộ chiếu là một khâu bắt buộc trong thủ tục xuất nhập cảnh ở Việt Nam cũng như tất cả các quốc gia khác trên Thế giới. Hiện nay, công việc này đang do nhân viên hải quan thực hiện một cách thủ công, dựa nhiều vào cảm tính. Từ quan sát trong thực tế đó, với mong muốn xây dựng một hệ thống bán tự động trợ giúp nhân viên hải quan, đề tài này tập trung vào xây dựng một hệ thống kiểm tra ảnh trong hộ chiếu dựa trên xác thực khuôn mặt.

Phần 2 của báo cáo giới thiệu về phương pháp trích chọn đặc trưng khuôn mặt sử dụng giải thuật phân tích thành phần cơ bản (Principal Component Analysis – PCA) [1].

Phần 3 trình bày chi tiết về các thành phần chức năng của hệ thống kiểm tra ảnh trong hộ chiếu dựa trên xác thực khuôn mặt, gồm có: thành phần đăng ký, thành phần xác thực và thành phần quản lý cơ sở dữ liệu.

Phần 4 trình bày các bước cài đặt hệ thống thử nghiệm, các kết quả thử nghiệm, đánh giá hệ thống đã xây dựng. Các đánh giá được thực hiện trên bộ cơ sở dữ liệu “The Indian Face Database” [6] và trên các ảnh chụp trong thực tế. Các hạn chế và phương hướng phát triển hệ thống cũng sẽ được đề cập đến.

Phần 5 nêu lên một số kết luận về kết quả đã đạt được và triển vọng của đề tài.

2. Trích chọn đặc trưng khuôn mặt sử dụng giải thuật phân tích thành phần cơ bản

Nguyễn Viết Thành Trung, sinh viên lớp Tin Pháp, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (diện thoại: 091 608 2617, e-mail: trungtpbk@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

Giải thuật phân tích thành phần cơ bản (Principal Component Analysis – thường được viết tắt là PCA) xuất phát từ một thực tế trong biểu diễn dữ liệu: liệu ta có thể tìm được một cơ sở khác là tổ hợp tuyến tính của cơ sở ban đầu và biểu diễn tốt nhất bộ dữ liệu đó. Bản chất của PCA là xác định một cơ sở mới có số chiều k nhỏ hơn số chiều của cơ sở ban đầu N sao cho khi ta chiếu dữ liệu trong cơ sở ban đầu lên cơ sở mới, phương sai lớn nhất nằm ở chiều thứ nhất (được gọi là thành phần cơ bản thứ nhất), phương sai lớn thứ hai nằm ở chiều thứ hai ... Giải thuật này đặc biệt có hiệu quả khi tập dữ liệu ban đầu đủ thừa về biểu diễn: các chiều biểu diễn dữ liệu phụ thuộc tuyến tính với nhau.

Về mặt toán học, bài toán có thể phát biểu như sau: gọi X là tập dữ liệu ban đầu, trong đó mỗi cột là một mẫu và mỗi hàng là một tiêu chí đo nào đó, gọi Y là cách biểu diễn mới của tập dữ liệu ban đầu. Khi Y là cách biểu diễn tối ưu của tập dữ liệu ban đầu, ma trận hiệp phương sai của Y trở thành ma trận đường chéo chính. Bài toán phân tích thành phần cơ bản có thể được phát biểu như sau [3]: tìm một ma trận trực giao P là ma trận chiều của phép biến đổi $Y = P^T X$ trong đó ma trận hiệp phương sai của Y là một ma trận đường chéo. Các hàng của ma trận P chính là các thành phần cơ bản của X .

Gọi C_X, C_Y lần lượt là ma trận hiệp phương sai của X và Y , ta có: $C_Y = \frac{1}{n} YY^T = \frac{1}{n} (PX)(PX)^T = \frac{1}{n} PXX^T P^T$
 $= P\left(\frac{1}{n} XX^T\right)P^T$
 $= PC_X P^T \quad (1)$

Do C_X là một ma trận đối xứng, nên ta có thể viết lại C_X như sau: $C_X = EDE^T$, trong đó E là ma trận có các cột là các vector riêng của C_X . Chọn ma trận P là ma trận chuyên vị của ma trận E . Do ma trận P trực giao nên ta có: $P^{-1} = P^T = E$. Thay vào công thức (1), ta có:

$$C_Y = P(EDE^T)P^T = P(P^T DP)P^T = (PP^T)D(P^T P)
= (PP^{-1})D(P^{-1}P) = D \quad (2)$$

Từ công thức (2) ta thấy rằng: để ma trận chiều P cần tìm là ma trận có các hàng là các vector riêng của ma trận X .

Việc áp dụng PCA vào trích chọn đặc trưng khuôn mặt được đề xuất bởi Matthew Turk và Alex Pentland vào năm 1991 [1]. Từ thực tế có thể thấy: khi biểu diễn lại một ảnh dưới dạng vector mà mỗi thành phần trong vector tương ứng với một điểm ảnh, chỉ có một số rất ít trong số các tổ hợp có thể có là biểu diễn ảnh khuôn mặt. Điều đó cho thấy không gian vector dùng để biểu diễn ảnh khuôn mặt dư thừa rất lớn. Turk và Pentland đã đề xuất áp dụng PCA để trích chọn đặc trưng khuôn mặt: giảm số chiều của không gian vector biểu diễn, chỉ giữ lại các thành phần cơ bản nhất. Cụ thể các bước trích chọn đặc trưng khuôn mặt sử dụng PCA như sau [1],[2]:

1. Xây dựng một tập các ảnh đầu vào có cùng kích thước. Mỗi ảnh được biểu diễn dưới dạng một vector cột. Tập hợp ảnh đầu vào được biểu diễn bằng một ma trận T có kích thước $N \times M$, với M là số ảnh trong tập hợp.

2. Tính giá trị ảnh trung bình Ψ . Chuẩn hóa tập ảnh đầu vào để có được một phân bố xác suất bằng cách lấy từng ảnh trong T trừ đi ảnh trung bình Ψ , ta thu được ma trận A .

3. Xác định ma trận chiếu P : đó là ma trận có các hàng là các vector riêng có trị riêng khác 0 của ma trận hiệp phương sai

$C = A^T A^T$. Các vector riêng này có kích thước bằng với kích thước ảnh trong tập đầu vào nên chúng cũng có thể coi là các ảnh khuôn mặt và còn được gọi là các *Eigenface*. Ma trận A có kích thước $N \times M$, trong đó $M < N$ nên thực tế ta chỉ có M vector riêng có trị riêng khác 0. Việc tính toán trực tiếp các vector riêng của ma trận C là rất khó khăn, vì N thường rất lớn. Do đó, trong thực tế ta thường áp dụng phương pháp như sau:

- Xét ma trận $L = A^T A$ có kích thước $M \times M$.
- Tìm các vector riêng v và trị riêng λ của L . Do $M \ll N$ nên việc tính toán này là dễ dàng hơn nhiều.
- Theo định nghĩa: $A^T A v = \lambda v$ (3). Nhân cả hai vế của biểu thức (3) với A ta có: $AA^T A v = A\lambda v = \lambda(Av)$, suy ra vector $u = Av$ là vector riêng của $C = AA^T$, ứng với trị riêng λ .

4. Để xác định đặc trưng của một ảnh khuôn mặt Γ : lấy ảnh đó trừ đi ảnh trung bình Ψ rồi chiếu lên không gian tạo bởi các *Eigenface* thu được ở trên: $\Omega = u^T (\Gamma - \Psi)$. Vector Ω chính là đặc trưng khuôn mặt cần tìm.

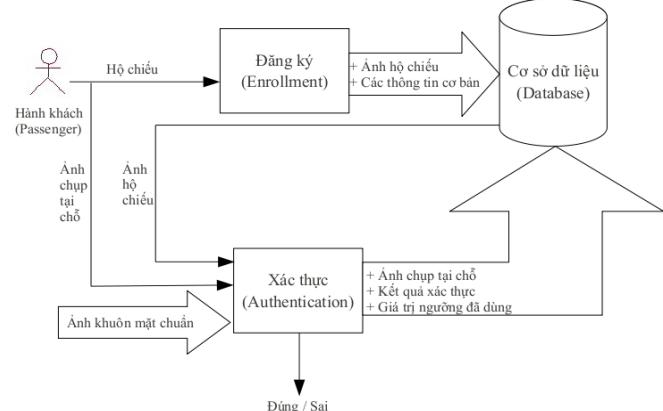
Bảng 1: Các bước trích chọn đặc trưng khuôn mặt sử dụng PCA

3. Hệ thống kiểm tra ảnh hộ chiếu sử dụng xác thực khuôn mặt

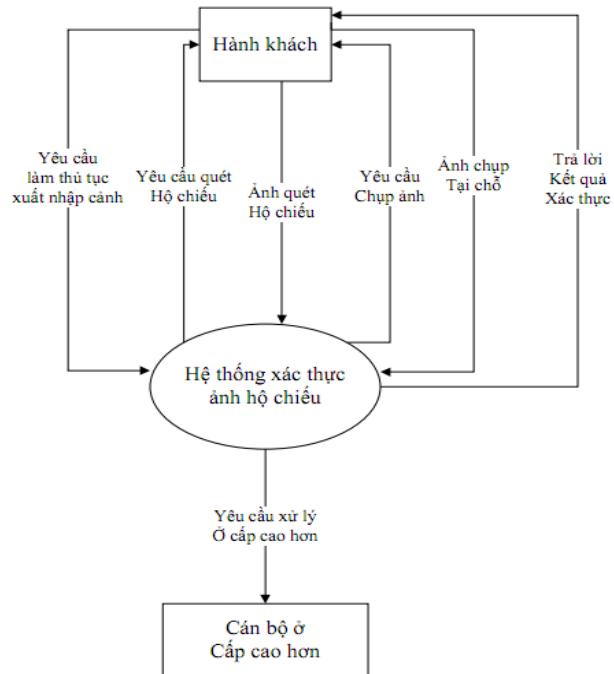
Cũng như các hệ thống xác thực sinh trắc học nói chung và xác thực khuôn mặt nói riêng, hệ thống này cũng bao gồm ba thành phần cơ bản, tương ứng với ba chức năng chủ yếu của hệ thống, đó là:

- Đăng ký (enrollment)
- Xác thực hành khách (authentication)
- Lưu trữ cơ sở dữ liệu

Tuy nhiên, do đặc thù của hệ thống này, nên ở mỗi khâu đều có những điểm khác biệt so với hệ thống truyền thống.



Hình 1: Sơ đồ khái niệm hệ thống kiểm tra ảnh trong hộ chiếu sử dụng xác thực khuôn mặt



Hình 2: Biểu đồ luồng dữ liệu mức khung cảnh

3.1. Thành phần đăng ký

Khác với các hệ thống xác thực thông thường, đòi hỏi người sử dụng phải đăng ký và cung cấp mẫu sinh trắc (ở đây là khuôn mặt) trước khi sử dụng, khâu đăng ký trong hệ thống kiểm tra ảnh hộ chiếu được thực hiện ngay trong khi người dùng bắt đầu sử dụng hệ thống.

Khi hành khách yêu cầu làm thủ tục đăng ký xuất nhập cảnh, họ phải xuất trình hộ chiếu. Hộ chiếu sẽ được quét (scan) bằng thiết bị chuyên dụng. Các thông tin cơ bản trong hộ chiếu như: mã hộ chiếu, mã quốc gia, họ và tên của người cầm hộ chiếu ... sẽ được lưu lại. Quá trình này tương ứng với quá trình khai báo thông tin khi đăng ký trong các hệ thống thông thường.

Ảnh khuôn mặt của hành khách trong hộ chiếu cũng cần được phát hiện. Ảnh này được sử dụng như là ảnh đăng ký của người dùng với hệ thống. Phương pháp phổ biến nhất hiện nay để xác định khuôn mặt trong ảnh là phương pháp do Paul Viola và Michael John đề xuất^[4], sử dụng thuật toán Ada Boost: kế hợp tuyển tính các bộ phân loại yếu (weak learner) để trở thành một bộ phân loại mạnh (strong learner). Bộ phân loại mạnh được tạo ra vừa thỏa mãn tính chính xác cao, lại vừa có tốc độ thực hiện nhanh do các thao tác tính toán đơn giản. Do đó, phương pháp này cho phép xác định khuôn mặt trong ảnh trong thời gian thực, thích hợp với các hệ thống đòi hỏi xử lý tức thì như hệ thống kiểm tra ảnh hộ chiếu này.

Ngoài ra, các thông tin bổ sung như ngày, giờ hành khách tiến hành thủ tục xuất nhập cảnh cũng được lưu lại để phục vụ quá trình kiểm tra, dù vết nứt như có sự cố xảy ra.

3.2. Thành phần xác thực

Quá trình xác thực được thực hiện ngay sau khi hoàn tất việc lưu các thông tin cơ bản của hành khách vào cơ sở dữ liệu. Việc xác thực được thực hiện thông qua đối sánh giữa ảnh trong hộ

chiếu của hành khách với ảnh của chính hành khách đó được chụp tại thời điểm tiến hành thủ tục xuất nhập cảnh.

Để đảm bảo tính chính xác của hệ thống, hành khách sẽ được yêu cầu chụp 03 ảnh. Việc chụp ảnh do nhân viên hải quan thực hiện và cần đảm bảo một số tiêu chí như: ảnh chụp thẳng (để việc so sánh với ảnh trong hộ chiếu là có ý nghĩa), toàn bộ các bộ phận cơ bản trên khuôn mặt như mắt, mũi, miệng... đều nằm trong ảnh.

Các ảnh này cùng với ảnh trong hộ chiếu của hành khách sau đó sẽ qua giai đoạn tiền xử lý để cải thiện chất lượng của ảnh và kết hợp với một số ảnh khuôn mặt đã được chuẩn hóa trước sẽ được dùng làm tập ảnh đầu vào cho phương pháp trích chọn đặc trưng khuôn mặt sử dụng PCA (đã được trình bày ở mục 2). Ta cần tính giá trị đặc trưng của ảnh trong hộ chiếu và 03 ảnh của hành khách được chụp tại thời điểm đăng ký. Giá trị đặc trưng của ảnh chụp tại chỗ được tính bằng trung bình cộng của 03 ảnh chụp được. Quá trình xác thực được thực hiện như sau:

- Tính khoảng cách Euclidean giữa đặc trưng của ảnh hộ chiếu và đặc trưng của ảnh chụp tại chỗ:

$$d(t1, t2) = \sqrt{\sum_{i=1}^k (t1[i] - t2[i])^2} \quad (4)$$

- Tính độ tương tự giữa đặc trưng của ảnh hộ chiếu và đặc trưng của ảnh chụp tại chỗ: $s(t1, t2) = \frac{1}{1 + \frac{d}{CONST}}$ (5)

trong đó s là độ tương tự giữa hai mẫu, d là khoảng cách Euclidean giữa hai mẫu đó, giá trị CONST là một hằng số để đảm bảo s không quá nhỏ.

- So sánh giá trị độ tương tự $s(t1, t2)$ với giá trị ngưỡng θ . Nếu $s(t1, t2) > \theta$ thì kết luận hành khách đúng là người trong hộ chiếu và cho phép hộ thực hiện xuất nhập cảnh. Nếu $s(t1, t2) < \theta$, cần bộ hải quan có thể cho phép hành khách thực hiện lại quá trình xác thực, tối đa là 03 lần. Nếu sau cả 3 lần hệ thống đều cho kết quả $s(t1, t2) < \theta$ (tức là đều kết luận hành khách và người trong hộ chiếu không phải là một người) thì cần bộ hải quan chuyển việc xử lý cho cấp cao hơn giải quyết.

Để phục vụ việc kiểm tra, thẩm định trong trường hợp xảy ra sự cố, hệ thống lưu lại 01 ảnh trong số 03 ảnh đã chụp. Kết quả xác thực và giá trị ngưỡng sử dụng cũng được lưu lại để phục vụ đánh giá, cải thiện chất lượng hệ thống.

3.3. Thành phần lưu trữ cơ sở dữ liệu

Như đã trình bày trong các mục 3.1 và 3.2, hệ thống cần lưu trữ lại các thông tin sau:

- Các thông tin cơ bản trong hộ chiếu: mã hộ chiếu, mã quốc gia, họ và tên của người cầm hộ chiếu.
- Ảnh quét hộ chiếu.
- Thời gian (ngày, giờ cụ thể) tiến hành thủ tục xuất nhập cảnh.
- 01 ảnh của hành khách chụp tại thời điểm làm thủ tục xuất nhập cảnh.
- Kết quả xác thực của hệ thống.
- Giá trị ngưỡng được dùng để đánh giá.

Các thông tin này cho phép dễ dàng theo vết một hành khách hoặc xử lý các cán bộ hải quan trong trường hợp xảy ra sự cố. Nó cũng cho phép những người điều hành có được thống kê về

hiệu quả hoạt động của hệ thống, từ đó có được sự điều chỉnh các thông số cho phù hợp.

Tuy nhiên, hiện nay lưu lượng người tham gia xuất nhập cảnh trong một ngày là rất lớn. Nếu ta tiến hành lưu trữ tất cả thông tin vào cơ sở dữ liệu hiện hành thì cơ sở dữ liệu rất lớn, dẫn đến việc thao tác dữ liệu trở nên chậm. Do đó, cần phải có một giải pháp xử lý theo lô: sau một khoảng thời gian nhất định hoặc khi số lượng hành khách đã lưu trữ vượt quá một số lượng nhất định, các thông tin về hành khách đang được lưu trữ sẽ được đóng gói và lưu lại trong kho chứa thông tin. Trong trường hợp cần thiết, những người có chức năng sẽ truy cập kho chứa để lấy ra các thông tin cần thiết. Việc lập lịch cho quá trình xử lý theo lô do người lãnh đạo của đơn vị và người quản trị hệ thống quyết định.

4. Đánh giá hệ thống thử nghiệm

Hệ thống thử nghiệm được xây dựng sử dụng bộ thư viện mã nguồn mở OpenCV 2.1 trên nền tảng Visual Studio 2008.

Quá trình xác định khuôn mặt trong ảnh được thực hiện nhờ hàm `cvHaarDetectObject` trong bộ thư viện OpenCV [5]. Hàm cài đặt thuật toán xác định khuôn mặt trong ảnh do John và Viola đề xuất. Do ảnh trong hộ chiếu và các ảnh của hành khách được chụp khi đăng ký đều là các ảnh chụp thẳng nên bộ dò tìm phân tầng thích hợp để sử dụng cho hiệu quả cao nhất là `haarcascade_frontalface_alt2.xml`, cũng đã được OpenCV hỗ trợ. Do vị trí của ảnh khuôn mặt trong hộ chiếu, cũng như trong ảnh được chụp lúc hành khách tiến hành làm thủ tục xuất nhập cảnh đều đã được căn chỉnh, nên phạm vi xác định ảnh khuôn mặt nhỏ, thời gian thực hiện nhanh và kết quả thu được có độ chính xác cao, đảm bảo cho quá trình trích chọn đặc trưng khuôn mặt và so khớp mẫu.

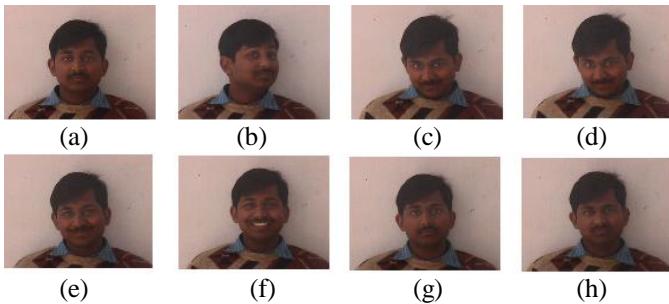
Ảnh khuôn mặt sau khi được phát hiện sẽ qua giai đoạn tiền xử lý, bao gồm chuyển từ ảnh màu sang ảnh xám, cân bằng histogram để làm giảm ảnh hưởng của sự thay đổi điều kiện ánh sáng, khung cảnh nền... lên kết quả trích chọn đặc trưng và xác thực.

Quá trình xác định ma trận vector chiếu được thực hiện nhờ hàm `cvCalcEigenObjects`. Hàm trả về các vector riêng và trị riêng của ma trận hiệp phương sai tương ứng với tập ảnh dữ liệu đầu vào. Giá trị các đặc trưng của ảnh khuôn mặt được tính bởi hàm `cvEigenDecomposite`.

Các đánh giá được thực hiện trên bộ cơ sở dữ liệu “The Indian Face Database” [6] và trên các ảnh chụp trong thực tế. Để ước tính tỉ lệ chấp nhận lỗi (False Accept Rate – viết tắt là FAR) ta so sánh ảnh trong hộ chiếu của một người với các ảnh chụp tại chỗ của 05 người khác cùng giới tính. Mỗi lần hệ thống trả lời Đúng thì tăng FAR lên 1 đơn vị. Để ước tính tỉ lệ từ chối lỗi (False Rejection Rate – viết tắt là FRR), ta sử dụng 03 trong số 07 ảnh của mỗi người để yêu cầu xác thực. Mỗi lần hệ thống trả lời Sai thì tăng FRR lên 1 đơn vị.

Bộ cơ sở dữ liệu “The Indian Face Database” bao gồm ảnh của 55 người (22 nữ và 33 nam). Các ảnh đều được chụp thẳng, khuôn mặt của người trong ảnh ở các góc độ khác nhau và biểu hiện cảm xúc khác nhau. Đối với mỗi người trong bộ cơ sở dữ liệu, ta sẽ lấy 01 ảnh làm ảnh hộ chiếu và 07 ảnh khác là các ảnh được chụp lúc tiến hành làm thủ tục xuất nhập cảnh.

Đối với các thử nghiệm với ảnh trong thực tế, hộ chiếu được quét bằng máy scan, còn ảnh hành khách khi đăng ký được chụp thông qua webcam của máy tính.

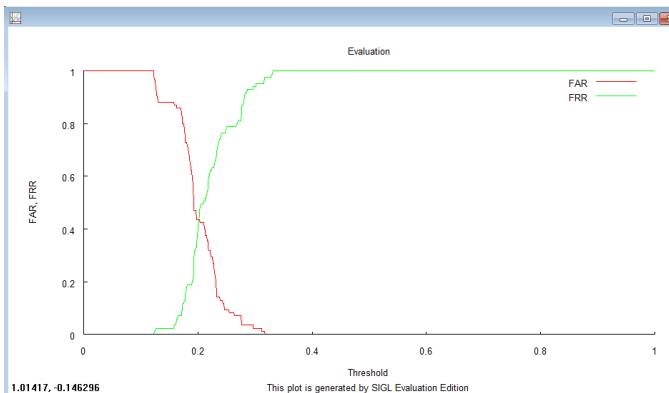


Hình 3: một ví dụ về ảnh trong bộ cơ sở dữ liệu “The Indian Face Database”. Ảnh (a) được dùng làm ảnh hộ chiếu, các ảnh (b), (c), (d), (e), (f), (g), (h) dùng làm các ảnh được chụp khi làm thủ tục xuất nhập cảnh.

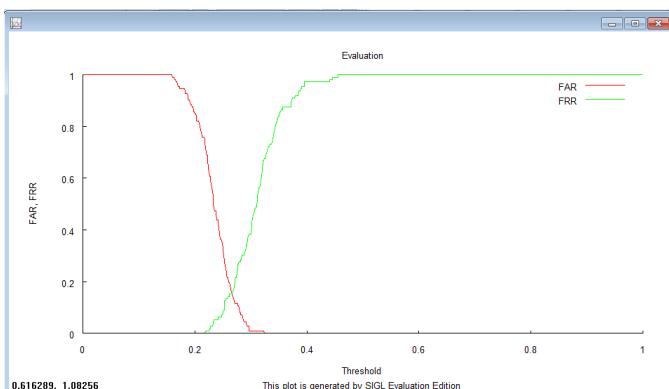
Kết quả thử nghiệm, đánh giá hệ thống xác thực ảnh hộ chiếu đã xây dựng được thể hiện ở bảng 2 và các hình 4, 5.

Cơ sở dữ liệu thử nghiệm	Bộ cơ sở dữ liệu “The Indian Face Database”	Ảnh chụp trong thực tế
Số lượng người trong CSDL	55 người: 22 nữ và 33 nam	17 người: 6 nữ và 11 nam
Số lượng mẫu thử	FAR: 275 mẫu FRR: 275 mẫu	FAR: 85 mẫu FRR: 85 mẫu
Error Equal Rate (EER)	0.266000	0.201333

Bảng 2: Kết quả thử nghiệm, đánh giá hệ thống



Hình 4: Đồ thị FAR và FRR trong thử nghiệm với ảnh chụp trong thực tế



Hình 5: Đồ thị FAR và FRR trong thử nghiệm với ảnh trong bộ cơ sở dữ liệu “The Indian Face Database”

Từ các bảng và hình vẽ trên, có thể thấy rằng có sự sai khác tương đối giữa kết quả thu được khi tiến hành thử nghiệm với cơ sở dữ liệu sẵn có (ảnh tĩnh, đã có cân chỉnh về độ sáng, khung cảnh nền ...) và với ảnh trong thực tế.

Với thử nghiệm trên ảnh tĩnh, hiệu năng của hệ thống là khá cao: tại giá trị EER (mà ta thường chọn làm giá trị ngưỡng), FAR và FRR xấp xỉ 15%. Tuy nhiên, với các thử nghiệm trên ảnh chụp trong thực tế, FAR và FRR tại giá trị EER lên đến gần 40%. Nguyên nhân của điều này là do sự “nhạy” của phương pháp trích chọn đặc trưng khuôn mặt sử dụng PCA với sự thay đổi điều kiện ánh sáng và khung nền. Để khắc phục, cải thiện chất lượng hệ thống, một số giải pháp có thể được áp dụng vào hệ thống:

- Giới hạn chặt chẽ điều kiện chụp ảnh hành khách khi làm thủ tục xuất nhập cảnh: điều kiện ánh sáng và khung nền khi chụp được giữ cố định. Với điều kiện hiện có ở các sân bay, yêu cầu này hoàn toàn có thể thực hiện.

- Thay vì đưa ra một giá trị ngưỡng cố định, hệ thống đưa ra một khoảng giá trị ngưỡng và cho phép nhân viên hải quan điều chỉnh trong phạm vi này. Trong hệ thống thử nghiệm, khoảng giá trị ngưỡng là giá trị $ERR \pm 0.05$.

Trong tương lai, để có thể triển khai hệ thống này vào trong hoạt động thực tế, hệ thống cần được bổ sung, sửa đổi thêm các tính năng:

- Cải tiến phương pháp trích chọn đặc trưng khuôn mặt sử dụng PCA.
- Sử dụng các camera có chức năng đánh giá chất lượng ảnh chụp và tự động căn chỉnh để ảnh khuôn mặt ở vị trí thích hợp.
- Bổ sung thêm thành phần cho phép tự động đọc các thông tin trong hộ chiếu (đặc biệt là mã hộ chiếu) và lưu vào cơ sở dữ liệu.

5. Kết luận

Hệ thống ứng dụng xác thực khuôn mặt vào kiểm tra ảnh hộ chiếu đã được xây dựng, thử nghiệm và bước đầu cho những kết quả khả quan. Tuy vẫn còn những điểm cần cải tiến, bổ sung, nhưng với những khảo sát cụ thể, đây đủ về yêu cầu nhiệm vụ của công tác làm thủ tục xuất nhập cảnh ở các đơn vị cụ thể, chắc chắn hệ thống sẽ có thể có triển vọng được ứng dụng, triển khai trong thực tế.

6. Lời tri ân

Báo cáo này được hoàn thành nhờ sự chi bảo, hướng dẫn tận tình của PGS.TS. Nguyễn Linh Giang, bộ môn Truyền thông và mạng máy tính, Viện Công nghệ thông tin và truyền thông, trường Đại học Bách Khoa Hà Nội. Em xin chân thành cảm ơn thầy và các thầy cô trong bộ môn Truyền thông và mạng máy tính đã truyền đạt cho em những kiến thức, kinh nghiệm quý báu. Em cũng xin cảm ơn các bạn lớp Tin Pháp K51 đã nhiệt tình giúp đỡ, cung cấp các ảnh chụp thử nghiệm trong thực tế.

Tài liệu tham khảo

- [1] Matthew Turk, Alex Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience Volume 3 Number 1, 1991.
- [2] Kwang-Baek Kim, Am-suk Oh, Young Woon Woo, "PCA - Based Face Verification and Passport Code Recognition Using Improved FKCN Algorithm", Eighth International Conference on Intelligent Systems Design and Applications, 2008.
- [3] Jonathon Shlens, "A Tutorial on Principal Component Analysis", version 3.01, April 22, 2009.
- [4] Paul Viola, Michael John, "Robust Real-time Object Detection", Second international workshop on statistical and computational theories of vision-modeling, learning, computing, and sampling, Vancouver, Canada, 2001.
- [5] "Object Recognition Reference", Université Claude Bernard Lyon 1,
http://www710.univ-lyon1.fr/~bouakaz/OpenCV-0.9.5/docs/ref/OpenCVRef_ObjectRecognition.htm
- [6] Vudit Jain, Amitabha Mukherjee. The Indian Face Database.
<http://vis-www.cs.umass.edu/~vidit/IndianFaceDatabase/>, 2002

Hệ thống thu thập tài liệu theo chủ đề cho tiếng Việt

Nguyễn Xuân Hòa

Tóm tắt - Trong khuôn khổ nghiên cứu của mình, bài báo này đề xuất một phương pháp phân tích và dự đoán các liên kết dẫn đến tài liệu đúng chủ đề mà không cần phải tải nội dung của nó về. Sau khi lựa chọn được những liên kết có chất lượng tốt nhất theo điểm số, một phương pháp lựa chọn động sẽ tiếp tục được áp dụng để lựa chọn một lần nữa ra các liên kết ưu việt nhất trong số các tài liệu “tốt nhất” đó. Các kết quả thử nghiệm đã cho thấy, phương pháp này đem lại hiệu quả vượt trội so với một số phương pháp thu thập tài liệu hiện tại.

Từ khóa - Dynamic Crawler, Reinforcement Learning, SVM, Vertical Search Engine, Vietnamese Word Segmentation.

1. GIỚI THIỆU

Ngày nay, hầu hết các thông tin đều đã được đưa lên mạng Internet. Điều này một phần giúp chúng ta có thể dễ dàng tiếp xúc với các nguồn thông tin này hơn, nhưng mặt khác, nó lại vô tình tạo ra một thách thức lớn cho chúng ta, những người đi tìm kiếm thông tin. Thực vậy, khi có quá nhiều nguồn thông tin, các khó khăn gặp phải trong việc tìm ra đâu là nguồn thông tin có ích nhất, đáng tin cậy nhất là khó tránh khỏi. Chính vì vậy, chúng ta luôn mong muốn có một máy tìm kiếm có tính hiệu quả cao, giúp chúng ta tìm đến nguồn thông tin mà ta cần một cách nhanh chóng và chính xác.

Trên thực tế, Google cũng đã làm rất tốt điều này, nhưng, do Google là một máy tìm kiếm đa mục đích (general purpose) và đa ngôn ngữ, nên những tài liệu mà nó thu về, thường có nội dung dàn trải, và, hiệu quả tìm kiếm cũng thay đổi phụ thuộc vào ngôn ngữ được sử dụng để tìm kiếm. Yếu tố đa mục đích được thể hiện qua ví dụ về từ khóa “book”, nếu chúng ta đưa vào từ khóa này, kết quả trả về có thể đem đến các tài liệu với hai chủ đề khác nhau, là đặt chỗ du lịch, và sách. Việc tìm kiếm cùng một chủ đề xác định trên các ngôn ngữ khác nhau, cũng đem lại hiệu quả tìm kiếm khác nhau. Ví dụ, với ngôn ngữ tìm kiếm là tiếng Anh và tiếng Việt, thường thì kết quả tìm kiếm tiếng Việt sẽ ít chính xác hơn.... Điều này là dễ hiểu, bởi tiếng Anh phổ biến hơn rất nhiều so với tiếng Việt, nhưng, nó lại vô tình gây khó khăn cho những người Việt khi họ muốn tìm kiếm thông tin...

Vấn đề đặt ra ở đây là, cần phải có một máy tìm kiếm theo chủ đề chuyên sâu, với độ chính xác cao và chi phí thấp dành riêng cho người Việt. Với mục đích giúp chúng ta xây

Công trình này được thực hiện dưới sự hướng dẫn của thầy giáo Trần Đức Khanh, giảng viên bộ môn Hệ Thống Thông Tin, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội

Nguyễn Xuân Hòa, sinh viên lớp Tin Pháp, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0987731789, e-mail: nxhoaf@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

dụng được một máy tìm kiếm tiếng Việt đáp ứng được các tiêu chí trên thông qua việc xây dựng một bộ thu thập tài liệu có hiệu quả và độ chính xác cao, nghiên cứu này đã được tiến hành.

Phần còn lại của bài báo được tổ chức như sau: Phần hai đề cập đến các nghiên cứu liên quan đã được tiến hành. Phần ba sẽ mô tả các kỹ thuật dự đoán tài liệu trả về. Trên cơ sở những kỹ thuật đó, phần bốn sẽ trình bày cách thức xây dựng hệ thống thu thập tài liệu, cùng với những hệ thống con của nó. Tiếp đến, trong phần năm, chúng ta sẽ tiến hành thử nghiệm và đánh giá các kết quả thu được của hệ thống, trên cơ sở các nhận định, so sánh với các hệ thống đã có để rút ra những ưu, nhược điểm của hệ thống. Và cuối cùng phần tổng kết sẽ đem đến cho người đọc những nhận định khái quát, tổng hợp nhất, cùng với những hướng phát triển tiếp theo của đề tài.

2. MỘT SỐ NGHIÊN CỨU LIÊN QUAN

Các nghiên cứu về máy tìm kiếm theo chủ đề đã được tiến hành từ trước đó.

Vấn đề về hệ thống thu thập tài liệu tập trung đã được Chakrabart et al [1] đề xuất trước đó vào năm 1999. Hệ thống này có ba thành phần chính là crawler, classifier và distiller. Nó sử dụng một tập tài liệu đã được phân loại sẵn để xác định các chủ đề có liên quan và các chủ đề không liên quan của tài liệu thu thập được. Tới năm 2002, phiên bản cải tiến với nhiều tính năng vượt trội hơn tiếp tục được ông đề xuất trong [2]. Ông cũng là người đưa ra công thức xác định hiệu quả thu thập được dùng đến khá phổ biến sau này.

$$\text{Harvest rate} = \frac{N_R}{N} \quad (1)$$

Hệ thống thu thập thông minh (Intelligent crawling) [3] là phương pháp thu thập tài liệu khác sử dụng kỹ thuật học tăng cường để đo tính liên quan tới chủ đề của các tài liệu trong quá trình thu thập. Hệ thống này cũng có khả năng dự đoán độ xa của các tài liệu tới nguồn tài liệu đúng chủ đề.

Liên quan đến cách thức cho điểm, tài liệu, Grigoriadis [4] đã nhận định, các tài liệu thường có liên quan chặt chẽ về mặt ngữ nghĩa tới những tài liệu mà nó trỏ tới. Dựa trên nhận định này, ông đã xây dựng một hệ thống cho điểm tài liệu mà chưa cần tải về, dựa trên điểm số của các trang web trỏ tới nó.

Với máy thu thập tài liệu cho chủ đề tiếng Việt, Phương pháp của Dinh-Thi Vu et al [7] cũng sử dụng kỹ thuật học tăng cường để dự đoán tài liệu đúng chủ đề. Mặt khác, do quá trình xử lý được thực hiện trên tiếng Việt, nên hệ thống còn thêm vào một số xử lý tiếng Việt khác nữa.

3. PHÂN TÍCH VÀ DỰ ĐOÁN URL

Khả năng phân tích và dự đoán các URL của hệ thống dựa trên hai yếu tố chính, đó là phân tích nội dung của trang web

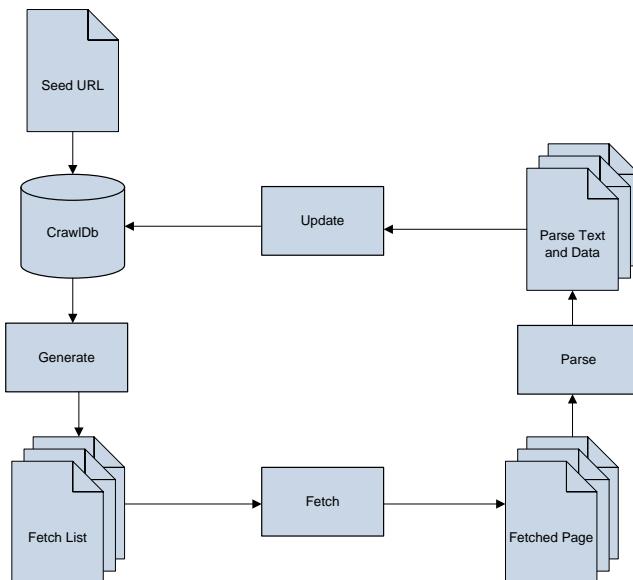
chứa URL đó, cũng như nội dung anchor text¹ của URL đó. Các nghiên cứu trước đó cũng đã chứng minh tính đúng đắn của lập luận này. Theo Grigoriadis [4], các tài liệu thường có liên quan chặt chẽ về mặt ngữ nghĩa tới những tài liệu mà nó trỏ tới. Hơn thế nữa, bản thân các anchor text của các link cũng mang một thông tin nào đó về nội dung trang web mà nó trỏ tới.

Ngoài ra, để biết nội dung đoạn văn bản đó là đúng, hay sai, hay cách tài liệu đúng chủ đề một khoảng bao nhiêu, hệ thống cần có một module phân loại văn bản. Module phân loại văn bản muốn có khả năng phân loại văn bản vào các chủ đề khác nhau thì cần phải có một tập tài liệu mẫu. Học tăng cường được sử dụng đến như một lời giải cho bài toán xây dựng tập tài liệu mẫu này. Và cuối cùng, do toàn bộ hệ thống hoạt động trên các trang web tiếng Việt, vốn có cấu trúc ngữ pháp và từ vựng đặc thù, nên nó cần được tích hợp thêm khả năng tách từ tiếng Việt.

Với những cơ sở lý thuyết này, cùng với các nhận định thực tế trước đó, chúng ta đã có những điều kiện cần, cả về thực tế lẫn lý thuyết để xây dựng nên một hệ thống hoàn chỉnh. Phần tiếp theo đây sẽ trình bày về vấn đề này.

4. HỆ THỐNG DYNAMIC CRAWLER

4.1. KIẾN TRÚC HỆ THỐNG



Hình 1: Tổng quan về hệ thống

Hình 1 mô tả kiến trúc tổng quan của hệ thống, hệ thống hoạt động theo các bước sau: Đầu tiên, các link hạt giống (Seed URL) sẽ được đưa vào để khởi tạo CSDL WebDB, tiếp theo, một danh sách các URL dự định tải về được sinh ra (fetchList) dựa trên các URL tốt nhất² có trong CrawlDB. Các URL này sau đó sẽ được chính thức tải về và phân tích nội dung. Cuối cùng chúng ta sẽ cập nhật vào CrawlDB những URL vừa thu được từ quá trình phân tích nội dung trên. Quá trình này sẽ được lặp đi lặp lại đến khi một điều kiện dừng nào

¹ Trong ví dụ sau đây thì anchor text là “My Web Page” và URL là “mywebpage.html”

My Web Page

² Đây cũng chính là mấu chốt của việc xây dựng bộ phân loại văn bản

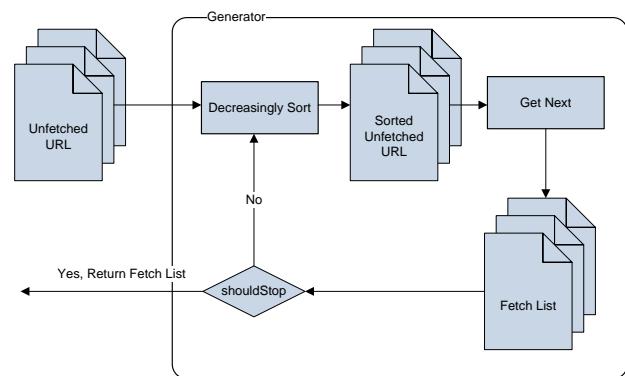
đó được thỏa mãn (đạt đến độ sâu cần thiết, hoặc không còn tài liệu trả về)

Về mặt kiến trúc hệ thống bao gồm ba thành phần cơ bản sau: Bộ tải dữ liệu (fetcher), một hàng đợi ưu tiên chứa các tài liệu sẽ được tải về (fetchList) và một hàm cho điểm (scoring function). Phần tiếp theo sẽ lần lượt mô tả chi tiết ba thành phần đó.

4.2. MODULE LỰA CHỌN CÁC URL TIỀM NĂNG

Chức năng: Module này có liên quan chặt chẽ nhất đến thành phần thứ hai: hàng đợi chứa các dữ liệu sẽ được tải về. Và nó cũng chính là phần quan trọng nhất quyết định hiệu năng của hệ thống, mặc dù phương pháp cho điểm chỉ là sự kết hợp của các phương pháp cho điểm trước đó.

Phân tích chi tiết: Hình 2 là module sinh danh sách các tài liệu trả về cho bước tiếp theo. Nhiệm vụ của module này là lấy dữ liệu đầu vào từ CrawlDb dưới dạng các tài liệu chưa được duyệt ở bước trước, sắp xếp các tài liệu đó theo thứ tự giảm dần của điểm số. Tới đây, chúng ta thu được một tập các tài liệu đã được sắp xếp. Tiếp theo, chúng ta sẽ thực hiện một vòng lặp, làm nhiệm vụ lần lượt lấy từ cao xuống thấp các phần tử trong Sorted Unfetched URL, cho vào fetchList, sau đó, kiểm tra xem, liệu có lấy thêm phần tử để cho vào fetchList nữa không. Hàm shouldStop kiểm tra một số điều kiện sau đây:



Hình 2: Lựa chọn các URL động

- Nếu điểm số vừa thu được vẫn ở trên ngưỡng điểm s_0 , thì ta sẽ tiếp tục thu thập tài liệu.
- Nếu điểm số vừa thu thập được ở dưới ngưỡng s_0 thì ta sẽ không lấy đủ topN phần tử mà tiến hành kiểm tra:
 - Nếu dung lượng hiện tại của fetchList (c) đã đạt một ngưỡng dung lượng c_0 (hiển nhiên, do ta lấy từ cao xuống thấp, nên trong fetchList lúc này sẽ toàn những tài liệu trên ngưỡng) thì ta sẽ không thu thập thêm tài liệu nữa, mà sẽ lấy fetchList được điều chỉnh $c\%$ này để di tài dữ liệu, với hi vọng số dữ liệu này lại đem lại nhiều tài liệu đúng chủ đề hơn.
 - Nếu dung lượng fetchList vẫn chưa đạt tới c_0 , chúng ta sẽ chỉ lấy tiếp cho tới khi nó đạt tới c_0 rồi dừng lại

³ Những điểm số cao hơn ngưỡng đảm bảo khả năng tài liệu thu thập về sẽ đúng chủ đề là rất cao

Với cách lựa chọn như trên, thì ngay cả trong trường hợp tồi nhất, số link mà ta chưa chắc là đem lại tài liệu đúng chủ đề cũng chỉ là $c_0\%$ mà thôi. Ở đây, có thể người đọc sẽ đặt câu hỏi là, liệu lấy ít như thế, thì có thể sau một vài bước, số link thu thập được sẽ ít dần đi và hết. Thực tế không phải vậy, bởi số link chưa được lấy vẫn nằm trong crawlDb và có thể sẽ được lấy về ở các bước sau. Quá trình của chúng ta chỉ **làm chậm lại quá trình thu thập tài liệu** khi chúng ta không chắc chắn các đường link tương ứng dẫn tới tài nguồn liệu đúng chủ đề.

Ta có đoạn pseudo code minh họa sau đây:

```

if (score > s0) then
    add link to fetchList
else
    if ( c < c0) then
        add link to fetchList
    else
        return
    endif
endif

```

s_0, c_0 được thiết lập trong quá trình chạy hệ thống. Các thử nghiệm ở phần 5 đặt $s_0 = 0.5$ và $c_0 = 0.15$, còn c được tính theo công thức:

$$C = \frac{n}{topN} * 100\% \quad (2)$$

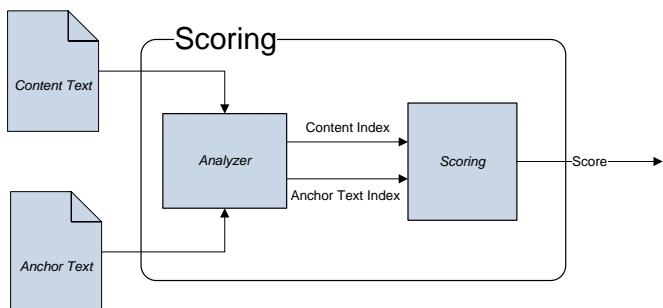
Ở đây:

c là ngưỡng dung lượng hiện tại, n là số tài liệu hiện tại trong fetchList
 $topN$ là số tài liệu tối đa mà fetchList có thể chứa được

Như vậy, với cách lựa chọn này, số lượng các tài liệu sẽ được tải về là **không cố định** cho từng chu kỳ khác nhau của hệ thống (xem Hình 1) với mong muốn đảm bảo rằng, tỉ lệ URL mà ta còn đang “nghi ngờ” về khả năng đem lại tài liệu đúng chủ đề luôn vượt quá một ngưỡng nào đó.

Thực nghiệm trên một số trang web cho thấy (xem thêm 5.2), trong quá trình dữ liệu được tải về, dung lượng của fetchList cũng biến thiên, không phải lúc nào cũng đạt 100%. Điều này càng chứng tỏ thêm tính đúng đắn của phương pháp.

4.3. MODULE CHO ĐIỂM



Hình 3: Module cho điểm

Chức năng: Module này tương ứng với thành phần thứ ba: hàm cho điểm (scoring function). Tiêu chí để đánh giá một URL là tốt hay không tốt chính là điểm số. Chính vì vậy, cách thức cho điểm ảnh hưởng rất lớn tới hiệu năng của toàn hệ thống. Một cách lựa chọn điểm số tốt sẽ đem lại khả năng thu về các tài liệu đúng chủ đề cao. Module này làm nhiệm vụ cho điểm các URL được tìm thấy trên trang web hiện tại, điểm số

được đánh giá dựa trên nội dung của văn bản hiện tại, và nội dung của anchor text của URL đó (xem thêm phần 3) với hi vọng rằng, những URL có điểm số cao sẽ đảm bảo khả năng thu về tài liệu đúng chủ đề cao.

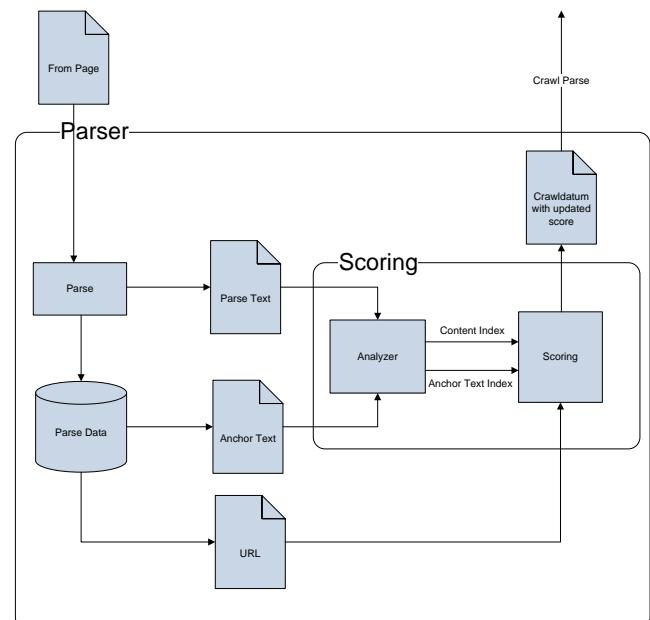
Ngoài ra, do được tích hợp trực tiếp vào module phân tích văn bản, nên module còn làm thêm nhiệm vụ loại bỏ những đường liên kết được dự đoán là có khả năng đem lại tài liệu đúng chủ đề nhưng qua quá trình phân tích thực tế nội dung, nó lại dẫn tới nhiều tài liệu không đúng chủ đề. Việc dự đoán trước và loại bỏ những hướng đi không tốt này, cũng góp phần cải thiện đáng kể thời gian thực thi của hệ thống. Điều này sẽ được trình bày chi tiết ngay dưới đây.

Phân tích chi tiết: Hình 3 là module cho điểm URL được tải về. Như đã phân tích ở trên, điểm số của một URL được tính theo công thức (3)

$$score_{URL} = \frac{\alpha_1 * score_{content} + \alpha_2 * score_{anchor}}{\alpha_1 + \alpha_2} \quad (3)$$

Ở đây:

- α_1, α_2 lần lượt là trọng số của các điểm số $score_{content}$ và $score_{anchor}$.
- $score_{content}$ và $score_{anchor}$ lần lượt là điểm của nội dung trang web hiện tại và điểm của anchor text của URL chứa trong trang web đó



Hình 4: Module cho điểm bên trong module phân tích văn bản

Trong số sẽ phụ thuộc vào tính chất của từng trang web khác nhau, bởi có những trang web thì nội dung của nó có quyết định nhiều hơn đến nội dung của các link mà nó trỏ đến (xem thêm các phân tích trong 5.2), nhưng có những trang web thì ngược lại, tức là, anchor text của link mà nó trỏ đến lại giàu thông tin hơn nội dung hiện tại của trang web.

Như đã trình bày ở trên, module cho điểm được tích hợp trực tiếp vào module phân tích nội dung văn bản trả về (xem Hình 4) nên chúng ta có thể sử dụng module này để quyết định

xem, liệu hướng đi mà ta đang phân tích có khả thi không, để đưa ra quyết định phù hợp: nên tiếp tục hay dừng lại.

Tại một thời điểm nào đó, khi đang phân tích nội dung của một trang web, module này sẽ ghi nhận lại số lần nó đánh điểm 0 tuyệt đối, theo (3), cho outlink của nó.

Dưới đây là giải thuật quyết định xem, liệu chúng ta có tiếp tục phân tích tiếp không.

```

if (scoreURL == 0) then
    zeroScoreCounter++;
if (shouldStop)
then
    Stop and Analyze other
Else
    Continue analyzing the current page

Boolean shouldStop()
{
    If (zeroScoreCounter) > topN * γ
        Return true
    Else
        Return false
}

```

Ngay sau khi gặp một outlink có điểm 0 tuyệt đối, module này kiểm tra xem liệu số điểm 0 đã vượt quá một ngưỡng cho phép nào đó chưa. Ngưỡng này được tính bằng tích của hệ số tỉ lệ và dung lượng cực đại của fetchList. Trong cài đặt, hệ thống được thiết lập với ngưỡng $\gamma = 0.2$, tức là, khi số điểm 0 tuyệt đối đã vượt quá 20% dung lượng cực đại của fetchList, chúng ta sẽ không tiếp tục phân tích theo hướng này nữa.

Hệ số γ được định nghĩa như sau:

$$\gamma = \frac{\text{zeroScoreCounter}}{\text{topN}} * 100\% \quad (4)$$

Và được gọi là *hệ số dừng*.

4.4. MODULE TẢI DỮ LIỆU VỀ

Chức năng: Module này lấy đầu vào là danh sách các URL có trong fetchList và làm nhiệm vụ tải về dữ liệu tương ứng với các URL này.

Phân tích chi tiết: Module này được cài đặt theo cơ chế song song. Điều đó giúp cho hệ thống có thể đồng thời tải nhiều URL về cùng một lúc. Module này lại được tách thành hai module con nữa, phối hợp hoạt động với nhau.

- Module cung cấp dữ liệu, nó làm nhiệm vụ đọc nội dung hiện tại của fetchList để lấy danh sách các URL và cho vào hàng đợi để module tải dữ liệu lấy ra và tải dữ liệu về.
- Module tải dữ liệu gồm nhiều luồng hoạt động đồng thời, chúng sẽ cùng nhau lấy dữ liệu ra khỏi hàng đợi, tiến hành tải về.

Tới đây, chúng ta đã đi từ việc phân tích yêu cầu thực tế dẫn đến sự ra đời của nghiên cứu này, cho đến những khía cạnh lý thuyết được đưa ra, rồi những hệ thống con được xây dựng trên những khía cạnh lý thuyết đó, để đáp ứng những yêu cầu đó. Và bây giờ sau khi các thành phần đã được ghép nối lại với nhau, tạo nên bức tranh cuối cùng: Hệ thống hoàn thiện, chúng ta sẽ tiến hành một số thử nghiệm.

5. THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ¹

Cáu hình được sử dụng để thử nghiệm là:

- CPU: 3.16GHz x 2
- RAM: 2GB
- OS: Linux Ubuntu 10.10

Các phương pháp thu thập tài liệu được sử dụng là

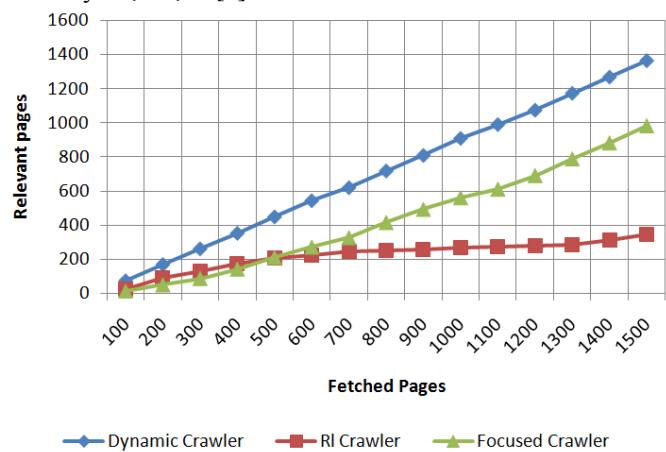
- Dynamic Crawler: là phương pháp thu thập tài liệu mà tôi đề xuất trong khuôn khổ nghiên cứu của mình.
- Reinforcement Crawler: là phương pháp thu thập tài liệu dựa trên học tăng cường được xây dựng và phát triển bởi [7]
- Focused Crawler: phương pháp thu thập tài liệu tập trung

Trên cơ sở đó, tôi tiến hành hai hình thức thử nghiệm chính, thứ nhất, thử nghiệm về khả năng thu thập tài liệu của hệ thống, và thứ hai, thử nghiệm để kiểm tra sự biêt thiên dung lượng của fetchList trong quá trình thu thập dữ liệu

5.1. KHẢ NĂNG THU THẬP TÀI LIỆU CHỦ ĐỀ BÓNG ĐÁ

Thử nghiệm đầu tiên trong loạt thử nghiệm về khả năng thu thập tài liệu được tiến hành với chủ đề về bóng đá trên trang web <http://vnexpress.net>. Hình 5 là kết quả của quá trình thử nghiệm.

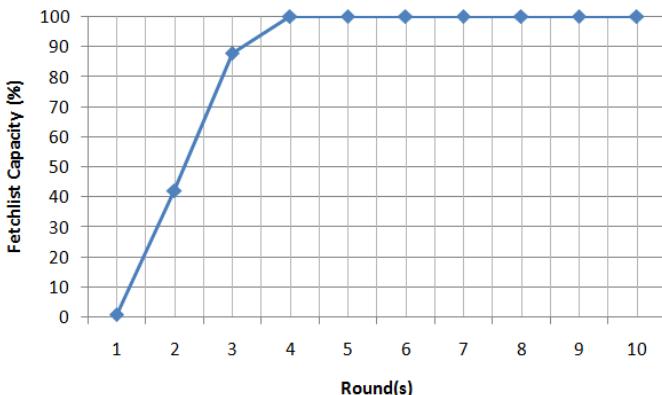
Trong suốt quá trình thử nghiệm, hiệu năng của hệ thống luôn tỏ ra vượt trội, với hiệu quả thu thập tính theo công thức (1) luôn giữ ở mức ổn định. Sự biến đổi của FetchList (xem Hình 6) cho thấy, sau khoảng thời gian ban đầu tìm kiếm, các crawler đã tìm được mỏ tài liệu. Tới đây, focused Crawler, bằng cơ chế đánh đồng điểm, đã tỏ ra vượt trội so với RI Crawler, đúng như những gì mà các tác giả của hệ thống tìm kiếm này nhận định [7].



Hình 5: Thử nghiệm trên chủ đề bóng đá

Còn đối với hệ thống của chúng ta, do nó được xây dựng dựa trên sự kết hợp giữa học tăng cường, cùng với cơ chế hai lần lựa chọn điểm số, thông qua cả điểm số tương đối giữa các URL với nhau (tìm topN url có điểm cao nhất) lẫn lựa chọn tuyệt đối với một điểm chuẩn cụ thể s_0 (xem phần 4.2), nó luôn đảm bảo khả năng đem lại tài liệu đúng chủ đề cao nhất.

¹ Để đảm bảo tính minh bạch, các kết quả thử nghiệm đều được lưu ra file log để tiện đối chứng.

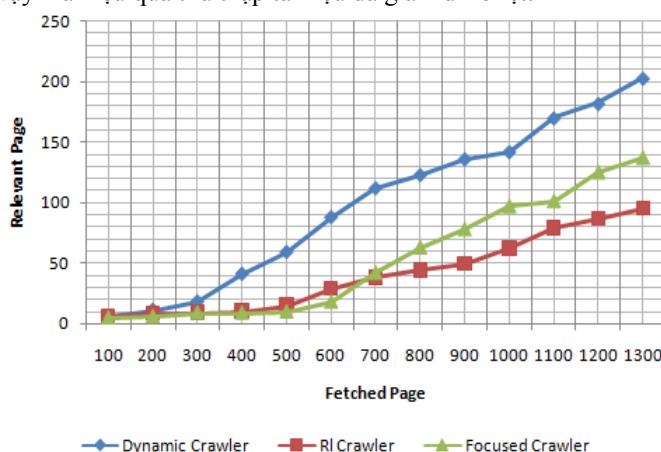


Hình 6: Dung lượng fetchList trong thử nghiệm về bóng đá

Sau khi tiến hành thử nghiệm đầu tiên với chủ đề là bóng đá, chúng ta tiếp tục tiến hành một thử nghiệm nữa về khả năng thu thập tài liệu của hệ thống với chủ đề về tuyên dụng CNTT. Đây chính là thử nghiệm chính, nó hứa hẹn sẽ đem lại cho người đọc nhiều nhận xét và khám phá thú vị...

CHỦ ĐỀ TUYÊN DỤNG CÔNG NGHỆ THÔNG TIN

Thử nghiệm thứ hai được thực hiện trên trang web <http://hn.24h.com.vn/> với chủ đề tuyên dụng CNTT, có thể thấy, đây là một chủ đề có tính nhập nhằng cao, hơn nữa, bản thân chủ trang web này thường xuyên có những đường link ít liên quan đến chủ đề theo kiểu thu hút người xem, chính vì vậy mà hiệu quả thu thập tài liệu đã giảm đi rõ rệt.



Hình 7: Thử nghiệm trên chủ đề tuyên dụng CNTT

Hình 7 ghi lại số tài liệu đúng chủ đề cho cả ba phương pháp : DynamicCrawler: là hệ thống đang được cài đặt, RICrawler, là hệ thống được cài đặt dựa trên kỹ thuật học tăng cường thuần túy, và FocusedCrawler, là hệ thống được cài đặt dựa trên kỹ thuật tìm kiếm theo chủ đề.

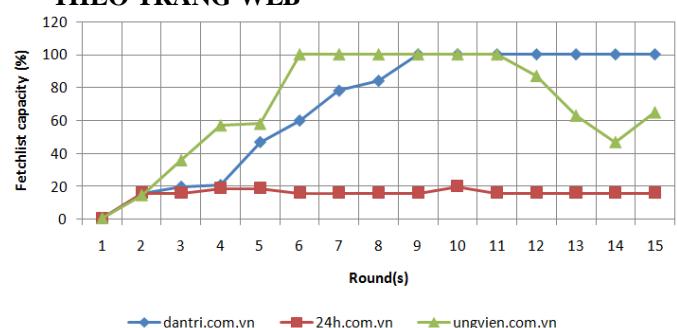
Trong suốt quá trình thử nghiệm, phần tử thu thập tài liệu động luôn thể hiện hiệu năng vượt trội của mình. Hiệu năng này thu được chính là do sự mềm dẻo trong dung lượng của fetchList (chi tiết xem thêm 5.2). Trong một môi trường mà nguồn tài liệu đúng chủ đề là ít, thì việc các phương pháp kia luôn sử dụng một fetchList cứng nhắc, với số lượng URL bên trong nó luôn là cực đại, tức là, chỉ quan tâm đến điểm số tương đối giữa chúng để chọn mà không quan tâm liệu những “tài liệu có điểm số cao nhất” ấy có thực sự đem lại một trang web đúng chủ đề hay không, thì kết quả thu được như trên là

điều dễ hiểu. Trong khi đó, với cơ chế động của tôi, phần tử thu thập tài liệu luôn đảm bảo rằng, ngoài những URL bên trong fetchList là có điểm số tốt nhất (loại bỏ tương đối giữa các URL với nhau), nó còn thực hiện việc loại bỏ một lần nữa chính các URL có điểm số tốt này, nhưng vẫn chưa vượt quá ngưỡng để đem lại tài liệu đúng chủ đề (loại bỏ tuyệt đối các URL còn lại với một ngưỡng số cụ thể như đã đề cập ở trên).

Hình 7 một lần nữa đã thể hiện rất đúng những gì được kết luận trong [7], đó là, hệ thống thu thập tài liệu dựa trên học tăng cường sau một thời gian hoạt động sẽ bị chính hệ thống thu thập tài liệu chuyên sâu vượt qua. Cần lưu ý là tại [7], những thử nghiệm đã được tiến hành trên chính trang web này, với những điều kiện tương tự.

Kết thúc loạt thử nghiệm thứ nhất, chúng ta sẽ tiếp tục tiến hành những thử nghiệm để kiểm tra sự biến thiên dung lượng của fetchList trong quá trình thu thập dữ liệu.

5.2. BIẾN THIỀN DUNG LƯỢNG CỦA FETCHLIST THEO TRANG WEB



Hình 8: Biến thiên dung lượng fetchlist theo trang web với chủ đề về tuyên dụng CNTT

Thử nghiệm này được tiến hành nhằm xem xét sự biến thiên dung lượng của fetchlist trong suốt quá trình hoạt động của hệ thống. Chủ đề được lựa chọn vẫn là tuyên dụng công nghệ thông tin, và các trang web là dantri.com.vn, 24h.com.vn và ungvien.com.vn.

Kết quả thử nghiệm được thể hiện trong Hình 8. Quan sát kết quả thử nghiệm, tôi xin rút ra một số nhận định sau:

Đối với trang ungvien.com.vn, do đây là một trang web liên quan trực tiếp đến đề tài tuyên dụng nên hệ thống thu thập tài liệu nhanh chóng tìm đến nguồn tài liệu đúng chủ đề, chính vì vậy mà dung lượng fetchList được làm đầy lên rất nhanh. Sau một thời gian ở trong mờ link, phần tử thu thập thông tin lại rời khỏi mờ link. Kết quả thu được từ file log tại vòng thứ 13 và 14 cho thấy, lý do kết quả giảm là do, các đường link đúng chủ đề thường dẫn tới các công ty tuyên dụng, và nội dung trang web của các công ty tuyên dụng này thường không hẳn chỉ có chủ đề tuyên dụng CNTT.

Đối với trang web dantri.com.vn, dung lượng của fetchList tăng lên rất từ từ, và sau đó giữ ổn định ở mức tối đa. Điều này cho thấy, trong giai đoạn đầu, phần tử thu thập thông tin (crawler) của hệ thống vẫn đang tiến hành thăm dò, và tìm ra mờ link, sau đó, nó đã tìm thấy mờ link, thông qua dấu hiệu dung lượng của fetchlist liên tục được làm đầy. Tính đến cuối bài thử nghiệm, nó hiện vẫn đang ở trong mờ link.

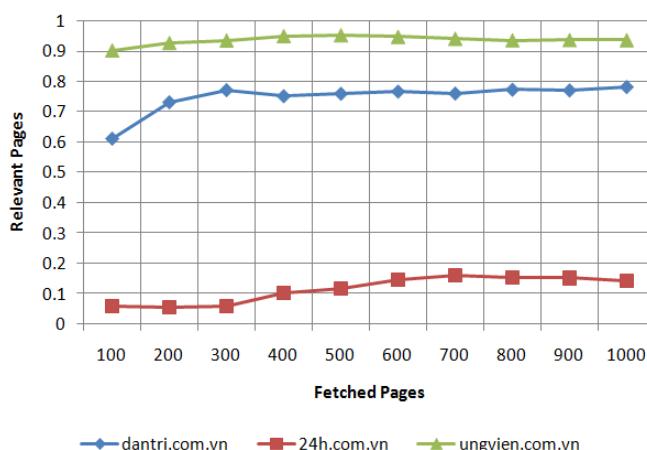
Đối với trang 24h.com.vn, trong suốt giai đoạn thử nghiệm, dung lượng của fetchList luôn được duy trì ở mức tối thiểu. Điều này chứng tỏ, crawler vẫn đang hoạt động một cách thăm dò, với hi vọng tìm ra mờ link ở những bước sau.

Thực tế khảo sát cho thấy, nguồn tài liệu đúng chủ đề chỉ cách trang web khoảng 3-4 đường link, nhưng phần tử thu thập thông tin vẫn chưa thể tìm ra. Lý do là, cấu trúc của trang web 24h.com.vn có nhiều định dạng (flv, mp3, jpg...) hơn hẳn các file khác. Hơn nữa, thực tế cho thấy, trang web này lại thường đặt những tiêu đề ít liên quan đến nội dung, điều này vô hình gây khó khăn cho phần tử thu thập thông tin của chúng ta.

Tới đây, người đọc có thể sẽ thắc mắc, liệu dung lượng của fetchList có ảnh hưởng gì đến hiệu quả thu thập theo công thức (1) hay không. Câu trả lời hoàn toàn ngược lại, hiệu quả thu thập không những luôn được giữ ở mức ổn định (xem Hình 9) mà sự ổn định này còn do sự biến thiên trong fetchList đem lại. Đây cũng chính là lý do mà tôi đặt tên cho phần tử thu thập tài liệu của mình là **"Phân tử thu thập tài liệu động – Dynamic crawler"**.

Như vậy, một đóng góp nữa của hệ thống, đó là thông qua sự biến thiên dung lượng của fetchList, chúng ta có thể biết được vị trí hiện tại của crawler. Qua đó, biết được liệu nó có đang ở trong vùng tài liệu đúng chủ đề hay không, và quan trọng hơn, trả lời được câu hỏi về cấu trúc liên kết của trang web mà ta đang kiểm tra. Thực vậy, bằng việc quan sát sự biến đổi của fetchList (xem thêm Hình 8) chúng ta sẽ dự đoán phần nào được cấu trúc liên kết trang web phía dưới. Ở đây, qua việc quan sát fetchList, ta có nhận xét như sau:

- Trang web ungvien.com.vn (vốn là một trang web tuyển dụng) cho thấy, các trang chứa chủ đề về tuyển dụng CNTT nằm khá gần so với trang chủ. Nhưng sau đó, khi dung lượng fetchList giảm đi, ứng với thực tế là, các tuyển dụng này thường chứa đường link tới công ty đăng tuyển.



Hình 9: Tỉ lệ tài liệu liên quan theo số lượng tài liệu trả về

- Với trang dantri.com.vn, các tài liệu đúng chủ đề nằm xa so với trang gốc một chút. Tuy nhiên, nguồn tài liệu đúng chủ đề vẫn đang khá ổn định. Điều này được lý giải là do, đây là một trang đưa tin, nên các đường link thường có xu hướng đi đến các thông tin tuyển dụng khác liên quan.
- Với trang 24h.com.vn, dung lượng fetchList luôn ở mức thấp, điều này cho thấy phần tử thu thập thông tin hoạt động không hiệu quả trên trang web này. Như đã nhắc đến ở trên, trang web này thường có nhiều định dạng file (flv, mp3, avi...) gây khó khăn

cho quá trình phân tích. Hơn nữa, các tiêu đề của nó cũng thường mang nội dung không sát với nội dung trang web đó (anchor text ít mang ý nghĩa).

Như vậy, trong suốt bài báo này tôi đã trình bày từ các yêu cầu thực tiễn dẫn đến sự ra đời của hệ thống, cho đến các cơ sở lý thuyết được sử dụng. Và tiếp theo đó là các cài đặt dựa trên cơ sở lý thuyết đó. Những thử nghiệm cũng đã được tiến hành để đánh giá tính năng hoạt động của hệ thống. Cuối cùng, tôi xin đưa ra những điểm tóm lược nhất của bài báo, đồng thời đề xuất hướng phát triển tiếp theo.

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo này đã trình bày với các bạn một phương pháp giải quyết cho bài toán thu thập tài liệu theo chủ đề cho những trang web tiếng Việt. Hai đóng góp chính của bài báo này, đó là:

- Đề xuất một phương pháp thu thập tài liệu mới với nhiều ưu điểm hơn so với các phương pháp thu thập tài liệu hiện có
- Đề xuất một phương pháp giúp chúng ta có thể hiểu hơn về cấu trúc liên kết của trang web phía dưới

Tuy nhiên, hệ thống hiện tại mới chỉ có chức năng thu thập các tài liệu trên bề mặt web (surface web). Trên thực tế, còn một nguồn thông tin rất lớn nữa mà nó chưa có khả năng thu thập (deep web). Trong quá trình nghiên cứu tiếp theo, tôi sẽ cố gắng đưa thêm module thu thập các nguồn thông tin deep web cho hệ thống. Hướng đi này hiện vẫn đang được tôi phát triển và cho kết quả khá quan.

Cuối cùng, tôi hi vọng đề tài nghiên cứu này như là một đóng góp vào những nghiên cứu liên quan đến tài về máy tìm kiếm theo chủ đề, qua đó, phần nào, đặt nền móng lý thuyết cho việc ra đời những sản phẩm có tính ứng dụng sau này.

7. TÀI LIỆU THAM KHẢO

- [1] Chakrabarti, S., van den Berg, M., and Dom, B. (1999). "Focused crawling: a new approach to topic-specific web resource discovery". Computer Networks, 31(11–16):1623–1640.
- [2] Chakrabarti, S., Punera, K., & Subramanyam, M. "Accelerated focused crawling through on-line relevance feedback". Proc. of 11th Intl. Conf. on World Wide Web, 148 – 159, 2002.
- [3] Aggarwal, F. Al-Garawi, and P. S. Yu. "Intelligent crawling on the world wide web with arbitrary predicates". In WWW '01: Proceedings of the 10th international conference on World Wide Web, 2001
- [4] Alexandros Grigoriadis, Georgio Paliouras. *Focused Crawling Using Temporal Diference-Learning*, In Proceedings of SETN. 2004
- [5] Tom White. *Hadoop: The definitive guide*. California: O'Reilly, 2009.
- [6] Chuck Lam: *Hadoop In Action*, Manning, 2010
- [7] Dinh-Thi Vu, Ngoc-Duc Nguyen, Dai-Duong Le, Duc-Khanh Tran. Efficiently Crawl Topical Vietnamese Web Pages using Machine Learning Techniques. IEEE-RIVF 2010, Vietnam. (Submitted)
- [8] Bin He - Mitesh Patel - Zhen Zhang - Kevin Chen-Chuan Chang, "Accessing the deep web", May 2007
- [9] Jayant Madhavan, David Ko, Lucja Kot, Vignesh Ganapathy, Alex Rasmussen, Alon Y. Halevy, Google's Deep-Web Crawl, 2008

Xây dựng thiết bị tích hợp dịch vụ phục vụ cho hệ thống mạng doanh nghiệp vừa và nhỏ

Bạch Hà Duy, Hoàng Xuân Nam

Tóm tắt - Sự bùng nổ của Internet trong thời đại hiện nay đã khiến nhu cầu kết nối mạng và sử dụng các dịch vụ mạng trở nên thiết yếu hơn bao giờ hết đối với mọi doanh nghiệp. Để giải quyết bài toán thiết kế mạng, nhiều hãng sản xuất thiết bị mạng trên thế giới như Juniper, Cisco, Planet... đã có những bộ sản phẩm và giải pháp đáp ứng nhu cầu của các doanh nghiệp. Tuy nhiên những giải pháp này lại chưa phù hợp với mô hình mạng các doanh nghiệp vừa và nhỏ tại Việt Nam. Mục tiêu của đề tài là chỉ ra những nhược điểm của hệ thống mạng và đưa ra được một giải pháp phù hợp với điều kiện cơ sở vật chất của các doanh nghiệp vừa và nhỏ tại Việt Nam. Đề tài đã giải quyết hai bài toán quan trọng nhất đối với mạng doanh nghiệp vừa và nhỏ. Đó là đưa ra một thiết kế mạng phù hợp và cài đặt thành công một thiết bị tích hợp dịch vụ tương ứng với thiết kế đó. Ngoài việc tích hợp các phần mềm mã nguồn mở có sẵn vào thiết bị thành một hệ thống thống nhất, chúng tôi còn xây dựng thêm một số module phục vụ cho giải pháp đã đưa ra.

Từ khóa - Mã nguồn mở, Mạng doanh nghiệp, Thiết kế mạng, Tích hợp dịch vụ.

1. CƠ SỞ LÝ THUYẾT

1.1. Tổng quan về mạng doanh nghiệp

Để có thể xây dựng được giải pháp như đã đề cập ở trên, việc đầu tiên chúng tôi quan tâm tới là sự phân chia quy mô của một doanh nghiệp. Cụ thể, chúng tôi thực hiện phân chia dựa trên quy mô hệ thống mạng của doanh nghiệp đó.

Do sự đa dạng trong nhu cầu của doanh nghiệp, cùng với sự đa dạng về công nghệ và kiến trúc mạng, việc phân chia quy mô mạng doanh nghiệp cho tới nay vẫn chưa có một tài liệu thiết kế chuẩn. Mặc dù vậy, dựa trên một số tài liệu thiết kế của các hãng khác nhau, chúng tôi có thể tạm thời phân chia quy mô doanh nghiệp dựa trên một số tiêu chí [3]: Số lượng người dùng và số lượng thiết bị; Tốc độ những đường truyền cung cấp kết nối cho mạng Local Area Network (LAN), Wide Area Network (WAN), hay Internet, và khả năng dự phòng cho chúng; Hệ thống mạng tổng thể với số lượng vùng cũng như quy mô cụ thể tại từng vùng. Dựa trên những tiêu chí như vậy, chúng tôi nhận thấy có thể phân

Bach Hà Duy, sinh viên lớp Truyền thông và mạng máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (e-mail: bachhaduy@gmail.com).

Hoàng Xuân Nam, sinh viên lớp Truyền thông và mạng máy tính, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (e-mail: hoang.xuan.nam262@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

chia quy mô mạng doanh nghiệp thành hai nhóm chính: Nhóm doanh nghiệp với quy mô mạng vừa và nhỏ và nhóm doanh nghiệp với quy mô mạng lớn.

Nhóm doanh nghiệp có quy mô mạng lớn có số lượng người dùng trên 600. Nhóm này sử dụng các kết nối Internet và WAN với tốc độ lên tới 1Gbps. Các kết nối mạng LAN trong doanh nghiệp có thể từ 10 Mbps lên đến 10 Gbps. Ngoài dự phòng đường truyền, hệ thống mạng còn dự phòng cứng nhiều thiết bị, đảm bảo cho các luồng thông tin quan trọng không bị gián đoạn. Về thiết kế hệ thống mạng, nhóm doanh nghiệp lớn thường có nhiều vùng chính với quy mô lớn, lên tới 300/500 người, và các vùng phụ với nhiều quy mô khác nhau, có thể đáp ứng 100/200 người. Ngoài ra, các doanh nghiệp có dữ liệu quan trọng có một vùng dự phòng riêng dành cho dữ liệu, với các công nghệ đồng bộ giữa Data Center (DC) và Disaster Recovery Center (DR).

Trong khi đó, nhóm doanh nghiệp có quy mô mạng vừa và nhỏ là các doanh nghiệp với số lượng người dùng không vượt quá 600. Họ sử dụng những đường truyền Internet và WAN có tốc độ nhỏ, khoảng 10/100 Mbps, thậm chí không có đường truyền WAN. Các kết nối mạng LAN trong doanh nghiệp có thể là 10/100/1000 Mbps. Thông thường nhóm này chỉ tính đến khả năng dự phòng các đường truyền, ít có dự phòng thiết bị. Về thiết kế hệ thống mạng, doanh nghiệp chỉ có 1 vùng chính, vùng phụ nếu có thì rất ít, và có thiết kế đơn giản hơn nhiều. Quy mô vùng phụ cũng rất nhỏ, chỉ phục vụ cho khoảng dưới 50 người.

Với những so sánh trên, chúng tôi đi sâu tìm hiểu các thiết kế dành cho mô hình mạng doanh nghiệp lớn, áp dụng vào bài toán đối với mô hình mạng doanh nghiệp vừa và nhỏ, từ đó đưa ra một giải pháp dành cho mô hình mạng này.

1.2. Mô hình mạng doanh nghiệp lớn

Trong thiết kế mô hình mạng doanh nghiệp lớn, chúng tôi nhận thấy nổi bật lên hai đặc điểm quan trọng: Tính phân lớp và tính module hóa.

Tính phân lớp ở đây là việc hệ thống mạng được phân chia thành các lớp rõ ràng, với yêu cầu về năng lực, kết nối, giao thức cũng như yêu cầu về các chính sách nâng cao khác. Việc tham khảo mô hình phân lớp của Juniper [6] cho thấy sự phân chia hợp lý hệ thống mạng thành ba lớp: Access; Aggregation và Core. Lớp Access phụ trách công việc cung cấp kết nối cho người dùng, phân nhóm người dùng, và đánh dấu các gói tin. Lớp Aggregation làm trung gian kết nối giữa Lớp Access và Lớp Core với nhiệm vụ chính bao gồm định tuyến cho các kết nối, thực thi chính sách chất lượng dịch vụ, và thực hiện các cơ chế

bảo mật cho hệ thống. Lớp Core tập trung thực hiện nhiệm vụ chuyển mạch với tốc độ cao. Qua đây ta có thể nhận thấy nguyên nhân của sự phân lớp không nằm ngoài các nhu cầu về chuyển mạch, bảo mật, dự phòng, khả năng mở rộng và việc thực thi các chính sách của hệ thống.

Tính module hóa của hệ thống là việc hệ thống được phân chia thành các khối, mỗi khối phục vụ một chức năng riêng, có tính độc lập cao. Tham khảo mô hình module hóa hệ thống theo tài liệu thiết kế Cisco [2], chúng tôi nhận thấy hệ thống có thể được phân chia làm một số module cơ bản như: Module User Access; Module Application Service; Module Internet; Module Management... Ngoài ra, còn một số module khác có thể xuất hiện phụ thuộc nhu cầu doanh nghiệp như: Module Data Center; Module Partner... Mục đích của việc thực hiện phân chia, module hóa hệ thống chính là: Đảm bảo cho các dịch vụ trong hệ thống hoạt động một cách hiệu quả; Độc lập cao và không ảnh hưởng lẫn nhau, nhất là khi xảy ra sự cố; Triển khai các chính sách và quản lý nhóm dịch vụ một cách dễ dàng.

1.3. Bài toán với mạng doanh nghiệp vừa và nhỏ

Nhu đã đề cập đến, mạng doanh nghiệp lớn được phân thành nhiều lớp, nhiều module, đồng thời sử dụng các thiết bị chuyên dụng phục vụ từng yêu cầu khác nhau của hệ thống. Tuy nhiên, với doanh nghiệp vừa, nhỏ, hoặc chi nhánh nhỏ, việc xây dựng một mô hình với nhiều thiết bị và nhiều lớp là điều lãng phí và không cần thiết. Với ý tưởng đó, chúng tôi hướng tới việc giữ được thiết kế về mặt logic mà vẫn phù hợp với điều kiện của mạng doanh nghiệp vừa và nhỏ.

Khi so sánh hệ thống mạng doanh nghiệp vừa và nhỏ với doanh nghiệp lớn, đề tài tập trung vào hai nhược điểm chính của mạng doanh nghiệp vừa và nhỏ: Vấn đề thiết kế mạng; Vấn đề lựa chọn thiết bị phù hợp.

Do không tiến hành thiết kế trước mà thường chỉ xây dựng thỏa mãn nhu cầu hiện tại, hệ thống mạng của doanh nghiệp vừa và nhỏ gặp nhiều vấn đề: Vấn đề mở rộng; Vấn đề bảo mật; Vấn

đề chất lượng dịch vụ; Vấn đề thực thi các chính sách cho hệ thống... Đối chiếu với thiết kế mạng doanh nghiệp lớn, đây chính là bài toán phân lớp cho mạng doanh nghiệp.

Nhu cầu dịch vụ cũng là một bài toán đối với mạng doanh nghiệp nhỏ. Do nhu cầu của doanh nghiệp nhỏ thay đổi theo thời gian, nên một số dịch vụ cần được bổ sung. Tuy nhiên, do muốn tiết kiệm chi phí hay chưa có sự quan tâm đúng mức mà khi thêm các dịch vụ vào thì doanh nghiệp không chú ý tới các vấn đề như: Sư độc lập của các dịch vụ; Khả năng tương thích của các dịch vụ với nhau... So sánh với thiết kế mạng doanh nghiệp lớn, đây chính là bài toán module hóa cho mạng doanh nghiệp.

Tổng kết lại, bài toán đặt ra cho doanh nghiệp nhỏ là: Cần một thiết kế đảm bảo cho doanh nghiệp nhỏ những tiêu chí như tính chịu lỗi, tính sẵn sàng, khả năng mở rộng, khả năng bảo mật...; Cần một hoặc một nhóm thiết bị có khả năng cung cấp dịch vụ một cách khoa học (theo kiểu module hóa) thỏa mãn nhu cầu của doanh nghiệp.

2. THỰC NGHIỆM VÀ KẾT QUẢ ĐẠT ĐƯỢC

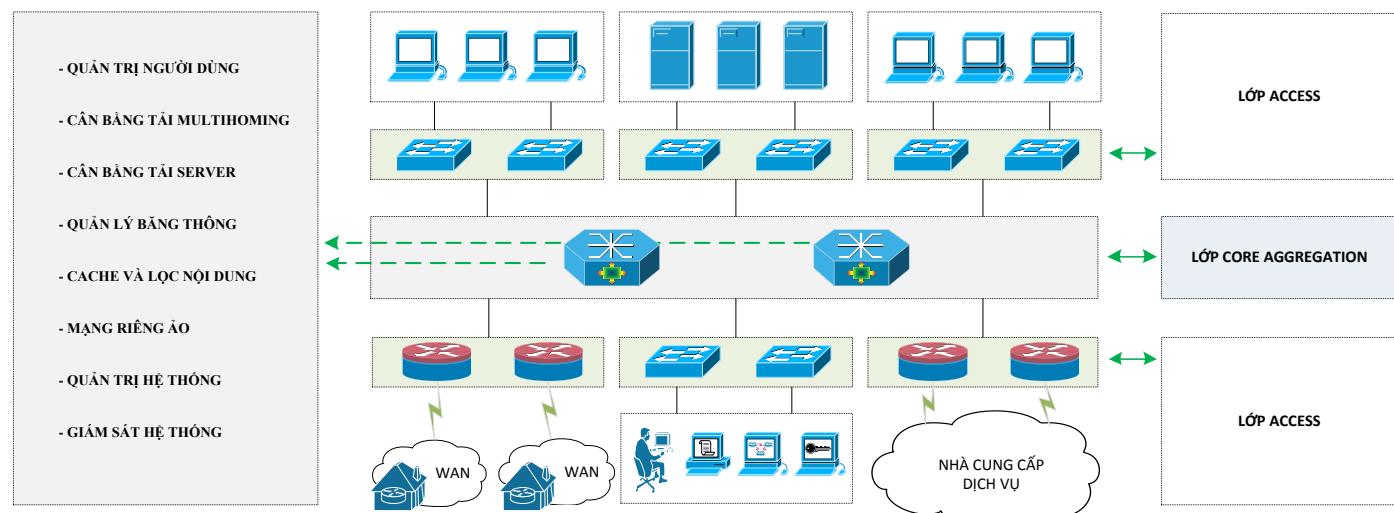
2.1. Đề xuất thiết kế phân lớp ở mạng doanh nghiệp vừa và nhỏ

Đối với doanh nghiệp có quy mô mạng nhỏ, việc triển khai một mô hình 3 lớp là lãng phí và không cần thiết. Chúng tôi thực hiện rút gọn mô hình ba lớp thành một mô hình hai lớp: Lớp Access; Lớp Core Aggregation.

Với vai trò tương tự như lớp Access ở mô hình mạng doanh nghiệp lớn, lớp Access mà thiết kế đưa ra phụ trách công việc cung cấp kết nối cho người dùng, phân nhóm người dùng, và đánh dấu các gói tin.

Lớp Core Aggregation là nơi thực hiện việc định tuyến cho các kết nối, thực thi chính sách chất lượng dịch vụ, và thực hiện các cơ chế bảo mật cho hệ thống.

Chúng tôi thực hiện gộp lớp Core và Distribution trong thiết kế mạng doanh nghiệp vừa và nhỏ vì hai nguyên nhân chính.



Hình 1 - Mô hình mạng và phương án tích hợp dịch vụ cho doanh nghiệp vừa và nhỏ.

Thứ nhất, mục đích của lớp Core là cung cấp khả năng chuyển mạch tốc độ cao cho các kết nối 1-10 Gbps với tổng băng thông có thể lên 400-500 Gbps [1]. Điều này là không cần thiết đối với hệ thống mạng doanh nghiệp nhỏ, vốn chỉ có các kết nối 10/100/1000 Mbps. Thứ hai, qua việc thu thập các thông số hiệu năng từ nhiều hãng, chúng tôi nhận thấy chúng ta hoàn toàn có thể đặt tại lớp Core Aggregation những thiết bị với khả năng chuyển mạch trong khoảng từ 30 tới 100 Gbps. Những thiết bị này là phù hợp với mô hình doanh nghiệp nhỏ vốn có số lượng người dùng không vượt quá 600, thậm chí ở Việt Nam là 200-300 người. Những thiết bị này vẫn có thể đảm đương được các công việc của thiết bị lớp Aggregation trong mô hình mạng doanh nghiệp lớn, đó là định tuyến, thực thi chính sách, và thực hiện bảo mật.

Có thể nói, qua mô hình mạng nêu trên, chúng tôi đã giải quyết cơ bản được vấn đề phân lớp trong thiết kế mạng doanh nghiệp vừa và nhỏ. Trọng tâm còn lại là phải giải quyết được nhu cầu dịch vụ dành cho doanh nghiệp.

2.2. Mô tả phương án tích hợp dịch vụ theo module

Trong một báo cáo về sử dụng thiết bị đa tính năng trong môi trường doanh nghiệp vừa và nhỏ, Gartner, một trong những công ty tư vấn và nghiên cứu thị trường IT nổi tiếng trên thế giới đã khuyến cáo [4]:

Mỗi truwong SMB (Small Medium Business) và các văn phòng chi nhánh cần quan tâm đến các thiết bị đa tính năng thay vì mua các thiết bị chuyên dụng đơn tính năng.

Thực vậy, do nhu cầu sử dụng nhiều dịch vụ khác nhau của doanh nghiệp, và giới hạn về mặt chi phí, việc tích hợp nhiều module dịch vụ lên trên một thiết bị là cần thiết. Tuy nhiên, cần tích hợp ở vị trí nào và như thế nào trong thiết kế mạng thì cần phải xem xét kỹ lưỡng.

Để có thể cung cấp dịch vụ cho hệ thống, chúng tôi đề xuất sử dụng tại lớp Core Aggregation một hoặc một cặp thiết bị không chỉ thực hiện các nhiệm vụ cơ bản của lớp, mà còn thực hiện cung cấp các dịch vụ theo nhu cầu của doanh nghiệp. Chúng tôi lựa chọn vị trí này để thực hiện tích hợp dịch vụ vì đây là vị trí tập trung các công việc xử lý gói tin của hệ thống mạng như: định tuyến, thực thi chính sách, bảo mật... Hơn nữa, lớp Core Aggregation thực sự đóng vai trò như một Gateway của hệ thống mạng phía dưới, nên nó cũng chính là nơi hội tụ của các dịch vụ.

Ngoài ý tưởng về thiết bị lớp Core Aggregation, việc tích hợp các dịch vụ lên trên cùng một thiết bị sẽ gây sinh một số vấn đề đáng quan tâm: Vấn đề cài đặt dịch vụ; Vấn đề tích hợp nhiều dịch vụ gây xung đột và giảm hiệu năng; Vấn đề bảo mật.

Về vấn đề cài đặt các dịch vụ, chúng tôi thực hiện cung cấp đủ các thư viện hỗ trợ cho dịch vụ, đồng thời dựa vào các phần mềm

quản lý như yum để quản lý sự cài đặt dịch vụ. Sau khi cài đặt dịch vụ, chúng tôi thực hiện các cấu hình cần thiết, và xây dựng các mô hình kiểm tra quá trình hoạt động của dịch vụ.

Về vấn đề tích hợp nhiều dịch vụ có thể gây xung đột, trong quá trình thực hiện đề tài, chúng tôi xây dựng một giao diện quản trị chung nhất cho tất cả các dịch vụ. Không chỉ thế, để giảm xung đột và tránh ảnh hưởng đến hiệu năng hệ thống, chúng tôi giải quyết dựa trên ý tưởng: Người quản trị sẽ chỉ kích hoạt những module cần thiết với mạng của doanh nghiệp. Các module không cần thiết sẽ tạm thời ngừng hoạt động.

Vấn đề bảo mật hệ thống được giải quyết theo hướng sử dụng hệ thống tường lửa iptables của Linux cùng với hệ thống antivirus mã nguồn mở.

Tóm lại, chúng tôi thực hiện tích hợp dịch vụ trên thiết bị lớp Core Aggregation. Lúc này, các thiết bị không chỉ nhận các nhiệm vụ cơ bản của một phân lớp, mà còn cung cấp thêm các dịch vụ cần thiết cho hệ thống, như cân bằng tải, quản lý người dùng, tăng tốc và lọc nội dung... Trong phần tiếp theo, chúng tôi sẽ mô tả chi tiết cấu hình thiết bị thử nghiệm và đánh giá các module sẽ sử dụng trong thiết bị đó.

2.3. Mô tả thiết bị và đánh giá các module sử dụng trong thiết bị

Việc cài đặt thực nghiệm của chúng tôi thực hiện trên thiết bị Network Appliance của Lanner với các thông số kỹ thuật như sau: Ví xử lý Dual Core Intel Atom (D510); RAM 2GB; 4 cổng 1Gbps; 2 cổng có chức năng bypass; 1 cổng console RJ45. Các cổng có chức năng bypass là các cổng đặc biệt phục vụ truy cập thiết bị gấp sự cố xảy, dòng dữ liệu được đẩy ra các cổng này, để thực hiện phương án dự phòng.

Về phía các module dịch vụ, chúng tôi đề xuất cài đặt những module như sau:

- Module dịch vụ mạng cơ bản: Cung cấp các dịch vụ cơ bản cho mạng nội bộ, như dịch vụ DHCP (Dynamic Host Configuration Protocol), dịch vụ DNS (Domain Name System), v.v...
- Module bảo mật: Cung cấp khả năng bảo mật từ lớp 1 tới lớp 3 trong mô hình OSI (Open Systems Interconnection) cho hệ thống mạng.
- Module quản trị người dùng: Khả năng quản trị và xác thực tập trung người dùng trong hệ thống.
- Module quản lý băng thông: Quản lý lưu lượng theo nhiều cách khác nhau, như gán lưu lượng cố định cho người dùng, cho một dải mạng, v.v...
- Module cân bằng tải máy chủ: Cân bằng tải cho các máy chủ dịch vụ trong hệ thống nếu doanh nghiệp có nhu cầu cung cấp dịch vụ ra ngoài Internet.
- Module tăng tốc và lọc nội dung: Tăng tốc độ truy cập thông qua hệ thống caching và lọc một số nội dung web.

- Module mạng riêng ảo: Cung cấp môi trường mạng riêng ảo cho các kết nối WAN, hay cho các kết nối của Home User.
- Module giám sát hệ thống: Giám sát hoạt động của hệ thống và đưa ra các cảnh báo cho người quản trị.
- Module quản trị hệ thống: Các Module trên được quản trị chung bởi hệ thống quản trị với giao diện web tiếng Việt để sử dụng. Hệ thống cần cho phép chúng tôi tích hợp và quản lý các dịch vụ trên một thiết bị và có cơ chế xây dựng các module mở rộng.
- Module quản lý dịch vụ: Giúp người quản trị nhận biết các module đang hoạt động, tắt và bật các module dịch vụ cần thiết cho hoạt động của doanh nghiệp.
- Module cân bằng tải multihoming: Chúng tôi nhận thấy trong một hệ thống thì việc cân bằng tải để tiết kiệm và tối ưu hóa tài nguyên là vô cùng cần thiết. Vì vậy, chúng tôi thực hiện xây dựng module cân bằng tải cho hệ thống. Module cung cấp khả năng phân tải, chạy dự phòng kết nối egress.

Việc tích hợp các module vào thiết bị không có nghĩa sẽ làm thiết bị nặng nề hơn và kém hiệu năng. Các module được tích hợp với ý tưởng là người quản trị sẽ chỉ kích hoạt những module tương ứng với nhu cầu của doanh nghiệp. Các module không cần thiết sẽ được người quản trị tắt để tránh ảnh hưởng đến hệ thống.

Trên đây là những module chúng tôi đã thực hiện tích hợp vào trong thiết bị. Tùy theo nhu cầu mà doanh nghiệp có thể lựa chọn những module cần thiết cho mô hình mạng của mình.

2.3. Tích hợp các module mã nguồn mở vào trong hệ thống

Dưới đây, chúng tôi mô tả chi tiết hơn việc thực hiện tích hợp các module vào trong hệ thống, bao gồm: Phần mềm mã nguồn mở được sử dụng; Các bài toán đã giải quyết; Và một số lưu ý khác khi chúng tôi thực hiện tích hợp module.

Module dịch vụ mạng cơ bản: Với module dịch vụ mạng cơ bản, chúng tôi sử dụng phần mềm mã nguồn mở *bind9* dành cho dịch vụ DNS và *dhcp_server* với dịch vụ DHCP. Với dịch vụ DNS, chúng tôi đã cài đặt giải quyết hai trường hợp: Phân giải tên miền nội bộ của hệ thống; Và phân giải tên miền trên Internet. Với dịch vụ DHCP, chúng tôi đã cấu hình tự động cấp phát địa chỉ IP, thông tin default gateway và một số thông số cơ bản khác của mạng cho các máy tính yêu cầu.

Module bảo mật: Với module bảo mật, chúng tôi sử dụng phần mềm mã nguồn mở *iptables*. Với phần mềm này, chúng tôi đã giải quyết một số bài toán như: Xây dựng các luật NAT (Network Address Translation); Quản lý và lọc các kết nối tcp, udp, icmp cho hệ thống; Nhận biết và đánh dấu các gói tin phục vụ cho chính sách bảo mật và chất lượng dịch vụ.

Module quản trị người dùng: Với module quản trị người dùng, chúng tôi sử dụng phần mềm mã nguồn mở *openldap* để thực hiện một số vấn đề như: Tích hợp *openldap* vào hệ thống hiện có thông qua công cụ migration; Quản lý thêm bớt người

dùng trong hệ thống; Thực hiện việc replicate thông tin người dùng qua một thiết bị khác nhằm phân tải và dự phòng.

Module mạng riêng ảo: Đối với module mạng riêng ảo, chúng tôi sử dụng hai phần mềm mã nguồn mở là *openswan* và *openvpn*. *Openswan* được cài đặt cho các kết nối site-to-site ở những doanh nghiệp có nhiều trụ sở. *Openvpn* được cài đặt phục vụ cho người làm việc từ xa muốn truy cập hệ thống.

Module quản lý băng thông: Đối với module quản lý băng thông, chúng tôi sử dụng phần mềm mã nguồn mở *HTB*. Đây là phần mềm cho phép quản lý băng thông hết sức linh hoạt. Với phần mềm này, chúng tôi đã cài đặt nhằm giải quyết các bài toán: Quản lý băng thông động cho nhiều dịch vụ; Phân chia băng thông cố định cho nhóm dịch vụ; Điều chỉnh các thông số ưu tiên cho dịch vụ.

Module cân bằng tải máy chủ: Đối với module cân bằng tải, chúng tôi sử dụng phần mềm mã nguồn mở *ldirector* và *heartbeat*. *Ldirector* là phần mềm dùng để giám sát và quản lý trạng thái của các máy chủ thuộc nhóm máy chủ cần cân bằng tải. Trong quá trình thử nghiệm, chúng tôi cài đặt *ldirector* trên hai thiết bị tạo thành kiến trúc LVS (Linux Virtual Server). *Heartbeat* được sử dụng để giám sát các hoạt động của *ldirector* giúp hai thiết bị có thể thực hiện chạy dự phòng khi có sự cố xảy ra.

Module tăng tốc và lọc nội dung: Nhằm thực hiện tăng tốc cho người dùng khi sử dụng các dịch vụ Web như http, https, ftp..., chúng tôi sử dụng phần mềm mã nguồn mở *squid* hoạt động như một proxy để thực hiện lưu trữ lại thông tin Web thường xuyên được sử dụng. *Squidguard* là phần mềm mã nguồn mở dành cho squid dùng để chặn các URL (Uniform Resource Locator) khi yêu cầu tới chúng đi qua *squid* proxy.

Module giám sát hệ thống: Chúng tôi sử dụng phần mềm mã nguồn mở *cacti* nhằm theo dõi, giám sát, lập báo cáo về trạng thái của thiết bị, của băng thông và của dịch vụ trong hệ thống. Đồng thời, hệ thống có cơ chế cảnh báo cho người quản trị qua âm thanh và qua mail khi xảy ra sự cố.

Module quản trị hệ thống: Với module quản trị hệ thống, chúng tôi sử dụng phần mềm mã nguồn mở *Webmin* [5]. Phần mềm có khả năng quản lý các cấu hình riêng rẽ của từng module khác trong hệ thống với giao diện web tiếng Việt. Ngoài ra, *Webmin* là nền tảng để chúng tôi xây dựng các module cần thiết theo thiết kế được đề cập đến ở phần tiếp theo.

2.4. Xây dựng và tích hợp module quản lý dịch vụ

Như phần 2.2 đã đề cập đến, để tránh làm ảnh hưởng tới hiệu năng của hệ thống, người quản trị sẽ chỉ kích hoạt những module cần thiết với mạng của doanh nghiệp. Các module không cần thiết sẽ tạm thời ngừng hoạt động.

Để thực hiện được điều này, chúng tôi sử dụng hai tiện ích của linux là service và chkconfig. Hai tiện ích này kiểm soát các đoạn mã điều khiển dịch vụ trên hệ thống. Tiện ích service giúp khởi động và tắt các dịch vụ lâm thời. Tiện ích chkconfig được dùng

để quản lý các dịch vụ được khởi động cùng với hệ thống.

Chúng tôi thực hiện tích hợp hai dịch vụ này vào trong hệ thống Webmin thông qua một module được lập trình sử dụng ngôn ngữ Perl. Module giúp người quản trị nhận biết các module đang hoạt động, tắt và bật các module dịch vụ cần thiết cho hoạt động của doanh nghiệp.

2.5. Xây dựng và tích hợp module cân bằng tải multihoming

Module cân bằng tải multihoming cung cấp khả năng chạy phân tải cho hệ thống. Module này thường được sử dụng trong các trường hợp doanh nghiệp muốn tận dụng tốt tất cả các đường truyền của mình trên mạng WAN, Internet.

Theo phương pháp truyền thống, việc cân bằng tải dựa hoàn toàn vào phần mềm mã nguồn mở iptables. Phương pháp cân bằng tải kiểu cũ chỉ có thể cân bằng tải theo gói tin nhờ một số module được cộng đồng mã nguồn mở viết ra như module nth, module random, module route... . Phương pháp này phải can thiệp vào kernel Linux và phải dịch lại kernel nên hết sức phiền phức. Trong đề tài này, chúng tôi sử dụng một phương pháp mới kết hợp hai phần mềm mã nguồn mở xtables (phiên bản mới của iptables) và iproute2 để thực hiện việc cân bằng tải cho hệ thống.

Đối với Module cân bằng tải Multihoming, chúng tôi nghiên cứu hai phương pháp cân bằng tải: Cân bằng tải theo gói tin (per-packet load balancing) và cân bằng tải theo phiên (per-session load balancing) mỗi phương pháp có cơ chế và ưu nhược điểm riêng.

Cân bằng tải theo gói tin là việc hệ thống tự động lựa chọn các gói tin theo tỉ lệ cho trước để đẩy tới các cổng đầu ra. Chúng tôi sử dụng iproute2 để khai báo các tỉ lệ cân bằng tải cho trước. Khi các gói tin đi tới thiết bị, chúng tôi phân loại sử dụng xtables. Các gói tin cần được phân tải sẽ được đánh dấu lại (thông qua trường mark) khi đi qua quá trình tiền định tuyến (pre-routing) của xtables. Sau đó, chúng được xtables trả về cho iproute2. Iproute2 thực hiện định tuyến các gói tin theo tỉ lệ đã khai báo. Sau khi qua iproute2, các gói tin được trả về cho xtables, thực hiện quá trình xử lý hậu định tuyến (post-routing) như NAT động hay NAT tĩnh. Sau đó, gói tin sẽ được đưa ra cổng được chỉ bởi iproute2.

Cân bằng tải theo phiên là việc hệ thống bảo toàn phiên làm việc của người dùng luôn đi qua một cổng đầu ra nhất định trong khi vẫn thực hiện phân tải theo tỉ lệ cho trước. Quá trình xử lý cũng tương tự như phương pháp cân bằng tải theo gói tin. Tuy nhiên, trong quá trình xử lý, chúng tôi thực hiện đánh dấu lại các gói tin khởi tạo phiên, theo dõi cổng đầu ra của chúng trong quá trình xử lý hậu định tuyến (post-routing). Khi các gói tin sau đó của phiên đi tới thiết bị, chúng tôi thực hiện phục hồi lại trường mark cho gói tin đó. Nhờ vậy, iproute2 nhận diện được gói tin của phiên cũ, và bảo toàn cổng đầu ra khi định tuyến cho gói tin.

Cân bằng tải theo phiên luôn có ưu điểm hơn cân bằng tải theo gói tin. Cân bằng tải theo phiên giúp tránh hai vấn đề quan trọng: Độ trễ không đồng đều; Nhận diện và cho phép kết nối qua

tường lửa. Vì thế, các thiết bị cân bằng tải tiên tiến hiện nay đều yêu cầu cân bằng tải theo phiên.

Với cơ chế như vậy, chúng tôi đã thiết kế trên Webmin module cân bằng tải multihoming sử dụng phương pháp mới kết hợp xtables và iproute2. Module được lập trình sử dụng ngôn ngữ Perl [7], và đã được chạy thử thành công qua các kịch bản thử nghiệm.

3. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN ĐỀ TÀI

Với đề tài này, chúng tôi đã giải quyết hai bài toán quan trọng nhất đối với mạng doanh nghiệp vừa và nhỏ. Đó là đưa ra một thiết kế mạng phù hợp và cài đặt thành công một thiết bị tích hợp dịch vụ tương ứng với thiết kế đó. Chúng tôi đã đưa ra mô hình thiết kế bao toàn sự phân lớp và tính logic về mặt module cho hệ thống. Chúng tôi cũng đã tích hợp thành công các module cần thiết trên một phần cứng cụ thể, và chạy thử nghiệm thiết bị với các kịch bản thử nghiệm khác nhau.

Hệ thống nói chung, cũng như từng module nói riêng đều có tính ứng dụng cao, có thể phát triển theo nhiều hướng. So sánh với các sản phẩm đa tính năng trên thế giới, hệ thống còn một số hạn chế về tối ưu hóa hiệu năng. Kiến trúc hệ thống vẫn là kiến trúc multi-pass, ảnh hưởng rất nhiều đến khả năng hoạt động của hệ thống. Hiện tại, có duy nhất một hãng trên thế giới là Paolo Alto đã phát triển thành công kiến trúc single-pass cho phép tận dụng tối đa hiệu năng của thiết bị. Chúng tôi dự định trong thời gian tới sẽ phát triển hệ thống theo kiến trúc single-pass này.

4. LỜI TRI ÂN

Chúng tôi xin chân thành cảm ơn PGS. TS. Đặng Văn Chuyết đã có những hướng dẫn và góp ý quý báu giúp chúng tôi hoàn thành nghiên cứu khoa học của mình.

Chúng tôi xin chân thành cảm ơn anh Hoàng Mạnh Cường, trưởng nhóm giải pháp công ty cổ phần công nghệ Sao Bắc Đẩu đã có những gợi ý hữu ích cho việc lựa chọn các phần mềm mã nguồn mở phù hợp với định hướng của đề tài.

5. TÀI LIỆU THAM KHẢO

- [1] Cisco Systems, *Cisco Catalyst 6500 and 6500-E Series Switch*, 2009.
- [2] Cisco Systems, *Cisco SAFE Reference Guide*, July 8, 2010.
- [3] Cisco Systems, *Smart Business Architecture for Midsize Networks Design Guide*, July 2009.
- [4] Gartner Inc, John Pescatore, Bob Walder, *Magic Quadrant for Unified Threat Management*, October 22, 2010.
- [5] Jamie Cameron, *Managing Linux Systems with Webmin System Administration and Module Development*, Prentice Hall, 2004.
- [6] Juniper networks, *Campus LAN Design Guide*, 2010.
- [7] Scott Guelich, Shishir Gundavaram, Gunther Birznieks, *CGI Programming with Perl, Second Edition*, O'Reilly, July 2000.

Phát hiện và theo vết đối tượng chuyển động

Phạm Đức Long, Trương Thị Tâm

Tóm tắt - Bài toán phát hiện và dò vết đối tượng chuyển động có nhiều ứng dụng trong đó yêu cầu phát hiện nhanh chóng các mục tiêu di chuyển trong chuỗi ảnh video và theo vết các đối tượng đó. Chuỗi ảnh do đối tượng chuyển động tạo nên được chia thành hai nhóm: chuỗi ảnh với nền tĩnh, và chuỗi ảnh với nền thay đổi. Đối với trường hợp nền không thay đổi thì sử dụng sai phân tạm thời hoặc trừ nền thích hợp. Trường hợp thứ hai xuất hiện khi camera di chuyển, cách tiếp cận là sử dụng phương pháp luồng quang học. Quá trình dò vết đối tượng sử dụng bộ lọc Kalman và luồng quang học. Báo cáo trình bày các nghiên cứu và thử nghiệm đối với bài toán phát hiện và theo vết đối tượng chuyển động trong cả hai trường hợp. Các phương pháp này được sử dụng để thử nghiệm và phát triển kỹ thuật StroMotion, trong đó đối tượng chuyển động được phát hiện bằng phương pháp nền và theo dõi sử dụng bộ lọc Kalman kết hợp với giải thuật MeanShift. Hệ thống thử nghiệm được phát triển sử dụng bộ công cụ mã nguồn mở OpenCV và cho phép phát hiện đối tượng chuyển động trong video có nền tĩnh và dò vết đối tượng trong video có nền thay đổi.

Từ khóa - background model, Kalman filter, Mean Shift, moving object detection and tracking.

1. GIỚI THIỆU

Phát hiện và theo vết đối tượng chuyển động của video là một chủ đề nghiên cứu quan trọng trong các kỹ thuật thị giác máy tính. Phát hiện đối tượng chuyển động nghĩa là phân đoạn và tách đối tượng chuyển động từ các ảnh video liên tục, đó là việc phân chia thành ảnh nội và ảnh nền. Trong khi đó theo vết chuyển động là xử lý mà nhờ đó một đối tượng hoặc vùng được theo vết sử dụng thông tin về hành vi của nó trong chuyển động. Hiện nay, nhiều phương pháp đã đưa ra để thực hiện hai công việc trên. Các phương pháp để phát hiện đối tượng chuyển động như phương pháp trừ nền thích hợp, phương pháp sai phân tạm thời, phương pháp luồng quang học. Mỗi phương pháp đều có ưu nhược điểm riêng và được ứng dụng phù hợp. Phương pháp trừ nền thích hợp là phương pháp phổ biến được áp dụng bởi tính đơn giản trong thực thi của nó, tính hiệu quả đối với những video có nền tĩnh, xong nhưng điểm là phụ thuộc mạnh vào sự thay đổi nền của ảnh. Do đó đối với những ảnh có nền thay đổi, phương pháp này tỏ ra không hiệu quả. Phương pháp sai phân tạm thời rất phù hợp với môi trường động,

nhưng nó phụ thuộc mạnh vào tốc độ của đối tượng di chuyển trong cảnh, và đây có thể xem như điểm yếu nổi bật nhất. Đối với trường hợp video có nền thay đổi, phương pháp luồng quang học là lựa chọn ưu tiên hơn cả. Tuy nhiên phương pháp này đòi hỏi tính toán phức tạp và không thể áp dụng cho các dòng video full-frame trong thời gian thực mà không có phần cứng chuyên biệt [1].

Trong số các giải thuật theo vết đối tượng, có giải thuật Mean Shift phát triển bởi Comaniciu, Ramesh và Meer. Mặc dù giải thuật theo vết đối tượng Mean Shift thực hiện tốt trên video với dịch chuyển đối tượng tương đối nhỏ, hiệu quả của nó không đảm bảo khi các đối tượng chuyển động nhanh hoặc biến dạng một phần, hoặc bị che lấp. Để khắc phục khó khăn của phương pháp theo dõi Mean Shift, một giải thuật theo dõi đối tượng Mean Shift cải tiến bởi việc khởi tạo Mean Shift với giá trị ước lượng của bộ lọc Kalman [4].

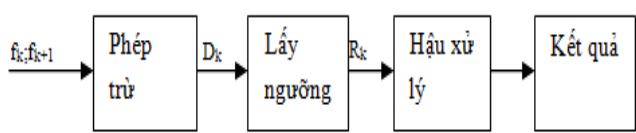
Trong báo cáo này sẽ đưa ra hai kịch bản đối với video sử dụng để phát hiện và theo vết chuyển động, đó là: video có nền cố định và video có nền thay đổi. Đối với video có nền cố định, phương pháp trừ nền thích hợp sẽ được sử dụng để phát hiện đối tượng chuyển động. Kết quả thu được là mặt nạ chuyển động của đối tượng, từ đó thực hiện kỹ thuật StroMotion. Đối với video có nền thay đổi, đối tượng chuyển động sẽ được người dùng lựa chọn bằng chuột, đối tượng này sẽ được theo vết sử dụng giải thuật Mean Shift kết hợp với Kalman Filter.

Phần còn lại của báo cáo sẽ được tổ chức như sau: phần 2 là sơ lược về nền tảng lý thuyết của phép trừ nền, giải thuật Mean Shift. Phần 3 nói về ứng dụng phát hiện và theo vết đối tượng vào kỹ thuật StroMotion đối với video có nền cố định, đối với video có nền thay đổi thì chỉ trình bày phương pháp theo vết đối tượng. Phần 4 là trình bày kết quả thử nghiệm và cuối cùng là kết luận.

2. CƠ SỞ LÝ THUYẾT

2.1 Phép trừ nền

Nguyên lý dựa trên phương pháp phép trừ nền trong [1] rất đơn giản, xử lý cơ bản được chỉ ra trong hình dưới đây:



Hình 1: Biểu đồ luồng phép trừ nền.

Phương pháp phép trừ nền giả thiết rằng nền là cố định, vì thế nền không thay đổi với nhiều frame. Đầu tiên, bằng cách sử dụng phương trình (1), tìm ra hiệu D_k giữa frame f_k và nền b_k .

$$D_k(x, y) = |f_k(x, y) - b_k(x, y)|, \quad (1)$$

Theo phương trình (2), lấy ngưỡng hiệu:

$$R_k(x, y) = \begin{cases} 1, & D_k(x, y) > T \\ 0, & D_k(x, y) \leq T \end{cases}, \quad (2)$$

Trong (2) T là một giá trị ngưỡng, độ chính xác của T được lựa chọn ảnh hưởng trực tiếp tới chất lượng của mặt nạ chuyển động R_k . Nếu ngưỡng T lựa chọn quá cao, vùng của đối tượng di chuyển nhảy qua sự tạo ngưỡng sẽ xảy ra hiện tượng đứt quãng, ngược lại, nếu ngưỡng T quá thấp, nó sẽ đưa đến nhiều nhiễu. Phương pháp phổ biến nhất để lựa chọn ngưỡng T là sử dụng biểu đồ histogram để tìm 2 điểm cao nhất hoặc nhiều hơn, chọn giá trị đáy giữa hai đỉnh là ngưỡng.

2.2 Giải thuật Mean Shift.

Trong phương pháp theo dõi đối tượng MS [4], mô hình mục tiêu được định nghĩa là histogram màu chuẩn hóa của nó $q = \{q_u\}_{u=1, \dots, m}$, với m là số lượng bin. Phân phối màu chuẩn hóa của một ứng viên mục tiêu $p(y) = \{p_u(y)\}_{u=1, \dots, m}$ xung quanh y trong frame hiện thời có thể được tính như sau:

$$p_u(y) = C_h \sum_{i=1}^{n_h} k \left(\left\| \frac{y - x_i}{h} \right\|^2 \right) \delta[b(x_i) - u], \quad (3)$$

Với $\{x_i\}_{i=1, \dots, n_h}$ là các vị trí điểm ảnh của ứng viên mục tiêu n_h trong vùng mục tiêu, δ là hàm Kronecker delta, $b(x_i)$ kết hợp điểm ảnh x_i với histogram bin, $k(x)$ là mặt cắt nhân với bề rộng h , và C_h là hằng số chuẩn hóa. Các phương trình tương tự được sử dụng để thu phân phối màu của mô hình mục tiêu q .

Hệ số Bhattacharyya đánh giá sự tương đồng của mô hình mục tiêu và mô hình ứng viên mục tiêu, được định nghĩa là:

$$\rho(y) = \rho[p(y), q] = \sum_{u=1}^m [p_u(y) q_u]^{1/2}, \quad (4)$$

Để tìm vị trí tương ứng với mục tiêu trong frame hiện thời, hệ số Bhattacharyya trong phương trình (4) đạt cực đại đối với y , nó có thể được giải quyết bằng cách chạy các vòng lặp Mean Shift. Chúng ta giả sử rằng việc tìm kiếm vị trí mục tiêu mới trong frame hiện thời bắt đầu tại vị trí y_0 . Tại mỗi bước của xử lý lặp, mục tiêu ước lượng chuyển động từ y_0 tới vị trí mới y_1 được định nghĩa như sau:

$$y_1 = \frac{\sum_{i=1}^{n_h} x_i w_i g\left(\|(y_0 - x_i)/h\|^2\right)}{\sum_{i=1}^{n_h} w_i g\left(\|(y_0 - x_i)/h\|^2\right)}, \quad (5)$$

Với

$$w_i = \sum_{u=1}^m \left[\frac{q_u}{p_u(y_0)} \right]^{1/2} \delta[b(x_i) - u], \quad (6)$$

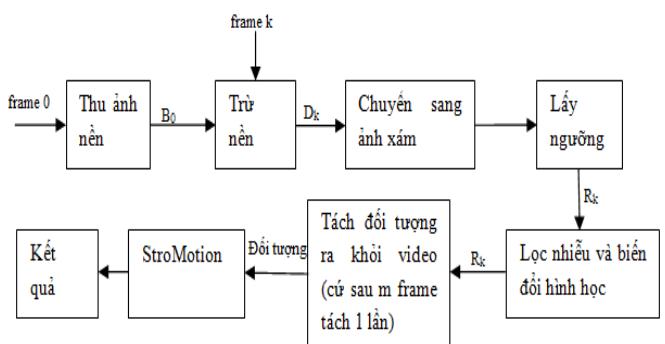
Và $g(x) = -k'(x)$.

3. ỨNG DỤNG PHÁT HIỆN VÀ THEO VÉT ĐỐI TƯỢNG VÀO KỸ THUẬT STROMOTION

3.1 Video có nền cố định.

StroMotion trong [6] là kỹ thuật hiển thị ảnh tĩnh của đối tượng chuyển động theo thời gian và không gian qua các frame. Để giữ tất cả các ảnh tĩnh, một ảnh mặt nạ, gọi là “clipboard”, được phát triển. Các điểm ảnh từ frame trước mà tương ứng với điểm ảnh của đối tượng trong clipboard sẽ được chép sang mọi frame khi frame đó xử lý. Khi thu được vị trí mới của đối tượng, giải thuật đảm bảo rằng ảnh trước của đối tượng không chồng lên ảnh đối tượng hiện tại.

Như đã đề cập ở trên, đối với video có nền cố định, phương pháp trừ nền được áp dụng để phát hiện đối tượng chuyển động. Dưới đây là chi tiết luồng xử lý để thu được kết quả của kỹ thuật StroMotion.



Hình 2: Luồng xử lý video có nền cố định

Từ frame đầu tiên khi chưa có đối tượng chuyển động xuất hiện, ta thu được ảnh nền B_0 . Mất khác với điều kiện nền không thay đổi cho nên ta không cần cập nhật nền và sẽ sử dụng ảnh nền B_0 cho các xử lý sau này. Áp dụng công thức (1), với $B_k=B_0$, $k=1 \dots n$, ta sẽ thu được D_k .

Do ảnh đầu vào là ảnh theo hệ màu (RGB), các xử lý được thực hiện trên thành phần độ chói Y (theo hệ màu YUV), cho nên ảnh D_k được chuyển sang ảnh xám. Sau bước này, áp dụng công thức (2) đối với D_k ta thu được mặt nạ chuyển động R_k . Tuy nhiên, mặt nạ thu được có nhiều, các thành phần của đối tượng bị rời rạc, không liên tục. Sử dụng bộ lọc trung vị (median) loại bỏ nhiễu, và phép biến đổi hình học (morphology) làm liền các thành phần của đối tượng. Sau các phép xử lý này, mặt nạ đã được cải thiện. Dựa vào mặt nạ chuyển động và frame xử lý hiện thời, ta thực hiện tách đối tượng từ frame hiện thời.

Để thực hiện kỹ thuật StroMotion thì cứ sau một số frame, số frame này phụ thuộc vào chuyển động nhanh chậm của đối tượng trong video, thì tách đối tượng một lần. Sau khi tách đối tượng

thì nhúng đối tượng vào video ban đầu ta sẽ thu được kết quả StroMotion.

3.2 Video có nền thay đổi.

Đối với video có nền thay đổi, đối tượng chuyển động được lựa chọn bằng chuột để tiến hành theo vết. Phương pháp theo vết được sử dụng là giải thuật Mean Shift kết hợp với một bộ lọc Kalman tương thích.

Bộ lọc Kalman

Chúng ta định nghĩa biến thời gian rời rạc t , vector trạng thái $X(t)$, vector phép đo $Z(t)$, ma trận chuyển tiếp trạng thái A , ma trận phép đo C , nhiễu trạng thái $\nu(t)$, và nhiễu phép đo $\mu(t)$. Hệ thống được biểu diễn như sau:

$$\begin{cases} X(t) = AX(t-1) + \nu(t-1), \\ Z(t) = CX(t) + \mu(t), \end{cases} \quad (7)$$

Chúng ta giả thiết rằng $\nu(t-1)$ và $\mu(t)$ là các biến ngẫu nhiên Gaussian với trung bình 0, do đó hàm mật độ xác suất của chúng là $N[0, Q(t-1)]$ và $N[0, R(t)]$, với ma trận hiệp phương sai $Q(t-1)$ và $R(t)$ được coi như là ma trận hiệp phương sai nhiễu chuyển tiếp và ma trận hiệp phương sai nhiễu phép đo.

Chúng ta định nghĩa một mô hình để dò vết đối tượng như sau. Vector trạng thái là $X = (x, v, a)^T$, với x, v và a tương ứng biểu diễn tâm (ngang hoặc dọc), vận tốc và gia tốc. Vector phép đo là $Z = x$. Ma trận chuyển tiếp trạng thái sẽ là:

$$A = \begin{bmatrix} 1 & \Delta t & 0.5\Delta t^2 \\ 0 & 1 & \Delta t \\ 0 & 0 & 1 \end{bmatrix}, \quad (8)$$

Với Δt là khoảng thời gian. Ma trận phép đo là $C = (1, 0, 0)$. Ma trận hiệp phương sai nhiễu chuyển tiếp là

$$Q(t-1) = \begin{bmatrix} \sigma_1^2(t-1) & 0 & 0 \\ 0 & 0.5\sigma_1^2(t-1) & 0 \\ 0 & 0 & 0.2\sigma_1^2(t-1) \end{bmatrix} \quad (9)$$

Và hiệp phương sai nhiễu phép đo là $R(t) = \sigma_2^2(t)$. Ước lượng của tham số $\sigma_1^2(t)$ và $\sigma_2^2(t)$ được mô tả sau.

Phương pháp dò xuất để theo vết đối tượng sử dụng giải thuật Mean Shift kết hợp với bộ lọc Kalman

Trong giải thuật Kalman Filter, hiệp phương sai nhiễu phép đo $R(t)$ và độ khuếch đại Kalman tỉ lệ nghịch. Khi ma trận hiệp phương sai $R(t)$ tiệm cận tới 0, độ khuếch đại Kalman tăng lên nhiều hơn. Trong trường hợp này, phép đo càng tin cậy hơn, trong khi kết quả dự đoán thì càng ít tin cậy. Ngược lại, khi hiệp phương sai nhiễu ước lượng $a-priori$ của Kalman Filter tiệm cận gần tới 0, khuếch đại Kalman giảm đi nhiều. Phép đo thực càng ít tin cậy, trong khi kết quả dự đoán ngày càng tin cậy hơn.

Do vậy, hệ thống sẽ đạt được một kết quả gần tối ưu nếu chúng ta có thể quyết định cái nào tin cậy. Trong bài này, Kalman Filter tương thích cho phép các tham số ước lượng $R(t)$ và $Q(t-1)$ điều chỉnh tự động tương ứng với hệ số Bhattacharyya của dò vết đối tượng Mean Shift.

Trong phương pháp dò vết Mean Shift, hệ số Bhattacharyya đánh giá sự tương đồng của mô hình của mục tiêu và ứng viên. Khi đối tượng được dò vết bị che bởi các đối tượng khác hoặc nền, hệ số Bhattacharyya sẽ giảm đột ngột. Do đó, chúng ta có thể định nghĩa một ngưỡng T_h để xác định sự che lấp có xảy ra hay không.

Giả thiết kết quả tìm kiếm của Mean Shift là \hat{y}_t là frame hiện tại t , hệ số Bhattacharyya $\rho(\hat{y}_t)$ đánh giá sự tương đồng giữa mô hình mục tiêu và mô hình ứng viên xung quanh \hat{y}_t . Khi kết quả của Mean Shift được sử dụng là phép đo của Kalman Filter, trong bước hiệu chỉnh hệ số Bhattacharyya được dùng để điều chỉnh các tham số ước lượng của Kalman Filter thích nghi. Nếu các hệ số Bhattacharyya $\rho(\hat{y}_t)$ lớn hơn ngưỡng T_h thì giá trị của $\sigma_1^2(t-1)$ được đặt là $\rho(\hat{y}_t)$ và $\sigma_2^2(t)$ là $1 - \rho(\hat{y}_t)$. Ngược lại, hãy đặt $\sigma_1^2(t-1)$ và $\sigma_2^2(t)$ tương ứng bằng 0 và vô cùng, do đó khuếch đại Kalman bằng 0. Để biến đổi thời gian đều đặn, các tham số được liên hệ với frame hiện tại được thu về thông qua phép lọc thời gian:

$$\begin{cases} \hat{\sigma}_1^2(t-1) = (1 - \lambda)\hat{\sigma}_1^2(t-1) + \lambda\sigma_1^2(t-2) \\ \hat{\sigma}_2^2(t) = (1 - \lambda)\hat{\sigma}_2^2(t) + \lambda\sigma_2^2(t-1) \end{cases} \quad (10)$$

Với

$$\hat{\sigma}_1^2(t-1) = \begin{cases} \rho(\hat{y}_t) & \text{nếu } \rho(\hat{y}_t) \geq T_h \\ 0 & \text{nếu ngược lại} \end{cases}, \quad (11)$$

$$\hat{\sigma}_2^2(t) = \begin{cases} 1 - \rho(\hat{y}_t) & \text{nếu } \rho(\hat{y}_t) \geq T_h \\ T & \text{nếu ngược lại} \end{cases}. \quad (12)$$

T là một hằng số lớn, do vậy ước lượng posterior của Kalman Filter xấp xỉ tới giá trị dự đoán của nó và $\lambda \in [0, 1]$ là nhân tố bỏ qua. Số λ càng nhỏ, cập nhật của $\sigma_1^2(t-1)$ và $\sigma_2^2(t)$ trở nên càng nhanh.

Tương ứng với hệ số Bhattacharyya, hệ thống Kalman Filter có thể được điều chỉnh tự động để ước lượng tâm của đối tượng theo vết. Để cho rõ ràng, dưới đây là nội dung giải thuật :

Đầu vào: vector trạng thái $X_x(t)$ của tâm theo chiều ngang của mục tiêu; vector trạng thái $X_y(t)$ của tâm theo chiều dọc của mục tiêu và mô hình mục tiêu $q = \{q_u\}_{u=1,\dots,m}$

Bước 1: dự đoán tâm chiều ngang và tâm chiều dọc của mục tiêu bằng cách sử dụng phương trình trạng thái của Kalman Filter tương ứng.

Bước 2: áp dụng Mean Shift được khởi tạo bởi giá trị được dự đoán của Kalman Filter để tìm kiếm tâm của đối tượng trong frame hiện tại thứ $t+1$, sau đó lấy kết quả tìm kiếm

$$\hat{y}_{t+1} = (\hat{x}_{t+1}, \hat{y}_{t+1}).$$

Bước 3: tính toán hệ số Bhattacharyya $\rho(\hat{y}_{t+1})$.

Bước 4: tương ứng với phương trình (8) đến (10), tính toán các tham số $Q(t)$ và $R(t+1)$.

Bước 5: sử dụng \hat{x}_{t+1} và \hat{y}_{t+1} như là phép đo của hai Kalman Filter, tính toán $X_x(t+1)$ và $X_y(t+1)$ bằng bước hiệu chỉnh của Kalman Filter tương ứng.

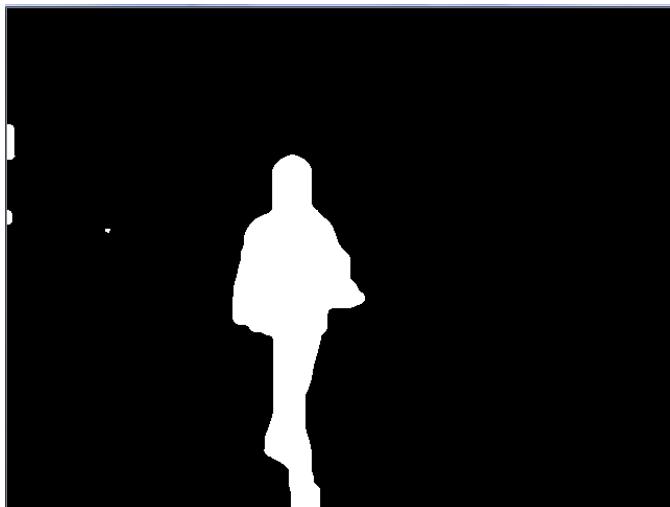
4. KẾT QUẢ THÍ NGHIỆM

Chương trình thực hiện sử dụng thư viện OpenCV 2.1, trên hệ điều hành Window.

4.1 Video có nền cố định.

Trong thí nghiệm chúng tôi sử dụng video có các thông số sau: kích thước 640x480 pixels, định dạng AVI, 30 frames/s, video có độ dài 5s, video được quay bằng camera SamSung PL20 trong điều kiện giữ camera cố định, cường độ ánh sáng môi trường tương đối ổn định.

Kết quả thí nghiệm như sau:



Hình 3: Mặt nạ chuyển động thu được sử dụng phương pháp trừ nền.

Trong hình 3 miêu tả mặt nạ chuyển động của đối tượng. Mặt nạ thu được có kết quả khá tốt: nhiễu đã được lọc gần hết, các phần của đối tượng được liền nết.

Sau bước này, khi duyệt video, cứ sau 15 frame chúng tôi tách đối tượng khỏi video 1 lần, và nhúng vào video ban đầu. Do vậy chúng tôi được kết quả StroMotion:



Hình 4: Kết quả StroMotion đối với video có nền tĩnh

4.2 Video có nền thay đổi.

Video sử dụng trong thí nghiệm có kích thước 320x240 pixels, 25 frames/s, định dạng AVI, độ dài video là 8s.

Chúng tôi lựa chọn một đối tượng chuyển động bằng chuột để tiến hành theo vết.



Hình 4: Kết quả theo vết đối tượng sử dụng giải thuật Mean Shift kết hợp với bộ lọc Kalman.

Trong hình 4 là kết quả theo vết đối tượng. Đối tượng theo vết được khoanh hình chữ nhật màu đỏ. Kết quả hiển thị từ trái sang phải, từ trên xuống dưới theo thứ tự là các frame 75, 100, 127, 158.

5. KẾT LUẬN

Những kết quả trong báo cáo này nhằm hỗ trợ phát triển kỹ thuật StroMotion. Báo cáo đã trình bày được một phương pháp áp dụng hiệu quả để phát hiện đối tượng, và áp dụng kỹ thuật StroMotion trong video có nền cố định và một phương pháp đề xuất để theo vết đối tượng trong video có nền thay đổi. Tuy nhiên, hạn chế của báo cáo là kỹ thuật StroMotion vẫn chưa thể áp dụng đối với trường hợp video có nền thay đổi. Nhưng chúng

tôi tin rằng, khó khăn này vẫn có thể khắc phục được. Do đó, hướng phát triển của đề tài này vẫn còn nhiều triển vọng vì tính thực tế và hiệu quả cao của của nó.

6. LỜI TRI ÂN

Để hoàn thành báo cáo này, trước hết chúng tôi xin chân thành cảm ơn PGS.TS Nguyễn Linh Giang đã tận tình hướng dẫn chúng tôi trong thời gian qua. Chúng tôi cũng xin cảm ơn tới gia đình và bạn bè đã động viên, hỗ trợ chúng tôi để chúng tôi có điều kiện thực hiện công việc này.

7. TÀI LIỆU THAM KHẢO

- [1] Liang Xiao and Tong-qiang Li, “Research on Moving Object Detection and Tracking”, 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010).
- [2] Massimo Piccardi, “Background subtraction techniques: a review *”, 2004 IEEE International Conference on Systems, Man and Cybernetics.
- [3] Xiaoli Zhao, “Practical Technique for Moving Object Detection and Tracking”.
- [4] Xiaohe li, Taiyi Zhang, Xiaodong Shen, and Jiancheng Sun, “Object tracking using an adaptive Kalman filter combined with mean shift”, OE Letters.
- [5] Gary Bradski & Adrian Kaehler, “Learning OpenCV: Computer Vision with the OpenCV Library”, O'Reilly.
- [6] Nhat H.Nguyen, “Understanding Tracking and StroMotion of Soccer Ball”.
- [7] Nguyễn Duy Nghĩa, “Nghiên cứu kỹ thuật xử lý video số, ứng dụng vào theo vết và phân loại đối tượng”.
- [8] http://opencv.willowgarage.com/documentation/motion_analysis_and_object_tracking.html

Hệ thống xác thực khuôn mặt hỗ trợ quản lý thẻ thư viện

Bùi Thị Minh Yên

Tóm tắt -- Các hệ thống xác thực và nhận diện khuôn mặt thực sự đã có rất nhiều ứng dụng trong thực tế như các hệ thống nhận dạng tội phạm hay là hệ thống giám sát ở nhà ga, sân bay hoặc là các hệ thống bảo mật ...

Trong các hệ thống thư viện hiện nay, việc xác thực thẻ thư viện với bạn đọc là rất thô sơ, rất khó quản lý và xác thực trong các trường hợp giả mạo. Vì vậy, em đã đề xuất một giải pháp sử dụng xác thực khuôn mặt trong hệ thống quản lý thẻ thư viện giúp việc quản lý bạn đọc của thư viện tốt hơn, tránh các trường hợp giả mạo người khác. Triển khai thành công hệ thống xác thực khuôn mặt sẽ đem lại một lợi ích và có hiệu quả cao.

Bài báo này trình các kết quả nghiên cứu ứng dụng xác thực ảnh khuôn mặt dựa trên thuật toán PCA. Mô hình ứng dụng hệ thống xác thực khuôn mặt được xây dựng và thiết kế triển khai nhằm hỗ trợ quản lý thẻ thư viện của bạn đọc.

Từ khóa - PCA, FAR, FRR, Eigenface, Euclidean.

1. GIỚI THIỆU

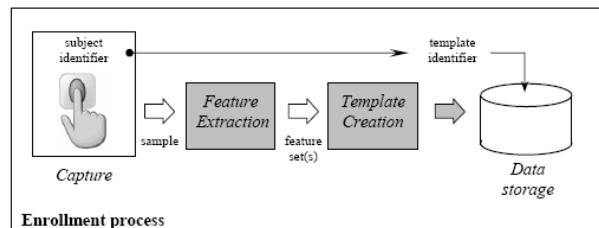
Xác thực và nhận dạng khuôn mặt có rất nhiều ứng dụng trong cuộc sống. Hệ thống xác thực sinh trắc nhằm xác định một người bằng việc đối sánh đặc điểm sinh trắc nhận được với mẫu sinh trắc (template) của người này đã được lưu trữ trước đó trong cơ sở dữ liệu. Hệ thống xác thực là các hệ thống đối sánh một-một, xác định xem đặc trưng sinh trắc thu được có phải là của người được chỉ định hay không.

Đề tài nghiên cứu đã thực hiện phân tích xây dựng hệ thống xác thực khuôn mặt ứng dụng hỗ trợ quản lý thẻ thư viện. Bài báo được trình bày gồm các nội dung sau: phần 1 giới thiệu vấn đề, phần 2 trình bày tổng quan về hệ thống xác thực sinh trắc, phần 3 trình bày về phương pháp xác thực khuôn mặt dựa trên diện mạo sử dụng phương pháp thống kê, phần 4 là nội dung chủ yếu của đề tài, tập trung trình bày xây dựng hệ thống xác thực khuôn mặt ứng dụng hỗ trợ quản lý thẻ thư viện và bài báo được kết thúc bằng phần 5 là kết luận.

2. HỆ THỐNG XÁC THỰC SINH TRẮC

Một hệ thống xác thực sinh trắc bao gồm 2 quá trình chính: đăng ký (Enrollment) và xác thực (Verification).

- **Đăng ký (Enrollment):** Quá trình này tương ứng với việc một người đăng ký vào cơ sở dữ liệu của hệ thống. Trong pha này, đầu tiên các thiết bị chuyên dụng được dùng để thu lại các thông tin về sinh trắc dưới dạng thông tin thô. Tiếp đến các thông tin thô này được kiểm tra trong quá trình tiền xử lý nhằm đảm bảo độ tin cậy và đưa vào khôi trích chọn đặc trưng để thu lại những thông tin thích hợp, ổn định (các đặc trưng sinh trắc), được dùng cho các quá trình đối sánh sau này. Tùy thuộc vào từng ứng dụng mà các đặc trưng sinh trắc có thể được lưu trong cơ sở dữ liệu của hệ thống dưới dạng các cặp (đặc trưng, định danh) hay trên các thẻ từ (magnetic card), thẻ thông minh (smartcard)... Các đặc trưng sinh trắc lưu trong cơ sở dữ liệu được gọi là các mẫu (*template*).

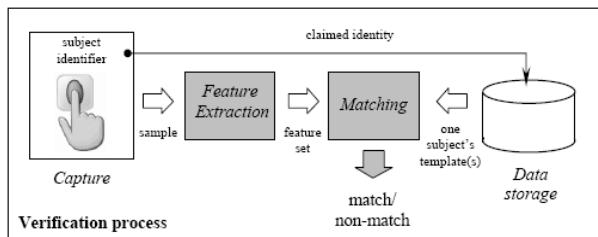


Hình 1: Quá trình đăng ký

- **Xác thực (Verification):** Xác thực là quá trình tương ứng với việc xác định xem một người có thể truy nhập vào hệ thống được hay không. Trong pha này, định danh (userid) hoặc PIN (Personal Identification Number) được người dùng cung cấp và hệ thống sẽ lấy đặc điểm sinh trắc của người này thông qua các thiết bị đọc sinh trắc tương ứng. Các thông tin này cũng phải trải qua quá trình trích chọn đặc trưng để lấy các thông tin thích hợp, sau đó đem đối sánh với template trong cơ sở dữ liệu. Kết quả của quá trình đối sánh cho biết người này có được phép truy nhập vào hệ thống hay không.

Bùi Thị Minh Yên, sinh viên lớp Truyền thông mạng, Khóa 51, Viện Công nghệ thông tin và Truyền thông, Trường Đại học Bách Khoa Hà Nội (Điện thoại: 0973.110.551, e-mail:minhyenbk@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.



Hình 2: Quá trình xác thực

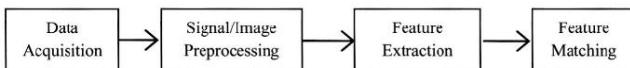
Nhìn chung các quá trình đăng ký và xác thực trong hệ thống xác thực sinh trắc đều phải trải qua 3 giai đoạn: thu nhận dữ liệu, tiền xử lý và trích chọn đặc trưng. Ở quá trình đăng ký, đặc trưng sinh trắc sau khi được lưu vào cơ sở dữ liệu dưới dạng template thì quá trình kết thúc. Còn đối với quá trình xác thực, có thêm một giai đoạn thứ 4 nữa là đối sánh.

Thu nhận dữ liệu (Data Acquisition): Dữ liệu sinh trắc (tín hiệu/ảnh) được lấy thông qua các thiết bị chuyên dụng.

Tiền xử lý (Preprocessing): Tăng cường chất lượng của dữ liệu thô bằng các biện pháp: phân đoạn, lọc nhiễu, quay và thực hiện các phép biến đổi...

Trích chọn đặc trưng (Feature Extraction): chỉ lấy các thông tin ổn định và duy nhất , dùng cho việc xác thực sau này.

Đối sánh (Feature Matching): hệ thống thực hiện so sánh giữa mẫu sinh trắc đầu vào với mẫu đã lưu trong cơ sở dữ liệu để đưa ra quyết định phù hợp.



Hình 3: Giai đoạn hoạt động của các quá trình đăng ký và xác thực

3. XÁC THỰC KHUÔN MẶT DỰA TRÊN DIỆN MẠO SỬ DỤNG PHƯƠNG PHÁP THỐNG KÊ

Hướng tiếp cận dựa trên diện mạo sử dụng phương pháp thống kê thực hiện việc xác thực một cách trực tiếp mà không cần thực hiện bước trích rút một số các đặc trưng trên khuôn mặt như mắt mũi, miệng... như hướng tiếp cận dựa trên các đặc trưng hình học. Mục đích chính của phương thức này đó là đó là đi tìm được một không gian có số chiều ít hơn không gian biểu diễn bức ảnh ban đầu, để thực hiện được mục tiêu này thì các phương pháp trong hướng tiếp cận này đều đi tìm một ma trận chiếu W rồi thực hiện việc chiếu toàn bộ tập huấn luyện cũng như là bức ảnh đầu vào lên không gian mới bằng ma trận chiếu W . Việc xác thực sẽ được thực hiện đối với các bức ảnh đã được chiếu lên trên không gian mới này. Ban đầu một bức ảnh khuôn mặt có kích thước $m \times n$ sẽ được biểu diễn bằng một vector có kích thước $I \times N$ với $N = m \times n$. Như vậy không gian ban đầu có N chiều. Thông thường thì N là rất lớn gây nên sự phức tạp cho tính toán cũng như là thời gian tính toán. Ví dụ như một bức ảnh khuôn mặt có kích thước là 128×128 thì $N = 16384$, một giá trị rất lớn cho việc tính toán các vector trên không gian R^N chiều. Với không gian mới

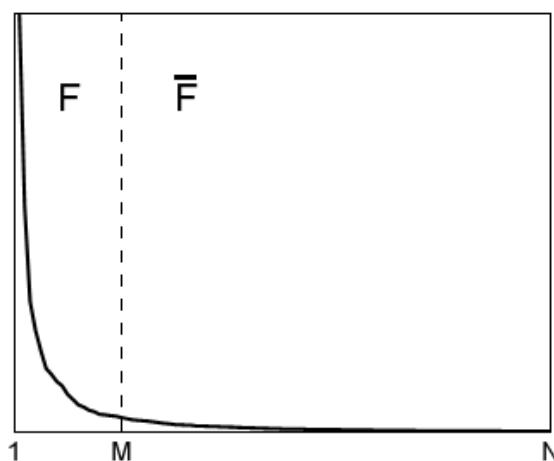
R^M có $M (<< N)$ chiều thì việc tính toán sẽ trở nên thuận lợi hơn rất nhiều. Không gian mới này còn được gọi là không gian khuôn mặt hay là không gian đặc trưng. Hơn nữa, điểm ảnh giữa các bức ảnh luôn có một độ tương quan nhất định, trong các bức một số vùng không có tác dụng trong việc xác thực, với việc chiếu lên không gian con này thì các thành phần dư thừa có thể được loại bỏ.

Các phương thức chủ yếu được sử dụng trong hướng tiếp cận này đó là PCA, SVD, ICA hay là LDA. Các phương thức này là các phương thức tuyến tính, hàm quyết định là hàm tuyến tính thường sử dụng sẽ thực hiện tính khoảng cách Euclidean.

Mục đích chính của PCA là thực hiện việc lấy tất cả các sự thay đổi trên tập huấn luyện các khuôn mặt và biểu diễn lại sự thay đổi đó chỉ với một vài tham số. Khi chúng ta làm việc với nhiều các bức ảnh, việc giảm được số chiều của không gian là rất quan trọng, nó giúp làm giảm thời gian cũng như là pharc tạp của việc tính toán.

Không gian các bức ảnh thường có sự dư thừa khi miêu tả các khuôn mặt. Đó là bởi vì mỗi điểm ảnh trong khuôn mặt có sự tương ứng cao đối với các điểm ảnh khác. Mục đích của PCA đó là giảm kích thước của không gian làm việc. Thậm chí để giảm kích thước thì một số thành phần chính (**principal component**) nên được bỏ qua. Điều này có nghĩa là một số thành phần có thể bị loại bỏ bởi vì chúng chỉ có một lượng thông tin rất ít liên quan đến đặc trưng của khuôn mặt.

Với phương pháp PCA thì không gian biểu diễn các bức ảnh khuôn mặt sẽ được biểu diễn thông qua các vector trực giao với nhau. Bất cứ một ảnh khuôn mặt nào cũng có thể là tổ hợp tuyến tính của các vector trực giao đặc trưng cho không gian đang xét. Dựa trên việc giảm các thông tin dư thừa hay các đặc trưng không cần thiết, số lượng các vector cần lưu trữ sẽ giảm đi đáng kể. Như hình trên chúng ta thấy F là phần chứa M vector đặc trưng có ý nghĩa nhất.



Hình 4: Phân chia không gian trực giao trong PCA

Thuật toán PCA giúp tính toán các thành phần cơ bản, chuyển không gian N chiều thành không gian M chiều dựa trên các vector riêng của ma trận $A^T A$ để suy ra vector riêng của ma trận hiệp phương sai $C = A A^T$. Tuy nhiên các vector tìm được này không phải là các vector đặc trưng nhất nên phương pháp PCA

thường đạt kết quả thấp hơn so với một số phương pháp khác.

Phương pháp thông thường cho kết quả tốt hơn nhưng đòi hỏi nhiều thời gian xử lý do phải tìm toàn bộ vector riêng và giá trị riêng của ma trận hiệp phương sai C và sắp xếp các giá trị riêng để tìm ra các đặc trưng tốt nhất phục vụ cho việc lưu trữ.

4. XÂY DỰNG HỆ THỐNG XÁC THỰC KHUÔN MẶT UNG DỤNG HỖ TRỢ QUẢN LÝ THẺ THƯ VIỆN

Về yêu cầu chức năng thì hệ thống cần đảm bảo 3 chức năng là chức năng đăng ký, chức năng xác thực và chức năng quản lý.

Chức năng đăng ký: Được thực hiện khi bạn đọc có yêu cầu làm thẻ thư viện.

Chức năng xác thực: Dùng khi người quản trị (hay thủ thư) muốn xác thực thông tin bạn đọc khi cần.

Chức năng quản lý: Thêm, xóa, update thông tin bạn đọc. Chức năng này được thực hiện bởi người quản lý.

Người quản lý là người có toàn quyền trong hệ thống. Với người quản lý thì ngoài quyền như một người dùng hệ thống còn có quyền xác nhận đăng ký từ người dùng, cập nhật hay tạo mới người dùng.

a) Xây dựng hệ thống:

Cài đặt OpenCV 2.1, MySQL 5.1, Matlab R2009a, Visual Studio 2008.

Xây dựng cơ sở dữ liệu cho hệ thống trên MySQL 5.1

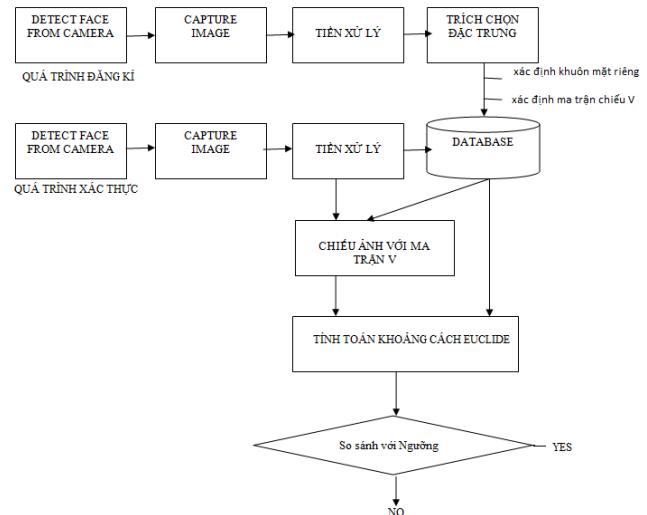
Bảng user:

Tên trường	Kiểu dữ liệu	Kích thước	Not Null	Mô tả
userid	varchar	10	x	ID người dùng(đuy nhất) có thể là số CMT
username	varchar	30	x	Tên người dùng
sex	Varchar	3	x	Giới tính người dùng
email	Varchar	50		Địa chỉ email
address	Varchar	50		Địa chỉ nơi ở(có thể là họ khẩu)
office	Varchar	50		Cơ quan

Hình 5: Cơ sở dữ liệu của chương trình

Xây dựng chương trình với ngôn ngữ sử dụng là C++ dùng thư viện mã nguồn mở OpenCV, hệ quản trị CSDL MySQL có 3 chức năng cơ bản là: Đăng ký người dùng, xác thực người dùng và quản lý người dùng.

Sơ đồ triển khai hệ thống:



Hình 6: Sơ đồ triển khai hệ thống

b) Kịch bản xác thực khuôn mặt dùng hỗ trợ quản lý thẻ thư viện.

- Trước tiên người dùng gửi yêu cầu đăng ký làm thẻ thư viện đến người quản lý bạn đọc của thư viện.
- Người quản lý lấy thông tin và đặc trưng sinh trắc người dùng để đăng ký người dùng đó vào hệ thống và cấp cho người dùng một userid – mã số thẻ thư viện
- Khi cần thiết người quản lý sẽ xác thực bạn đọc dùng chức năng xác thực.
- Khi người dùng muốn thay đổi thông tin, yêu cầu lên người quản lý, người quản lý sau khi xác thực người dùng này thì cho phép người dùng update thông tin mới.
- Khi thẻ hết hạn người quản lý thực hiện xóa người dùng khỏi hệ thống. userid này có thể dùng cho bạn đọc mới.

c) Chức năng quản lý người dùng

Cho phép:

- + Thêm 1 user vào CSDL
- + Xóa 1 người dùng trong CSDL
- + Lấy thông tin tất cả các user trong CSDL
- + Update thông tin user
- + Kiểm tra xem 1 user đã tồn tại trong CSDL hay chưa
- + Đếm xem hệ thống hiện có bao nhiêu user

d) Chức năng đăng ký

5 ảnh đầu vào tương ứng với 5 ma trận sẽ được chuyển sang dạng vector có N phần tử $N = m \times n$ với (m, n là kích thước ảnh đầu vào – 48×48 trong tích hợp của chương trình).

Các ảnh đầu vào này sẽ được ghép với ảnh của những người đăng ký trước đó. Ta giả sử trước đó có $K - 1$ người, thì sau thời điểm hiện tại sẽ có $M = K + 5$ bức ảnh được lưu trong một thư mục $D:\Database$. Ta tính ra ma trận m_face chứa tất cả các ảnh của các user trong CSDL.

Đưa ma trận m_face này vào hàm $cvSVD$ để tính ra các ma trận đặc trưng của thuật toán U, S, V.

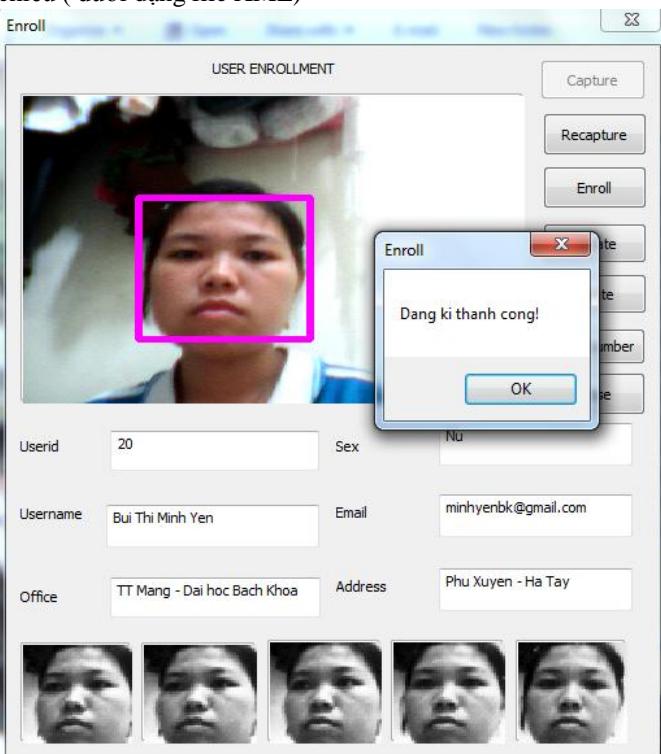
Tính toán tập của các vector trọng số bằng cách $PC = U.S$.

Giảm số chiều của ma trận PC và V . Như vậy hai ma trận này sẽ có kích thước mới tương ứng là $M \times \text{NUM_DIM}$ và $N \times$

NUM_DIM với NUM_DIM = 50 (với NUM_DIM = 50 là có thể lấy hầu hết các đặc trưng của khuôn mặt).

Mỗi một người trong ma trận PC sẽ được đại diện bởi 5 vector trọng số khác nhau tương ứng với số ảnh đầu vào cho mỗi người. Ta tính *template* ($1 \times \text{NUM_DIM}$) được chỉ gồm các vector trọng số trung bình của 5 vector trên. Như vậy mỗi một người sẽ chỉ có một vector duy nhất đại diện, ta gọi đó là khuôn mặt riêng (*eigenface*).

Tiến hành lưu trữ các ma trận thu được vào trong CSDL. Mỗi thư mục con lưu 5 ảnh đăng ký và template của người đó và toàn bộ hệ thống lưu ma trận chiếu V ($N \times \text{NUM_DIM}$) đã giảm số chiều (dưới dạng file XML).



Hình 7: Giao diện chức năng đăng kí

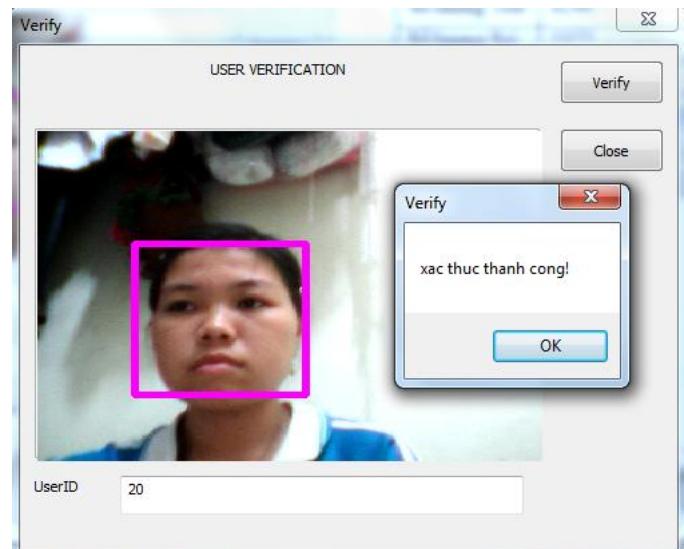
e) Chức năng xác thực:

Ảnh đầu vào I có kích thước $m \times n$ sẽ được biến đổi về vector với $N = m \times n$ phần tử (với m, n là kích thước ảnh đầu vào $m=n=48$).

Nhân vector trên với ma trận chiếu V đã lưu trong hệ thống để thu được vector đặc trưng feature.

So sánh vector đặc trưng feature thu được với template tương ứng có trong CSDL bằng việc tính khoảng cách Euclidean.

Khoảng cách này sẽ được so sánh với một n gưỡng cho trước để đưa ra quyết định xác thực có thành công hay không. Nguồn này tính dựa vào đồ thị far và frr.



Hình 8: Giao diện chức năng xác thực

f) Đánh giá kết quả thử nghiệm

Thử nghiệm đối với ảnh CSDL

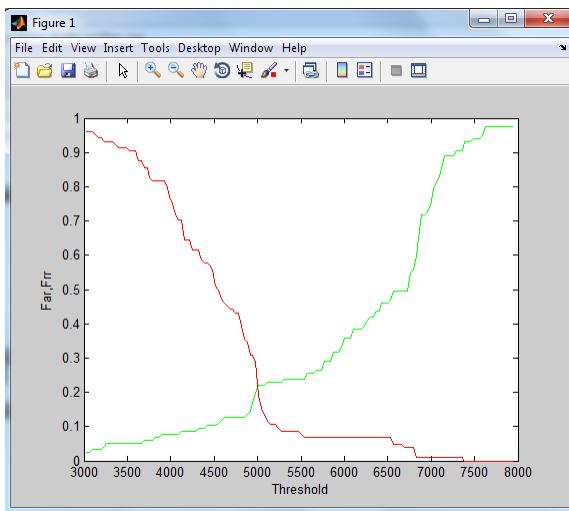
Dùng CSDL ảnh khuôn mặt ORL. Đây là một CSDL ảnh khuôn mặt gồm có 40 người mỗi người có 10 bức ảnh khuôn mặt, các bức ảnh này đều có kích thước là 48×48 . Các ảnh đã được qua một số các phép xử lý cơ bản như loại bỏ các phần khung cảnh thừa, nâng cao chất lượng của khuôn mặt.

10 bức ảnh này chủ yếu thể hiện ở các góc chụp khác nhau nhằm mục đích có vị trí tương ứng đối với các ảnh đầu vào. Một CSDL càng nhiều bức ảnh ở các tư thế khác nhau thì hiệu năng của việc nhận dạng càng tăng lên.

Trong thực nghiệm so sánh hiệu năng xác thực của thuật toán, em sử dụng 40 người làm CSDL, mỗi người lấy ảnh thứ 1, 2, 3, 4, 5, 6 như vậy CSDL của chúng ta có 240 bức ảnh dùng tính toán template. 160 bức ảnh còn lại sẽ được đưa vào hệ thống kiểm tra.

Kết quả thu được:

	Chấp nhận sai (FAR)	Từ chối sai (FRR)
Số lượng Test	6240	160
Số lượng Sai	1372	19
Tỉ lệ	21,98%	11,88%



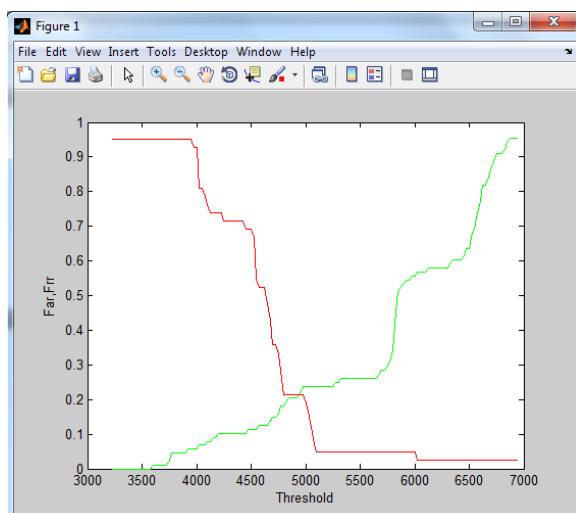
Hình 9: Đường cong FRR,FAR đối với ảnh CSDL

Thử nghiệm đối với ảnh thu nhận trực tiếp

Để kiểm tra độ chính xác của hệ thống, em đã dùng CSDL gồm 18 người, bao gồm cả nam và nữ. Mỗi người được chụp 5 ảnh dùng để tính template lưu vào CSDL. Thực hiện quá trình xác thực để tính tỉ lệ FRR là 54 lần (mỗi người 3 lần) và trường hợp để tính FAR là 306 lần (18×17).

Kết quả thu được:

	Chấp nhận sai (FAR)	Tù chối sai (FRR)
Số lượng test	306	54
Số lượng sai	69	11
Tỉ lệ	22.55%	20.37%



Hình 10: Đường cong FRR, FAR đối với ảnh sống

g) Ưu khuyết điểm.

Ưu điểm:

Thuật toán đơn giản, quá trình thực hiện xác thực nhanh

Kết quả xác thực đối với ảnh sống so với CSDL mẫu là tương tự nhau. Sai số ở mức chấp nhận được.

Khuyết điểm

Độ chính xác của thuật toán không cao.

Các phép tiền xử lý áp dụng cho hệ thống vẫn còn đơn giản và hiệu quả chưa thật cao.

CSDL ảnh sống kích thước còn nhỏ (18 người) nên các thử nghiệm còn hạn chế.

5. KẾT LUẬN

Trên đây đã trình bày nghiên cứu phát triển ứng dụng xác thực khuôn mặt hỗ trợ quản lý thẻ thư viện. Toàn bộ hệ thống đã được thiết kế cài đặt dùng Cài đặt OpenCV 2.1, MySQL 5.1, Matlab R2009a, Visual Studio 2008 dựa trên khảo sát yêu cầu ứng dụng. Các kết quả thử nghiệm trình bày trong các bảng cho thấy hướng tiếp cận của giải pháp theo phương pháp trên diện mạo sử dụng phương pháp thống kê có triển vọng phát triển cho các ứng dụng hỗ trợ quản lý thẻ thư viện không đòi hỏi khắt khe về thời gian thực và độ chính xác. Tuy nhiên giải pháp này cần tiếp tục nghiên cứu nâng cao hiệu năng hơn để đưa ra ứng dụng thực tế.

6. LỜI TRI ÂN

Đề tài được nghiên cứu và triển khai thực hiện tại phòng thí nghiệm Bộ môn Truyền thông và Mạng máy tính dưới sự hướng dẫn của **PGS.TS Nguyễn Thị Hoàng Lan**. Nhân đây cho phép em gửi lời cảm ơn chân thành nhất đến các thầy cô trong viện CNTT&TT, Bộ môn Truyền thông và Mạng máy tính và đặc biệt là **PGS.TS Nguyễn Thị Hoàng Lan** cùng **ThS Trần Quang Đức** và **các thầy trên PTN chuyên đề liên mạng** đã tận tình chỉ bảo và có những đánh giá, đóng góp quý báu giúp em hoàn thành đề tài nghiên cứu này.

7. TÀI LIỆU THAM KHẢO

- [1] Face Recognition using Eigenface Approach -Seminar Report - Submitted in partial fulfillment of the requirements for the degree of (Aditya Kelkar - Department of Computer Science and Engineering Indian Institute of Technology, Bombay Mumbai
- [2] "Face Recognition" Edited by Kresimir Delac and Mislav Grgic
- [3] "Eigenfaces for Recognition" by Matthew Turk and Alex Pentland
- [4] FaceRecognitionusingPrincipleComponent Analysis – Kyungnam Kim (Department of Computer Science University of Maryland,CollegePark MD20742,USA)
- [5] "Learning OpenCV" by Gary Bradski and Adrian Kaehler
- [6] "Đồ án tốt nghiệp" – Trần Quang Đức Bộ môn truyền thông và mạng máy tính – Khoa Công nghệ thông tin – Đại học Bách Khoa Hà Nội
- [7] Lê Minh Quang, Nghiên cứu các phương pháp và xây dựng ứng dụng cho bài toán nhận dạng khuôn mặt, Đồ án tốt nghiệp KSTN-K49 Khoa Công nghệ thông tin-Trường Đại học Bách Khoa-Hà Nội, 2009
- [8] <http://opencv.willowgarage.com/wiki/>
- [9] www.face-rec.org/algorithms

Xây dựng ứng dụng tổng đài nội bộ thoại và hội nghị VOIP trên nền Asterisk

Nguyễn Văn Nhẫn, Nguyễn Trung Hiếu

Tóm tắt - Báo cáo này trình những kết quả nghiên cứu về xây dựng ứng dụng thử nghiệm hệ thống tổng đài thoại dựa trên phần mềm nguồn mở Asterisk với các loại thiết bị đầu cuối khác nhau: máy tính cá nhân, điện thoại di động, điện thoại cố định. Trên nền hệ thống tổng đài Asterisk này, mô hình ứng dụng và triển khai các loại hình hội nghị VoIP (hội thảo, họp trực tuyến) đã được thiết kế và cài đặt trong môi trường mạng phòng thí nghiệm.

Từ khóa - Asterisk, hội nghị, tổng đài thoại, VoIP Conference

1. GIỚI THIỆU CHUNG

Một hệ thống tổng đài điện thoại mềm hoàn toàn thực hiện được các chức năng mà một tổng đài thông thường có thể làm và đồng thời nó có được những đặc tính vượt trội mà tổng đài điện thoại thông thường không có được bao gồm sự mềm dẻo, linh hoạt, dễ dàng mở rộng. Asterisk là một phần mềm mã nguồn mở và miễn phí đang được nhiều nơi quan tâm và nghiên cứu triển khai tổng đài thoại mềm. Hiện nay Asterisk là giải pháp khả thi và có khả năng đạt hiệu quả cao và đang được rất nhiều doanh nghiệp triển khai ứng dụng phù hợp với môi trường mạng nội bộ các công ty. Triển khai thành công hệ thống Asterisk sẽ đem lại một lợi ích và có hiệu quả cao phù hợp với các đơn vị vừa và nhỏ.

Đề tài nghiên cứu đã thực hiện phân tích xây dựng hệ thống tổng đài thoại trên mạng nội bộ, xây dựng và cài đặt ứng dụng các mô hình hội nghị VoIP trên nền Asterisk trong môi trường mạng phòng thí nghiệm. Bài báo được trình bày bao gồm các phần sau: phần 1 giới thiệu chung, phần 2 giới thiệu khái quát về phần mềm mở Asterisk, phần 3 trình bày thiết kế và xây dựng hệ thống tổng đài nội bộ thoại, các dịch vụ của hệ thống tổng đài được trình bày trong phần 4, phần 5 tập trung trình bày về phân tích thiết kế và xây dựng các ứng dụng hội nghị VOIP và bài báo được kết thúc bằng phần kết luận.

2. KHÁI QUÁT VỀ NGUỒN MỞ ASTERISK

Asterisk ra đời vào năm 1999 bởi một sinh viên sinh năm 1977 tên là Mark Spencer. Asterisk là hệ thống chuyển mạch mềm, là phần mềm nguồn mở được viết bằng ngôn ngữ C chạy trên hệ điều hành linux thực hiện tất cả các tính năng của tổng đài PBX

Nguyễn Văn Nhẫn, sinh viên lớp IS3, khóa 51, Dự án Việt Nhật - Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0975.522.818, e-mail: nhanhedsip@gmail.com).

Nguyễn Trung Hiếu, sinh viên lớp truyền thông mạng, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0914.991.699, e-mail: hieutn2008@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

và còn nhiều hơn thế nữa.

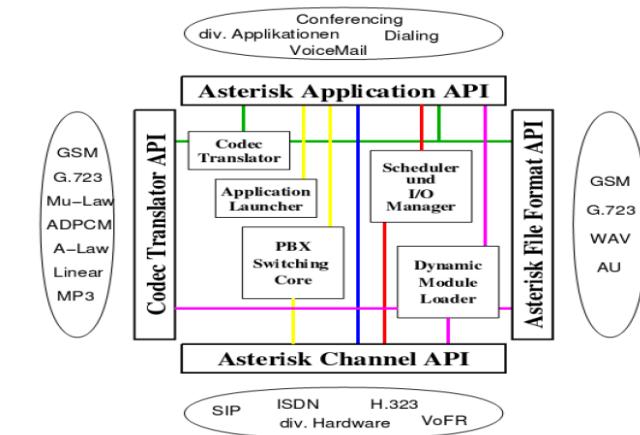
Asterisk có 4 khôi chức năng chính:

Codec translator API: các hàm đảm nhiệm thực thi và giải nén các chuẩn khác nhau như G711, GMS, G729...

Asterisk Channel API : Giao tiếp với các kênh liên lạc khác nhau, đây là đầu mối cho việc kết nối các cuộc gọi tương thích với nhiều chuẩn khác nhau như SIP, IAX, H323, Zaptel...

Asterisk file format API : Asterisk tương thích với việc xử lý các loại file có định dạng khác nhau như Mp3, wav, gsm...

Asterisk Application API : Bao gồm tất cả các ứng dụng được thực thi trong hệ thống Asterisk như voicemail, callerID...



Hình 1. Kiến trúc Asterisk [3]

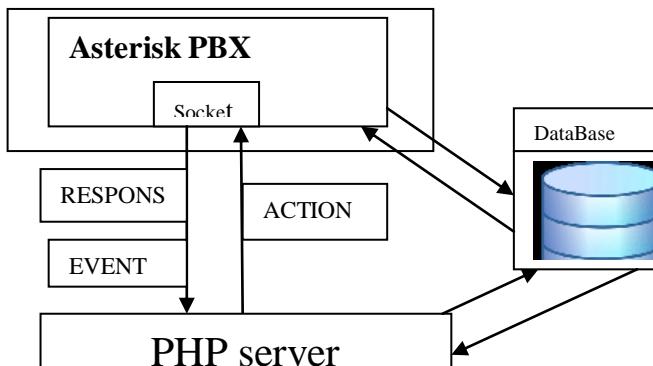
Cấu hình hệ thống Asterisk được thực hiện qua các file: sip.conf, extensions.conf, voicemail.conf ... Sau đó để thực hiện cấu hình có hiệu lực phải thực hiện reload lại toàn bộ hệ thống. Chẳng hạn khi thêm hay xóa một thuê bao thì cũng phải cấu hình trong các file tương ứng rồi reload lại hệ thống. Điều này rất bất tiện. Kiến trúc thời gian thực của Asterisk ra đời đã giải quyết được vấn đề trên. Ở kiến trúc thời gian thực thay vì phải cấu hình trong các file dạng text thì việc cấu hình sẽ được thực hiện trong database. Một khía cạnh khác của kiến trúc thời gian thực cũng đáp ứng nhu cầu người dùng lớn vì không cần nạp tất cả user cùng một lúc ngay khi khởi động asterisk. tiết kiệm chi phí tài nguyên hệ thống rất nhiều. Về hoạt động của kiến trúc thì asterisk không nạp danh sách các sip peer/sip user vào bộ nhớ. Khi có một yêu cầu đăng ký từ một thuê bao thì asterisk tìm kiếm trong bảng của database đã được định hướng trong các file extconfig.conf, res_mysql.conf rồi nạp thuê bao ấy vào bộ nhớ. Còn khi có một cuộc gọi tới thì cũng tương tự như vậy, asterisk sẽ nạp cả dòng

(record) liên quan trong database chứa extension vào bộ nhớ để thiết lập được cuộc gọi tương ứng.

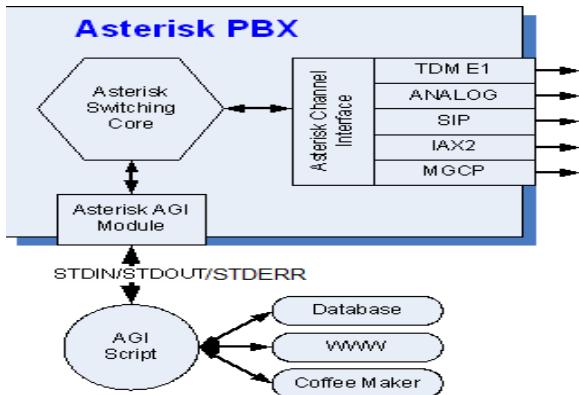
Asterisk manager API (Asterisk Manager Interface - AMI) là một cơ chế giao tiếp cho phát triển ứng dụng được Asterisk cung cấp. AMI cho phép một client kết nối đến Asterisk để phát lệnh hoặc đọc sự kiện trên giao thức TCP/IP.

AMI là một chuẩn giao tiếp để quản lý trực tiếp với server Asterisk. AMI hoạt động trên cổng 5038 cấu hình trên file/etc/asterisk/manager.conf. AMI có thể phát triển ứng dụng trên nhiều ngôn ngữ lập trình khác nhau: C, Shell, Perl, Python, php, C#.

AMI có khả năng kiểm tra trạng thái hoạt động của hệ thống Asterisk, kiểm tra hộp thư, thực thi các lệnh đối với kênh truyền thông, nhận các sự kiện xảy ra trong hệ thống và thực thi lệnh tương ứng...



Hình 2. Asterisk Manager API



Hình 3. Dịch vụ gia tăng dựa vào AGI

Asterisk Gateway Interface (AGI) là một chuẩn giao tiếp với Asterisk. AGI cho phép Asterisk gọi thực thi một chương trình ngoài để mở rộng nhiều chức năng của Asterisk như điều khiển các kênh thoại, phát âm thanh, liên kết với cơ sở dữ liệu.... Các chương trình ngoài được gọi là AGI Script, ta có thể lập trình ra các AGI Script bằng nhiều ngôn ngữ lập trình khác nhau như Perl, PHP, C, C#, Java.

3. THIẾT KẾ VÀ XÂY DỰNG HỆ THỐNG TỔNG ĐÀI THOẠI NỘI BỘ

Về yêu cầu chức năng thì hệ thống cần đảm bảo 2 chức năng là chức năng truyền thông và chức năng quản lý.

Chức năng truyền thông: Các cuộc gọi diễn ra theo đúng kịch bản đưa ra và đảm bảo chất lượng cuộc gọi tốt.

Chức năng quản lý (quản lý qua giao diện web): Tách biệt vai trò của người dùng chưa đăng ký và người dùng hệ thống, người quản lý.

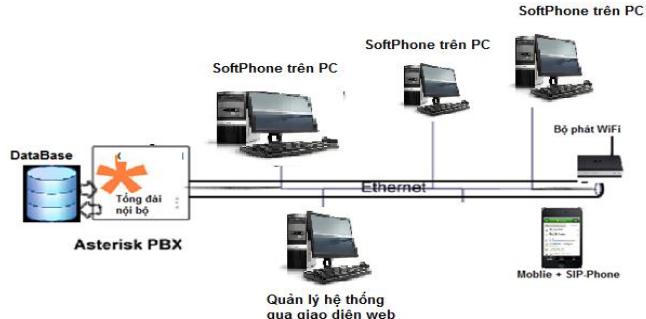
Người dùng chưa đăng ký là người dùng chưa login vào hệ thống. Với người dùng chưa đăng ký thì có thể vào mục đăng ký thông tin và dịch vụ muốn sử dụng, xem danh bạ điện thoại.

Người dùng hệ thống là người đã được cấp số thuê bao, dịch vụ và đã được người quản lý xác nhận. Với người dùng hệ thống thì có thể cập nhật thông tin cá nhân, cập nhật dịch vụ, xem thông tin về dịch vụ cung cấp, xem thông tin về những cuộc gọi của bản thân cũng như cước phí sử dụng.

Người quản lý là người có toàn quyền trong hệ thống. Với người quản lý thì ngoài quyền như một người dùng hệ thống còn có quyền xác nhận đăng ký từ người dùng, cập nhật hay tạo mới người dùng, thuê bao, extensions. Xem thông tin cước phí hay chi tiết cuộc gọi của các người dùng, xem thông tin những người dùng đang online, xem thông tin những cuộc gọi đang được thực hiện cũng như có quyền ngắt cuộc gọi.

Ngoài ra, hệ thống đưa ra một số thuê như số 100 (của người quản lý) để phục vụ việc giải đáp thông tin ở đây sử dụng cơ chế phân phối cuộc gọi tự động theo hàng đợi. Số 101 được liên kết tới AGI để cung cấp cho các thuê bao có thể thay đổi mật khẩu hay thay đổi một số dịch vụ như voicemail, nhạc chờ. Số 102 là số dành cho dành cho voicemail.

Với yêu cầu về tính năng và thiết bị hiện có, hệ thống được triển khai theo mô hình sau:



Hình 4. Mô hình xây dựng

Xây dựng hệ thống tổng đài:

Cài đặt Asterisk cùng một số gói phụ thuộc khác như Asterisk-Addon, DAHDI, LibPRI.

Xây dựng cơ sở dữ liệu và định hướng Asterisk đến cơ sở dữ liệu tương ứng.

Xây dựng Website AsteriskManager với ngôn ngữ sử dụng là PHP, có 2 chức năng cơ bản là: quản lý người dùng và cấu hình hệ thống thông qua giao diện web. Cộng đồng Asterisk cũng đã phát triển một website cho phép cấu hình và quản lý Asterisk

thông qua giao diện web đó là FreePBX. Tuy nhiên nó khá cứng nhắc và chức năng chủ yếu là để cấu hình Asterisk qua giao diện web.

Sau đây, chúng tôi trình bày so sánh giữa AsteriskManager và FreePBX:

Chức năng	AsteriskManager	FreePBX
Tin kết nối	X	
Peers Online	X	X
Quản lý người dùng	X	
Quản lý thuê bao SIP	X	X
Quản lý cước phí và nhạc chờ	X	
Người dùng đăng ký thông tin và dịch vụ	X	
Quản lý voicemail	X	X
Nghe Voicemail trên web		X
Hội nghị	X	
Tình trạng bộ nhớ CPU sử dụng		X

Bảng 1. So sánh AsteriskManager và FreePBX

Để xác định trạng thái của hệ thống như hiện tại đang có cuộc gọi giữa những user nào và cuộc gọi được bắt đầu từ thời điểm nào thì phải dựa vào những event mà Asterisk đưa ra thông qua AMI. Mỗi khi trạng thái hệ thống có thay đổi thì Asterisk sẽ sinh ra những event tương ứng. Chẳng hạn khi có kết nối giữa 2 kênh thì Asterisk sinh ra event là ‘bridge’, còn khi kết nối hủy bỏ thì event sẽ là ‘unlink’... Việc bắt được các event ‘bridge’, ‘unlink’ cùng một số tham số để cập nhật vào cơ sở dữ liệu tương ứng sẽ xác định hiện trạng cuộc gọi của hệ thống.

Để xác định trạng thái voicemail của thuê bao hay yêu cầu ‘hangup’ một kênh nào đấy thì thông qua AMI ta gửi request như ‘VoicemailStatus’, ‘Hangup’ cùng với thông tin về số thuê bao hay kênh thoại yêu cầu.

Xây dựng qui trình nghiệp vụ sử dụng hệ thống:

- Trước tiên người dùng cần đăng ký thông tin cá nhân và dịch vụ yêu cầu thông qua website.
- Người quản lý xác nhận thông tin đăng ký của người dùng đăng ký.
- Sau khi đã được xác nhận thì người dùng trở thành người dùng hệ thống. Người dùng có thể đăng nhập vào website để xem số thuê bao được cấp phát cũng như cập nhật thông tin khác.
- Người dùng cấu hình thiết bị VoIP tương ứng với số thuê

bao được cấp. Có thể gọi điện đến số dịch vụ 101 để cập nhật lại mật khẩu.

4. XÂY DỰNG MỘT SỐ DỊCH VỤ TRONG HỆ THỐNG

Dịch vụ ‘call center’ cho số tổng đài ‘100’: Với mỗi số thuê bao gọi đến số ‘100’ đều được đưa vào hàng đợi rồi lần lượt được kết nối đến số tổng đài, thời gian tối đa nằm trong hàng đợi được đặt là 600s.

Dịch vụ “nhạc chờ”: Khi một thuê bao gọi đến một thuê bao khác mà chủ nhân có cài đặt nhạc chờ thì thay vì nghe tiếng tút người gọi có thể nghe nhạc mà chủ nhân của số bị gọi thiết lập. Ngoài ra xây dựng giao diện để người quản lý có thể cập nhật được dữ liệu nhạc chờ và người dùng có thể xem thông tin những bài hát nhạc chờ.

Dịch vụ ‘voicemail’: Khi thuê bao được gọi không bắt máy hoặc khi không online thì những cuộc gọi đến sẽ được chuyển vào hộp thư.

Mặt khác, Dựa trên cơ sở Asterisk cung cấp AGI để mở rộng chức năng của Asterisk. Xây dựng số trung tâm dịch vụ để người dùng có thể gọi đến để có thể tự động cập nhật dịch vụ của bản thân như đăng ký hay hủy bỏ dịch vụ voicemail, dịch vụ nhạc chờ, thay đổi bài hát nhạc chờ...

5. PHÂN TÍCH THIẾT KẾ VÀ XÂY DỰNG CÁC ỨNG DỤNG HỘI NGHỊ VOIP

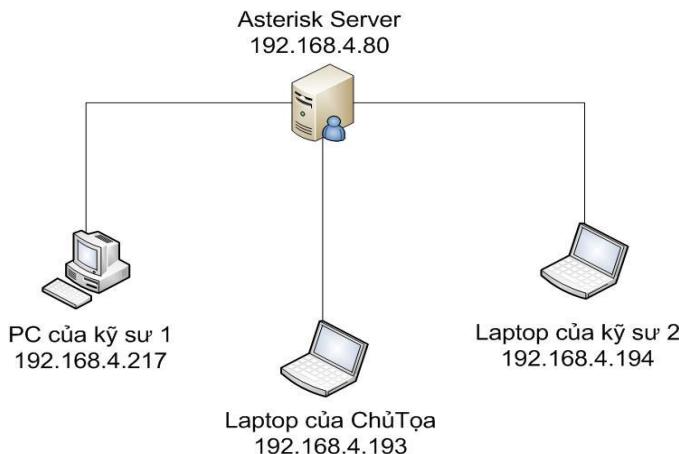
Trong vài năm trở lại đây, nhờ sự cố gắng của cộng đồng [4], Asterisk đang ngày một phát triển hơn. Đặc biệt là từ phiên bản Asterisk-1.6.x trở lên, ứng dụng hội nghị đã trở nên khá hoàn thiện, hoàn toàn đủ khả năng thực hiện những cuộc họp quy mô nhỏ, với chi phí rất thấp, và dần dần được sử dụng rộng rãi tại các doanh nghiệp vừa và nhỏ. Có rất nhiều ứng dụng hội nghị được phát triển trên nền tảng Asterisk (có thể kể đến: Meetme, Conference, ConfBridge, Konference...), nhưng được sử dụng nhiều nhất vẫn là ứng dụng Meetme kèm theo bộ điều khiển hội nghị MeetmeAdmin vì khả năng xây dựng và quản lý hội nghị tương đối hoàn chỉnh. Sau đây chúng tôi xin được trình bày nguyên tắc ứng dụng lõi Meetme và cách khai thác sử dụng để xây dựng ứng dụng hội nghị VOIP trên môi trường mạng LAN.

Meetme[5] có thể hiểu là một phòng họp. Phòng họp này được xác định qua 2 yếu tố: tên phòng họp (confno) và mã số (PIN). Nếu có 2 thành viên bất kỳ cùng nhập đúng confno và PIN từ điện thoại của họ, khi đó họ đã đang ở trong cùng một phòng, nghĩa là có thể nói và nghe được đối phương.

MeetmeAdmin[6] thì lại như người chủ tịch hội nghị, có toàn quyền quyết định trong hội nghị (cho phép phát biểu, mời thành viên khác ra khỏi cuộc họp ...)

Chỉ với hai ứng dụng trên, chúng ta đã có đầy đủ cơ sở để xây dựng phần lõi của một hội nghị VOIP. Tuy nhiên, hội nghị thực tế thì không đơn giản chỉ có vậy. Sau đây, chúng tôi xin trình bày thiết kế một kịch bản hội nghị Asterisk. Kịch bản này dựa trên thực tế xây dựng và cài đặt thử nghiệm trong môi trường mạng

cục bộ tại Phòng thí nghiệm (PTN) Bộ môn Truyền thông và Mạng, Viện CNTT&TT.



Hình 5. Sơ đồ mạng thực tế triển khai thử nghiệm hội nghị.

Kịch bản hội nghị cục bộ trong môi trường PTN.

Giai đoạn chuẩn bị hội nghị:

- Chủ tọa và các thành viên coi như đã đăng ký và có tài khoản trên server.
- Chủ tọa đăng nhập vào website AsteriskManager đặt trên Server, chủ tọa có thể thực hiện mời họp theo 2 cách: chủ động gửi lời mời đến từng thành viên(conference), hoặc gửi thông báo cho toàn thể các thành viên(hội thảo)
- Khi nhận được lời mời, các thành viên sẽ xác nhận có hoặc không tham gia hội nghị.
- Thông tin về hội nghị (confno, PIN, PIN của chủ tọa, thời gian họp) sẽ được lưu trong cơ sở dữ liệu, và phòng họp sẽ được tự động tạo ra trước thời điểm họp 1 tiếng, các thành viên sẽ được vào phòng họp trước thời điểm họp 5 phút.
- Tất cả các thành viên đều có thể xem danh sách các thành viên đang trong cuộc họp bằng cách thao tác trên giao diện Web.

Giai đoạn bắt đầu hội nghị

- Các thành viên gọi điện đến số 103 và thực hiện theo hướng dẫn (nhập confno và PIN – thông tin này được gửi tự động đến hộp thư của mọi thành viên)

Giai đoạn tiến hành hội nghị

- Trong cuộc họp mọi thành viên (kể cả chủ tọa) có thể bấm phím * để vào menu điều khiển của mình, và có thể điều chỉnh theo hướng dẫn
- Khi chủ tọa vào phòng họp, tất cả thành viên mất quyền nói, và chỉ có chủ tọa được phát biểu.

3. Sau khi phát biểu, tùy vào nội dung của cuộc họp, chủ tọa có thể ấn phím # để vào menu quản trị, và thực hiện theo hướng dẫn. Ví dụ: có thể cho phép thành viên nào đó phát biểu bằng cách ấn phím#, 1, [số điện thoại thành viên đó – xem trên Web].

4. Nếu sắp hết thời gian đăng ký hội nghị, hệ thống sẽ thông báo hội nghị sẽ kết thúc trong vòng 2 phút tới. Lúc này, các thành viên có thể chủ động bấm # để rời khỏi phòng họp, hoặc chủ tọa có thể bấm #, 4 để mời tất cả mọi thành viên còn lại ra khỏi phòng họp. Cuộc họp kết thúc.

Ưu khuyết điểm.

Ưu điểm:

- Người dùng(chủ tọa và các thành viên tham dự) có thể sử dụng hệ thống một cách dễ dàng bằng cách sử dụng đồng thời Website (AsteriskManager) và softphone (X-lite 4.0)

- Sử dụng Website (AsteriskManager-phần hội nghị) người dùng có thể (mời họp, đọc thông báo họp, kiểm tra danh sách thành viên đăng ký, danh sách thành viên đã tham gia vào cuộc họp..)

- Sử dụng softphone(X-lite 4.0), người dùng có thể thay đổi tương tác của mình với hệ thống (thoát phòng họp, thay đổi âm lượng..) chủ tọa có thể quản lý các thành viên tham gia cuộc họp (cho/ không cho phép nói, yêu cầu ra khỏi phòng họp..) tất cả điều thao tác trên 12 phím điện thoại.

- Hội nghị đã được phân luồng: tùy từng giai đoạn của cuộc họp mà số lượng người được nói khác nhau. Thông thường chỉ có 1 – 2 người cùng nói, nên chất lượng âm thanh tốt, rõ ràng, trễ ít. Khi có nhiều người cùng nói (khi thảo luận) chất lượng âm thanh sẽ bị giảm đi rõ rệt.

- Mọi cuộc họp đều phải được mời và chấp thuận thông qua hệ thống Website AsteriskManager, nên có ưu điểm về mặt lưu trữ và tra cứu thông tin (thời điểm họp, số người tham gia ..)

- Hội nghị được tạo ra một cách hoàn toàn tự động tại đúng thời điểm bắt đầu hội nghị, vừa tránh lãng phí tài nguyên, vừa giảm bớt phức tạp cho người điều khiển hội nghị.

Khuyết điểm:

Tuy nhiên, hội nghị sử dụng Meetme còn có những khuyết khuyết như sau:

- Hiện giờ chưa hỗ trợ các cuộc họp có truyền hình (Video).
- Hội nghị chưa được xử lý gì thêm về kỹ thuật truyền tín hiệu (echo, trễ ..) nên phải dựa vào đường truyền.
- Ngoài softphone, chưa có thiết bị nào khác được sử dụng để thử nghiệm hội nghị.

6. KẾT LUẬN

Trên đây đã trình bày các kết quả nghiên cứu và triển khai ứng dụng công nghệ VOIP của nhóm sinh viên. Toàn bộ hệ thống tổng đài đã được phân tích thiết kế xây dựng và triển khai cài đặt thử nghiệm trên môi trường mạng LAN tại phòng thí nghiệm. Các kết quả bước đầu đã đạt các yêu cầu đề ra về hệ thống tổng đài nội bộ và về các ứng dụng về hội nghị VOIP bao gồm 2 loại

hình: Họp trực tuyến và Hội thảo trực tuyến. Các kết quả thử nghiệm đạt được có triển vọng phát triển tốt phù hợp với các yêu cầu tổng đài thoại và các ứng dụng của các đơn vị vừa và nhỏ, triển khai trên nền mã nguồn mở Asterisk có tính kinh tế và khả thi trên thực tế.

7. LỜI TRI ÂN

Nội dung đề tài được nghiên cứu và triển khai thực hiện tại phòng thí nghiệm Bộ môn Truyền thông và Mạng máy tính dưới sự hướng dẫn của PGS.TS Nguyễn Thị Hoàng Lan. Chúng em gửi lời cảm ơn chân thành nhất đến các thầy cô trong Viện CNTT&TT, Bộ môn Truyền thông và Mạng và đặc biệt là **PGS.TS Nguyễn Thị Hoàng Lan** cùng **ThS. Trần Quang Đức, KS. Đào Vũ Hiệp, ThS. Trần Nguyên Ngọc** đã tận tình chỉ bảo và có những đánh giá, đóng góp quý báu giúp chúng em hoàn thành đề tài nghiên cứu này.

7. TÀI LIỆU THAM KHẢO

- [1] Jim Van Megelen, Jared Smith, and Leif Madsen, “Asterisk The Future of Telephony:”
- [2] Alan B. Johnston “Internet Communications Using SIP - Delivering VoIP and Multimedia Services with Session Initiation Protocol” - Second Edition Henry Simreich
- [3] Lê Quốc Toản “Asterisk”
- [4] <http://www.asterisk.org>
- [5] <http://www.voip-info.org/wiki/view/Asterisk+cmd+MeetMe>
- [6] <http://www.voip-info.org/wiki/view/Asterisk+cmd+MeetmeAdmin>
- [7] <http://phpagi.sourceforge.net/>
- [8] <http://Asterisk wikipedia.org>

Bộ thu thập trang Web ẩn theo chủ đề

Vũ Thành Đô, Bùi Anh Đức

Tóm tắt – Web ẩn hay còn gọi là Deep Web là một hướng đi vô cùng mới mẻ trong lĩnh vực tìm kiếm thông tin. Trên cơ sở phát triển máy tìm kiếm Tiếng Việt theo chủ đề, đề tài nghiên cứu xây dựng một bộ thu thập nhằm giúp máy tìm kiếm này có khả năng tiếp cận thông tin Deep Web. Đề tài áp dụng phương pháp Surfacing do Google giới thiệu và thực hiện những cải tiến nhằm phục vụ cho mục đích tìm kiếm theo chủ đề. Đầu tiên cần phải có phương pháp để bóc tách thông tin của HTML Form. Tiếp theo khi thực hiện phương pháp Surfacing cần có một giải thuật lựa chọn các input phù hợp, đồng thời cần phải có một giải thuật sinh ra các từ khóa phù hợp để điền vào textbox trong Form, sao cho tối đa khả năng tìm ra các tài liệu đúng chủ đề.

Từ khóa – Crawling, Deep Web, Surfacing, Topical

1. GIỚI THIỆU

Cùng với sự phát triển mạnh mẽ của Internet thì nhu cầu tìm kiếm thông tin cũng ngày càng cao. Máy tìm kiếm Tiếng Việt theo chủ đề ứng dụng phương pháp học tăng cường [6] đã rất thành công trong việc tìm ra những tài liệu đúng chủ đề với hiệu năng cao. Tuy nhiên máy tìm kiếm này vẫn chỉ đơn thuần crawl các liên kết tĩnh, tức là các liên kết nằm trong nội dung HTML của trang web. Chúng ta biết rằng còn có những liên kết được sinh ra từ những HTML Form. Đó là những liên kết chi sinh ra khi ta gửi thông tin của Form, tạo thành những câu truy vấn để truy cập vào cơ sở dữ liệu. Bởi vậy khái niệm Deep Web hay Hidden Web để chỉ những trang web động được tạo ra từ những truy vấn cụ thể với Form tìm kiếm. Khái niệm này trái ngược với Surface Web là những nội dung Web có thể crawl bởi các liên kết tĩnh thông thường.

Theo những số liệu nghiên cứu [2] thì có khoảng 43000 đến 96000 trang Deep Web chứa 7500 Terabytes dữ liệu, gấp 500 lần Surface Web. So sánh một cách tương đối thì Surface Web chỉ chiếm 30% lượng thông tin trên Internet trong khi Deep Web chiếm tới 70%, một lượng thông tin khổng lồ ẩn giấu trong các cơ sở dữ liệu.

Các website ngày nay thường có một cơ sở dữ liệu nằm bên

Đề tài này được thực hiện dưới sự hướng dẫn của TS Trần Đức Khanh, bộ môn Hệ thống thông tin, Viện Công nghệ thông tin và truyền thông, trường Đại học Bách Khoa Hà Nội.

Vũ Thành Đô, sinh viên lớp Hệ thống thông tin, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (e-mail: dovuthanh@gmail.com).

Bùi Anh Đức, sinh viên lớp AS1-Việt Nhật, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (e-mail: anhduc.k51hedsp@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

trong, chẳng hạn với một website tuyển dụng thì đó là dữ liệu về thông tin tuyển dụng, yêu cầu tuyển dụng, mức lương... hay một website bán hàng thì đó là những thông tin chi tiết về các sản phẩm mà website cung cấp. Những website đó thường có một giao diện truy vấn hay chính là các Form tìm kiếm để người dùng có thể tìm kiếm thông tin. Để tiếp cận được những thông tin nằm trong cơ sở dữ liệu của trang Web đó, người sử dụng phải tiến hành điền vào Form và submit Form. Khi mà các máy tìm kiếm truyền thống hoạt động theo nguyên tắc lần theo các liên kết trên trang Web thì những nội dung từ phía sau Form tìm kiếm hoàn toàn là ẩn đối với chúng.

Trên cơ sở phát triển máy tìm kiếm Tiếng Việt theo chủ đề, đề tài này tập trung vào việc xây dựng một bộ thu thập trang Web ẩn theo chủ đề nhằm giúp máy tìm kiếm này có thể tiếp cận được nội dung Deep Web.

Phản tiếp theo của báo cáo bao gồm những phần sau: Phần 2 giới thiệu những nghiên cứu có liên quan, Phần 3 mô tả phương pháp xử lý Form, Phần 4 mô tả phương pháp lựa chọn trường nhập liệu hay là các input phù hợp, Phần 5 mô tả phương pháp sinh từ khóa để điền vào textbox, Phần 6 là kết quả thực nghiệm và cuối cùng là những kết luận trong phần 7.

2. NGHIÊN CỨU LIÊN QUAN

Một trong những nghiên cứu nổi bật về Deep Web Crawling là phương pháp Surfacing do Google giới thiệu [1]. Phương pháp này tính toán cách submit HTML form phù hợp nhất để lấy được kết quả trả về là các URL phù hợp với nội dung tìm kiếm. Ưu điểm của nó là tận dụng được nền tảng của các máy tìm kiếm phổ thông, chỉ khác biệt ở điểm lấy ra các URL thông qua việc submit Form. Bởi thế nội dung Deep Web có thể được ghép lẫn vào kết quả crawl như những nội dung thông thường khác.

Việc tiến hành điền thông tin và submit Form thì giống như việc tạo ra những truy vấn và gửi đi. Mục tiêu của phương pháp Surfacing là tìm ra một tập hợp các truy vấn phù hợp để tiến hành submit Form. Để thực hiện phương pháp này chúng ta vấp phải hai trở ngại: lựa chọn trường nhập dữ liệu nào để tạo ra truy vấn và tìm ra giá trị thích hợp để điền vào những trường nhập liệu đó. Đối với vấn đề thứ nhất, giải pháp là xây dựng một phương pháp Test Informative để lựa chọn ra tập hợp các trường nhập liệu phù hợp, sao cho với số lượng truy vấn không quá nhiều vẫn lấy được phần lớn dữ liệu nằm trong cơ sở dữ liệu của website. Đối với vấn đề thứ hai, những trường nhập liệu có sẵn giá trị lựa chọn (select menu, radio button...) thì việc lấy được giá trị để điền vào là rất đơn giản. Tuy nhiên với trường nhập liệu dạng text box thì bắt buộc phải có một giải pháp sinh ra những từ khóa để điền vào

thì mới có thể tiến hành submit. Giải pháp là phương pháp thăm dò lặp [3] để sinh ra tập hợp từ khóa cho text box tương ứng.

Dựa vào những nghiên cứu trên, đề tài thực hiện phương pháp Surfacing với những cải tiến như:

- Thực hiện quá trình xử lý phân loại Form trước khi thực hiện quá trình Surfacing. Từ đó tiến hành Test Infomative hoặc sinh từ khóa tùy theo dạng Form tìm kiếm.
- Khi tiến hành sinh từ khóa cho text box lựa chọn những từ khóa có trọng số cao đối với chủ đề tìm kiếm nhằm tăng khả năng tìm thấy tài liệu đúng chủ đề.

3. XỬ LÝ FORM

Đối với mỗi Deep Website thì Form tìm kiếm có thể coi là lối vào duy nhất để tiếp cận được cơ sở dữ liệu bên trong. HTML Form được định nghĩa bên trong thẻ `<form>`. Chúng được dùng với mục đích chung là để người dùng có thể nhập dữ liệu và gửi đến server nhằm thực hiện một xử lý nào đó.

Ví dụ với một HTML Form như Hình 1 bao gồm 2 textbox để điền thông tin Họ tên, Lớp của sinh viên và một select menu để chọn thông tin khóa học. Khi nhấn nút submit trên Form những thông tin này sẽ được chuyển tới server để xử lý. Form có các thành phần là các input. Input có nhiều loại như textbox, select menu, check box, các nút radio và nút submit. Mỗi một input có một thuộc tính là name chính là tên của tham số ứng với input đó. Những tham số này sẽ được truyền tới server khi người dùng ấn nút submit.

Form có 2 thuộc tính quan trọng là method và action. Action chỉ ra đường dẫn tới một địa chỉ trên server nhằm xử lý các thông tin hay chính là những tham số từ Form truyền tới. Method quy định cách thức truyền tham số tới server. Có 2 cách thức truyền là GET và POST:

- Với phương thức GET, tham số được gắn vào với action và là một phần của URL trong HTTP Request. Với Form như hình 1 thì URL đó có thể là:
`http://abc.com/find?name=duc&class=AS1&course=1&s=Gửi"`
- Với phương thức POST, tham số nằm trong thân nội dung của HTTP Request và URL chỉ đơn giản là action:
`http://abc.com/find`

Như vậy URL sinh ra bởi phương thức GET là phân biệt trong khi phương thức POST sẽ chỉ sinh ra một URL. Máy tìm kiếm đánh chỉ mục cho các trang web dựa trên URL của chúng, bởi thế nó không thể đánh chỉ mục cho trang kết quả trả về bằng phương thức POST. Như vậy giới hạn của đề tài sẽ chỉ quan tâm đến những Form có phương thức GET.

Trong một website, ngoài Form tìm kiếm còn có thể chứa rất nhiều loại Form khác như Form đăng nhập, đăng ký, Form gửi lời bình, đóng góp... Đó là những loại Form không giúp ta lấy được dữ liệu từ CSDL. Bởi vậy cần có một quy trình lọc, loại bỏ

nhiều loại Form này không đưa vào xử lý. Hướng giải quyết ở đây là loại bỏ những Form đăng nhập, đăng ký bằng những trường input có dạng password hoặc những input có yêu cầu thông tin cá nhân như username. Đối với những Form dạng gửi lời bình, ý kiến đóng góp ta lọc những trường input dạng textarea.



```
<form method="GET" action="http://abc.com/find">
    Họ tên: <input type="text" name="name" /><br />
    Lớp: <input type="text" name="class" /><br />
    Khóa: <select name="course"><option value="1">K51</option>
          <option value="2">K52</option>
          <option value="3">K53</option>
          <option value="4">K54</option>
        </select><br />
    <input type="submit" name="s" value="Gửi" />
</form>
```

Hình 1 – Ví dụ Form

Khi đã có được Form tìm kiếm thỏa mãn tất cả các điều kiện, ta cần trích rút các thông tin cần thiết của Form để lưu trữ cho các xử lý sau này. Ngoài những thuộc tính quan trọng của Form (action, method) thì thông tin về các trường input của Form cần được xử lý chi tiết. Input của Form bao gồm rất nhiều loại như:

- Text Input: Nhập dữ liệu dạng text trên 1 dòng
- Select menu: Một menu sổ xuống có nhiều lựa chọn với các giá trị cho sẵn
- Check box: Một hoặc một nhóm các ô. Khi tích vào ô nào tức là ta lựa chọn giá trị ở ô đó. Có thể coi đây là lựa chọn một hoặc nhiều.
- Nút Radio: Một nhóm các ô tròn, chỉ được phép chọn một trong số những giá trị (không quá nhiều). Đây là lựa chọn một trong nhiều.
- Và một số các loại input khác

Trước tiên cần xác định những thông tin nào của input cần lưu trữ. Như ta đã biết Form với phương thức GET truyền dữ liệu trực tiếp trong URL bằng những tham số chính là thuộc tính "name" của input. Bởi vậy chúng ta cần lưu trữ thuộc tính này của từng Form Input. Đối với những input có tập giá trị cho trước như select menu, checkbox, nút radio thì cần phải lấy được tập giá trị này và lưu trữ cho từng input tương ứng. Ngoài ra mỗi input trong Form thường có một nhãn nằm phía bên trái hoặc bên trên chỉ ra tên gọi của input đó. Nhãn này thường không giống với thuộc tính name của input. Cần lưu ý nhãn để người dùng nhận biết input còn name phục vụ cho việc truyền tham số, bởi vậy name thường do người tạo ra form tự quy ước còn nhãn sẽ mang thông tin chính xác về kiểu input. Bởi vậy cần phải có xử

lý để lấy được nhãn của các input trong Form.

Tuy nhiên cấu trúc các thẻ trong Form thường chen lẫn những thẻ mang tính chất trình bày như `<table><tr><td>` `<dd>` `<dt>` ... khiến cho cấu trúc thực của các thẻ trong Form vô cùng đa dạng. Phương pháp hiện tại để lấy được nhãn của input là duyệt ngược cấu trúc DOM của Form bắt đầu từ thẻ `<input>` cho đến khi lấy được nội dung text. Khi đó ta hi vọng nội dung đó sẽ là nhãn của input.

4. LỰA CHỌN TRƯỜNG NHẬP LIỆU

Form tìm kiếm chính là giao diện truy vấn của Deep Website. Có thẻ khảng định như vậy vì mỗi một Form thường ứng với một cơ sở dữ liệu hoặc một bảng cơ sở dữ liệu và việc submit Form chính là thực hiện các truy vấn trên bảng cơ sở dữ liệu đó. Ta có thể mô hình hóa ứng với Form tìm kiếm f_D là một cơ sở dữ liệu gồm một bảng duy nhất D với m thuộc tính. f_D dùng để truy vấn đến D và có n input X_1, X_2, \dots, X_n . Mỗi một lần submit là một lần thực hiện một truy vấn có dạng:

```
SELECT * FROM D WHERE P
```

Trong đó P là điều kiện chỉ ra bởi giá trị các input của Form.

Như vậy vấn đề Surfacing trở thành vấn đề chọn ra một tập hợp các truy vấn phù hợp để submit Form, hay chính là việc lựa chọn một tập hợp input phù hợp để tiến hành điền giá trị và tạo ra các truy vấn. Từ đó ta đưa ra định nghĩa Query Template:

* *QueryTemplate là một tập con của tập hợp các Form Input*

Những input trong Query Template được gọi là binding input, những input còn lại là những input tự do. Bằng việc gán các giá trị khác nhau cho những binding input thì tương ứng là rất nhiều lần submit form. Như vậy một Query Template tương đương với tất cả những truy vấn SQL có dạng:

```
SELECT * FROM D WHERE P_B
```

Trong đó P_B là những điều kiện chỉ ra bởi các binding input nằm trong Query Template.

Có hai thách thức chính trong việc lựa chọn các template:

Thứ nhất, chúng ta mong muốn lựa chọn các template không chứa các input dạng trình bày, bởi các điều kiện này chỉ có tác dụng bố trí nội dung kết quả trả về mà không hề làm việc với CSDL bên dưới. Những input này thường là dạng sắp xếp kết quả trả về theo tiêu chí nào đó. VD: Sắp xếp tăng dần, Sort by Name...

Thứ hai, chúng ta phải lựa chọn một template có kích thước hợp lý. Việc lựa chọn một template có kích thước lớn nhất có thể sẽ đảm bảo độ phủ tối đa bằng việc tạo ra tất cả các truy vấn có thể. Tuy nhiên, phương pháp này sẽ tiêu tốn nhiều tài nguyên, đồng thời trả về khá nhiều tập kết quả rỗng. Nếu chúng ta lựa chọn các template có kích thước nhỏ hơn, chúng ta sẽ chỉ phải submit Form với số lần ít hơn.

Phương pháp chúng ta sử dụng để lựa chọn Query Template phù hợp là Infomativeness Test (phương pháp kiểm tra tính thông tin). Phương pháp này sẽ đánh giá Query Template dựa trên tính phân biệt của các trang web kết quả thu được từ những lần submit form với tập giá trị khác nhau. Như vậy cần ước tính số lượng các trang web phân biệt mà template tạo ra bằng cách phân cụm chúng dựa trên tính tương đồng nội dung giữa chúng. Nếu số cụm là nhỏ so với số lần submit Form, lý do có thể là:

- Template chưa đầu vào input dạng trình bày, như vậy sẽ có nhiều lần submit cho số lượng kết quả trả về là giống nhau

- Kích thước của template là quá lớn với CSDL phía dưới, và

```
Get Informative Query Templates (W: WebForm) {
    L: Set of Input = Get Candidate Inputs (W)
    candidates: Set of Template = { T:Template | T.binding
        = {I}, IεL }
    informative: Set of Template = φ

    while (candidates ≠ φ) {
        newcands: Set of Template = φ
        for each (T: Template in candidates) {
            if ( Check Informative (T, W) ) {
                informative = informative ∪ { T }
                newcands = newcands ∪ Augment (T, I)
            }
        }
        candidates = newcands
    }
    return informative
}

Augment (T:Template, I:Set of Input) {
    return { T' | T'.binding = T.binding ∪ { I }, IεP, I ∈ T.binding}
}
```

Hình 2 – Giải thuật ISIT

do đó có phần lớn kết quả trả về là "không tìm thấy dữ liệu", các trang web thông báo kết quả này là hoàn toàn giống nhau.

- Có lỗi xảy ra trong Template, khiến một số lượng lớn các trang web thông báo lỗi sẽ được tạo ra, các trang này cũng tương tự nhau.

Chúng ta gọi một Template là informative nếu nó tạo ra các trang kết quả có nội dung phân biệt, ngược lại, chúng ta sẽ gọi là uninformative. Sau khi tiến hành phân cụm các trang web mà ta thu được từ quá trình submit Form, chúng ta sẽ gán cho một template là uninformative nếu nó sinh ra ít cụm so với số lượng submit. Trên cơ sở đó, chúng ta có định nghĩa sau:

Định nghĩa Informative Query Template

Gọi T là một Query Template và K là số cụm của các trang HTML. G là tập tất cả các form submission (hay là các URL) có thể được sinh ra bởi T .

Khi đó, chúng ta nói rằng T là một Informative Template nếu $K|G| \geq \tau$. Trái lại, ta nói T là uninformative.

Tỉ số $K|G|$ được gọi là hệ số phân biệt (distinctiveness fraction)

Với một Form có n input thì để tránh việc phải vét cạn không gian $2^n - 1$ template, chúng ta phát triển một giải thuật sao cho nó chỉ kiểm tra những template có nhiều khả năng là informative. Ý tưởng cơ bản của giải thuật này là tìm kiếm trên không gian template theo cách tiếp cận bottom-up, bắt đầu từ những template có một input. Nếu Template T có kích thước k, và không có template nào mà nó kế thừa là informative, thì hiển nhiên, nó cũng phải là uninformative. Đó cũng chính là ý tưởng của giải thuật ISIT (**Error! Reference source not found.**).

Trong giải thuật ISIT, chúng ta bắt đầu với các ứng viên là các template có kích thước 1.

Hàm *GetCandidateInputs* lựa chọn tập các đầu vào trên một form cụ thể:

- Select menu
- Những đầu vào khác được thiết lập giá trị mặc định

Hàm *CheckInformative* kiểm tra các ứng viên xem liệu chúng có phải là informative không theo định nghĩa.

Hàm *Augment* xây dựng ứng viên có kích thước bằng kích thước của ứng viên hiện tại cộng thêm 1 phần tử. Do đó, các ứng viên được sinh ra sau phải có cha là các ứng viên đã được xếp vào loại informative, chúng được sinh ra bằng cách kết hợp ứng viên cha với một phần tử trong tập đầu vào.

Giải thuật sẽ kết thúc khi không còn informative template nữa. Có một trường hợp giải thuật này không đúng, đó là khi tất cả các template có kích thước là 1 đều không thỏa mãn, như vậy giải thuật sẽ dừng lại luôn. Tuy nhiên, vẫn có trường hợp khi các template có kích thước 1 không thỏa mãn những các template có kích thước 2 lại thỏa mãn. Ví dụ, ta có hai thuộc tính *a* và *b*. Thuộc tính *a* được thiết lập sao cho nếu để mặc định giá trị của nó, và chỉ chọn *b* thì sẽ không có kết quả nào được trả về, và ngược lại, nếu chỉ chọn *a* và để mặc định *b* thì cũng không có kết quả trả về, kết quả chỉ thực sự được trả về khi ta chọn *a* và *b* một cách phù hợp. Điểm yếu này có thể được giải quyết bằng việc kiểm tra thêm các mẫu có kích thước 2 khi không một mẫu nào có kích thước 1 thỏa mãn.

ISIT cho phép chúng ta tìm ra tập hợp informative query template trong khi số lần kiểm tra khá là ít.

Sau khi quá trình tìm kiếm kết thúc, chúng ta thu được một tập các informative template, chúng ta có thể lấy các URL thu được từ tập các informative template này làm đầu vào cho việc đánh chỉ số, qua đó tận dụng được cơ sở hạ tầng của các máy tìm kiếm hiện tại.

5. SINH TỪ KHÓA

Một số lượng lớn các HTML Form có text box. Hơn thế nữa, một số Form có chứa các select menu đòi hỏi giá trị trong text

box phải chuẩn xác trước khi chúng có thể trả về bất cứ một kết quả nào. Text box được sử dụng theo hai cách khác nhau:

- Cách thứ nhất, từ khóa bên trong text box được sử dụng để tìm kiếm tất cả các tài liệu trong cơ sở dữ liệu chứa các từ khóa đó.
- Thứ hai, từ khóa bên trong text box được sử dụng như là giá trị của một thuộc tính trong mệnh đề **WHERE** của truy vấn SQL tới cơ sở dữ liệu của website. Giá trị này thường thuộc một trong hai dạng sau:
 - Một tập hữu hạn được định nghĩa trước, ví dụ zip code, tên viết tắt của các thành phố...
 - Là một thể hiện của một kiểu dữ liệu liên tục, ví dụ như giá cả, hay một bộ ba ngày tháng năm.

Chính vì vậy, chúng ta cũng phân textbox ra làm hai loại tương ứng là Generic Textbox và Typed Textbox. Đối với loại thứ hai, một giá trị đầu vào sai thường dẫn đến lỗi, do vậy, điều quan trọng là chúng ta cần phải xác định chính xác kiểu dữ liệu để truyền vào. Ngoài ra, việc lựa chọn từ khóa sai trong Generic text box cũng có thể dẫn đến việc trả về một vài kết quả, như vậy, có một thách thức đặt ra, đó là chúng ta cần phải xác định được tập các từ khóa phù hợp để thu được các trang kết quả một cách phong phú nhất.

Như vậy vấn đề ở đây là cần xác định được 2 loại textbox và đưa ra những xử lý khác nhau. Đối với Generic Textbox thì giá trị nhập vào là tùy ý nên cần có một phương pháp nhầm sinh ra các keyword – từ khóa – để điền vào. Đối với Typed Textbox thì cần xác định rõ dạng dữ liệu nhầm điền được những giá trị thỏa mãn mới có thể thực hiện submit Form thành công.

Hướng giải quyết duy nhất cho đến hiện tại là xác định loại Typed Textbox thông qua nhãn của chúng (Mã tinh, Giá, Zipcode, Price...). Tuy nhiên việc xây dựng một giải thuật toàn diện tổng quát với mọi loại Typed Textbox là vô cùng khó khăn, bởi thế trong giới hạn của đề tài này những Typed Textbox tạm thời chưa được xử lý và chúng được để giá trị mặc định ban đầu khi thực hiện submit Form.

Để có thể làm việc hiệu quả với Generic Textbox, chúng ta cần phải có một tập các từ khóa tốt. Một cách cảm tính, chúng ta có thể thiết kế một tập các từ khóa của nhiều lĩnh vực khác nhau để nhập vào text box. Tuy nhiên, số lĩnh vực là rất nhiều, và số từ khóa đặc thù cho lĩnh vực đó cũng không ít. Hơn thế nữa, đối với Generic Textbox, ngay cả khi chúng ta xác định ra các đầu vào trên hai form khác nhau, và các đầu vào này liên quan đến cùng một quan niệm trên cùng một lĩnh vực, chúng ta cũng không nhất thiết phải nhập vào từ khóa trong hai trường hợp là giống hệt nhau để có thể thu được kết quả là như nhau. Vì vậy, cách tiếp cận này là không hợp lý.

Giải thuật được đề xuất ở đây là áp dụng phương pháp thăm dò lặp (Iterative probing) để tìm ra tập các từ khóa cho text box. Cần có khái niệm rõ ràng về thể nào là từ khóa của một Textbox: đó phải là từ khi điền vào Textbox sẽ trả về ít nhất một kết quả.

Bảng 1 - Kết quả ứng dụng Test Infomative (ngưỡng $\tau = 0.3$)

Tên trang web	Số lượng Template	Số Infomative Template	Số url sinh ra	Tổng số bản ghi có trên trang	Số bản ghi lấy được
ungvien.com.vn	16	1	515	2904	2904
dangcv.com	33	2	1580	4389	4389

Bảng 2 - Kết quả ứng dụng sinh từ khóa

Tên trang web	Số từ khóa	Số vòng lặp	Số url sinh ra	Tổng số bản ghi có trên trang	Số bản ghi lấy được
tuyendung.org.vn	1500	2	1500	~15000	~13000
vinahr.com	1500	2	1500	~5000	~4700

Nếu chưa xác định được điều này chúng ta sẽ gọi là những từ ứng viên (candidate word # keyword).

Đầu tiên, chúng ta sẽ cần một tập các từ ứng viên hạt giống là giá trị đầu vào cho text box. Với mục đích tối đa khả năng tìm ra các tài liệu đúng chủ đề nên những từ hạt giống này sẽ được trích từ một loạt các trang web đúng chủ đề. Những từ hạt giống sẽ được lựa chọn từ các trang web bằng chỉ số TF-IDF. Những từ có trọng số TF-IDF cao sẽ được chọn làm ứng viên.

*Tần suất từ (term frequency – TF)

Trọng số của một từ là tần suất xuất hiện của từ đó trong tài liệu. Cách định trọng số này nói rằng một từ là quan trọng cho một tài liệu nếu nó xuất hiện nhiều lần trong tài liệu đó.

$$TF(w, p) = \frac{n_{w, p}}{N_p} \quad (1)$$

Trong đó: $n_{w, p}$ là số lần từ đó xuất hiện trong trang web.

N_p là số từ trong trang web.

*TFIDF (term frequency – inverse document frequency)

Trọng số của một từ là tích của tần suất từ TF và tần suất tài liệu nghịch đảo IDF của từ đó.

Tần suất tài liệu nghịch đảo được xác định bằng công thức:

$$IDF(w) = \log \frac{D}{d_w} \quad (2)$$

Trong đó: D là kích thước của tập tài liệu

d_w là số tài liệu mà từ xuất hiện trong đó.

Trọng số $TFIDF$ là sự kết hợp của TF và IDF :

$$TFIDF(w, p) = TF(w, p) * IDF(w) \quad (3)$$

Rõ ràng, khi một từ xuất hiện trong càng ít tài liệu thì khả năng sử dụng từ đó để phân biệt càng cao.

Từ đó, chúng tiến hành điền lần lượt các từ hạt giống vào Form và submit để lấy kết quả và lại trích ra được các từ ứng viên mới từ kết quả thu về, đồng thời sẽ cập nhật chỉ số TF-IDF của tập hợp các từ khóa.

Từ ứng viên nào khi điền vào có kết quả trả về sẽ trở thành từ khóa. Từ khóa mới lại được sử dụng để cập nhật ứng viên cho text box.

Chúng ta cứ lặp lại quá trình cho đến khi:

- Không thể tách thêm từ khóa được nữa
- Đạt tới một điều kiện dừng nào đó, chẳng hạn như số lượng từ khóa đã đạt 1000 từ.

Như vậy, khi thuật toán kết thúc, chúng ta đã thu được một tập các từ khóa làm đầu vào cho text box.

6. THỰC NGHIỆM

Hệ thống tiến hành chạy thử nghiệm trên 4 trang web tuyển dụng Tiếng Việt: ungvien.com.vn, dangcv.com, tuyendung.org.vn và vinahr.com.

Nhận xét ban đầu cho thấy những website này đều cung cấp 2 lựa chọn tìm kiếm là tìm kiếm nhanh và tìm kiếm nâng cao. Với Form tìm kiếm nhanh thì số lượng input là rất ít (thông thường nhỏ hơn 3 input) đồng thời luôn chứa một text box để nhập từ khóa tìm kiếm. Vì thế việc thực hiện lựa chọn Query Template trên những Form tìm kiếm nhanh là không cần thiết. Đối với Form tìm kiếm nâng cao thì số lượng input là khá nhiều, nhưng chủ yếu là select menu. Với những nhận định trên thì phương pháp kiểm thử sẽ như sau:

- Với trang ungvien.com.vn và dangcv.com chỉ thực hiện lựa chọn Query Template trên Form tìm kiếm nâng cao.
- Với trang tuyendung.org.vn và vinahr.com chỉ thực hiện sinh từ khóa trên Form tìm kiếm nhanh.

Kết quả thực nghiệm trên 4 trang web tuyển dụng cho thấy tính khả quan của phương pháp Surfacing trên những cơ sở dữ liệu tương đối nhỏ. Với những Form tìm kiếm nâng cao, kết quả ở bảng 1 cho thấy có thể lấy được gần như toàn bộ số bản ghi trong cơ sở dữ liệu. Số lượng Template cần phải test là 16 và 33 trên $2^6 - 1 = 63$ Template có thể sinh ra. Trong đó chỉ có 1 đến 2 Template là Infomative. Với những Form tìm kiếm nhanh, kết

quả ở bảng 2 cho thấy với số lần lặp là 2 chúng ta đã có thể lấy được 1500 từ khóa. Tỉ lệ số bản ghi lấy được / số bản ghi ước lượng của website là từ 80 – 90%.

Hệ thống hiện tại chưa được tích hợp với máy tìm kiếm theo chủ đề nên chưa có kết quả thực nghiệm với việc tìm kiếm theo chủ đề. Tuy nhiên với khả năng tiếp cận được số lượng lớn tài liệu ẩn thì việc tìm ra các tài liệu đúng chủ đề là rất khả quan.

7. KẾT LUẬN

Deep Web là một lĩnh vực còn vô cùng mới mẻ và đầy thử thách nhưng cũng ẩn chứa rất nhiều tiềm năng, mở ra một hướng đi hoàn toàn mới cho lĩnh vực tìm kiếm ở Việt Nam. Có thể nói những thông tin mà Deep Web mang lại là không hề nhỏ và ngày càng giúp thỏa mãn nhu cầu tìm kiếm thông tin số hiện nay.

Vì là một hướng đi mới nên những cải tiến trong tương lai là vô cùng cần thiết. Đó là những cải tiến để tăng tính khả dụng của hệ thống như: tiến hành xử lý cả những Form có sử dụng javascript, xây dựng được phương pháp đánh chỉ mục cho các URL sinh ra từ Form có dạng POST ... Ngoài ra việc tích hợp hoàn chỉnh với hệ thống Máy tìm kiếm theo chủ đề tạo ra một máy tìm kiếm không chỉ tìm kiếm không những hiệu quả mà còn đem lại nhiều kết quả chuẩn xác hơn cho người sử dụng cũng là một vấn đề cần quan tâm thực hiện trong tương lai.

8. TÀI LIỆU THAM KHẢO

- [1] Jayant Madhavan, David Ko, Lucja Kot, Vignesh Ganapathy, Alex Rasmussen, Alon Halevy. Google's Deep-Web Crawl. PVLDB '08, August 23-28, 2008, Auckland, New Zealand Copyright 2008 VLDB Endowment, ACM 978-1-60558-306-8/08/08.
- [2] B.He, M.Patel, Z.Zhang and K.C.-C.Chang. Accessing the Deep Web: A survey. Communications of the ACM, 50(5):95–101, 2007.
- [3] L.Barbosa and J.Freire. Siphoning hidden-web data through keyword-based interfaces. In SBBD, 2004.
- [4] Ntoulas, P.Zerfos, and J.Cho. Downloading Textual Hidden Web Content through Keyword Queries. In JCDL, pages 100–109, 2005.
- [5] Rajaraman, Y.Sagiv and J.D.Ullman. Answering Queries Using Templates with Binding Patterns. In PODS, 1995.
- [6] Duc-Khanh Tran, Dinh-Thi Vu, Ngoc-Duc Nguyen, Dai-Duong Le. Efficiently Crawl Topical Vietnamese Web Pages using Machine Learning Techniques, IEEE-RIVF 2010.
- [7] HtmlUnit API Document.
<http://htmlunit.sourceforge.net/apidocs/index.html>

Kỹ thuật định vị dựa trên wifi và ứng dụng

Chu Bảo Trung, Phạm Hữu Hoàng

Tóm tắt - Cung cấp thông tin hữu ích và phù hợp với người dùng tuỳ thuộc mỗi quan tâm, vị trí, bối cảnh của người dùng trở thành vấn đề quan trọng trong các dịch vụ cung cấp thông tin [1]. Thông tin vị trí là một trong những chìa khoá để làm điều đó. Chúng tôi lựa chọn cách tiếp cận khai thác cường độ sóng wifi để định vị người dùng. So với các phương pháp định vị khác như dựa trên mạng viễn thông di động, bluetooth, RFID, GPS,... định vị dựa vào wifi có thể tận dụng hạ tầng mạng không dây phổ biến mà không cần đầu tư thêm thiết bị chuyên dụng. Kỹ thuật này phù hợp với các bài toán định vị trong nhà đòi hỏi độ chính xác tương đối cao.

Chúng tôi đã nghiên cứu và cài đặt hai phương pháp định vị dựa vào mạng không dây, đó là: (i) phương pháp “*Đặc tam giác*” – Trilateration [6] và (ii) phương pháp “*Đối sảnh mẫu*”- Pattern matching [7]. Áp dụng kỹ thuật định vị này, chúng tôi xây dựng ứng dụng “*Hướng dẫn tham quan*” tại Viện bảo tàng Lịch sử Việt Nam(số 1 Phạm Ngũ Lão, Hà Nội). Các kết quả thử nghiệm và đánh giá, phân tích sẽ được trình bày cụ thể trong báo cáo.

Từ khóa - location based services, wifi-based positioning, trilateration, pattern matching.

1. GIỚI THIỆU

Trong thời đại bùng nổ thông tin như ngày nay, yêu tố quyết định của nhiều vấn đề thường chỉ nằm ở chỗ có hay không có thông tin. Thông tin vị trí là một trong những tài nguyên quý giá đó. Các “*dịch vụ dựa vị trí*”-location based services đang ngày càng trở nên phổ biến và đang đi dàn vào cuộc sống hàng ngày của chúng ta. Các hoạt động như: tìm kiếm địa điểm vui chơi, giải trí gần nhất, truy tìm đồ vật bị đánh cắp, tìm kiếm cứu nạn,...đang là đối tượng hướng tới của lĩnh vực này.

Để có được thông tin về vị trí, trong khoảng 50 năm trở lại đây, đã có rất nhiều kỹ thuật được nghiên cứu và đưa vào triển khai, đó là: hệ thống định vị toàn cầu GPS của Hoa Kỳ, hệ thống GLONASS của Nga hay hệ thống GNSS của châu Âu. Ban đầu các hệ thống này đều được nghiên cứu cho mục đích quân sự, tuy nhiên hiện nay nó đang dần được dân sự hoá. Các hệ thống vừa nêu trên là những hệ thống định vị có phạm vi toàn cầu, kỹ thuật được sử dụng ở đây là dựa vào mạng lưới vệ tinh địa tĩnh (có vị

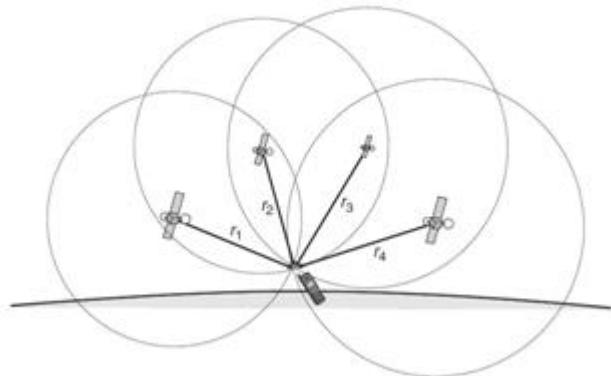
Công trình này được thực hiện dưới sự hướng dẫn của TS. Vũ Tuyết Trinh, cùng với sự giúp đỡ của cán bộ, nhân viên viện bảo tàng Lịch Sử Việt Nam.

Chu Bảo Trung, sinh viên lớp IS1-chương trình Việt Nhật, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (diện thoại: 01234 47 1988, e-mail: trungchubao@gmail.com).

Phạm Hữu Hoàng, sinh viên lớp IS1-chương trình Việt Nhật, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (diện thoại: 0904 628 865, e-mail: phamhuuhoang@gmail.com).

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

trí tương đối so với trái đất là cố định). Dựa vào chênh lệch giữa thời điểm truyền sóng từ vệ tinh và thời điểm nhận lại sóng từ thiết bị di động, từ đó tính ra khoảng cách từ vệ tinh tới thiết bị di động, và khi có được thông tin như vậy từ 4 vệ tinh đã biết vị trí ta có thể định vị được thiết bị di động như dưới hình 1 [2,4]. Đó là nguyên tắc chung cho các hệ thống trên, mặc dù chi tiết kỹ thuật có thể khác.



Hình 1: Định vị bằng GPS

Tuy nhiên những hệ thống này chỉ hoạt động tốt trong môi trường ngoài trời, nơi không có hoặc có ít chướng ngại vật. Vì vậy đã xuất hiện những kỹ thuật riêng áp dụng cho việc định vị trong nhà, những khu vực nhỏ hơn, nhiều chướng ngại vật hơn. Trong các kỹ thuật định vị trong nhà thì kỹ thuật dựa vào sóng wifi là một kỹ thuật có độ chính xác tương đối cao, chi phí đầu tư thấp.

Trong nghiên cứu này, chúng tôi tập trung vào kỹ thuật định vị dựa trên cường độ sóng wifi và lựa chọn hai phương pháp: “*Đặc tam giác*”[5,6] và “*Đối sảnh mẫu*”[7] để nghiên cứu và triển khai ứng dụng.

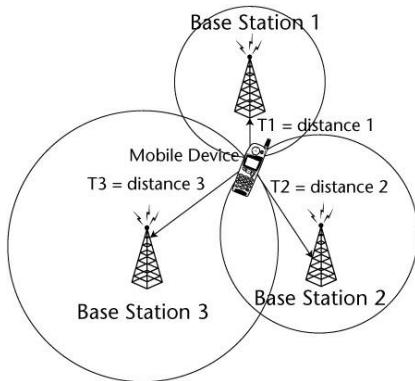
Phản tiếp theo của báo cáo gồm các nội dung sau:

- Mục 2: Phương pháp đặc tam giác
- Mục 3: Phương pháp đối sảnh mẫu
- Mục 4: Hệ thống tham quan bảo tàng
- Mục 5: Các công việc liên quan
- Mục 6: Kết luận và hướng phát triển
- Mục 7: Lời tri ân

2. PHƯƠNG PHÁP ĐẶC TAM GIÁC[5,6]

2.1. Phương pháp

Nguyên tắc của phương pháp “*đặc tam giác*” là dựa trên khoảng cách từ thiết bị di động tới những điểm đặt access point(AP) – điểm truy nhập mạng không dây có vị trí xác định.



Hình 2: Phương pháp đặc tam giác

Như mô tả trên hình 2, phương pháp này thực hiện giống như hệ thống GPS.

Để tính được khoảng cách từ thiết bị di động tới các AP, ta dựa vào cường độ sóng wifi mà thiết bị di động thu được. Mỗi liên hệ giữa cường độ này với khoảng cách từ thiết bị tới AP được thể hiện qua một hàm số. Bằng đo đặc thông kê ta có thể suy ra được hàm số đó.

2.2. Thực nghiệm

Trong các thiết bị di động có khả năng bắt wifi, card mạng có đo một giá trị thể hiện độ mạnh yếu của sóng không dây thu được. Đó là RSSI (Received Signal Strength Indication). Trong chuẩn mạng không dây IEEE 802.11, RSSI là một giá trị tương đối, không có đơn vị và nằm trong khoảng từ 0 – 255 (ở một số thiết bị thì khoảng này âm). Chuẩn IEEE 802.11 không định nghĩa bắt cứ mối quan hệ tương quan nào giữa RSSI và các đơn vị đo lường cường độ khác như mW hay dBm, mà sự tương quan này do nhà sản xuất phần cứng định nghĩa. Do đó với phần cứng khác nhau thì RSSI đo được tại cùng 1 vị trí có thể khác nhau.

Pha 1: Nội suy hàm liên hệ

Trong pha này, chúng tôi đã tiến hành đo đặc để suy ra được hàm liên hệ giữa cường độ sóng và khoảng cách từ thiết bị di động tới AP.

- Thiết bị:
 - + 3 AP cùng loại D-LINK 108G
 - + iPhone 4, bộ nhớ 32GB

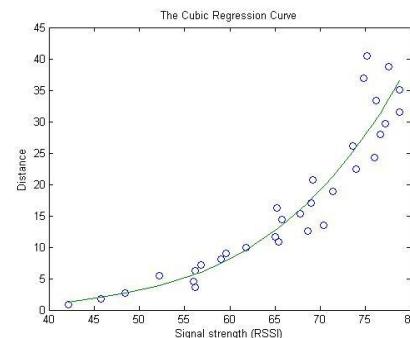
- Triển khai đo đặc:

- + AP được đặt tại một vị trí cố định
- + Bắt đầu từ 0.9m, cứ 0.9m lại dùng điện thoại iPhone đo cường độ sóng. Tại mỗi điểm như vậy đo 5 lần. Bảng 1 là một phần dữ liệu thu được. Cường độ sóng biểu diễn dưới dạng các số âm.
- + Đo tới khi điểm đo cách AP ~ 30m thì dừng.

Khoảng cách(m)	L1	L2	L3	L4	L5	Trung bình
0.9	-39	-45	-41	-45	-41	-42.2
1.8	-43	-45	-49	-46	-46	-45.8
2.7	-49	-46	-47	-46	-54	-48.4
3.6	-57	-55	-56	-55	-58	-56.2

Bảng 1: Dữ liệu đo cường độ sóng wifi

- Nội suy hàm:



Hình 3: Nội suy hàm số bằng matlab

+ Nhập dữ liệu đo được vào matlab (hình 3), chúng tôi thu được hàm liên hệ như dưới đây.

$$D = 0.00057670 \times RSSI^3 - 0.07379443 \times RSSI^2 + 3.36703189 \times RSSI - 52.81736996 \quad (1)$$

Pha 2: Đạc tam giác

- Phương pháp:

- + Chia khu vực định vị thành lưới kích thước 1m x 1m.
- + Xét mỗi ô lưới, khoảng cách từ tâm ô tới 3 AP là d_1, d_2, d_3 .
- + Khoảng cách thu được sau khi áp dụng (1) là r_1, r_2, r_3 .
- + Tìm 3 ô lưới có giá trị của biểu thức sau là nhỏ nhất:

$$\Delta = (r_1 - d_1)^2 + (r_2 - d_2)^2 + (r_3 - d_3)^2 \quad (2)$$

- + Trung bình theo trọng số của 3 ô đó là toạ độ định vị được (trọng số là nghịch đảo của Δ)

Kết quả:

Toạ độ thật	Toạ độ định vị	Δ_x	Δ_y
(1;1)	(3.5; 0.5)	2.5	0.5
(6;1)	(6.5; 0.5)	0.5	0.5
(13.5;1)	(20.5; 1.49)	7	0.49
(11.4; 6.6)	(6.5; 0.5)	4.9	6.1
(6; 12.5)	(14.5; 12.5)	8.5	0
(1;11.3)	(0.5; 11.5)	0.5	0.2
(6;7)	(3.17; 2.16)	2.83	4.84
(18.6; 6.6)	(20.5; 11.5)	1.9	4.9
(18.6; 5.4)	(20.5; 1.4)	1.9	4
(18.6; 1)	(20.5; 1.4)	1.9	0.4

Đơn vị: mét

Bảng 2: Kết quả định vị bằng đặc tam giác

Từ kết quả trên suy ra:

$$\bar{\Delta}_x = 3.243(m); \bar{\Delta}_y = 2.193(m);$$

$$\bar{\Delta} = \sqrt{\bar{\Delta}_x^2 + \bar{\Delta}_y^2} = 3.915(m)$$

2.3. Nhận xét

Phương pháp “đặc tam giác” có sai số trung bình là 3.91 m, tuy nhiên có những trường hợp sai số này là xấp xỉ 10m. Đối với định vị trong nhà thì đây là một sai số khá lớn.

Tuy nhiên ưu điểm của phương pháp này là:

- Cách thực hiện khá đơn giản, chi phí nhỏ.
- Chỉ cần đo đặc một lần duy nhất để xác định hàm liên hệ.

3. PHƯƠNG PHÁP ĐỐI SÁNH MẪU[7]

3.1. Phương pháp

Phương pháp đối sánh mẫu dựa vào nhận xét sau: “*Hai vị trí khác nhau sẽ có cường độ sóng tới các AP là khác nhau, ít nhất là tại một AP nào đó, khi số lượng AP >= 3*”. Bởi vì cường độ sóng thu được đặc trưng cho khoảng cách từ AP tới thiết bị di động, hay nói cách khác, 2 điểm có cùng cường độ sóng tới một AP thì 2 điểm đó cách đều AP đó, hay nó cùng nằm trên đường tròn tâm là AP đó. Vậy mà, nếu $>= 3$ đường tròn nhiều nhất chỉ có thể giao nhau tại 1 điểm, vậy nên, không thể có 2 điểm ở vị trí khác nhau mà lại cùng nằm trên $>= 3$ đường tròn khác nhau.

Áp dụng nhận xét trên, phương pháp đối sánh mẫu sẽ được thực hiện bằng cách đo “mẫu” tại các điểm mốc định sẵn, khi định vị, lấy cường độ sóng thu được so sánh với các “mẫu”, điểm mốc nào có “mẫu” khớp thì đó chính là vị trí của thiết bị di động

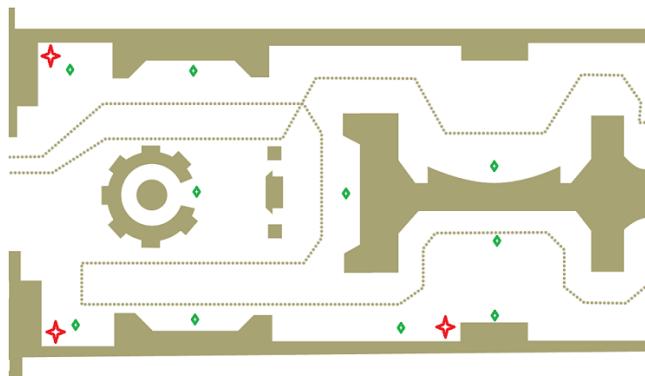
3.2. Thực nghiệm

Về mặt thiết bị thì phương pháp này cũng dùng các thiết bị của phương pháp đặc tam giác.

Khi thực hiện phương pháp “đối sánh mẫu” trải qua 2 pha.

Pha 1: Đào tạo (training)

Hình 4 là bản đồ của khu vực thực nghiệm tại viện bảo tàng Lịch Sử Việt Nam. Những vị trí đánh dấu ngôi sao màu đỏ là vị trí của các AP, các điểm đánh dấu hình thoi màu xanh là vị trí của các hiện vật được chọn thử nghiệm, và cũng là những điểm thực hiện lấy mẫu



Hình 4: Bản đồ lấy “mẫu”

- Đo và xác định toạ độ các điểm lấy mẫu
- Tại mỗi điểm lấy mẫu, đo cường độ sóng theo 4 hướng (trái \Leftrightarrow

0 độ, dưới \Leftrightarrow 90 độ, phải \Leftrightarrow 180 độ, trên \Leftrightarrow 270 độ), mỗi hướng đo 3 lần sau đó lấy trung bình

- Nhập toạ độ của các điểm lấy mẫu và mẫu tại đó vào cơ sở dữ liệu.
- Bảng 3 là ví dụ về mẫu đo được (mẫu gồm 3 cường độ sóng đo được từ 3 AP)

Toạ độ(m)	Hướng	AP1	AP2	AP3
(0;0)	0°	45	75	65
	90°	45	74.3	74
	180°	45	76.3	66.3
	270°	45.6	74.6	62
(3;0)	0°	59	75.6	70.6
	90°	55.6	77.6	73
	180°	56.6	74.6	64.6
	270°	57	70	67.3

Bảng 3: “mẫu” cường độ sóng

Pha 2: Đối sánh mẫu (matching)

Khi có dữ liệu về mẫu trong cơ sở dữ liệu rồi ta tiến hành so sánh sóng thu được với các mẫu. Cụ thể các bước như dưới đây:

- Lấy cường độ sóng thiết bị di động nhận được từ các AP, lần lượt là ss_1, ss_2, ss_3 .
 - Với mỗi mẫu i , tính biểu thức dưới đây:
- $$\Delta_{ss}[i] = |SS_1[i] - ss_1| + |SS_2[i] - ss_2| + |SS_3[i] - ss_3| \quad (3)$$
- Tìm $\Delta_{ss}[i]$ có giá trị nhỏ nhất.
 - Nếu $\min(\Delta_{ss}[i]) < \varepsilon (\varepsilon = 10)$ thì vị trí thiết bị di động chính là vị trí của điểm mốc thứ i . Còn ngược lại thì không định vị được thiết bị di động.

Kết quả:

Toạ độ thật	Toạ độ định vị	Δ_x	Δ_y
(1;1)	(1; 1)	0	0
(6;1)	Không xác định	-	-
(13.5;1)	(13.5; 1)	0	0
(11.4; 6.6)	Không xác định	-	-
(6; 12.5)	Không xác định	-	-
(1;11.3)	(1; 11.3)	0	0
(6;7)	Không xác định	-	-
(18.6; 6.6)	(18.6; 6.6)	0	0
(18.6; 5.4)	(18.6; 5.4)	0	0
(18.6; 1)	(18.6;1)	0	0

Đơn vị: mét

Bảng 4: Kết quả định vị bằng đối sánh mẫu

Từ bảng 4 không thể suy ra $\bar{\Delta}_x, \bar{\Delta}_y$ và $\bar{\Delta}$ được.

3.3. Nhận xét

Phương pháp đối sánh mẫu, với những điểm gần với mẫu luôn cho kết quả chính xác, tuy nhiên với những vị trí khác thì không

đưa ra được kết quả định vị. Vì vậy sai số của phương pháp này không thể xác định được.

- **Ưu điểm:**

- + độ chính xác khi thiết bị di động ở gần điểm mốc là gần như tuyệt đối.
- + phương pháp khá đơn giản, chi phí thấp

- **Nhược điểm:**

- + không thể định vị được khi thiết bị di động ở xa mốc.
- + muốn tăng khả năng định vị, cần phải tăng số lượng mốc.
- + mỗi khi có thay đổi về bài trí phòng, xây dựng mới sẽ yêu cầu do đặc lại các mốc.

4. HỆ THỐNG HƯỚNG DẪN THAM QUAN BẢO TÀNG

4.1. Mục đích

Bảo tàng Lịch sử Việt Nam nói riêng và các bảo tàng khác nói chung, đều có rất nhiều thông tin về các hiện vật. Họ khó có thể trưng bày tất cả các thông tin này cùng hiện vật. Vì thế có một số bảo tàng có những hướng dẫn viên, giúp khách tham quan tiết kiệm thời gian mà vẫn có thông tin đầy đủ. Tuy nhiên cũng có những bảo tàng không có hướng dẫn viên như bảo tàng Lịch sử Việt Nam. Với mục đích thay thế những người hướng dẫn viên, giúp khách tham quan có những thông tin đầy đủ và phù hợp với mong muốn của họ, đồng thời cũng gợi ý cho họ lộ trình tham quan bảo tàng, chúng tôi đã nghĩ đến việc áp dụng kỹ thuật định vị đã tìm hiểu ở trên để xây dựng hệ thống hướng dẫn tham quan bảo tàng.

4.2. Kỹ thuật định vị

Hai kỹ thuật định vị dựa vào sóng wifi có những ưu và nhược điểm khác nhau, mà chúng có thể hỗ trợ cho nhau trong hệ thống hướng dẫn tham quan bảo tàng của chúng tôi.

- Kỹ thuật “đặc tam giác”: có sai số khá lớn (từ 3 đến 10m), vì vậy nó khó có thể dùng để phân biệt giữa các hiện vật trong quá trình khách đi tham quan bảo tàng. Tuy nhiên nhược điểm này có thể được khắc phục bởi phương pháp “đối sảnh mốc”, phương pháp này đặc biệt hữu ích trong việc xác định hiện vật mà khách đang tham quan.
- Kỹ thuật “đối sảnh mốc”: có sai số rất nhỏ nếu thiết bị di động ở gần các điểm mốc (các hiện vật được chọn làm điểm mốc). Tuy nhiên khi ở xa điểm mốc một chút, thì phương pháp này không đưa ra được kết quả định vị. Tuy nhiên nhược điểm này có thể được giảm bớt bằng phương pháp “đặc tam giác”.

Như vậy, quá trình định vị ở đây sẽ như sau:

+ Khi thiết bị di động ở gần điểm mốc, “đối sảnh mốc” định vị được, lúc đó ta dùng kết quả của phương pháp này làm vị trí định vị được.

+ Khi thiết bị di động không ở gần điểm mốc nào, “đối sảnh mốc” không đưa ra được vị trí của thiết bị di động. Thì lúc này ta dùng phương pháp “đặc tam giác” để định vị. Kết quả mặc dù có sai số khá lớn nhưng để dẫn đường khách tham quan thì vẫn có thể được.

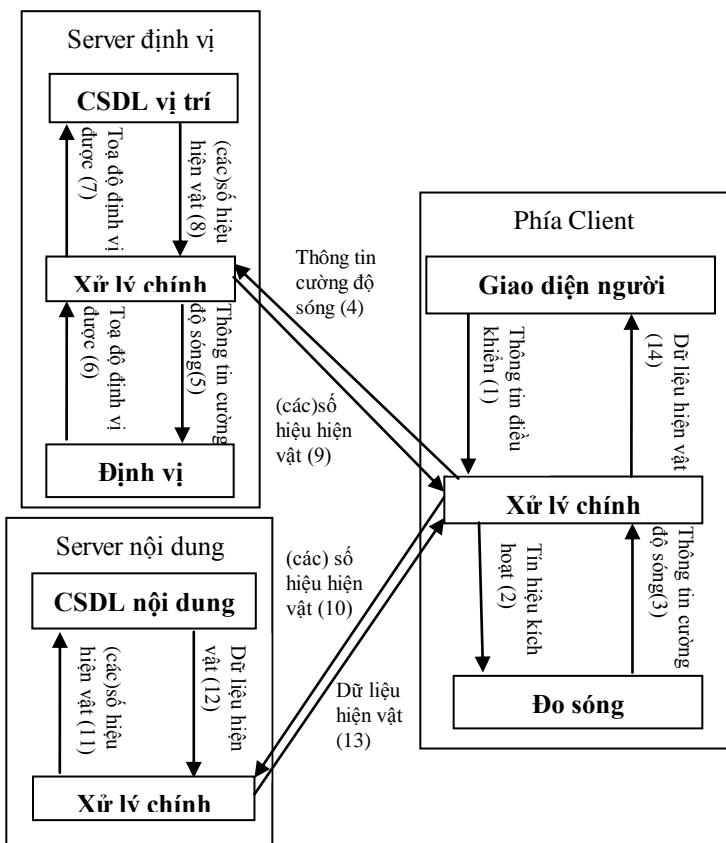
4.3. Kiến trúc hệ thống

Sử dụng hai kỹ thuật “đặc tam giác” và “đối sảnh mốc” hỗ trợ cho nhau, chúng tôi xây dựng hệ thống có kiến trúc được mô tả như trong hình 5.

Hệ thống sẽ gồm có hai server và một client.

Hai server được viết bằng ngôn ngữ PHP, chạy trên hệ điều hành MAC OS Snow Leopard 10.6.4, client là iPhone4.

Luồng xử lý của hệ thống được thực hiện theo số thứ tự đã đánh như trong hình 5.



Hình 5: Kiến trúc hệ thống hướng dẫn tham quan bảo tàng

5. CÁC CÔNG VIỆC LIÊN QUAN

Trong nghiên cứu này đã sử dụng những kết quả của các nghiên cứu thuộc phòng nghiên cứu “định vị và dẫn đường vệ tinh”-SNAP thuộc trường đại học New South Wales, Úc. Hiện nay họ đã tiến hành cải tiến hai kỹ thuật này lên khá nhiều nhằm khắc phục nhược điểm của mỗi phương pháp.

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết luận:

Hai kỹ thuật “đặc tam giác” và “đối sảnh mốc” đều có những ưu điểm và nhược điểm riêng, với mục tiêu xây dựng hệ

thống hướng dẫn tham quan bảo tàng thì hai phương pháp này đã hỗ trợ tốt cho nhau.

Sự thành công trong việc triển khai hệ thống này chính là bước khởi đầu cho việc áp dụng những kỹ thuật định vị bằng sóng wifi vào các ứng dụng thực tế. Với lợi thế về chi phí đầu tư, phương pháp này sẽ là lựa chọn hàng đầu cho các “*dịch vụ dựa vị trí*” có phạm vi hoạt động nhỏ.

Hướng phát triển:

Hướng phát triển đầu tiên mà ta có thể nghĩ đến chính là cải tiến khắc phục nhược điểm của hai kỹ thuật trên:

- Phương pháp “*đặc tam giác*”: bằng cách tính toán sự ảnh hưởng của vật cản lên cường độ sóng, ta có thể tăng độ chính xác lên.
- Phương pháp “*đối sánh mẫu*”: nhược điểm của phương pháp này là việc phải đo đặc nhiều, ta có thể tìm cách giảm số điểm phải đo đặc, thực tế thì tại phòng nghiên cứu của SNAP họ đã nghiên cứu phương pháp nội suy cường độ sóng tại những điểm chưa đo từ những điểm đã đo, đây là một hướng để cải tiến phương pháp này.

Ngoài việc cải tiến kỹ thuật định vị, ta còn có thể cải tiến theo cách kết hợp những thông tin khác, ngoài thông tin vị trí, như là: hồ sơ của người sử dụng, phần cứng thiết bị di động, bằng cách này ta có thể cung cấp các dịch vụ thông tin linh hoạt hơn, đưa ra được những thông tin thích hợp hơn với người sử dụng.

7. LỜI TRI ÂN

Trong suốt quá trình thực hiện nghiên cứu tốt nghiệp chúng em đã luôn nhận được sự hướng dẫn tận tình của TS. Vũ Tuyết Trinh, vì vậy chúng em muốn dành lời cảm ơn chân thành sâu sắc nhất tới cô.

Tiếp theo chúng em muốn gửi lời cảm ơn chân thành tới những cán bộ, nhân viên viện bảo tàng Lịch Sử Việt Nam, những người đã tạo điều kiện thuận lợi và giúp đỡ chúng em trong quá trình thử nghiệm, xây dựng hệ thống này.

8. TÀI LIỆU THAM KHẢO

- [1] “*Context-aware mobile and ubiquitous computing for enhanced usability: adaptive technologies and applications*”, Dragan Stojanovic, University of Nis, Serbia, Information Science Reference, 2009.
- [2] “*Location based services: fundamentals and operation*”, Axel Küpper, Ludwig Maximilian University Munich, Germany, John Wiley & Sons Ltd, 2005.
- [3] “*Mobile Location Services: The Definitive Guide*”, Andrew Jagoe, Prentice Hall PTR, 2002.
- [4] “*Location-based services*”, Jochen Schiller & Agne's Voisard, Morgan Kaufman Publishers, 2004.
- [5] “*An indoors wireless positioning system based on wireless local area network infrastructure*”, Y.Wang, X. Jia, H.K.Lee, G.Y.Li, presented at SatNav 2003.
- [6] “*Two new algorithms for indoor wireless positioning system (WPS)*”, Y. Wang, X.Jia, Chris Rizos, School of Surveying and Spatial Information Systems, University of New South Wales, Sydney, Australia, 2004.
- [7] “*A new method for yielding a database of location fingerprints in WLAN*”, Binghao Li, Y.Wang, H.K.Lee, Andrew Dempster, Chris Rizos, SNAP, University of New South Wales, 2005.

Nghiên cứu, đánh giá và cải tiến hiệu quả sử dụng năng lượng và hiệu suất truyền gói tin của các giao thức định tuyến trong mạng cảm biến không dây

Nguyễn Sơn Thủy, Nguyễn Đình Minh

Tóm tắt — Ngày nay, các ứng dụng của mạng cảm biến không dây (WSNs) được sử dụng rộng rãi trong rất nhiều lĩnh vực như công nghiệp, quân sự, môi trường, y tế, nhà thông minh hay trong giao thông vận tải v.v...[1]. Tuy nhiên trong WSNs, các nút mạng bị giới hạn về khả năng truyền phát tín hiệu, khả năng tính toán, bộ nhớ hạn chế cũng như nguồn năng lượng cung cấp cho các nút mạng là có hạn[2]. Và các giao thức định tuyến trong WSNs có một vai trò rất quan trọng vì nó ảnh hưởng tới chất lượng của mạng và năng lượng của các nút mạng. Trong bài báo này chúng tôi đã nghiên cứu và phân tích các thuật toán định tuyến cho mạng cảm biến không dây. Sau đó sẽ tiến hành cài đặt và mô phỏng một số thuật toán định tuyến như : Flooding, DSR, RIP, GF... để đánh giá hiệu quả của các thuật toán trong với các tiêu chí như tính tin cậy, độ tiêu thụ năng lượng và hiệu suất truyền gói tin. Sau đó chúng tôi sẽ đề xuất một số cải tiến cho các giao thức trên. Phần mềm mô phỏng WSNET và WSNETStudio sẽ được sử dụng để tiến hành cài đặt, mô phỏng các thuật toán

Từ khóa — Routing protocol, Wireless Sensor Network (WSN) – chắc chắn tiếng Việt

1. GIỚI THIỆU

Mạng cảm biến không dây được cấu tạo bởi rất nhiều các thiết bị có kích thước nhỏ, chúng được gọi là các nút mạng. Mỗi nút mạng thường có bộ phận cảm biến, bộ xử lý, bộ truyền tín hiệu, bộ định vị và nguồn cung cấp năng lượng thường là pin [3][4]. Ngày nay do khả năng triển khai dễ dàng và giá thành của thiết bị ngày càng rẻ nên hiện nay các ứng dụng của mạng cảm biến không dây được sử dụng ngày càng rộng rãi trong rất nhiều các lĩnh vực như quân sự, công nghiệp, y tế và trong cuộc sống sinh hoạt hàng ngày. Ví dụ như : Hệ thống đo lường và điều khiển số trong các nhà máy, hệ thống quan trắc môi trường, hệ thống nhà thông minh,...Những hệ thống này đem lại hiệu quả hết sức to lớn, giúp tăng năng suất lao động và cải thiện chất lượng cuộc sống của con người. Vì vậy mạng cảm biến không dây được đánh giá là một trong những công nghệ giúp thay đổi cuộc sống con người trong tương lai sắp tới[5]. Tuy nhiên, trong các mạng cảm

Công trình này được thực hiện dưới sự hướng dẫn của TS. Nguyễn Kim Khánh, cùng với sự giúp đỡ của thạc sĩ Phạm Văn Thuận, viện CNTT, Đại học Bách Khoa Hà Nội

Nguyễn Sơn Thủy, sinh viên lớp IS1-chương trình Việt Nhật, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0983 212 885, e-mail: nguyensonthuy@gmail.com).

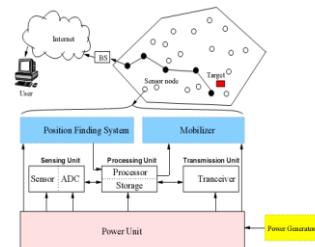
Nguyễn Đình Minh, sinh viên lớp AS1-chương trình Việt Nhật, khóa 51, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (điện thoại: 0975 111 688, e-mail: akaryuvn@gmail.com).

**© Viện Công nghệ thông tin và Truyền thông, trường
Đại học Bách Khoa Hà Nội.**

bien không dây hiện nay vẫn còn tồn tại những vấn đề cần phải khắc phục :

- Đảm bảo độ tin cậy trong quá trình truyền thông tin vì ở đây, phương tiện truyền tin là sóng radio rất dễ bị ảnh hưởng bởi môi trường xung quanh.
- Mô hình mạng có thể thay đổi khá nhiều theo thời gian do sự thay đổi mới, thay thế hoặc hỏng hóc của các nút mạng.
- Mô hình mạng phức tạp nên cần có phương pháp định tuyến (routing protocol) phù hợp để các gói tin đến được đích với quãng đường và thời gian cho phép.
- Các nút mạng thường là các máy tính nhúng có giới hạn về tài nguyên cũng như giới hạn về khả năng tính toán, thường sử dụng nguồn pin nên cần phải thiết kế phương pháp định tuyến đơn giản nhưng hiệu quả và đảm bảo đạt được mục tiêu tiết kiệm năng lượng...[6][7][8]

Vì vậy việc phát triển, lựa chọn những giao thức định tuyến phù hợp cho mạng cảm biến không dây đóng một vai trò rất quan trọng, nó quyết định chất lượng của mạng cũng như thời gian hoạt động của các nút mạng. Hiện nay các nghiên cứu về giao thức định tuyến trong mạng cảm biến không dây đang rất được quan tâm, cả trên thế giới cũng như tại Việt Nam. Trong nội dung bài báo sẽ trình bày nghiên cứu, phân tích các giao thức định tuyến trong mạng cảm biến không dây. Phần tiếp theo sẽ tiến hành cài đặt mô phỏng một số giao thức định tuyến như Flooding, RIP, GF, Gossiping ... để đánh giá khả năng tự cấu hình khi đồ thị mạng thay đổi, độ tin cậy, năng lượng tiêu thụ và hiệu suất truyền gói tin trong mạng. Sau khi có kết quả mô phỏng sẽ là một số cải tiến cho các giao thức trên. Cuối cùng là phần kết luận.



Hình 1: Cấu tạo mạng cảm biến không dây

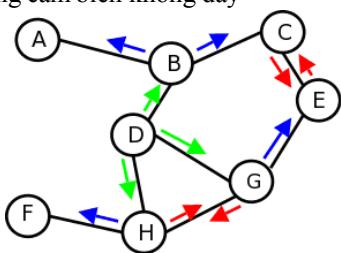
2. CÁC GIAO THỨC CHO MẠNG CẢM BIẾN KHÔNG DÂY

Thông thường trong mạng cảm biến không dây, các giao thức định tuyến thường được lựa chọn sao cho phù hợp với ứng dụng của mạng. Hiện nay có rất nhiều các giao thức định tuyến đã và đang được phát triển cho mạng cảm biến không dây. Các giao thức định tuyến có thể được phân loại như sau [9]:

1. **Các giao thức định tuyến dựa trên vị trí (Location - based Protocols)** với các giao thức như: MECN, SMECN, GAF, GEAR, Span, TBF, BVGF, GeRaF, GF, GEOStatic
2. **Các giao thức định tuyến theo nguyên tắc dữ liệu tập trung (Data-centric Protocols)**: SPIN, COUGAR , ACQUIRE, EAD
3. **Các giao thức định tuyến phân cấp (Hierarchical protocols) : LEACH, PEGASIS, HEED, TEEN, APTEEN**
4. **Các giao thức định tuyến hướng di động (Mobility - based Protocols) : SEAD, TTDD, Data MULES, Dynamic Proxy Tree-Base Data Dissemination**
5. **Các giao thức định tuyến đa hướng (Multipath-based Protocols) : Flooding, Sensor - Disjoint Multipath, Braided Multipath**
6. **Các giao thức định tuyến không đồng nhất (Heterogeneity - based Protocols) : IDSQ, CADR, CHR**
7. **Các giao thức định tuyến hướng chất lượng dịch vụ (QoS-based protocols) : SAR, SPEED, Energy-aware routing**

Trong bài báo này, hướng nghiên cứu tập trung chủ yếu vào một số giao thức định tuyến phổ biến có khả năng đa đường truyền, thiết kế cho mạng không dây nhiều tầng (multi-hop wireless network) , đó là các giao thức : Flooding, GF, RIP và giao thức DSR.

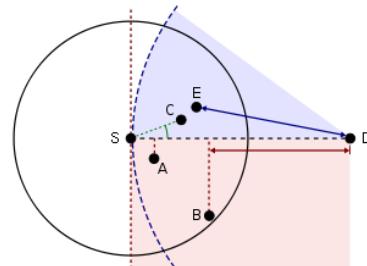
1.Giao thức Flooding : là giao thức đơn giản nhất, không phụ thuộc vào cấu hình mạng. Trong giao thức Flooding, gói tin được gửi quảng bá từ một nút tới tất cả các nút lân cận của nó cho tới khi tới đích. Vì thế nó tạo ra quá nhiều gói tin dư thừa, chiếm dụng hết băng thông mạng.Flooding thường chỉ được sử dụng để so sánh, đánh giá hiệu năng hoạt động của các giao thức khác trong mạng cảm biến không dây



Hình 2 : Giao thức Flooding

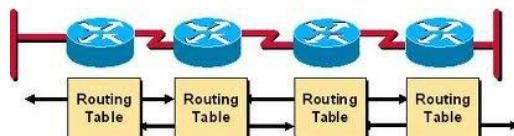
2.Giao thức định tuyến chuyển tiếp tham lam GF (Geographic greedy forwarding) [10] là một thuật toán định tuyến hiệu quả dựa trên vị trí của các nút mạng và được sử dụng trong mạng ad hoc không dây có quy mô lớn. Trong giao

thức định tuyến GF một nút mạng không cần lưu thông tin về đường đi mà nó sẽ gửi gói tin tới một nút mạng lân cận của mình khi nó thấy nút mạng đó có vị trí gần nút đích nó muốn gửi đến nhất. Và nút mạng trung gian khi nhận được gói tin cũng chuyển tiếp gói tin dựa vào cơ chế trên. Giao thức GF cũng có khả năng tự cấu hình do cứ sau một thời gian nhất định nút mạng sẽ gửi gói tin Hello đến các nút xung quanh để cập nhật lại thông tin.



Hình 3 : Giao thức GF

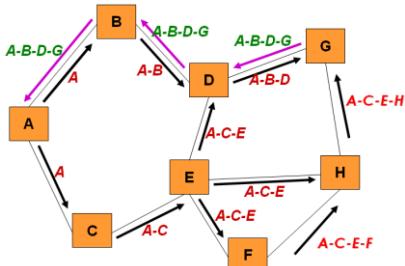
3.Giao thức định tuyến RIP[11] :RIP là giao thức được sử dụng trong mạng Ethernet, được đề cử thử nghiệm để từ đó có thể cải tiến dùng trong mạng WSN được hiệu quả. Đặc điểm của giao thức RIP là dựa trên định tuyến theo vectơ khoảng cách. Giao thức này thực hiện cơ chế cập nhật định kỳ bảng định tuyến bằng cách gửi và nhận thông tin với các nút lân cận. Việc cập nhật định kỳ này giúp trao đổi thông tin khi cấu trúc mạng thay đổi. Khi một gói tin cần chuyển tới đích, nó sẽ dựa vào bảng thông tin định tuyến của nó để chọn ra nút đi tiếp theo. Ưu điểm của giao thức này là không có gói tin dữ liệu dư thừa khi tới đích. Nhược điểm là không thể dùng để định tuyến cho mạng có đồ hình lớn do số hop tối đa của RIP là 16, thời gian hội tụ (thời gian xây dựng và cập nhật bảng định tuyến khi có thay đổi đồ hình mạng) là khá chậm. RIP thích hợp dùng cho mạng nhỏ và ổn định.



Hình 4 : Giao thức RIP

4.Giao thức định tuyến nguồn động (DSR - Dynamic Source Routing)[12] cho phép có thể tự tổ chức, tự cấu hình mà không cần phải có bất cứ thông tin nào sẵn có về hạ tầng mạng hay quản trị mạng. DSR dựa vào hai cơ chế chính đó là cơ chế tìm đường đi và cơ chế bảo trì đường đi. Cơ chế dò tìm đường đi chỉ hoạt động khi chưa thấy đường đi trong bộ nhớ của nút gửi. Vì thế mà nó không cần gửi gói tin quảng bá hay định kỳ gửi gói tin tìm các nút lân cận như một số giao thức trên, nên tiết kiệm được băng thông và năng lượng của mạng. Mỗi khi phát sinh ra cơ chế dò đường, thì gói tin dò đường sẽ được đánh số định danh duy nhất. Nhờ số định danh này mà khi nút mạng nhận được một gói tin dò đường có chỉ số định danh trùng với gói tin trước đó thì gói tin này bị huỷ. Ngoài ra, nếu nút nào nhận được gói tin dò đường mà nút này nằm trong danh sách các nút trung gian đi qua của gói tin, thì gói tin này cũng bị huỷ. Điều này

tránh được việc lặp các gói tin. Cơ chế bảo trì đường đi phát hiện ra đường đi cũ từ nút nguồn tới nút đích đã gặp lỗi. Nút đích sẽ tìm trong bộ nhớ xem có đường đi khác không, nếu không nó sẽ khởi động quá trình dò tìm đường đi. Chính vì giao thức này dựa vào chế độ chỉ thực thi khi có yêu cầu, nên làm giảm đáng kể thông lượng của toàn bộ hệ thống.



Hình 5 : Giao thức DSR

3. MÔ PHỎNG VÀ ĐÁNH GIÁ

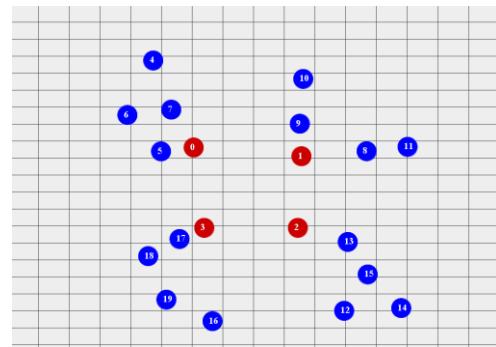
A.Cài đặt, mô phỏng :

Để tiến hành phân tích, đánh giá và cải tiến các giao thức định tuyến, nhóm tác giả đã sử dụng phần mềm WSNET[13] để tiến hành cài đặt mô phỏng các giao thức định tuyến. WSNET là một phần mềm mô phỏng hướng sự kiện cho mạng cảm biến không dây. Wsnet cung cấp các khả năng mô phỏng: mô phỏng các nút mạng, mô phỏng môi trường xung quanh nút mạng, mô phỏng môi trường truyền.

Kịch bản thử nghiệm các giao thức như sau : các nút mạng được triển khai để giám sát, theo dõi thông tin trong một tòa nhà. Trong đó các mục tiêu giám sát bao gồm :

- Thu thập thông tin về nhiệt độ
- Cảm biến hồng ngoại thu thập thông tin về số lượng người ra, vào tòa nhà cũng như ra vào từng phòng của tòa nhà.

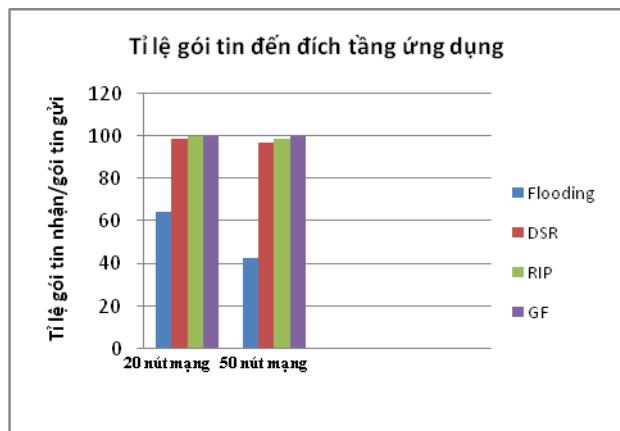
Ở đây giả thiết là hệ thống được đặt tại tầng 6 thư viện Tạ Quang Bửu. Thử nghiệm sẽ được tiến hành trong vòng 1 ngày (thời gian mô phỏng) với hai lần thử nghiệm 20 nút mạng và 50 nút mạng. Tất cả thông tin truyền/nhận sẽ được ghi ra file log và phân tích số liệu. Số liệu được phân tích trên hai góc độ: tầng ứng dụng và phân định tuyến của tầng mạng. Dữ liệu dùng để phân tích bao gồm nồng độ ô nhiễm, số gói tin gửi, nhận, chuyển tiếp và các gói tin bị hủy trên các nút mạng.



Hình 6 : Đồ hình mạng 20 nút

Trong đồ hình trên có 4 nút mạng chính (gateway) có nhiệm vụ thu thập, xử lý dữ liệu, các nút cảm biến sẽ gửi thông tin về cho nút gateway.

B.Kết quả mô phỏng



Hình 7 : Biểu đồ thống kê gói tin tầng ứng dụng

Bảng 1: Thống kê gói tin với 50 nút mạng

Bảng thống kê gói tin tầng mạng

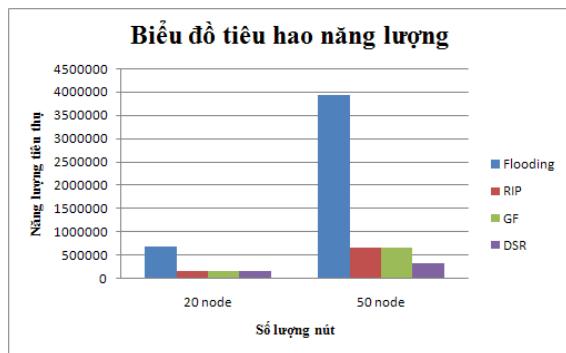
	Gửi	Gửi data	Nhận	Nhận data	Chuyển tiếp	Hủy
Flooding	6758	6758	116538	4377	59969	52192
RIP	20124	7152	14255	7152	4933	106
GF	19647	7059	15431	7059	4576	78
DSR	14540	7256	25985	7256	14982	857

Bảng 2 : Thống kê gói tin với 20 nút mạng

Qua biểu đồ và bảng thống kê gói tin tầng mạng ta thấy :

Tỉ lệ gói tin đến đích của giao thức Flooding là khá thấp đặc biệt là khi số lượng nút mạng tăng lên giảm xuống dưới 50% , đi kèm với đó là số lượng gói tin phải chuyển tiếp và hủy là rất lớn bằng 10 tới 25 lần số lượng gói tin chuyển đi và gấp 20 tới 50 lần số gói tin đến được đích. Hai giao thức RIP và GF có tỉ lệ gói tin đến đích rất cao gần như 100% tuy nhiên số lượng gói tin gửi đi tăng do nhu cầu gửi thông tin cập nhật bằng định tuyến (RIP) cũng như nhu cầu cập nhật thông tin về các nút lân cận (GF). Trong các giao thức thì giao thức DSR tỏ ra hiệu quả hơn cả về độ giảm thiểu gói tin dư thừa trong mạng. Tuy hiệu suất thành công của gói tin không cao bằng giao thức RIP và giao thức GF nhưng tỉ lệ vẫn rất cao xấp xỉ 97% với đồ hình 50 nút mạng.

Độ tiêu hao năng lượng của các giao thức :



Hình 8: Độ tiêu hao năng lượng

Ta nhận thấy rằng, với số lượng nút mạng ít (20 nút), thì 3 giao thức RIP, GF, DSR có mức tiêu hao năng lượng là gần giống nhau và giao thức Flooding tiêu tốn gấp khoảng 5 lần so với 3 giao thức còn lại. Còn với đồ hình mạng có số nút tăng lên, thì giao thức Flooding tiêu tốn khá nhiều năng lượng mạng, điều này cũng chứng minh được độ tiêu hao năng lượng tỷ lệ thuận với số nút mạng, làm tăng số gói tin dư thừa, giảm băng thông toàn mạng. Hai giao thức RIP và GF có năng lượng tiêu hao gần là như nhau. Đặc biệt giao thức DSR tiêu tốn ít năng lượng nhất, điều này cũng chứng tỏ được rằng, do nút mạng chỉ gửi gói tin tìm đường khi có nhu cầu, và giao thức này không phải gửi định các gói tin như giao thức RIP và GF.

C. Đánh giá, đề xuất cải tiến các giao thức

Flooding : Qua quá trình cài đặt và thử nghiệm ta thấy rằng hiệu suất truyền gói tin của giao thức Flooding không cao bên cạnh đó lại tạo ra số lượng gói tin dư thừa rất lớn số lượng gói tin trong mạng tăng cao gây nghẽn mạng và tiêu tốn năng lượng của nút mạng. Vì vậy trong thực tế Flooding thường chỉ được sử dụng

để làm mốc so sánh, đánh giá độ hiệu quả của các giao thức định tuyến khác.

Cải tiến : Chúng ta có thể cải tiến Flooding bằng cách lưu vết đường đi của các gói tin qua các nút mạng để tránh hiện tượng gửi trùng lặp gói tin. Thay cơ chế quảng bá gói tin bằng cách chọn ngẫu nhiên nút lân cận để gửi hoặc có thể kết hợp giao thức toán định tuyến nguồn

RIP : Giao thức RIP là giao thức định tuyến được sử dụng cho mạng có dây với số lượng nút mạng nhỏ, tuy nhiên nếu chúng ta giữ nguyên cơ chế hoạt động của RIP khi áp dụng vào mạng cảm biến không dây thì sẽ xảy ra một số bất cập thứ nhất là RIP tiêu tốn năng lượng do đều đặn sinh ra và gửi đi các gói tin chứa băng định tuyến của mình cho các nút mạng lân cận, thứ hai là khi mạng có số lượng nút lớn thì gói tin không đến được đích do hop_limit chỉ là 16, thứ ba là RIP chỉ hoạt động tốt trong các mạng có tần số ổn định cao nó cập nhật đường đi chậm khi đồ hình mạng có sự thay đổi trong khi sự thay đổi đồ hình mạng trong mạng cảm biến không dây là khá cao nhất là các mạng cảm biến có nút mạng chuyển động. Chúng ta có thể cải tiến bằng cách :

Cải tiến : Thay vì định kì gửi gói tin tìm đường các nút mạng chỉ tìm đường khi có yêu cầu gửi gói tin, cơ chế này giống với cơ chế của giao thức AODV, bên cạnh đó chúng ta cũng phải tăng hop_limit để gói tin có thể truyền đi xa hơn.

GF: hoạt động khá hiệu quả về khả năng truyền gói tin đến đích, tuy nhiên chúng cũng làm tăng dung lượng gói tin trong mạng do phải định kì gửi và các gói tin cập nhật thông tin của nút lân cận. Cũng giống như RIP, GF là một giao thức có độ tin cậy cao và có

	Gửi	Gửi data	Nhận	Nhận data	Chuyển tiếp	Hủy
Flooding	17634	17634	533657	7502	402583	1233541
RIP	61795	17525	65985	17524	42260	2930
GF	62497	18009	71671	18009	33306	874
DSR	31106	19455	53563	19445	26919	5458

khả năng tự cấu hình mạng vì nó cập nhật thông tin về đồ hình mạng định kì nên khi có sự thay đổi trong mạng các giao thức này vẫn có thể tìm được đường đi để gói tin đến được đích. GF có khả năng cập nhật nhanh hơn RIP vì nó chỉ cập nhật thông tin của các nút lân cận chứ không phải thông tin đường đi như RIP. Độ hiệu quả của thuật toán định tuyến GF trong chương trình mô phỏng là rất tốt nhưng trong thực tế lại khó có thể áp dụng lên tất cả các mạng cảm biến không dây bởi GF chỉ hoạt động được khi các nút mạng biết được vị trí của mình, của các nút lân cận và trong header của gói tin phải có thông tin về vị trí nút đích mà trong thực tế bài toán xác định vị trí cho các nút mạng trong mạng cảm biến không dây hiện cũng là một vấn đề không dễ giải quyết.

Cải tiến : Với những mạng có độ ổn định cao, các nút mạng cố định hoặc hầu như không di chuyển thì chúng ta có thể cải tiến GF bằng cách cho các nút mạng cập nhật thông tin các nút lân cận khi bắt đầu triển khai mạng, sau đó sẽ dùng các thông tin đó để chuyển tiếp gói tin chứ không phải cập nhật thường xuyên 30

giây một lần, như vậy sẽ giảm đáng kể lưu lượng gói tin trong mạng.

DSR : Là một giao thức khá linh động, bởi nút mạng sẽ tự cập nhật đường đi khi chưa có thông tin đường đi lưu trong bộ nhớ. Do đó DSR có thể áp dụng cho rất nhiều mạng có đồ hình khác nhau hay có các mạng có đồ hình hay thay đổi. DSR cũng tiêu thụ năng lượng rất ít, do chỉ gửi gói tin dò đường khi chưa biết đường đi. Nếu có đường đi, thì nó sẽ gửi theo định tuyến nguồn và tới đích dễ dàng. Tuy nhiên còn một số khuyến điểm là nếu như sử dụng giao thức DSR để gửi tới số lượng nút đích nhiều, thì mỗi nút sẽ lưu rất nhiều đường đi trong bộ nhớ của nó, hoặc cũng xảy ra trường hợp phải gửi gói tin dò đường nhiều lần do gói tin trả lời không về được nút nguồn.

Cải tiến : Thay vì lưu nhiều đường đi, mỗi nút sẽ chỉ lưu đường đi từ nó đến nút đích của gói tin dò đường từ nút nguồn đến. Hạn chế lưu quá nhiều đường đi từ nút đó đến các nút khác mà không phải nút đích. Các nút trung gian cũng có thể cập nhật thông tin đường đi thông qua các gói tin chuyển qua nó.

4. KẾT LUẬN

Trong mạng cảm biến không dây, sự hiệu quả của các thuật toán định tuyến phụ thuộc khá nhiều vào đồ hình mạng cũng như ứng dụng của mạng. Vì vậy để chọn ra một thuật toán định tuyến tốt cho mọi cảm biến không dây là rất khó. Trong thực tế khi triển khai một mạng cảm biến không dây ta phải phân tích rõ yêu cầu chức của ứng dụng cũng như đồ hình mạng để lựa chọn những thuật toán định tuyến cho phù hợp.

Trong nội dung báo cáo chúng tôi đã giới thiệu và phân tích một số thuật toán định tuyến cho mạng cảm biến không dây dựa trên cơ sở lý thuyết cũng như trên kết quả mô phỏng. Một số cải tiến đề nghị giúp nâng cao hiệu quả truyền gói tin trong mạng cho các giao thức Flooding, RIP, DSR, GF cũng đạt được những kết quả nhất định trên chương trình mô phỏng. Tuy nhiên cần phải được kiểm chứng trên môi trường thực tế là các nút mạng thật thi mới có thể đánh giá chính xác được sự hiệu quả của các giao thức cũng như của các cải tiến.

Trên cơ sở những những phần đã đạt được cùng với khả năng phát triển của đề tài, định hướng nghiên cứu tiếp theo sẽ là :

- Tiếp tục tìm hiểu, cài đặt, phân tích và đánh giá thêm các giao thức định tuyến khác, từ đó tìm ra những ưu, nhược điểm để áp dụng và cải tiến.
- Cài đặt trên thiết bị thực tế là các nút mạng các giao thức đã cho kết quả tốt trên phần mềm mô phỏng để đánh giá hiệu năng thực tế của giao thức.

5. LỜI TRI ÂN

Trong thời gian thực hiện đề án cùng với sự cố gắng, nỗ lực của bản thân, chúng tôi còn nhận được sự giúp đỡ và hướng dẫn tận tình của Tiến sĩ Nguyễn Kim Khánh và Thạc sĩ Phạm Văn Thuận đã giúp chúng tôi hoàn thành đề tài này. Chúng tôi muốn gửi lời cảm ơn chân thành và sâu sắc tới hai thầy. Bên cạnh đó chúng tôi cũng xin cảm ơn những người bạn đã ủng hộ động viên và cho chúng tôi những ý kiến đóng góp để chúng tôi có thể hoàn thành tốt hơn đề tài này.

6. TÀI LIỆU THAM KHẢO

- [1] Jun Zheng and Abbas Jamalipour, “Wireless Sensor Networks: A Networking Perspective”, a book published by A John & Sons, Inc, and IEEE, 2009.
- [2] Jamal Al-Karaki, and Ahmed E. Kamal, “Routing Techniques in Wireless Sensor Networks: A Survey”, IEEE Communications Magazine, vol 11, no. 6, Dec. 2004, pp. 6-28.
- [3] Ian F. Akyildiz, Ilyas Su, Yogesh Sankarasubramaniam, and Erdal Cayirci. “A survey on Sensor Networks”. IEEE Communications Magazine, August 2002.
- [4] Jamal N. Al-Karaki Ahmed E. Kamal - Routing Techniques in Wireless Sensor Networks: A Survey
- [5] www.technologyreview.com
- [6] Pham Van Thuan – Optimizing routing protocol for Wireless Sensor Network – Ms Thesis – Ha Noi University of Science and Technology - 2009
- [7] W. R. Heinzelman, A. Chandrakasan and H. Balakrishnan. “Energy - efficient Communication Protocol for Wireless Microsensor Networks ”. Proceedings of the IEEE Hawaii International Conference on System Sciences (HICSS), Vol. 8. (2000) 1-10
- [8] R. C. Shah and J. M. Rabaey. “Energy aware routing for low energy adhoc sensor networks”. Proceedings of the IEEE Wireless Communication and Networking Conference (WCNC) (2001)
- [9] Karl, H.; Willig, A. Protocols and Architectures for Wireless Sensor Networks. John Wiley & Sons: Chichester, West Sussex, UK, 2005
- [10] A. Kermarrec and G. Tan - Greedy Geographic Routing in Large-Scale Sensor Networks: A Minimum Network Decomposition Approach - MobiHoc 2010
- [11] C. Hendrik, [RFC 1058](#), Routing Information Protocol, The Internet Society (June 1988)
- [12] Y. Hu – UIUC – D. Maltz - The Dynamic Source Routing Protocol (DSR) Rice University – Microsoft Research - February 2007
- [13] WSNET - An event-driven simulator for large scale wireless sensor networks - <http://wsnet.gforge.inria.fr>

Nghiên cứu lý thuyết và xây dựng hệ thống phát hiện xâm nhập

Nguyễn Xuân Quang

Tóm tắt: Hiện nay cùng với sự phát triển của mạng máy tính thì cũng xuất hiện những nguy cơ về mất an toàn thông tin cũng như an ninh mạng. Với xu thế đó việc nghiên cứu về các giải pháp để tăng cường an ninh cho hệ thống thông tin cũng ngày càng được chú ý. Trong đó nghiên cứu về các hệ thống phát hiện xâm nhập dựa trên bắt thường mạng cũng ngày càng được quan tâm. Nội dung đề tài này trình bày về thuật toán TCM-KNN và ứng dụng của nó trong việc phát hiện xâm nhập. Các kết quả kiểm tra với bộ dữ liệu KDD cup 99 cho kết quả phát hiện tốt tốt cùng với độ phức tạp của thuật toán không quá cao mở ra khả năng ứng dụng thuật toán trong việc xây dựng hệ thống phát hiện xâm nhập trong thực tế.

Từ khóa: hệ thống phát hiện xâm nhập, TCM-KNN, an ninh mạng.

1. Giới thiệu.

Ngày nay khi công nghệ thông tin ngày càng phát triển ngoài những tiện ích của nó như hỗ trợ việc giao tiếp cũng như làm việc của con người. Mạng internet phát triển cũng mang lại nhiều rủi ro như vấn đề an toàn thông tin, tấn công trên mạng. Do đó việc bảo vệ an toàn thông tin và giữ an ninh cho các hệ thống mạng ngày càng trở thành vấn đề cấp thiết cần được nghiên cứu. Một trong những chủ đề được đề cập cũng như được nghiên cứu rất nhiều là nghiên cứu và xây dựng các hệ thống phát hiện xâm nhập (một thiết bị không thể thiếu trong việc bảo đảm an toàn cho các mạng thông tin).

Phát hiện xâm nhập dựa trên bắt thường mạng [8] là một phần quan trọng trong việc nghiên cứu về an ninh mạng nói chung cũng như về các hệ phát hiện xâm nhập nói riêng. Hiện tại có rất nhiều phương pháp được đưa ra trong việc phát hiện xâm nhập dựa vào bắt thường mạng như: thuật toán KNN, SVMs (support vector machines), Neutron.... Nhưng các thuật toán này còn gặp phải nhiều hạn chế như xác suất phát hiện đúng còn thấp, đòi hỏi dữ liệu đầu vào rất khắt khe. Do đó ứng dụng trong thực tế của những thuật toán này còn chưa cao.

Đề tài được thực hiện dưới sự bảo trợ của trường Đại học Bách Khoa Hà Nội, sử dụng các kết quả nghiên cứu của giáo sư Yang Li học viện khoa học Bắc Kinh.

Nguyễn Xuân Quang, sinh viên lớp HTT&TT-KSCLC-K51 trường Đại học Bách Khoa Hà Nội (e-mail:quangnx@vnsecurity.vn)

© Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội.

Trong khuôn khổ đề tài đưa ra một thuật toán phát hiện xâm nhập mới có khả năng ứng dụng tốt TCM-KNN (Transductive Confidence Machines for K-Nearest Neighbors). Trên quan điểm kế thừa các đặc điểm nổi bật của thuật toán KNN (k-nearest neighbor) cũng như lý thuyết thống kê. Thuật toán TCM-KNN đã tỏ ra có nhiều điểm nổi bật như khả năng phát hiện rất tốt, không đòi hỏi quá khắt khe đối với dữ liệu đầu vào. Với những kết quả đạt được khả năng ứng dụng thuật toán trong thực tế là rất lớn.

2. Cơ sở lý thuyết và cơ bản về thuật toán TCM-KNN.

2.1. Các đại lượng ngẫu nhiên và phân phối xác suất trong việc phát hiện xâm nhập.

Đại lượng ngẫu nhiên hay biến ngẫu nhiên là giá trị thực tùy thuộc vào kết quả ngẫu nhiên của phép thử. Để xác định đặc trưng cho các đại lượng ngẫu nhiên người ta sử dụng hàm phân phối xác suất. Với một biến ngẫu nhiên liên tục ta có hàm phân phối xác suất của nó là:

$$F(x) = P(X < x) \quad (1)$$

Với x là một biến số bất kỳ. Hàm phân phối ngẫu nhiên này có thể sử dụng với các biến ngẫu nhiên rời rạc. Hàm phân phối xác suất này đặc trưng cho xác suất xuất hiện của các đại lượng nhỏ hơn biến số x bất kỳ trong không gian mẫu. Với một không gian mẫu xác định xuất phát từ 0 nó có thể đặc trưng cho khả năng tồn tại của một biến số bất kỳ x trong một không gian mẫu xác định trước. Nhờ vào tính chất này ta có thể xác định được một giá trị bất kỳ có thuộc vào một lớp cho trước không.

Các dữ liệu về mạng cũng là một đại lượng ngẫu nhiên (đại lượng ngẫu nhiên liên tục) nhưng do việc lấy mẫu theo chu kỳ nên có thể coi việc tính toán như đối với biến ngẫu nhiên rời rạc. Những nghiên cứu của Martin-Löf đã chỉ ra rằng tồn tại một phương pháp để xác định đặc điểm của một chuỗi dữ liệu. Nhưng không thể tính toán được mà phải dựa tính toán gần đúng dựa vào một giá trị p (xác suất xuất hiện). p đặc trưng cho xác suất xuất hiện hay quan sát được của một điểm trong một mẫu các biến ngẫu nhiên. Giá trị p này được sử dụng nhưng một giá trị để đánh giá khả năng tồn tại hay không tồn tại của một điểm trong một tập dữ liệu cho trước. Giá trị p lớn cho ta thấy khả năng điểm ta xem xét thuộc lớp dữ liệu ta xem xét là cao, giá trị p càng nhỏ cho ta thấy khả năng điểm ta xem xét thuộc lớp các dữ liệu ta có là nhỏ. Về tính toán ta có:

$$p(x) = P(X > x) \quad (2)$$

Với tính chất của mình giá trị p được dùng trong các phương pháp phân lớp và chia đặc tính rất tốt. Trong môi trường mạng mỗi một thời điểm có thể được chia vào các trạng thái (lớp) khác nhau như: bình thường, ddos, Việc sử dụng xác suất p trong việc xác định bắt thường mạng đạt hiệu quả tốt. Ta có với các dữ liệu về mạng cùng trạng thái như sau: (x_1, x_2, \dots, x_n) ta sẽ có xác suất ứng với một giá trị ngẫu nhiên x sẽ được tính như sau:

$$p(x) = \frac{\#(x \leq x_n)}{n} \quad (3)$$

Với $\#(x \leq x_n)$ là phép đếm các giá trị ($x \leq x_n$). Việc sử dụng

hàm xác suất p này cho ta một phương pháp xác định tốt hơn việc xác định giá trị thuộc mẫu xác định sử dụng vọng số như sau :

$$x \leq (X \pm \varepsilon) \quad (4)$$

Với ε là một giá trị nhỏ thường được dùng bằng 1 hoặc bằng 2 lần độ lệch chuẩn phương của tập các biến ngẫu nhiên có cùng đặc tính.

Với các nghiên cứu về đại lượng ngẫu nhiên cho ta một phương pháp có tính chính xác cao và tính toán được dùng trong việc phát hiện các bất thường trong một trường mạng. Giá trị xác suất p được thuật toán TCM-KNN sử dụng như một kết quả để xác định và phân lớp các trạng thái của mạng cũng như xác định trạng thái của một điểm cần xem xét.

2.2. Thuật toán KNN (*k-nearest neighbor*).

Thuật toán KNN là thuật toán với độ phức tạp thấp sử dụng trong việc phân lớp hay là chia đặc tính với các dữ liệu ta cần phân tích. Thuật toán KNN sử dụng phương pháp là dựa và điểm khác biệt của điểm ta cần xét với những hàng xóm có cùng đặc điểm, từ đó đưa ra dữ liệu mà từ đó làm căn cứ để xác định một điểm ta xem xét có cùng đặc tính với những điểm mà ta đã có không. Trong thuật toán KNN với mỗi điểm ta xét ta chỉ quan tâm đến những điểm gần nó nhất. Việc này làm cho việc tính toán đơn giản hơn nhưng vẫn đưa ra được khả năng phân lớp cũng như xác định sự khác biệt tốt.

Điểm hạn chế cơ bản của thuật toán KNN là do việc sử dụng duy nhất các điểm gần với điểm ta xem xét, do đó có các khuyết điểm như sau:

- Các điểm lân cận có thể không giống nhau với điểm mà ta xem xét.
- Do đặc điểm trên nên khả năng phân lớp cũng như phát hiện bất thường mạng không cao.

Để hạn chế những đặc điểm này trong thuật toán TCM-KNN thay vì chỉ tính toán khác biệt của điểm ta xem xét với các điểm gần nhất giá trị này được tính toán bằng sự khác biệt của điểm ta xem xét với các điểm gần giống nó nhất (bằng phép sắp xếp tập khoảng cách giữa các điểm).

2.3. Thuật toán TCM-KNN.

Thuật toán TCM-KNN được sử dụng để phân lớp dữ liệu và được phát triển để phát hiện các bất thường trong mạng. Giả sử chúng ta có một nhóm các dữ liệu mẫu đầu vào (training set) với n điểm $\{(x_1, y_1), \dots, (x_n, y_n)\}$ trong đó $x_i = \{x_i^1, \dots, x_i^m\}$ là tập các thông tin về trạng thái của mạng như (số lượng connection, số bytes gửi nhận ...) và y_i là lớp của điểm đó thuộc vào có giá trị từ $(1, 2, \dots, c)$ như (DDOS, Sql injection ...). Quá trình kiểm tra của thuật toán sử dụng một tập hợp các điểm với cùng các đặc tính giống như dữ liệu mẫu đầu vào và tiến hành phân lớp các điểm thuộc nhóm này.[3]

Để loại bỏ các hạn chế của thuật toán KNN khác với thuật toán KNN (K-Nearest Neighbors) với thuật toán TCM ta sẽ sắp xếp tập các khoảng cách của các điểm trong nhóm (được tính toán sử dụng các phương pháp tính khoảng cách ví dụ như Euclid Distance). Ta có: D_i^y là khoảng cách từ điểm i đến các điểm cùng loại đã được sắp xếp, D_{ij}^{-y} là khoảng cách ngắn thứ j.

D_i^y là khoảng cách từ điểm i đến các điểm khác loại. Các dữ liệu này được sử dụng để tính toán một giá trị mà ta gọi là giá trị xác định sự khác biệt. Giá trị này xác định sự khác biệt của điểm

ta đang xét với các điểm khác có quan hệ với nó. Trong trường hợp này ta sử dụng thuật toán KNN (chỉ quan tâm đến các điểm gần nó) có giá trị xác định sự khác biệt được tính như sau:

$$a_{iy} = \frac{\sum_{j=1}^k D_{ij}^{-y}}{\sum_{j=1}^k D_{ij}^y} \quad (5)$$

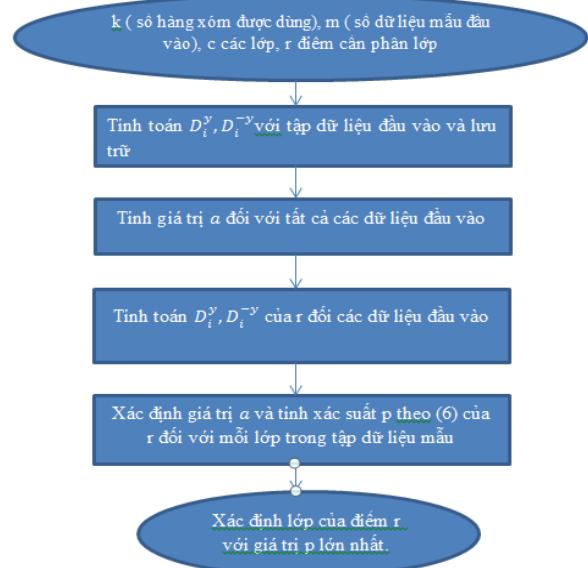
Với k là số các điểm hàng xóm được. Giá trị này được tính toán bằng thương của tổng khoảng cách của điểm đó với k điểm cùng lớp với tổng khoảng cách của điểm đó với k điểm khác nhau (tổng này đã được sắp xếp). Giá trị này sẽ tăng khi khoảng cách từ điểm đó đến các điểm cùng lớp tăng lên và khoảng cách từ nó đến các điểm khác lớp ngắn lại.

Với giá trị gần đúng đã đề cập ở phần trước để làm căn cứ cho các tính toán đối với các dữ liệu ngẫu nhiên ta có hàm p tính từ các giá trị khác biệt (strangeness) của một điểm được đưa ra như sau:

$$p(a_{new}) = \frac{\#\{i:a_i > a_{new}\}}{n+1} \quad (6)$$

Với a_{new} là giá trị xác định sự khác biệt của điểm ta đang xem xét thuộc tập dữ liệu cần phân lớp (test set). Các dữ liệu dùng để so sánh với a_{new} là tập các giá trị đánh giá sự khác biệt của tập hợp các điểm dùng cho quá trình học tập (training phase). Số lượng các giá trị đánh giá sự khác biệt của tập dữ liệu học tập lớn hơn a_{new} là j thì ta sẽ có giá trị xác suất hiện của điểm ta cần xem xét được tính bằng thương của j chia cho $(n+1)$. Hay giá trị p là xác định xác suất của giá trị ta xem xét trong tập hợp gồm bao gồm các điểm thuộc nhóm dữ liệu học tập và điểm ta cần xem xét. Thuật toán TCM-KNN được xây dựng dựa trên hàm tính toán giá trị p phương trình (6) này.

Thuật toán được trình bày dưới dạng sơ đồ khối như sau:



Hình 1.a. Sơ đồ khối thuật toán TCM-KNN.

Ta thấy thuật toán như trên để phân lớp các dữ liệu ta cần dữ liệu training đầy đủ, thêm nữa thuật toán này có độ phức tạp tính toán và tốn bộ nhớ rất lớn. Vì những lý do trên thuật toán trên chỉ có giá trị về mặt lý thuyết và cần được thay đổi để áp dụng được trong thực tế.

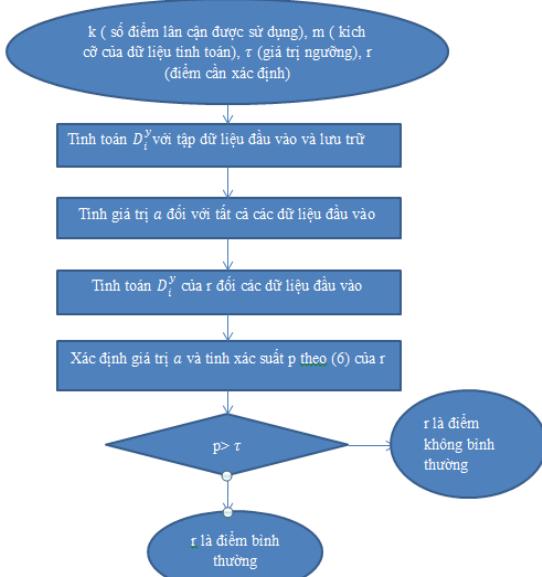
2.4. Thuật toán TCM-KNN áp dụng trong thực tế.

Để giải quyết các hạn chế của thuật toán TCM-KNN được nêu ra ở trên để thuật toán sử dụng được với bộ dữ liệu đầu vào “không đầy đủ” và có thời gian tính toán đủ thấp. Thuật toán được áp dụng một số thay đổi cơ bản như sau:

- Các điểm chỉ được chia làm 1 lớp là “bình thường” (normal) và “không bình thường” (unormal).
- Do chỉ có hai lớp nên giá trị của a_{iy} được thay bằng a_i và được tính toán:

$$a_i = \sum_{j=1}^k D_{ik} \quad (7)$$

Những thay đổi như vậy giúp cho thuật toán thích hợp hơn khi xử lý trong thực tế, với độ phức tạp có thể chấp nhận được cho việc tính toán trong thời gian thực. Thuật toán được miêu tả sơ đồ khái niệm như sau:



Hình 1.b. Sơ đồ khái niệm thuật toán TCM-KNN dùng trong thực tế.

Thuật toán như trên để phân lớp các dữ liệu ta cần dữ liệu training đầy đủ, thêm nữa thuật toán này có độ phức tạp tính toán và tốn bộ nhớ rất lớn. Vì những lý do trên thuật toán trên chỉ có giá trị về mặt lý thuyết và cần được thay đổi để áp dụng được trong thực tế. Các thay đổi trên khiến cho độ phức tạp của thuật toán giảm xuống đến mức có thể chấp nhận được để sử dụng trong thực tế, trong khi đó vẫn duy trì được khả năng phát hiện cao.

2.5. Đánh giá độ phức tạp và sử dụng tài nguyên của thuật toán.

Khi dữ liệu đầu vào đã được chuẩn hóa về số duy nhất thì phép toán phức tạp và được sử dụng nhiều nhất trong thuật toán là phép toán sắp xếp dữ liệu (được sử dụng n lần). Do đó độ phức tạp của thuật toán phụ thuộc nhiều vào thuật toán sắp xếp và ta lựa chọn. Vì tính phụ thuộc về thời gian của thuật toán vào phép so sánh này nên thuật toán sắp xếp lựa chọn phải là một trong các thuật toán hiệu quả ví dụ như: quick sort, heap sort, merger sort.... Với việc sử dụng các thuật toán sắp xếp với độ phức tạp thấp mức $O(nlogn)$ thì độ phức tạp của thuật toán sẽ là $c O(n^2 logn)$ vì phép toán sắp xếp được sử dụng n lần trong thuật toán ở pharse tranning (đây là pharse đòi hỏi thời gian lớn nhất trong thuật toán).

Thời gian để xác định tính chất của một giá trị instance đầu vào phụ thuộc vào việc tính toán Distance của nó với các dữ liệu training bao gồm các khâu tính toán:

- Xác định khoảng cách với các giá trị trong training data set (độ phức tạp $O(n)$)
- Sắp xếp các giá trị (độ phức tạp $O(nlogn)$ với thuật toán sắp xếp hiệu quả)
- So sánh giá trị tính được để xác định p (độ phức tạp $O(n)$).

Do đó thời gian để tính toán với một giá trị instance đầu vào sẽ có độ phức tạp $O(nlogn)$ với khối lượng training data set đủ nhỏ thời gian tính toán này là chấp nhận được.

Đánh giá về bộ nhớ sử dụng của thuật toán: các giá trị D được lưu trữ với khối lượng lớn nhất $c O(n^2)$ nên bộ nhớ sử dụng chủ yếu khi tính toán và xử lý là bộ nhớ dành cho việc lưu trữ các giá trị khoảng cách và việc sắp xếp các giá trị này. Vì vậy việc sử dụng tài nguyên sẽ là mức n^2 độ lớn dữ liệu của D.

Bảng 1. Đánh giá độ phức tạp của thuật toán.

Phrase	Phép toán chủ yếu	Độ phức tạp
Training	Phép toán sắp xếp	$O(n^2 logn)$
detect	Phép toán sắp xếp	$O(nlogn)$

Đánh giá trong trường hợp phép tính distance có độ phức tạp đủ nhỏ.

Dữ liệu	Số lượng	Khối lượng dữ liệu
D	n^2	n^2 độ lớn của D
a	n	n * độ lớn của a
p	n	n * độ lớn dữ liệu
sắp xếp	Phụ thuộc vào thuật toán sắp xếp	Phụ thuộc vào thuật toán sắp xếp lựa chọn và n

Bảng 2. Đánh giá mức độ sử dụng bộ nhớ.

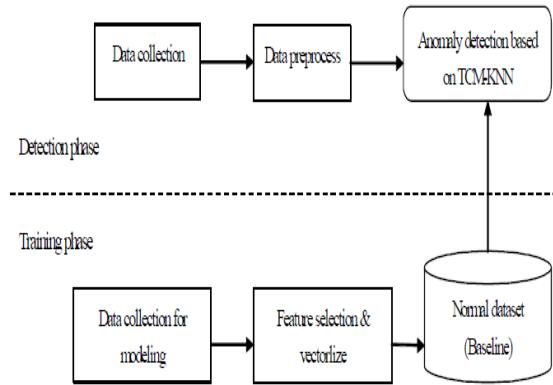
Bộ nhớ sử dụng phụ thuộc nhiều vào các giải thuật cũng như dạng dữ liệu mà ta lựa chọn

3.1. Mô hình thiết kế hệ thống.

Hệ thống bao gồm các thành phần chính:

- Thu thập dữ liệu.
- Phân tích dữ liệu (dùng trong quá trình học tập)
- Tiền xử lý dữ liệu.
- Thuật toán lõi TCM-KNN.

Sau đây là mô hình của hệ thống được thể hiện dưới dạng sơ đồ khái niệm bao gồm đầy đủ các thành phần cũng như bố trí của nó theo trình tự tính toán.



Hình 2. Mô hình thiết kế.[2]

Thành phần thu thập dữ liệu được dùng chung cho cả 2 quá trình là quá trình học và phát hiện, thành phần này sẽ thu thập các yếu tố đầu vào của mạng như số connections, trung bình thời gian của các connections,... Các dữ liệu này sẽ là dữ liệu cho quá trình xử lý về sau của các thành phần khác. Chất lượng của thành phần hay module này ảnh hưởng nhất nhiều đến khả năng phát hiện của thuật toán vì vậy cần những công cụ có tính chính xác cao.

Thành phần phân tích dữ liệu được dùng trong quá trình học, thành phần này sẽ phân tích các dữ liệu “bình thường” được thu thập bởi thành phần thu thập dữ liệu và đưa ra các thông tin như: dữ liệu nào được lựa chọn để đưa và bước tính toán sau, tính rank cho tham số đầu vào mà quá trình thu thập dữ liệu thu thập được. Quá trình này nếu được thực hiện tốt sẽ tăng xác suất phát hiện cho hệ thống mà ta xây dựng.

Thành phần tiền xử lý dữ liệu được sử dụng cho cả 2 trạng thái của hệ thống. Thành phần này dùng các thông tin mà thành phần phân tích dữ liệu đưa ra để xử lý với các điểm đầu vào đưa ra kết quả cho thuật toán TCM-KNN. Thành phần này góp phần giảm chi phí tính toán của thuật toán TCM-KNN khiến nó có nhiều khả năng ứng dụng trong thực tế hơn.

Trong khuôn khổ đề tài nghiên cứu thành phần phân tích dữ liệu và thuật toán lõi TCM-KNN là thành phần được nghiên cứu và xem xét chính.

3.2. Dữ liệu dùng để đánh giá thuật toán.

Dữ liệu sử dụng là tập dữ liệu kddcup99 [5] là bộ dữ liệu hay được sử dụng để kiểm tra các thuật toán máy học dùng cho việc phát hiện bất thường mạng

Dữ liệu đưa ra bao gồm 42 tham số cơ bản, được tách ra để kiểm tra thuật toán như sau:

Training Phase:

Clean data

Dữ liệu là 4000 điểm ở trạng thái bình thường.

Unclean data

Dữ liệu là 4000 điểm (3950 trạng thái “bình thường” 50 “không bình thường”)

Detecting Phase:

Dữ liệu là 8000 điểm (với 2463 ở trạng thái “không bình thường” và 5537 ở “trạng thái bình thường”).

3.3. Phân tích dữ liệu lựa chọn trọng số và ranking.

Quá trình phân tích và lựa chọn tham số cũng như ranking là một khâu rất quan trọng ảnh hưởng trực tiếp đến khả năng phát hiện bất thường của thuật toán. Dựa vào bất đẳng thức Trébursép:

$$P(|X - a| < \varepsilon) \geq 1 - \frac{D(X)}{\varepsilon^2} \quad (8)$$

Để cân bằng các tham số của mỗi điểm xem xét với một giá trị ε xấp xi bằng nhau ta sẽ tính toán giá trị $D(X)$ ứng với mỗi tham số trong miền các điểm dùng cho training phase.

Sau đó sắp xếp các giá trị $D(X)$ ta có giá trị D_{max} là giá trị lớn nhất của $D(X)$. Việc lựa chọn trọng số được thực hiện như sau:

- Chọn một giá trị ngưỡng τ (trong này ta lấy là 1% hay 0.01).
- Với $D(X) > \tau * D_{max}$ ta có ranking cho $D(X)$ là $\frac{D_{max}}{D(X)}$.
- Các trường hợp còn lại $D(X) \leq \tau * D_{max}$ ranking là $\frac{1}{\tau}$.

Giá trị thu được gọi là $r_D(X)$.

Một giá trị nữa sử dụng cho quá trình đánh ranking là số lượng các giá trị khác nhau trong mẫu đối với mỗi tham số (giá trị đã được làm tròn theo $(10 * \tau)\%$ của $D(X)$) thu được một giá trị mà ở đây gọi là $f_n(X)$. Quy đồng $f_n(X)$ với ranking theo $D(X)$ bằng cách chia với giá trị $n * \tau$ (n là số điểm sử dụng trong training phase). Giá trị thu được là $r_f(X)$.

Với 2 giá trị đã tính ở trên bằng việt đánh giá tương đương giữa chúng ta có trọng số ứng với mỗi tham số sẽ được tính toán như sau.

$$r(X) = \frac{r_D(X)}{r_f(X)} \quad (9)$$

Với việc đánh ranking như trên và khi đưa vào bộ tiền xử lý dữ liệu ta có thể đưa ra dữ liệu đầu ra với mỗi điểm cần xét là một tham số duy nhất. Như vậy cùng với việc nâng cao chất lượng của bộ phân tích và đánh giá dữ liệu thì cũng tương ứng với khả năng nâng cao chất lượng phát hiện cũng như giảm độ phức tạp của thuật toán.

3.4. Thủ nghiệm và các kết quả thu được.

Thuật toán và thành phần phân tích dữ liệu được lập trình bằng ngôn ngữ python dữ liệu đầu vào dưới dạng csv, và được lưu thành file. Quá trình kiểm tra thực hiện với bộ dữ liệu đầu vào là clean và unclean, với dữ liệu để kiểm tra giống nhau gồm 8000 điểm đã được lọc ra. Thực hiện tính toán với mỗi bộ dữ liệu với giá trị k thay đổi trong các giá trị (50, 100, 150, 200) và giá trị ngưỡng τ được giữ không đổi là 0.05 hay 5% cho kết quả rất đáng khâm quan.

Với dữ liệu clean :

Bảng 3. Kiểm tra với dữ liệu clean

	TP	FP
K=50 $\tau = 0.05$	99.41%	2.97%
K=100 $\tau = 0.05$	99.14%	3.44%
K=150 $\tau = 0.05$	98.98%	3.8%
K=200 $\tau = 0.05$	99.21%	3.86%

Với dữ liệu unclean

Bảng 4. Kết quả kiểm tra với dữ liệu unclean.

	TP	FP
K=50 $\tau = 0.05$	98.94%	3.61%
K=100 $\tau = 0.05$	98.74%	4.31%
K=150 $\tau = 0.05$	98.82%	4.67%
K=200 $\tau = 0.05$	98.32%	4.72%

Các kết quả thu được cho thuật toán có khả năng phát hiện cho thấy thuật toán có khả năng phát hiện rất tốt với xác suất phát hiện chính xác lớn cũng như xác suất báo sai không quá cao. Thêm vào đó kết quả kiểm tra với các dữ liệu clean và các dữ liệu unclean (có nhiều) cho thấy rằng thuật toán không bị phụ thuộc nhiều vào chất lượng của dữ liệu đầu vào. Quá trình phân tích dữ liệu và lựa chọn tham số khiến cho thuật toán có độ phức tạp và thời gian tính toán giảm đi rất nhiều do việc không phải sử dụng các thuật toán tính distance với cả 42 tham số của dữ liệu đầu vào. Kết quả thu được cũng cho ta thấy rằng với k có giá trị nhỏ thì thuật toán đạt được độ chính xác cao hơn, việc này rất thuận lợi cho quá trình tính toán vì với k thấp thì độ phức tạp của thuật toán sẽ giảm đi nhiều.

Qua việc kiểm tra thuật toán ta thấy thuật toán TCM-KNN là một thuật toán có độ phức tạp chấp nhận được để áp dụng trong thực tế, cũng như với xác suất phát hiện tần công chính xác rất cao. Điều này cho thấy khả năng ứng dụng của thuật toán trong việc xây dựng các hệ thống phát hiện xâm nhập thực tế.

4. Kết luận và hướng phát triển trong tương lai.

Từ những kết quả đạt được cho ta thấy khả năng ứng dụng của thuật toán TCM-KNN trong thực tế. Để áp dụng thuật toán vào thực tế cần những nghiên cứu thêm trong tương lai về các phương pháp để giảm độ phức tạp tính toán của thuật toán cũng như tăng khả năng phát hiện và giảm xác suất báo sai. Trong thời gian tới tôi sẽ xây dựng sản phẩm dựa trên thuật toán này và thử nghiệm trong thực tế để kiểm tra hiệu năng và khả năng áp dụng của nó vào môi trường mạng.

Tài liệu tham khảo

- [1] An Effective TCM-KNN Scheme for High-Speed Network Anomaly Detection, Yang Li Chinese Academy of Sciences, Beijing China.
- [2] TCM-KNN Algorithm for Supervised Network Intrusion Detection, Yang Li , Bin-Xing Fang , Li Guo , and You Chen.
- [3] Proactive Detection of DDoS Attacks Utilizing k-NN Classifier in an Anti-DDos Framework, Hoai-Vu Nguyen and Yongsun Choi.
- [4] Mitigating Distributed Denial of Service Attacks Using a Proportional-Integral-Derivative Controller, Marcus Tylutki and Karl Levitt.
- [5]<http://kdd.ics.uci.edu/databases/kddcup99/task.html>
- [6] Anderson, James P., "Computer Security Threat Monitoring and Surveillance," Washington, PA, James P. Anderson Co., 1980.
- [7] Denning, Dorothy E., "An Intrusion Detection Model," Proceedings of the Seventh IEEE Symposium on Security and Privacy, May 1986, pages 119–131.
- [8] Vaccaro, H.S., and Liepins, G.E., "Detection of Anomalous Computer Session Activity," The 1989 IEEE Symposium on Security and Privacy, May, 1989.
- [9] Lunt, Teresa F., "Detecting Intruders in Computer Systems," 1993 Conference on Auditing and Computer Technology, SRI International.
- [10] Dowell, Cheri, and Ramstedt, Paul, "The ComputerWatch Data Reduction Tool," Proceedings of the 13th National Computer Security Conference, Washington, D.C., 1990.
- [11] Winkeler, J.R., "A UNIX Prototype for Intrusion and Anomaly Detection in Secure Networks," The Thirteenth National Computer Security Conference, Washington, DC., pages 115–124, 1990.

HỘI NGHỊ
SINH VIÊN NGHIÊN CỨU KHOA HỌC
LẦN THỨ XXVIII
NĂM HỌC 2010 – 2011

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

NHÀ XUẤT BẢN BÁCH KHOA – HÀ NỘI

Số 1 – Đại Cồ Việt – Quận Hai Bà Trưng – Hà Nội

Điện thoại: 04. 38684569; Fax: 04.38684570

Chịu trách nhiệm xuất bản:

Giám đốc – Tổng biên tập Phùng Lan Hương

Chịu trách nhiệm nội dung:

Viện CNTT&TT, Trường Đại học Bách Khoa Hà Nội

Biên tập:

Đỗ Bá Lâm, Trần Tuấn Vinh

Trình bày bìa:

Nguyễn Xuân Cương

In 80 cuốn khổ 21 x 29.7 cm tại xưởng in Nhà xuất bản Bách Khoa Hà Nội
Giấy xác nhận đăng ký kế hoạch xuất bản số:
In xong và nộp lưu chiểu quý II năm 2011