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 Hiểu

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

Hà Nội, Ngày 7 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 (và vô cùng quan trọng) 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

Danh sách hình vẽ

Danh sách bảng

Bảng ký hiệu và từ viết tắt

Ký hiệu	**Ý nghĩa**
MAP	Modified Adaptive PageRank
HITS	Hypertext Induced Topic Search
CCP	Connected Component in PageRank
SEOs	Search Engine Optimizes

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 (MUC–1) (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 (http://www.epispider.org/), Metro Monitor (http://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. Ngoài ra, một hệ thống theo dõi tin tức cũng được nhắc tới trong chương này.

Chương 2. *Một số phương pháp trích xuất sự kiện* đưa ra 3 phương pháp trích xuất sự kiện phổ biến và có độ chính xác cao. Hơn nữa, chúng tôi cũng phân tích những thuận lợi của từng phương pháp và cách áp dụng chúng vào mô hình giải quyết của mình để đạt được hiệu quả tốt hơn.

Chương 3. *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 4. *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.

1

Giới thiệu

Thế giới đang thay đổi rất mạnh với sự tham gia của các phương tiện truyền thông xã hội. Mọi thông tin tới với người dùng nhanh, từ nhiều nguồn khác nhau. 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 thóng 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 có tên *Hệ thống xử lý tiếng Việt tự động ePi* này được đặt tại www.baomoi.com được người dùng biết đến với tên Báo mới

1.1 Động lực nghiên cứu

1.2 Đặt vấn đề

1.2.1 Phát biểu bài toán

1.2.2 Câu hỏi nghiên cứu

1.2.3 Thách thức

1.3 Ý nghĩa

1.3.1 Ý nghĩa khoa học

1.3.2 Ý nghĩa thực tiễn

1.4 Nghiên cứu liên quan

1.4.1 Một số nghiên cứu liên quan ở nước ngoài

Ở trong nước vieetj nam yeue quy

1.4.2 Một số nghiên cứu liên quan ở trong nước

2

Một số phương pháp thực hiện

2.1 Final aim

Our ultimate goal is...

2.2 Preliminary aims

There will be several preliminary scientific targets to be accomplished on the way...

3

Phương pháp giải quyết

Tài liệu tham khảo

Tiếng Việt

[MAP] Đỗ Thị Diệu Ngọc, Nguyễn Hoài Nam, Nguyễn Thu Trang, Nguyễn Yến Ngọc. *Giải pháp tính hạng trang Modified Adaptive PageRank trong máy tìm kiếm.* Chuyên san "Các công trình nghiên cứu về CNTT và Truyền thông", Tạp chí Bưu chính Viễn thông, 14 , 4-2005, 65-71.

[CCP] Nguyễn Hoài Nam, Nguyễn Thu Trang, Nguyễn Yến Ngọc, Nguyễn Trung Kiên, Bùi Việt Hải. *Sử dụng thành phần liên thông nâng cao hiệu năng tính toán PageRank trong máy tìm kiếm Vinahoo.* Hội thảo quốc gia lần thứ VIII "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông", Hải Phòng, 24-26/8/2005

Tiếng Anh

[RB96] Ralph Grishman, Beth Sundheim. *Message Understanding Conference - 6: A Brief History.* MUC-6, 1996, WWW2005, 2005. 1