Phát hiện và theo dõi các sự kiện truyền thông xã hội trực tuyến cho tiếng Việt

Nguyễn Xuân Tới*, Phạm Bảo Sơn**, Phan Xuân Hiếu**

* Học viện An ninh nhân dân, ** Đại học Công nghệ – Đại học Quốc gia Hà Nội

Giới thiệu

Ngày nay, mạng internet cung cấp cho người dùng một lượng lớn thông tin và tri thức. Đặc biệt, trong những năm gần đây, số lượng người dùng mạng xã hội (social network) ngày càng tăng. Họ có thể dễ dàng trao đổi các kinh nghiệm, thông tin, sự kiện về thế giới thực và những điều họ quan tâm trên mạng này. Bởi sự tiện lợi, dễ dàng của nó mà những thông tin, sự kiện này thường được đăng tải trên các mạng xã hội ngay khi nó xảy ra. Trong khi đó các trang tin tức khác trên mạng thường đăng tải các thông tin này chậm hơn. Thậm chí, nhiều thông tin được đăng tải trên các trang mạng xã hội nhưng không được đăng tải qua các trang tin tức khác. Tin tức, thông tin, sự kiện sẽ có giá trị cao khi nó được truyền tải đến người dùng nhanh và chính xác, đặc biệt các thông tin, sự kiện về sự kiện liên quan đến bảo vệ an ninh quốc gia, thông tin về cứu hộ, cứu nạn… Vậy làm thế nào để tự động phát hiện, tập hợp nhanh các sự kiện đó và trả lời được các câu hỏi "sự kiện gì? xảy ra ở đâu? thời gian nào? diễn biến sự kiện?..." cho người dùng?

Trong nghiên cứu này, chúng tôi muốn xây dựng một hệ thống phát hiện và theo dõi các sự kiện thông qua việc sử dụng dữ liệu là các tin tức (news) trên các trang mạng, các bình luận (comment), các bài (post) trên mạng xã hội Facebook và ngôn ngữ của các dữ liệu này là tiếng Việt.

Phương pháp

Đầu tiên, chúng tôi biểu diễn các tài liệu dưới dạng các vector trong một không gian mà ở đó các chiều là những từ khác nhau trong tập hợp từ. Trong các hệ thống của mình, chúng tôi sẽ sử dụng trọng số TF-IDF và độ đo cosine (cosine similarity) để tính toán, tìm sự giống nhau giữa các tài liệu.

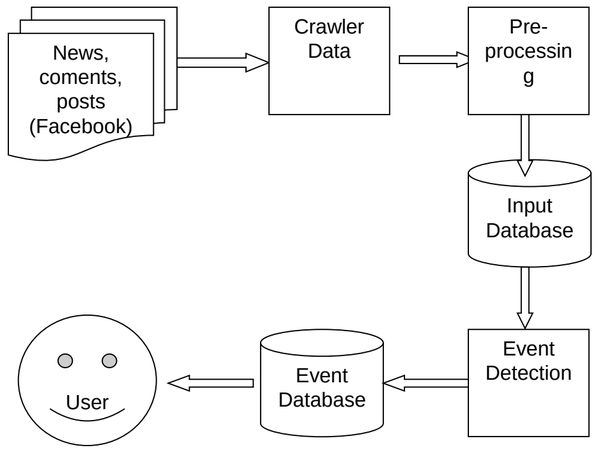
Để quyết định một tài liệu *d* có mô tả một sự kiện mới hay không, chúng tôi tính toán ***độ mới*** – Ns(d) của tài liệu trong tập hợp các tài liệu đã có D. Giá trị Ns(d) được tính theo công thức sau:

$Ns(d) = 1 - \max sim(d,d'),$ *với mọi* d' *thuộc* D.

Nếu giá trị này lớn hơn một ngưỡng *th* nào đó thì hệ thống quyết định đây là sự kiện mới. Ngược lại, hệ thống quyết định đây không là sự kiện mới. Khi nó không phải là sự kiện mới chúng tôi sẽ cập nhật các thông tin này cho sự kiện đã phát hiện trước đó mà nó đề cập. Giúp cho người sử dụng theo dõi được quá trình của sự kiện.

Kết quả

Tổng quan hệ thống phát hiện và theo dõi sự kiện



Đầu vào của hệ thống là các tin tức (từ các website tin tức), các comments (từ website tin tức, facebook) và các posts (trên facebook) sau khi được thu thập nó sẽ qua quá trình tiền xử lý để loại bỏ những dữ liệu dư thừa và chuẩn hóa, dữ liệu sẽ qua "Event Detection" để tính toán, tìm kiếm, phát hiện sự kiện Kết quả đầu ra của hệ thống là các sự kiện, thời gian, địa điểm diễn ra…

Ngoài việc xây dựng được kho dữ liệu về sự kiện nhằm phục vụ cho các nghiên cứu sau này, chúng tôi sẽ xây dựng hệ thống tìm kiếm, theo dõi sự kiện cho dữ liệu tiếng Việt.

Nghiên cứu, tìm kiếm để đưa ra phương pháp xử lý tốt cho dữ liệu lấy từ nhiều nguồn khác nhau và được viết bằng tiếng Việt, ví dụ như việc cùng một sự kiện nhưng lại được viết bởi nhiều người khác nhau, dùng các từ vựng khác nhau.

Chúng tôi sẽ xây dung hệ thống có khả năng phát hiện được nhiều loại sự kiện khác nhau chứ không tập trung vào một loại sự kiện nào, hệ thống phải phát hiện nhanh và có độ chính xác cao.

Chúng tôi cũng sẽ nghiên cứu để giải quyết được vấn đề xử lý dữ liệu lớn bởi hàng ngày, hàng giờ có rất nhiều dữ liệu được sinh ra từ nguồn internet. Đặc biệt vấn đề nhiễu trong dữ liệu cũng đặt ra nhiều thách thức, bởi rất nhiều dữ liệu từ các trang mạng xã hội không liên quan đến sự kiện mà người dùng quan tâm.

Kết luận

Với xu thế phát triển nhanh chóng của các hệ thống thông minh như hiện nay thì hệ thống này thực sự thiết thực, khoa học và đáp ứng được các yêu cầu của sự phát triển công nghệ thông tin.

Kết quả luận án có thể áp dụng cho các nhà làm báo để tổ chức các nhóm tin tức liên quan đến nhau; giúp cho các nhà đầu tư chứng khoán tìm được các thông tin nhanh nhất về những vấn đề mình quan tâm; giúp cho công việc bảo vệ an ninh quốc gia, cứu hộ, cứu nạn bởi dễ dàng tìm được sự kiện cụ thể nào đó phát sinh từ đâu như bạo loạn, động đất…

Kết quả luận án có thể còn áp dụng cho nhiều lĩnh vực, công việc khác như kết hợp chiến lược giảm phương sai (variance reduction strategy) và kỹ thuật tìm kiếm hàng xóm sấp sỉ gần nhất cho việc phát hiện nhanh các ngoại lệ trong tập dữ liệu lớn. sự kết hợp paraphrase với LSH có thể giúp giảm ảnh hưởng của lỗi từ trong các hệ thống IR quy mô lớn.

Tài liệu tham khảo

1. Agarwal, M. K., Ramamritham, K., and Bhide, M. Real time discovery of dense clusters in highly dynamic graphs: Identifying real world events in highly dynamic environments. In *Proceedings of the VLDB Endowment*, 5(10):980–991, 2012.

2. Petrovi ´ c, S., Osborne, M., and Lavrenko, V. Using paraphrases for improving first story detection in news and Twitter. In *Proceedings of Human Language Tech-nologies: Conference of the North American Chapter of the Association for Com-putational Linguistics, pages 338–346. Association for Computational Linguistics*, 2012.

3. Van Canneyt, S. Crommenlaan, G. Feys, M. Schockaert, S. Demeester, T. Develder, C and Dhoedt, B. Detecting Newsworthy Topics in Twitter. In *Proceedings of the SNOW 2014 Data Challenge*, 2014.

4. Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. Event detection and tracking in social streams. In *Proceedings of International Conference on Weblogs and Social Media* (ICWSM), 2009.

5. Van Canneyt, S. Crommenlaan, G. Feys, M. Schockaert, S. Demeester, T. Develder, C and Dhoedt, B. Detecting Newsworthy Topics in Twitter. In *Proceedings of the SNOW 2014 Data Challenge*, 2014