

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ



CÔNG TRÌNH DỰ THI
HỘI NGHỊ SINH VIÊN NGHIÊN CỨU KHOA HỌC

**MỘT PHƯƠNG PHÁP LAI TRÍCH XUẤT SỰ KIỆN
VÀ ÁP DỤNG VÀO HỆ THỐNG THEO DÕI TIN TỨC
TRỰC TUYẾN NewSOMoni**

Sinh viên thực hiện
Nguyễn Minh Hoàng
Nguyễn Sỹ Quân
Ngô Quang Hiếu

Cán bộ hướng dẫn
TS. Phan Xuân Hiếu
ThS. Trần Mai Vũ

Hà Nội, Ngày 21 tháng 3 năm 2012

Tóm tắt nội dung

Trích chọn thông tin luôn là vấn đề có vai trò cốt yếu khi xây dựng một hệ thống khai phá dữ liệu, đặc biệt trong các hệ thống theo dõi/giám sát thông tin, hệ thống tư vấn tin tức, hệ hỗ trợ ra quyết định. Một trong những bài toán cơ bản của trích chọn thông tin là trích xuất sự kiện trên dữ liệu lớn. Sự kiện được lấy ra đúng đắn từ kho dữ liệu lớn sẽ giúp các hệ thống khai phá dữ liệu dễ dàng hơn trong việc thực thi nhiệm vụ của mình. Nghiên cứu này sẽ tập trung xem xét một phương pháp trích xuất sự kiện hiệu quả dành cho tiếng Việt với lượng dữ liệu lớn và cách thức áp dụng vào hệ thống theo dõi tin tức trực tuyến cùng những đánh giá để cho thấy phương pháp đưa ra có khả quan. Nhóm tác giả hy vọng kết quả của nghiên cứu sẽ góp phần vào sự phát triển của các hệ thống xử lý tin tức dành cho tiếng Việt.

Mục lục

Tóm tắt nội dung	ii
Mục lục	iii
Danh sách hình vẽ	iv
Danh sách bảng	v
Danh sách ký hiệu và từ viết tắt	vi
Lời nói đầu	1
1 Giới thiệu bài toán trích xuất sự kiện	2
1.1 Động lực nghiên cứu	2
1.2 Vấn đề nghiên cứu	4
1.2.1 Bài toán	4
1.2.2 Câu hỏi nghiên cứu	6
1.3 Ý nghĩa	6
1.3.1 Ý nghĩa khoa học	6
1.3.2 Ý nghĩa thực tiễn	7
1.4 Thách thức	7
1.5 Nghiên cứu liên quan	8
1.5.1 Một số nghiên cứu liên quan ở nước ngoài	8
1.5.2 Một số nghiên cứu liên quan ở trong nước	11
1.6 Tóm lược chương	11

2	Mô hình hệ thống theo dõi tin tức trực tuyến NewSOMoni	12
2.1	Phương pháp đánh giá hiệu quả trích xuất sự kiện	12
2.2	Hệ thống theo dõi tin tức trực tuyến <i>NewSOMoni</i>	12
2.2.1	Kho dữ liệu	14
2.2.2	Thu thập dữ liệu	15
2.2.3	Phân loại sự kiện	16
2.2.4	Trích xuất sự kiện	16
2.2.5	Trực quan hóa dữ liệu	17
3	Thực nghiệm hệ thống	18
3.1	Yêu cầu phần cứng	18
A	Các nguồn cung cấp dữ liệu	20
	Tài liệu tham khảo	26

Danh sách hình vẽ

2.1	Mô hình hệ thống NewSOMoni	13
2.2	Khuôn dạng tin tức lấy qua kênh RSS	15

Danh sách bảng

2.1	Dữ liệu thô	16
-----	-----------------------	----

Bảng ký hiệu và từ viết tắt

Ký hiệu	Ý nghĩa
ACE	Automatic Content Extraction
DARPA	Defense Advanced Research Project Agency
MUC	Message Understanding Conferences
SIGIR	Special Interest Group on Information Retrieval
SIGKDD	International Conference on Knowledge Discovery and Data Mining
TDT	Topic Detection and Tracking

Lời nói đầu

Được cộng đồng nghiên cứu khoa học trên toàn thế giới quan tâm rất sớm, trích xuất sự kiện được xem là một bài toán quan trọng trong lĩnh vực trích chọn thông tin (Information Extraction). Từ năm 1987, trích xuất sự kiện đã trở thành đề tài chủ chốt tại hội nghị *Message Understanding Conferences* ngay lần tổ chức đầu tiên [RB96]. Từ đó đến nay, nhiều phương pháp trích xuất sự kiện đã được đưa ra và áp dụng trong các hệ thống thực tế như BioCaster (<http://born.nii.ac.jp/>), HealthMap (<http://healthmap.org>), EpiSpider (www.epispider.org/), Metro Monitor (www.metromonitor.com/), ...

Công trình nghiên cứu **Một phương pháp lai trích xuất sự kiện và áp dụng vào hệ thống theo dõi tin tức trực tuyến NewSOMonI** khảo sát một số phương pháp trích xuất sự kiện tiêu biểu có hiệu quả tốt, đang được sử dụng trong nhiều hệ thống theo dõi thông tin. Dựa trên cơ sở đó, chúng tôi nghiên cứu và đề xuất một phương pháp lai nhằm mục đích trích xuất sự kiện trên miền tin tức tiếng Việt và thử nghiệm trên hệ thống theo dõi tin tức trực tuyến NewSOMonI. Phương pháp được đề xuất là sự kết hợp của phương pháp học máy Maximum Entropy và phương pháp trích xuất dựa trên luật với những cải tiến khi áp dụng cho dữ liệu tiếng Việt. Qua tiến hành thực nghiệm, chúng tôi đã thu được kết quả tương đối tốt và ổn định. Điều này chứng tỏ tính đúng đắn của phương pháp đề xuất cũng như tính thực tiễn trong hệ thống theo dõi tin tức trực tuyến, góp phần đưa thông tin đến với người dùng chính xác, kịp thời.

Báo cáo bao gồm bốn chương được mô tả như dưới đây.

Chương 1. *Giới thiệu* khái quát chung về động lực thực hiện nghiên cứu, mô tả về bài toán trích xuất sự kiện và cũng nêu một số nghiên cứu liên quan ở trong và ngoài nước.

Chương 2. *Trích xuất sự kiện dựa trên luật kết hợp học máy và hệ thống theo dõi tin tức* trình bày phương pháp trích xuất sự kiện dựa trên luật kết hợp với phương pháp học máy Maximum Entropy—phương pháp chính trong mô hình giải quyết của nghiên cứu này. Đồng thời, mô hình hệ thống theo dõi tin tức cũng sẽ được nêu rõ và phân tích chi tiết.

Chương 3. *Thực nghiệm phương pháp trên hệ thống theo dõi tin tức* trình bày quá trình xây dựng hệ thống giám sát tin tức trên cơ sở áp dụng phương pháp đã đề xuất ở Chương 3. Kết quả thực nghiệm và đánh giá hiệu quả sẽ được mô tả kỹ lưỡng trong chương này.

Phần kết luận tổng kết, tóm lược nội dung của nghiên cứu và hướng phát triển tiếp theo.

Giới thiệu bài toán trích xuất sự kiện

1.1 Động lực nghiên cứu

Thế giới đang thay đổi rất nhanh với sự tham gia của các phương tiện truyền thông xã hội. Mọi thông tin đều có thể đến với người dùng theo nhiều nguồn khác nhau. Tuy nhiên, sử dụng phương tiện truyền thông xã hội riêng lẻ khó có thể cập nhật được kịp thời và chính xác thông tin. Để đáp ứng nhu cầu đó, những hệ thống tổng hợp tin tức lần lượt ra đời giúp cho con người có thể dễ dàng nắm bắt thông tin. Khởi đầu bởi <tên hệ thống đầu tiên trên thế giới>, tiếp sau đó là <tên hệ thống khác 1>, <tên hệ thống khác 2>, ... Vào năm 2005, hệ thống tổng hợp tin tức tự động đầu tiên của Việt Nam ra đời dựa trên thành tựu nghiên cứu *Hệ thống thu thập và tách thông tin ICPS* của hai tác giả Nguyễn Thành Long và Nguyễn Phú Bình đạt giải nhì cuộc thi Trí Tuệ Việt Nam 2002. *Hệ thống xử lý tiếng Việt tự động ePi* được người dùng biết đến với tên BÁO MỚI ¹ và nhanh chóng trở thành trang tin tức tổng hợp được nhiều người sử dụng bởi tính tiện lợi và cập nhật. Mặc dù có những ưu điểm như vậy, một hệ thống tổng hợp tin tức vẫn có những yếu điểm chưa thể khắc phục. Thứ nhất, thông tin được thu thập từ những nguồn tin định trước dựa trên giao diện cập nhật của nguồn tin, chưa phân tích sâu về ý nghĩa và tính chất của sự kiện chứa đựng trong thông tin. Thứ hai, tin tức không được trực quan hóa theo xu hướng quan tâm của người dùng. Thông thường, độ ưu tiên quan tâm của người dùng là: thời gian (when) > địa điểm (where)

¹www.baomoi.com

> thông tin gì(what). Hơn nữa, hệ thống tổng hợp tin tức xem xét tất cả các tin từ nguồn tin, sau đó phân lớp vào một lớp đã định nghĩa trước. Bởi tính phong phú của dạng thông tin, tính chính xác của quá trình phân lớp là một câu hỏi lớn chưa có lời giải đáp thỏa đáng!

Giải quyết nhược điểm của hệ thống tổng hợp tin tức tự động cần có một phương pháp trích xuất sự kiện phù hợp với tiếng Việt và hoạt động ổn định. Từ rất sớm, trích xuất sự kiện đã được cộng đồng khoa học máy tính đầu tư công sức nghiên cứu. Tiêu biểu có thể kể đến hội nghị **Message Understanding Conferences (MUC)** ¹ tổ chức lần đầu tiên năm 1987 dưới sự hỗ trợ của DARPA (Quỹ nghiên cứu bộ quốc phòng Hoa Kỳ). Một trong những đóng góp quan trọng của hội nghị MUC là đưa ra phương pháp trích xuất sự kiện theo khung mẫu (*scenario template*) với mục đích chính là lấy ra được sự kiện cùng các thông tin liên quan: tổ chức, đối tượng tham gia (người, sự vật, sự việc). Độ chính xác và độ hồi tưởng của các nghiên cứu tham dự MUC nằm trong khoảng 50% tới 60 %. Ngoài ra, chương trình nâng cao hiệu quả trích xuất sự kiện **Automatic Content Extraction (ACE)** ² của Đại học Pennsylvania (Hoa Kỳ) cũng là một chương trình nổi tiếng, thu hút được nhiều nhóm nghiên cứu về trích xuất sự kiện tham gia và có những kết quả rất tích cực. Tuy nhiên, trích xuất sự kiện là một vấn đề mang đặc trưng ngôn ngữ học. Ngôn ngữ ảnh hưởng rất lớn tới hiệu quả của một phương pháp trích xuất. Theo tìm hiểu của chúng tôi, trích xuất sự kiện trên dữ liệu tiếng Việt chưa có nhiều nghiên cứu. Bởi vậy, phương pháp trích xuất sự kiện dành cho tiếng Việt vẫn còn hạn chế cả về chất lượng lẫn số lượng.

Một yếu tố khác đưa chúng tôi đến với đề tài nghiên cứu này là sự thú vị trong xử lý dữ liệu lớn. Theo xu hướng phát triển Công Nghệ Thông Tin hiện đại, thi hành hệ thống với dữ liệu lớn là tất yếu. Các công ty hàng đầu thế giới về Công Nghệ như Microsoft ³, Google ⁴, Oracle ⁵, Facebook ⁶ đều có những chiến lược phát triển lâu dài về xử lý dữ liệu lớn. Cùng với đó, những trường đại học hàng đầu thế giới về khoa học máy tính đều đưa vào trường trình đào tạo của mình khoa học về xử lý dữ liệu lớn như Đại

¹http://www-nlpir.nist.gov/related_projects/muc

²<http://projects.ldc.upenn.edu/ace>

³www.microsoft.com

⁴www.google.com

⁵www.oracle.com

⁶www.facebook.com

học Princeton ¹ (Hoa Kỳ) , Đại học Stanford ² (Hoa Kỳ) , Đại học Carnegie Mellon ³ (Hoa Kỳ) hay Đại học tổng hợp Zurich ⁴ (Thụy Sĩ). Sự hỗ trợ tuyệt vời về dữ liệu và kỹ thuật từ phía ThS. Trần Mai Vũ đã giúp chúng tôi có thêm động lực và quyết tâm hoàn thành đề tài.

1.2 Vấn đề nghiên cứu

1.2.1 Bài toán

Những vấn đề phân tích ở phần 1.1 đã đưa nhóm nghiên cứu hướng tới ý tưởng đưa ra phương pháp trích xuất sự kiện phù hợp khi xử lý với dữ liệu tiếng Việt và xây dựng nên một hệ thống theo dõi tin tức trực tuyến mà trong đó trích xuất sự kiện là yếu tố trung tâm. Nghiên cứu đóng góp ở cả hai nội dung: khoa học và ứng dụng. Ý nghĩa của việc giải quyết vấn đề này được trình bày chi tiết ở mục 1.3.

Đầu vào của bài toán là một bản ghi tin tức về một trong ba lĩnh vực: tai nạn giao thông, hình sự, cháy nổ. Mỗi bản ghi bao gồm các thông tin: tiêu đề, tóm tắt nội dung, toàn văn tin tức. Gần 4 triệu ⁵ tin tức thu thập từ trang các nguồn cung cấp tin tức (phụ lục A) thông qua trang tổng hợp tin tức BÁO MỚI ⁶ là lượng dữ liệu mà hệ thống sẽ sử dụng.

Kết quả mong muốn của bài toán là có hay không có sự kiện trong bản ghi tin tức. Nếu có thì phải đưa ra được các thông tin liên quan tới sự kiện gồm có: tên sự kiện, thời gian, địa điểm, người, sự vật, sự việc. Sự kiện thu được cũng phải được trực quan hóa trên hệ thống theo dõi tin tức trực tuyến.

Vậy, sự kiện là gì? Theo Allan, tin tức được cho là phản ánh một sự kiện nếu nó có đủ ba yếu tố: chủ thể, thời gian, địa điểm [JRV98]. Chủ thể có thể là con người, sự vật hoặc sự việc. Cũng theo công bố này, để định nghĩa rõ ràng thế nào là sự kiện rất khó

¹<http://www.cs.princeton.edu/courses/archive/spr02/cs493>

²<http://www.stanford.edu/class/cs246>

³<http://www.cs.cmu.edu/neill/courses/90866.html>

⁴<http://las.ethz.ch/courses/datamining-s12>

⁵3.842.137 bài báo tin tức tổng hợp được thu thập trong 1 tháng, từ 01/12/2011 đến 01/01/2012 sử dụng bộ CRAWLER của tác giả Trần Mai Vũ

⁶www.baomoi.com

bởi tính nhập nhằng liên quan tới các yếu tố ngữ cảnh, ngôn ngữ, văn hóa. Ví dụ, *Chiều ngày 5/3/2012, tai nạn giao thông tại ngã tư Khuất Duy Tiến làm 2 người tử vong* là một sự kiện nói về tai nạn giao thông. Nhưng *Theo báo cáo của cảnh sát giao thông Hà Nội chiều nay, số người chết vì tai nạn giao thông giảm 30% so với cùng kỳ năm ngoái* lại không phải là một sự kiện dù có đủ 3 yếu tố kể trên. Trong phạm vi giải quyết bài toán trích xuất sự kiện, việc định nghĩa rõ ràng sự kiện mà nghiên cứu quan tâm luôn là yêu cầu trước tiên. Ban đầu hội nghị MUC chỉ quan tâm các sự kiện về hoạt động quân sự. Sau đó, tới lần tổ chức thứ 3 mở rộng thêm các sự kiện về khủng bố, đầu tư mạo hiểm, tai nạn máy bay, ... Các thuộc tính cần phải có của một sự kiện mà MUC yêu cầu gồm có: tác nhân, thời gian, địa điểm và các tác động của nó. Ở chương trình ACE, dạng sự kiện và các thuộc tính về sự kiện được quy định chặt chẽ hơn với tám dạng sau: LIFE (sự sống–chết), MOVEMENT (sự di chuyển), TRANSACTION (giao dịch), BUSINESS (kinh tế), CONFLICT (xung đột), CONTACT (giao thiệp, gặp gỡ), PERSONNEL (nhận–đuổi việc), JUSTICE (pháp lý). Mỗi dạng sự kiện lại có phân biệt từng dạng con. Ví dụ như LIFE có các dạng sự kiện con BE-BORN (chào đời), INJURE (bị thương), DIE (chết) hay PERSONNEL có START-POSITION (vị trí khi nhận việc), END-POSITION (vị trí trước khi bị thôi việc), NOMINATE (bổ nhiệm), ELECT (bầu chọn), ...

Hầu hết những nghiên cứu được trích dẫn trong báo cáo này đều chỉ tập trung vào một lĩnh vực cụ thể. [MM09], [YKW09] khai thác các sự kiện trên trang cá nhân. [CVJ09], [CHR04] tập trung vào sự kiện y sinh học. [HJM08], [JHP07] thực hiện trích xuất sự kiện thảm họa, mối nguy hiểm đe dọa. Ngoài ra, sự kiện về giải thưởng Nobel [FHH06], sự kiện về chứng khoán [FHD02], sự kiện về đầu tư tài chính [CM00] hay các sự kiện về chính trị [FK08], [CM00] cũng được quan tâm. Nghiên cứu này thực hiện trích xuất sự kiện từ các bản tin thông báo hằng ngày cho các loại sự kiện nói về tai nạn giao thông, các vi phạm hình sự, các vụ cháy nổ. Một cách tường minh, sự kiện được định nghĩa rằng phải có đủ ba thuộc tính: chủ thể, thời gian và địa điểm và bắt buộc thuộc ba dạng: TAI NẠN GIAO THÔNG, HÌNH SỰ, CHÁY NỔ.

Thế nào là trích xuất sự kiện? Trước hết, trích xuất sự kiện là một lĩnh vực con thuộc trích chọn thông tin (Information Extraction). Tự động nhận biết và tách được thông tin về sự kiện trong các tài liệu không có cấu trúc là định nghĩa tổng quát nhất về trích xuất sự kiện. Chi tiết hơn, trích xuất sự kiện tập trung nhận dạng sự kiện

thuộc một miền lĩnh vực cụ thể biết trước, đồng thời đưa ra được tập các tham số—là các thông tin xung quanh sự kiện đó, bao gồm: tác nhân, thời gian, địa điểm, ... Trong [RG10], Grishman cho rằng trích xuất sự kiện là một bài toán khó, bởi gặp nhiều vấn đề về xử lý ngôn ngữ tự nhiên cũng như khảo sát dữ liệu rất mất thời gian.

1.2.2 Câu hỏi nghiên cứu

Nghiên cứu sẽ trả lời ba câu hỏi.

Thứ nhất thế nào là trích xuất sự kiện tin tức và những phương pháp thường được sử dụng để làm điều đó?

Thứ hai tồn tại những khó khăn nào khi áp dụng những phương pháp từ câu hỏi trên vào dữ liệu tiếng Việt và cách giải quyết những khó khăn này?

Và cuối cùng một hệ thống theo dõi tin tức có khả thi không?

1.3 Ý nghĩa

1.3.1 Ý nghĩa khoa học

Về mặt khoa học, chúng tôi đề xuất phương pháp trích xuất sự kiện dựa trên luật ngữ nghĩa kết hợp với học máy để thu được sự kiện xảy ra hằng ngày thông qua dữ liệu tin tức tiếng Việt thu thập từ một số nguồn thông tin tin cậy dưới sự cho phép của Bộ Thông Tin và Truyền Thông ¹(xem danh sách trang báo điện tử cung cấp tin tức tại phụ lục A, trang 20). Luật ngữ nghĩa và học máy Maximum Entropy đều là những phương pháp đã được sử dụng trong các công bố quốc tế như [CVJ09], [RDA05], [MD04]. Mỗi phương pháp đều có những ưu, nhược điểm riêng. Để nâng cao hiệu quả trích xuất và rút ngắn thời gian thực hiện, kết hợp hai phương pháp trên là cách tiếp cận hợp lý. Tuy nhiên trên thế giới chưa có nghiên cứu nào đi theo hướng tiếp cận này. Trong bối cảnh vấn đề trích xuất sự kiện ở trong nước chưa có nhiều nghiên cứu, công trình của chúng tôi sẽ góp phần thôi thúc đề tài thú vị này được quan tâm nhiều hơn bởi lẽ đây là vấn đề tương đối mới mẻ, có khả năng ứng dụng thực tiễn cao và còn rất nhiều lĩnh vực cần quan tâm. Một số ví dụ như sự kiện Y-SINH, sự kiện KINH TẾ-ĐẦU TƯ, sự kiện CHÍNH TRI.

¹<http://mic.gov.vn/vbqppl/Lists/Luat-cong-nghe-thong-tin>

1.3.2 Ý nghĩa thực tiễn

Xét tới phương diện ứng dụng, chúng tôi tiến hành xây dựng một hệ thống theo dõi thông tin trực tuyến. Như đã nói ở mục 1.1, một hệ thống tổng hợp tin tức tự động chưa đủ thông minh để đáp ứng nhu cầu ngày càng cao của người dùng. Bởi thế, trong nghiên cứu này chúng tôi muốn xây dựng một hệ thống theo dõi, giám sát thông tin sự kiện. Bởi quy mô của một công trình sinh viên nghiên cứu khoa học, nhóm chúng tôi tập trung vào ba loại sự kiện thường xảy ra hằng ngày: tai nạn giao thông, hình sự và cháy nổ. Một cách rõ ràng nhất, sự kiện thuộc ba dạng trên sẽ được trích xuất theo các thông tin: tên sự kiện, thời gian/địa điểm diễn ra sự kiện, các nhân tố tham gia sự kiện. Sau đó, sự kiện được trực quan hóa trên bản đồ giúp cho người sử dụng dễ dàng theo dõi. Theo khảo sát của nhóm nghiên cứu, một hệ thống như đã mô tả chưa xuất hiện ở Việt Nam. Đề tài nghiên cứu đóng góp vào việc phổ biến hình thức nắm bắt tin tức mới dễ dùng và trực quan hơn so với các hệ thống cung cấp tin tức truyền thống.

1.4 Thách thức

Mặc dù được các nhà khoa học quan tâm nghiên cứu từ rất sớm, trích xuất sự kiện vẫn còn những khó khăn cần phải vượt qua.

Trích xuất sự kiện liên quan mật thiết tới các nghiên cứu về ngôn ngữ học. Lĩnh vực xử lý ngôn ngữ tự nhiên nói chung và xử lý tiếng Việt nói riêng tương đối rộng, tồn tại nhiều bài toán chưa được giải quyết triệt để mà trong đó có xử lý nhập nhằng ngữ nghĩa (Word Sense Disambiguation), bài toán đồng tham chiếu (Co-references) hay việc nhận dạng tính đa hình cấu trúc ngữ pháp trong tiêu đề tin tức (Syntactically Ambiguous Headlines). Ba bài toán trên là những khó khăn cơ bản nhất mà chúng tôi phải giải quyết để đưa ra được phương pháp trích xuất sự kiện phù hợp.

Tính tới thời điểm thực hiện công trình, Việt Nam chưa có nghiên cứu nổi bật về trích xuất sự kiện. Bởi vậy, nhóm nghiên cứu không được kế thừa những công trình, những kinh nghiệm khi thực hiện với dữ liệu tiếng Việt. Nhóm cần nhiều thời gian hơn để thử nghiệm và đánh giá phương pháp nào là tốt, phù hợp với mục tiêu đề ra.

Ngoài ra, khó khăn trong xử lý dữ liệu lớn cũng là một thách thức mà nhóm nghiên cứu phải đối mặt. Để có thể trích chọn được sự kiện từ tập dữ liệu lớn cần phải tối ưu thuật toán đảm bảo rằng hệ thống có thể hoạt động tốt trong điều kiện tài nguyên cho phép.

1.5 Nghiên cứu liên quan

1.5.1 Một số nghiên cứu liên quan ở nước ngoài

Kể từ hội nghị MUC lần đầu tiên (1987) cho tới nay, hàng ngàn nghiên cứu về trích xuất sự kiện đã được công bố trong những hội nghị, chương trình có uy tín cao như MUC, SIGKDD¹, ACM SIGIR², TDT³, ACE. Theo Hogenboom, F và các cộng sự, tựu chung lại các công bố này có thể phân loại theo ba hướng tiếp cận chính: phân tích ngữ nghĩa (còn gọi là hướng theo nội dung), học máy-thống kê (hướng theo dữ liệu) và cuối cùng là kết hợp hai cách tiếp cận trên [FFU11].

Giai đoạn cuối thập niên tám mươi, đầu thập niên chín mươi, sự kiện được trích xuất chủ yếu dựa trên các mẫu được tạo sẵn (*scenario template*) [BS92]. Mẫu là các bản ghi còn thiếu thông tin sự kiện. Thông tin về sự kiện còn thiếu này sẽ được bổ sung từ dữ liệu căn cứ vào những thông tin đã định nghĩa trên mẫu. Một cách thuần túy thì đây là bài toán tìm kiếm các từ được định nghĩa trước rồi lấy thông tin đi kèm với chúng để điền vào mẫu. Độ chính xác của phương pháp này ở mức trung bình nằm trong khoảng 50%–60% [MW11]. Cách giải quyết bài toán hết sức đơn giản mà về sau, trong các chương trình nghiên cứu TDT hay ACE vẫn còn sử dụng nhưng với những định nghĩa mẫu tổng quát và trên nhiều miền lĩnh vực khác nhau. Hơn nữa, đây cũng là sự khởi đầu của các phương pháp đi theo hướng tiếp cận đầu tiên kể ở trên: sử dụng luật phân tích ngữ nghĩa.

Trong nghiên cứu của Nishihara và cộng sự, ba thông tin: địa điểm, đối tượng, hành vi của sự kiện được lấy ra từ trang cá nhân⁴ [YKW09] sử dụng các *luật lexico-syntactic*⁵ để tìm kiếm các câu chứa sự kiện trong từng bài viết⁶. Cùng với cách tiếp cận này, Aone.C và Ramos.M đã trích chọn các sự kiện về tài chính và chính trị. Hai tác giả tập trung đưa ra các luật biểu diễn quan hệ giữa sự kiện với các thông tin xung quanh nhằm mục đích khai thác tối đa thuộc tính của sự kiện, và giữa các sự kiện để lấy được tập các sự kiện liên quan tới nhau [CM00]. Nghiên cứu của Xu và cộng sự cũng sử dụng các *luật lexico-syntactic* trên dữ liệu bản tin về sự kiện giải thưởng Nobel. Nhưng thay

¹International Conference on **K**nowledge **D**iscovery and **D**ata Mining

²**S**pecial **I**nterest **G**roup on **I**nformation **R**etrieval

³**T**opic **D**etection and **T**racking

⁴*blog*

⁵*Luật lexico-syntactic* là sự kết hợp giữa biểu thức chính quy với từ vựng thuộc miền lĩnh vực và các quy tắc ngữ pháp của ngôn ngữ để sinh luật

⁶*entry*

vì các luật được áp dụng ngay trên dữ liệu, một tập luật được tạo ra sau đó sử dụng học máy không giám sát để huấn luyện tập luật này trên tập các bản tin đã được gán nhãn. Sau đó mô hình học sẽ được áp dụng với các bản tin còn lại [FHH06].

Một điểm yếu của *luật lexico-syntactic* là không thể phủ hết được trạng thái quan hệ giữa các sự kiện, có nghĩa là không thể nhận biết hai sự kiện có trùng nhau hay không. Do đó, giám sát quá trình tiến triển của một sự kiện là tương đối khó khi sử dụng cách tiếp cận này. Nhằm khắc phục điều này, *luật lexico-semantic*¹ được đề xuất. Nghiên cứu của Li và đồng nghiệp chú trọng đưa ra các luật lấy sự kiện về giá cổ phiếu qua các bản tin chứng khoán [FHD02]. Một tập dữ liệu bản tin chứng khoán được gán nhãn bởi từ điển ngữ nghĩa chứa tên công ty, tập đoàn mà phần nhiều là tên vị trí địa lý. Ngoài ra, lĩnh vực y sinh cũng được nhiều nhà nghiên cứu quan tâm. Nghiên cứu của nhóm do Cohen chủ trì tập trung xây dựng bộ trích xuất nội dung có nhiệm vụ trích chọn sự kiện y tế bằng từ điển thuật ngữ y sinh và quan tâm tới nghĩa của các cụm từ [CVJ09]. Cùng sử dụng cách làm này, Vargas-Vera và Celjaska đã phát triển hệ nhận dạng sự kiện trên các bài báo của Knowledge Media Institute² [MD04].

Những phương pháp đã trình bày ở trên chủ yếu xây dựng luật dựa trên tri thức về ngôn ngữ. Chúng có một số lợi điểm có thể kể tới. Thứ nhất, thông tin muốn có được hoàn toàn có thể theo ý định của người nghiên cứu, và trên bất cứ lĩnh vực cụ thể nào. Thứ hai, không cần phải xem xét một tập dữ liệu quá lớn. Một luật chủ yếu dựa trên tri thức ngôn ngữ và sự khảo sát của người thực hiện. Tuy nhiên, các phương pháp này cũng có những điểm yếu cần phải khắc phục. Bởi luật được sinh ra cho từng dạng sự kiện cụ thể nên chúng ta không thể sử dụng lại luật cho trường hợp khác. Nếu trích xuất sự kiện trong lĩnh vực rộng thì áp dụng luật không thể bao quát toàn bộ không gian dữ liệu. Hơn nữa, việc khảo sát và sinh luật bằng tay là một công việc rất mất thời gian và tốn kém. Cách tiếp cận hướng dữ liệu sẽ cho chúng ta một cái nhìn cụ thể hơn khi giải quyết những vấn đề tồn đọng của phương pháp tiếp cận hướng nội dung. Đối với cách tiếp cận hướng dữ liệu, các nhà nghiên cứu thường sử dụng các phương pháp học máy: học giám sát (SVM), học bán giám sát, học không giám sát (phân cụm) hay là các phương pháp thống kê như trọng số IF-IDF. Năm 2009, Okamoto cùng cộng sự xây dựng một hệ thống phát hiện và trích xuất sự kiện trong một phạm vi địa lý

¹*luật lexico-semantic* là sự kết hợp giữa biểu thức chính quy, tập từ vựng thuộc miền lĩnh vực và vai trò ngữ nghĩa của từ vựng trong ngôn ngữ để sinh luật

²<http://kmi.open.ac.uk/>

sử dụng kỹ thuật phân cụm phân cấp với dữ liệu là các bài viết trên trang cá nhân ¹ [MM09]. Phân cụm cũng là kỹ thuật được sử dụng nhiều trong các nghiên cứu khác như công trình của nhóm Liu [MYL08], nhóm Tanev [HJM08]. Ở công trình thứ nhất, một cụm sự kiện liên quan tới tin tức hằng ngày hình thành sẽ được sắp xếp theo thứ tự nhờ sử dụng đồ thị vô hướng phân đôi. Công trình thứ hai lại sử dụng một tập dữ liệu đã được gán nhãn tự động để phân cụm sự kiện nói về mối nguy hiểm, thảm họa. Phương pháp máy vector hỗ trợ ² được Lei và cộng sự thử nghiệm trên hệ thống phát hiện sự kiện tin tức của họ [LWZ05]. Brants và cộng sự cải tiến cách tính trọng số TF-IDF để nhận dạng một sự kiện thông qua một sự kiện khác đã biết. Độ tương đồng giữa hai sự kiện quyết định bởi hai yếu tố: độ tương đồng giữa từ khóa của hai bản tin, độ tương đồng giữa hai nguồn cung cấp bản tin [TFA03]. Tiếp cận hướng dữ liệu vẫn còn tồn tại một số nhược điểm: không quan tâm đến ngữ nghĩa, và lượng dữ liệu phải khá lớn. Hướng tiếp cận này không thể nào trích xuất được quan hệ giữa các sự kiện cũng như quan hệ giữa các thuộc tính của sự kiện. Bởi sử dụng chủ yếu các phương pháp học máy, thống kê nên dữ liệu cần thiết là khá lớn. Xây dựng được kho dữ liệu đủ lớn cũng là một yêu cầu không đơn giản.

Như những dẫn chứng ở trên, cả hai cách tiếp cận hướng nội dung và hướng dữ liệu đều có những điểm mạnh và điểm yếu riêng. Một cách tự nhiên, kết hợp hai cách tiếp cận này với nhau sẽ giúp chúng hỗ trợ, bổ xung cho nhau. Nghiên cứu của Jungermann và Morik kết hợp *luật lexico-syntactic* với trường điều kiện ngẫu nhiên ³ để trích xuất sự kiện từ văn bản các phiên họp toàn thể của nghị viện Đức [FK08]. Trong [JHP07], các luật được học giám sát kết hợp với phân cụm nhằm trích xuất sự kiện có tính cảnh báo. Chun cùng cộng sự trích xuất sự kiện y học qua bằng hai phương pháp: sử dụng *luật lexico-syntactic* và thống kê từ khóa đồng xuất hiện [CHR04]. Tất cả những phương pháp trên đều cho độ chính xác và độ hồi tưởng cao. Tuy giúp hai hướng tiếp cận trên phụ trợ nhau, nhưng việc kết hợp chúng làm cho hệ thống trích xuất sự kiện trở nên phức tạp và khó xây dựng hơn.

¹blog

²Support Vector Machine

³Conditional Random Fields

1.5.2 Một số nghiên cứu liên quan ở trong nước

1.6 Tóm lược chương

2

Mô hình hệ thống theo dõi tin tức trực tuyến NewSOMoni

Chương này chúng tôi sẽ trình bày về hệ thống theo dõi tin tức trực tuyến có tên NewSOMoni^{1 2} cùng phương pháp lai giữa luật và học máy Maximum Entropy trong pha trích xuất sự kiện của hệ thống vừa được nhắc tới.

2.1 Phương pháp đánh giá hiệu quả trích xuất sự kiện

Mọi phương pháp trích xuất sự kiện được đề xuất đều phải đánh giá hiệu quả trích xuất. Tuy vậy, để đánh giá hiệu quả trích xuất một cách tự động là công việc khó. Điều đó có nghĩa rằng luôn cần có sự can thiệp của con người trong quá trình đánh giá. Hội nghị MUC ...

2.2 Hệ thống theo dõi tin tức trực tuyến NewSOMoni

Công trình nghiên cứu này chúng tôi đã xây dựng một hệ thống theo dõi tin tức trực tuyến. Nhiệm vụ chính của hệ thống là quan sát tức mới được đưa lên các nguồn cung cấp tin tức (phụ lục A, trang 20), phân loại và nhận dạng sự kiện thuộc ba lĩnh vực: TAI NẠN GIAO THÔNG, HÌNH SỰ, CHÁY NỔ. Cuối cùng là trực quan hóa trên bản đồ

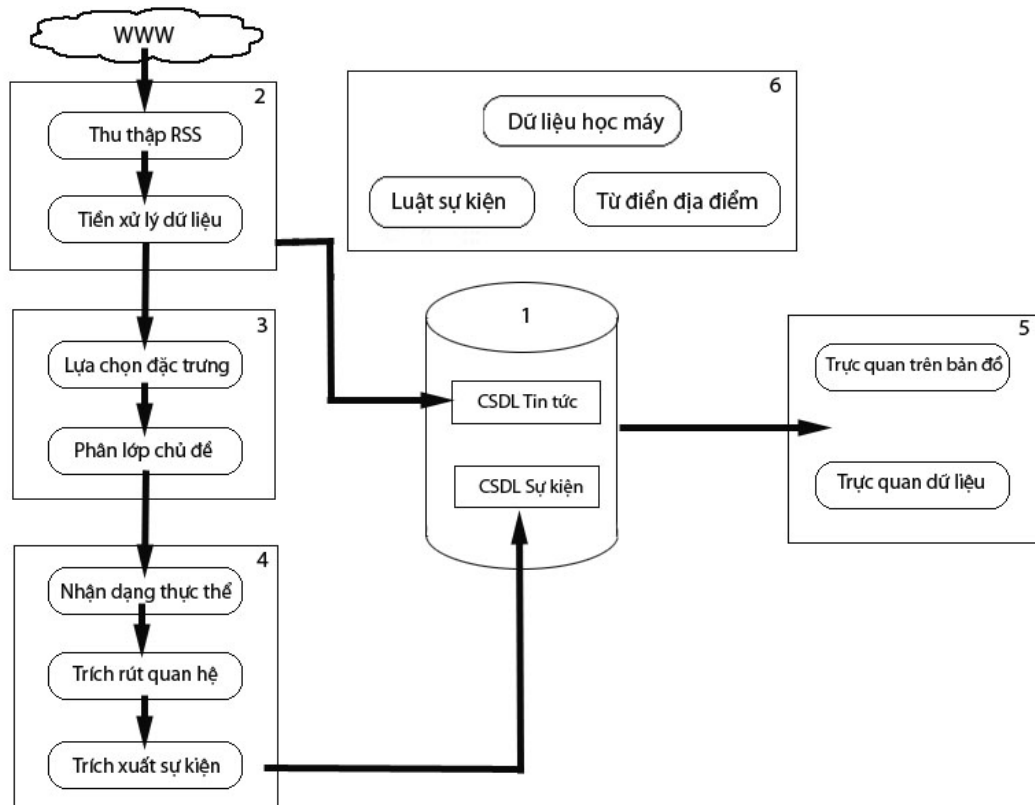
¹[nju: 'sʌm mʌni]

²New $\underbrace{\text{Smartness}}_S$ $\underbrace{\text{Orientation}}_O$ $\underbrace{\text{Magnitude}}_M$ *oni* viết tắt của News Online Monitoring

2.2 Hệ thống theo dõi tin tức trực tuyến NewsOMoni

cho người dùng dễ dàng theo dõi, cập nhật.

Mô hình của hệ thống được thể hiện rõ ở hình 2.1. Hệ thống NewsOMoni có năm phần chính:



Hình 2.1: Mô hình hệ thống NewsOMoni

- **Kho dữ liệu** cơ sở dữ liệu không ràng buộc, hướng tài liệu (MongoDB), lưu trữ lượng lớn dữ liệu tin tức
- **Thu thập dữ liệu** thu thập dữ liệu tự động và tiến xử lý dữ liệu
- **Phân lớp chủ đề** đưa tin tức thu thập được vào hai dạng: SỰ KIỆN, KHÔNG PHẢI SỰ KIỆN
- **Trích xuất sự kiện** thực hiện các bước cần thiết để trích xuất sự kiện
- **Trực quan hóa dữ liệu** có nhiệm vụ tương tác với cơ sở dữ liệu để hiển thị thông tin cho người dùng

Mỗi thành phần của hệ thống sẽ được diễn giải chi tiết dưới đây.

2.2.1 Kho dữ liệu

Hệ thống phải xử lý dữ liệu lớn nên cần lựa chọn kiểu lưu trữ cũng như thiết kế cơ sở dữ liệu phù hợp. Riêng chỉ lượng dữ liệu thu thập ngoại tuyến gồm 3.842.137 tin tức điện tử phục vụ cho quá trình sinh luật và học mô hình phân lớp ban đầu đã có dung lượng gần 60GB. Hơn nữa, hệ thống chạy trực tuyến mỗi ngày nhận khoảng 1500 bài báo điện tử. Do vậy, cần thiết một hệ cơ sở dữ liệu có khả năng truy xuất dữ liệu nhanh cũng như có khả năng mở rộng về sau. Qua khảo sát, nhóm nghiên cứu nhận thấy các hệ cơ sở dữ liệu không quan hệ (NoSQL) phù hợp với tiêu chí đề ra. NoSQL không tồn tại các ràng buộc giữa các bảng lưu trữ. Điều này giúp cho tốc độ truy vấn tốt hơn hẳn so với các hệ cơ sở dữ liệu quan hệ truyền thống. Thứ nữa, NoSQL là hệ cơ sở dữ liệu phân tán, có thể mở rộng theo chiều ngang, nghĩa là các yếu tố phần cứng như bộ nhớ ngoài (HDD), bộ nhớ trong (RAM) có thể tăng thêm bằng cách kết hợp nhiều thành phần phần cứng nhỏ hơn với nhau. Với khoảng 15 năm phát triển¹, đã có rất nhiều hệ cơ sở dữ liệu thuộc họ NoSQL ra đời. Ví dụ, lưu trữ dạng tài liệu có MongoDB, CouchDB, BaseX; lưu trữ dạng đồ thị có Neo4j, OrientDB, Sones GraphDB; lưu trữ dạng key-value có BigTable, Cassandra, Redis². BigTable được Google³ sử dụng, Cassandra là sản phẩm của mạng xã hội Facebook⁴, mạng xã hội của giới lập trình viên Github⁵ sử dụng Redis, MongoDB được dùng bởi mạng xã hội Foursquare⁶ là những cái tên nổi bật hơn cả. Trong nghiên cứu này, chúng tôi lựa chọn hệ cơ sở dữ liệu MongoDB làm thành phần lưu trữ dữ liệu bởi khả năng truy vấn dữ liệu nhanh, tự động dàn trải dữ liệu và dễ dàng phân tán.

Kho dữ liệu gồm hai phần: cơ sở dữ liệu tin tức, cơ sở dữ liệu sự kiện.

Cơ sở dữ liệu tin tức

Đầu vào: tin tức từ bộ thu thập dữ liệu sau khi đã tiền xử lý dữ liệu (Pha 2).

¹NoSQL lần đầu tiên được giới thiệu vào năm 1998 bởi Carlo Strozzi

²<http://en.wikipedia.org/wiki/NoSQL>

³www.google.com

⁴<http://www.facebook.com/>

⁵<https://github.com/>

⁶<http://foursquare.com/>

Cơ sở dữ liệu sự kiện

Đầu vào: sự kiện và các thông tin về sự kiện đó từ pha trích xuất sự kiện (Pha 4).

2.2.2 Thu thập dữ liệu

Hiện nay, hầu hết các trang tin tức đều cung cấp cơ chế chia sẻ tin RSS. Tận dụng tính năng này, một bộ thu thập dữ liệu qua RSS được xây dựng.

Thu thập tin tức RSS

Tin tức từ các kênh RSS của các trang tin tức điện tử theo dạng XML như hình 2.2 được tự động thu thập qua bộ RSSFeeder.

```
<?xml version="1.0" encoding="utf-8"?>
<rss xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema" version="2.0">
  <channel>
    <item>
      <title>
        <![CDATA[ Xe tải tông nát taxi, 3 người chết ]]>
      </title>
      <description>
        <![CDATA[ Một vụ tai nạn kinh hoàng đã xảy ra trưa ngày
          20/3 tại Hà Tĩnh, đã có 3 người chết và 1 người
          bị thương.]]>
      </description>
      <link>
        http://vietnamnet.vn/vn/xa-hoi/64932/xe-tai-tong-nat-taxi--3-nguoi-chet.html
      </link>
      <pubDate>
        3/20/2012 2:27:26 PM
      </pubDate>
      <image>
        http://imgs.vietnamnet.vn/Images/2012/03/20/14/20120320142843\_tn.jpg
      </image>
    </item>
  </channel>
</rss>
```

Hình 2.2: Khuôn dạng tin tức lấy qua kênh RSS

Tiền xử lý dữ liệu

Sau khi bộ RSSFeeder lấy tin tức về, dữ liệu cần phải lọc ra những thông tin cần thiết. Hai lý do cần thiết để làm việc này. Một là giảm dung lượng dữ liệu lưu trữ trên hệ thống. Hai là giúp cho các bước xử lý sau dễ dàng hơn.

Đầu vào là các bản tin có định dạng như hình 2.2

2.2 Hệ thống theo dõi tin tức trực tuyến NewsOMoni

Bảng 2.1: Dữ liệu thô

Tên trường	Giá trị
Tiêu đề	Xe tải tông nát taxi, 3 người chết
Tóm tắt	Một vụ tai nạn kinh hoàng đã xảy ra trưa ngày 20/3 tại Hà Tĩnh, đã có 3 người chết và 1 người bị thương.
Đường dẫn	vietnamnet.vn/vn/xa-hoi/64932/xe-tai-tong-nat-taxi-3-nguoi-chet.html
Ngày đăng tin	3/20/2012 2:27:26 PM

Đầu ra là các thông tin bao gồm: tiêu đề, tóm tắt, đường dẫn tới bài báo và ngày đăng tin. Bảng 2.1 là một ví dụ cho nội dung tin tức thể hiện ở hình 2.2. Dữ liệu sau khi tiền xử lý được lưu trữ trong cơ sở dữ liệu tin tức.

2.2.3 Phân loại sự kiện

Pha này sẽ giải quyết vấn đề nhận dạng sự kiện. Tin tức thu được từ pha thu thập dữ liệu sẽ được quyết định có chứa sự kiện hay không. Qua khảo sát dữ liệu, chúng tôi nhận thấy hầu hết tiêu đề tin tức thể hiện rõ được nội dung tin tức có nói về sự kiện. Bởi vậy, bài toán đưa về phân lớp nhị phân mức câu. Hai bước cần phải làm trong pha này. Đầu tiên, tập đặc trưng được lựa chọn. Sau đó, các đặc trưng sẽ được trích chọn trên tập tin tức đã được thu thập trước (dữ liệu ngoại tuyến). Cuối cùng, mô hình phân lớp được học bằng phương pháp học máy Maximum Entropy.

Lựa chọn và trích chọn đặc trưng

Phân lớp chủ đề

2.2.4 Trích xuất sự kiện

Trích chọn thực thể

Trích xuất quan hệ thực thể

Trích xuất sự kiện

2.2.5 Trực quan hóa dữ liệu

3

Thực nghiệm hệ thống

3.1 Yêu cầu phần cứng

Tổng kết

Phụ lục A

Các nguồn cung cấp dữ liệu

24h.com.vn	http://www21.24h.com.vn
VietnamNet (2Sao)	http://2sao.vietnamnet.vn
aFamily	http://afamily.channelvn.net
Alobacsi.vn	http://alobacsi.vn
Báo An Ninh Thủ Đô (ANTĐ)	http://www.anninhthudo.vn
Báo An Ninh Thế Giới (ANTG)	http://antg.cand.com.vn
Báo Công An Nhân Dân (ANTGCT)	http://antg.cand.com.vn
Archi.vn (Archi)	http://archi.vn
ATPVietnam	http://atpvietnam.com
Autonet	http://www.autonet.com.vn
AutoPro	http://autopro.channelvn.net
Báo Biên phòng	http://www.bienphong.com.vn
Báo Bóng Đá	http://http://www.baobongda.com.vn
Báo Công Lý	http://congly.com.vn
Báo Công Thương	http://baocongthuong.com.vn
Báo Đất Việt	http://www.baodatviet.vn
Báo Đất Việt (Báo Đất Việt Khoa học)	http://khoahoc.baodatviet.vn
Báo Đất Việt (Báo Đất Việt Quốc phòng)	http://quocphong.baodatviet.vn
Báo Giáo dục Việt Nam	http://giaoduc.net.vn
Báo Giao Thông Vận Tải (Báo GTVT)	http://giaothongvantai.com.vn
Báo Khoa học Phát triển	http://khoahocphattrien.com.vn
Báo Người cao tuổi	http://nguoi.caotui.org.vn

Báo Nông nghiệp VN	http://nongnghiep.vn
Báo Phụ Nữ (Báo Phụ Nữ Online)	http://www.phunuonline.com.vn
Báo Thể giới Việt nam	http://www.tgvn.com.vn
Báo Tia sáng	http://www.tiasang.com.vn
Thông Tấn Xã VN (Báo Tin tức)	http://baotintuc.vn
Báo Thể thao Văn Hóa (Báo TTVH)	http://thethaovanhoa.vn
Báo Thể Thao VN (Báo TTVN)	http://www.thethaovietnam.com.vn
Báo Văn hóa	http://www.baovanhoa.vn
Báo Khoa học Đời sống (Bee.net.vn)	http://bee.net.vn/
Bóng Đá 24H	http://www.bongda24h.vn
Bóng đá số	http://www.bongdaso.com
YTT (Bongda.com.vn)	http://www.bongda.com.vn
CafeF	http://cafef.vn
Báo Công An Nhân Dân (CAND Portal)	http://www.cand.com.vn
Công An TP.HCM (CATPHCM)	http://www.congan.com.vn
CTTĐT Chính phủ (Chinhphu.vn)	http://baodientu.chinhphu.vn
Báo Công An Nhân Dân (CSTC)	http://cstc.cand.com.vn
Báo Đại đoàn kết (Đại Đoàn Kết)	http://baodaidoanket.net
Báo Dân Trí (Dân Trí)	http://www.dantri.com.vn
Báo Nông thôn ngày nay (Dân Việt)	http://www.danviet.vn
Báo Đầu Tư Chứng Khoán (Đầu tư CK)	http://www.tinnhanhchungkhoan.vn
Báo Điện tử Đảng cộng sản VN (ĐCSVN)	http://www.cpv.org.vn
Địa ốc Online	http://www.diaonline.vn
Báo Diễn Đàn Doanh Nghiệp (Diễn đàn Doanh nghiệp)	http://www.dddn.com.vn
Điện Tử Tiêu Dùng	http://dientutieudung.vn
Doanh nhân 360	http://doanhnhan360.com
Doanh nhân Sài Gòn	http://doanhnhansaigon.vn
Báo Đời sống Pháp luật (Đời sống Pháp luật)	http://www.doisongphapluat.com.vn
Dothi.net	http://dothi.net
Báo Doanh nhân Việt Nam toàn cầu (DVT.vn)	http://dvt.vn
VnExpress (eBank)	http://ebank.vnexpress.net
eFinance	http://www.taichinhdientu.vn
24H.COM.VN (Eva.vn)	http://www.eva.vn

VnExpress (eVăn)	http://evan.vnexpress.net
Gafin.vn	http://gafin.vn
Game4V	http://news.game4v.vn
GameK	http://gamek.channelvn.net
Gamethu.net	http://gamethu.net
Báo Gia đình Xã hội (Giadinh.net)	http://giadinh.net.vn
Báo GDĐT (Giáo dục Thời đại)	http://giaoducthoidai.vn
Báo Hà Nội Mới (Hà Nội Mới)	http://hanoimoi.com.vn
Báo Hoa Học Trò (HHT)	http://www.hoahoctro.vn
Báo Bưu Điện (ICTNews)	http://ictnews.vn
ICTPress	http://ictpress.vn
Infonet	http://infonet.vn
InfoTV	http://infotv.vn
VnExpress (iOne.net)	http://ione.net
Kênh 14	http://kenh14.channelvn.net
KhoaHoc.com.vn	http://khoahoc.com.vn
Báo Kinh tế Đô Thị (KTĐT)	http://www.ktdt.com.vn
KTNT	http://kinhtenongthon.com.vn
LandToday	http://landtoday.net/
Báo Lao Động (Lao Động)	http://laodong.com.vn
Báo Mực Tím (Mực tím)	http://muctim.com.vn
MUST.vn	http://www.must.vn
NDHMoney.vn	http://ndhmoney.vn
Ngoisao.net	http://ngoisao.net
Báo Người Lao Động (Người Lao Động)	http://nld.com.vn
Báo Đời sống Pháp luật (Nguoiduatin.vn)	http://nguoiduatin.vn
Nhà báo Công luận	http://congluan.vn
Báo Nhân Dân (Nhân dân)	http://www.nhandan.com.vn
Nhịp Cầu Đầu Tư	http://nhipcaudautu.vn
Tạp chí PCWorld VN (PCWorld VN)	http://pcworld.com.vn
Petrotimes.vn (Petrotimes)	http://www.petrotimes.vn
Báo Pháp luật Xã hội (Pháp luật Xã hội)	http://www.phapluatxahoi.vn
Báo Pháp luật TPHCM (Pháp luật TPHCM)	http://www.phapluattp.vn

Pháp luật VN	http://www.phapluatvn.vn
Báo Đời sống Pháp luật (Phunutoday.vn)	http://phunutoday.vn
Báo Quân Đội Nhân Dân (QĐND)	http://www.qdnd.vn
Saga.vn	http://www.saga.vn
SaigonNews	http://www.saigonnews.vn
SaigonTimes Online (SaigonTimes)	http://www.thesaigontimes.vn
Sàn OTC	http://news.sanotc.com
Báo Sài Gòn Giải Phóng (SGGP)	http://sggp.org.vn
Báo Sài Gòn Tiếp Thị (SGTT)	http://www.sgt.com.vn
Sohoa.net	http://sohoa.net
StockBiz	http://stockbiz.vn
Truyền thông Tài chính StoxPlus (StoxPlus)	http://stox.vn/
Báo Sức khỏe Đời sống (Sức Khỏe Đời Sống)	http://suckhoedoisong.vn
Sức Sống Mới	http://www.sucsongmoi.net
Sinh viên Việt Nam (SVVN)	http://www.svv.vn
Tamnhin.net	http://www.tamnhin.net
Tạp chí ĐẸP Online (Tạp chí ĐẸP)	http://www.dep.com.vn
Tạp chí Hoạt Động Khoa Học (Tạp chí HĐKH)	http://www.tchdkh.org.vn
Tạp chí Tài chính	http://tapchitaichinh.vn/
Báo Thanh Niên (Thanh Niên)	http://www.thanhnien.com.vn
Báo Thanh Niên (Thanh Niên Thể thao)	http://www.thanhnien.com.vn/thethao
Báo Thanh Niên (Thanh Niên Tuần san)	http://www.thanhnien.com.vn/tnotuanan
Báo Thanh Niên (Thanh niên WC 2010)	http://www.thanhnien.com.vn/worldcup2010
Báo điện tử Thế Giới Điện Ảnh (Thế Giới Điện Ảnh)	http://thegioidienanh.vn
Thiennhien.net	http://www.thiennhien.net
Thongtinduan.vn	http://thongtinduan.vn
Báo Tiền Phong (Tiền Phong)	http://www.tienphongonline.com.vn
Tiin.vn	http://tiin.vn
Tin tức Du lịch	http://www.dulichvn.org.vn
Tin Tức Online	http://tintuonline.com.vn
Tinhte.com	http://tinhte.com
YTT (TinTheThao)	http://www.tinthethao.com.vn
Báo Tổ Quốc (Tổ quốc)	http://www.toquoc.gov.vn

Thông Tin Công Nghệ (TTCN)	http://www.thongtincongngh.com
Tuần Vietnamnet (Tuần Việt Nam)	http://tuanvietnam.net
Báo Tuổi Trẻ (Tuổi Trẻ)	http://tuoitre.vn
Báo Tuổi Trẻ (Tuổi Trẻ Địa Ốc)	http://diaoc.tuoitre.vn
Báo Tuổi Trẻ (Tuổi trẻ Du lịch)	http://dulich.tuoitre.com.vn
Báo Tuổi Trẻ (Tuổi trẻ Nhịp sống số)	http://nhipsongso.tuoitre.com.vn
Báo Tuổi Trẻ (Tuổi Trẻ Thể Thao)	http://thethao.tuoitre.vn
Báo Tuổi Trẻ (Tuổi trẻ Tuyển sinh)	http://chuyentrang.tuoitre.vn/tuyensinh/
Báo Tuổi Trẻ (Tuổi trẻ Việc làm)	http://vieclam.tuoitre.vn/
Báo Tuổi Trẻ (Tuổi trẻ WC 2010)	http://chuyentrang.tuoitre.vn/WorldCup2010
VietnamNet (VEF)	http://vef.vn
Vietnam Economic News Online (VEN)	http://ven.org.vn
Thông Tấn Xã VN (Vietnam Plus)	http://www.vietnamplus.vn
VietnamNet	http://vietnamnet.vn
VietnamNet (VietnamNet Thể Thao)	http://thethao.vietnamnet.vn
Vietstock	http://www.vietstock.com.vn
Báo Đầu Tư (VIR)	http://www.baodautu.vn
Vitinfo	http://vitinfo.com.vn
Báo Công An Nhân Dân (VNCA)	http://vnca.cand.com.vn
Thời báo Kinh Tế (VnEconomy)	http://vneconomy.vn
VnExpress	http://vnexpress.net
VnMedia	http://vnmedia.vn
VnRock	http://vnrock.com/
Đài Tiếng Nói TP.HCM (VOH)	http://voh.com.vn
Đài Tiếng Nói VN (VOV Online)	http://vov.vn
VTC News (VTC)	http://vtc.vn
VTC News (VTC Bạn đọc)	http://vtc.vn/trangbandoc/
VTC News (VTC Bảo vệ NTD)	http://vtc.vn/bvntd/
VTC News (VTC Công nghệ)	http://vtc.vn/congngh/
VTC News (VTC Games)	http://vtc.vn/thegioigame
Đài TH VN (VTV)	http://www.vtv.vn
Vzone	http://vzone.vn
Tạp chí Xã Hội Thông Tin (XHTT)	http://xahoithongtin.com.vn

Xinhxinh.com.vn <http://xinhxinh.com.vn>
XZone <http://xzone.vn>
Zing <http://www.zing.vn>
Zing (Zing Thể thao) <http://thethao.zing.vn/news>

Tài liệu tham khảo

Tiếng Việt

Tiếng Anh

- [FFU11] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska De Jong. *An Overview of Event Extraction from Text*. Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web, DeRiVE, 2011. 8
- [MW11] Martin Wunderwald. *NewsX–Event Extraction from News Articles*. Master Thesis. Dresden University of Technology, Germany, 2011. 8
- [RG10] Ralph Grishman. *The Impact of Task and Corpus on Event Extraction Systems*, International Conference on Language Resources and Evaluation, 7th, LREC, 2010. 6
- [YKW09] Yoko Nishihara, Keita Sato, Wataru Sunayama. *Event Extraction and Visualization for Obtaining Personal Experiences from Blogs*. Human Interface and the Management of Information. Information and Interaction. LNCS, vol. 5839, Springer–Verlag, 2009. 5, 8
- [MM09] Masayuki Okamoto, Masaaki Kikuchi. *Discovering Volatile Events in Your Neighborhood: Local-Area Topic Extraction from Blog Entries*. Asia Information Retrieval Symposium on Information Retrieval Technology, 5th, AIRS09, 2009. 5, 10
- [CVJ09] K. Bretonnel Cohen Karin Verspoor, Helen L. Johnson, Chris Roeder, Philip V. Ogren, William A. Baumgartner, Jr., Elizabeth White, Hannah Tipney,

- Lawrence Hunter. *High-precision biological event extraction with a concept recognizer*. BioNLP09 Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, 2009. 5, 6, 9
- [MYL08] Mingrong Liu, Yicen Liu, Liang Xiang, Xing Chen, Qing Yang. *Extracting Key Entities and Significant Events from Online Daily News*. International Conference on Intelligent Data Engineering and Automated Learning, 9th, IDEAL08, 2008. 10
- [HJM08] Hristo Tanev, Jakub Piskorski, Martin Atkinson. *Real-Time News Event Extraction for Global Crisis Monitoring*. International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems, 13th, NLDB08, 2008. 5, 10
- [FK08] Felix Jungermann, Katharina Morik. *Enhanced Services for Targeted Information Retrieval by Event Extraction and Data Mining*. International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems, 13th, NLDB08, 2008. 5, 10
- [JHP07] Jakub Piskorski, Hristo Tanev, Pinar Oezden Wennerberg. *Extracting violent events from on-line news for ontology population*. International Conference on Business Information Systems, 10th, BIS07, 2007. 5, 10
- [FHH06] Feiyu Xu, Hans Uszkoreit, Hong Li. *Automatic Event and Relation Detection with Seeds of Varying Complexity*. AAAI Workshop on Event Extraction and Synthesis, 2006. 5, 9
- [LWZ05] Zhen Lei, Ling-da Wu, Ying Zhang, Yu-chi Liu, *A System for Detecting and Tracking Internet News Event*. Pacific-Rim Conference on Multimedia, 6th, PCM05, 2005. 10
- [RDA05] Ralph Grishman, David Westbrook, Adam Meyers. *NYU's English ACE 2005 System Description*. ACE Program, 2005. 6
- [MD04] Maria Vargas-Vera, David Celjusk. *Event Recognition on News Stories and Semi-Automatic Population of an Ontology*. IEEE/WIC/ACM International Conference on Web Intelligence, WI04, 2004. 6, 9

- [CHR04] Hong-woo Chun, Young-sook Hwang, Hae-Chang Rim. *Unsupervised event extraction from biomedical literature using co-occurrence information and basic patterns*. International Joint Conference on Natural Language Processing, 1st, IJCNLP04, 2004. 5, 10
- [TFA03] Thorsten Brants, Francine Chen, Ayman Farahat. *A System for new event detection*. Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, 26th, SIGIR03, 2003. 10
- [FHD02] Fang Li, Huanye Sheng, Dongmo Zhang. *Event Pattern Discovery from the Stock Market Bulletin*. International Conference on Discovery Science, 5th, DS02, 2002. 5, 9
- [CM00] Chinatsu Aone, Mila Ramos-Santacruz. *REES: a large-scale relation and event extraction system*. Applied Natural Language Processing Conference, 6th, ANLP00, 2000. 5, 8
- [YTJ98] Yiming Yang, Tom Pierce, Jaime Carbonell . *A study of retrospective and on-line event detection*. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 21st, SIGIR98, 1998.
- [JRV98] James Allan, Ron Papka, Victor Lavrenko. *On-line new event detection and tracking*. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 21st, SIGIR98, 1998. 4
- [RB96] Ralph Grishman, Beth Sundheim. *Message Understanding Conference - 6: A Brief History*. MUC-6, 1996. 1
- [BS92] Beth Sundheim. *Overview of the fourth message understanding evaluation and conference*. MUC-4, 1992. 8