BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

TRỊNH CÔNG MINH QUÂN

XÂY DỰNG HỆ THỐNG PHÁT HIỆN NHỮNG XU HƯỚNG NỔI LÊN TRÊN MẠNG XÃ HỘI SỬ DỤNG TIẾNG VIỆT

LUẬN VĂN THẠC SĨ

Chuyên ngành : Công Nghệ Thông Tin

Mã số ngành: 06480201

TP. HỒ CHÍ MINH, tháng 10 năm 2015

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

TRỊNH CÔNG MINH QUÂN

XÂY DỰNG HỆ THỐNG PHÁT HIỆN NHỮNG XU HƯỚNG NỔI LÊN TRÊN MẠNG XÃ HỘI SỬ DỤNG TIẾNG VIỆT

LUẬN VĂN THẠC SĨ

Chuyên ngành : Công Nghệ Thông Tin

Mã số ngành: 06480201

CÁN BỘ HƯỚNG DẪN KHOA HỌC: PGS. TS QUẢN THÀNH THƠ

TP. HỒ CHÍ MINH, tháng 10 năm 2015

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

Cán bộ hướng dẫn khoa học : PGS.TS. Quản Thành Thơ.

Luận văn Thạc sĩ được bảo vệ tại Trường Đại học Công nghệ TP. HCM ngày 17 tháng 10 năm 2015

Thành phần Hội đồng đánh giá Luận văn Thạc sĩ gồm:

TT	**Họ và tên**	**Chức danh Hội đồng**
1	TS. Võ Đình Bảy	Chủ tịch
2	PGS.TSKH. Nguyễn Xuân Huy	Phản biện 1
3	TS. Trần Đức Khánh	Phản biện 2
4	TS. Lư Nhật Vinh	Ủy viên
5	TS. Nguyễn Thị Thúy Loan	Ủy viên, Thư ký

Xác nhận của Chủ tịch Hội đồng đánh giá Luận sau khi Luận văn đã được sửa chữa.

Chủ tịch Hội đồng đánh giá LV

TRƯỜNG ĐH CÔNG NGHỆ TP. HCM **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**

PHÒNG QLKH – ĐTSĐH **Độc lập – Tự do – Hạnh phúc**

TP. HCM, ngày..03 *tháng* ..*04*.. *năm 20*15

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên: Trịnh Công Minh Quân .....Giới tính:Nam.....

Ngày, tháng, năm sinh: .26/01/1990.....Nơi sinh:An Giang.....

Chuyên ngành: .Công nghệ thông tin.....MSHV:1341860049.....

I- Tên đề tài:

Xây dựng hệ thống phát hiện những xu hướng nổi lên trên mạng xã hội sử dụng tiếng Việt .....

II- Nhiệm vụ và nội dung:

Đưa ra những cơ sở lý thuyết và hướng tiếp cận mới từ đó hình thành nên phương pháp xây dựng một hệ thống phát hiện những xu hướng nổi lên trên mạng xã hội.

Trong đề tài này, tôi chỉ tập trung xây dựng mô hình phát hiện xu hướng dựa vào tập dữ liệu của một cơ sở dữ liệu được thu thập từ một mạng xã hội sử dụng ngôn ngữ tiếng Việt.

III- Ngày giao nhiệm vụ: 03/04/2015

IV- Ngày hoàn thành nhiệm vụ: 17/09/2015

V- Cán bộ hướng dẫn: PGS.TS Quản Thành Thơ

CÁN BỘ HƯỚNG DẪN **KHOA QUẢN LÝ CHUYÊN NGÀNH**

(Họ tên và chữ ký) (Họ tên và chữ ký)

PGS.TS Quản Thành Thơ

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong Luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện Luận văn này đã được cảm ơn và các thông tin trích dẫn trong Luận văn đã được chỉ rõ nguồn gốc.

Học viên thực hiện Luận văn

Trịnh Công Minh Quân

LỜI CÁM ƠN

Tôi xin gửi lời cảm ơn chân thành và sâu sắc nhất đến PGS TS Quản Thành Thơ, Thầy đã tận tình hướng dẫn tôi trong suốt quá trình thực hiện đề cương và luận văn cao học, tạo mọi điều kiện để tôi có thể hoàn thành tốt luận văn này.

Tôi xin gửi lời biết ơn chân thành đến các Thầy Cô trong khoa công nghệ thông tin trường Đại Học Công Nghệ TPHCM. Các Thầy Cô đã rất tận tình chỉ dạy, trang bị cho tôi những kiến thức quý báu trong suốt thời gian tôi học cao học tại trường.

Tôi xin gửi lời cảm ơn gia đình, bạn bè và các đồng nghiệp nơi tôi làm việc đã động viên và tạo mọi điều kiện thuận lợi giúp tôi hoàn thành luận văn.

Mặc dù đã cố gắng hết sức có thể để hoàn thành tốt nhất luận văn trong phạm vi và khả năng cho phép nhưng chắc chắn luận văn sẽ không tránh khỏi những thiếu sót, kính mong nhận được sự chỉ bảo tận tình của quý Thầy Cô và các bạn.

Tp. HCM, ngày 03 tháng 04 năm 2015

Học viên

Trịnh Công Minh Quân

TÓM TẮT

Sự phát triển lớn mạnh của mạng xã hội trong thời gian gần đây đã đưa đến nhiều cơ hội cũng như thách thức cho các công ty quản lý dòng dữ liệu truyền thông này. Thông tin được chia sẻ trên mạng xã hội ngày càng trở nên khổng lồ, khó kiểm soát và phân loại. Chính những khó khăn đó đã thúc đẩy sự phát triển mạnh mẽ của các nghiên cứu liên quan đến khai phá dữ liệu trong các mạng xã hội. Một trong những hướng nghiên cứu và phát triển nổi bật hiện nay của khai phá dữ liệu trên mạng xã hội là phát hiện những xu hướng nổi lên.

Các công việc về phát hiện xu hướng và thông tin nổi lên trên mạng xã hội đang thật sự thu hút nhiều sự quan tâm của các nhà nghiên cứu. Nghiên cứu này cung cấp một hướng đi mới: sử dụng phương pháp gom cụm trong khai phá dữ liệu kết hợp với thông tin thời gian để phát hiện những xu hướng nổi lên trên mạng xã hội.

ABSTRACT

The growth of social network in recently years has brought many opportunities and challenges to the companies which manage social media data. Information shared on social network became bigger and bigger so it's really not easy to control and classify them. But these difficulties have promoted the development of research relative to data mining in social network, one of them is detection of emerging trends.

Nowadays, detection trends and emerging information in social network is attracting many researchers. This research provides a new approach: using clustering method in data mining combine with temporal information to detect emerging trends in social network.

Mục Lục

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

Chương 1: GIỚI THIỆU

1.1 Lý do chọn đề tài

Trong những năm gần đây mạng xã hội ngày càng phát triển mạnh mẽ ở Việt Nam cũng như trên thế giới. Theo một số liệu thống kê[1] tháng 10 năm 2012 có gần 30 triệu người ở Việt Nam tham gia mạng xã hội, đến tháng 1 năm 2014 lên đến gần 40 triệu người ở Việt Nam tham gia mạng xã hội. Đây không những là nơi để con người trò chuyện, giải trí, kết nối bạn bè mà còn là một kênh cung cấp và chia sẽ thông tin giữa người sử dụng hoặc các doanh nghiệp, công ty muốn quảng cáo sản phẩm của mình.

Sự phát triển nhanh chóng của mạng xã hội cũng kéo theo sự bùng nổ dữ liệu: khối lượng dữ liệu trực tuyến, thông tin chia sẽ trên mạng xã hội ngày càng trở nên khổng lồ. Đây là một nguồn thông tin rất hữu ích, được cập nhật liên tục. Với thực tế trên, vấn đề đặc ra là làm thế nào để có thể khai thác được những thông tin hữu ích này từ mạng xã hội. Các nguồn thông tin này phải được xử lý như thế nào để người dùng có thể phát hiện được những chủ đề được thảo luận phổ biến trên mạng xã hội.

Việc xác định những chủ đề được thảo luận phổ biến của các thành viên trong một mạng xã hội và từ đó phát hiện ra những xu hướng nổi lên trong một mạng xã hội có ý nghĩa thật sự quan trọng trong việc giúp chúng ta có thể hiểu tốt hơn những mối quan tâm của xã hội. Một hệ thống phát hiện xu hướng nổi lên trên mạng xã hội sử dụng tiếng Việt là hết sức cần thiết. Nó giúp các công ty có những chiến lược quản cáo hiệu quả nhất và nắm bắt xu hướng người dùng một cách nhanh chóng.

1.2 Mục tiêu của đề tài

Đưa ra những cơ sở lý thuyết và hướng tiếp cận mới từ đó hình thành nên phương pháp xây dựng một hệ thống phát hiện những xu hướng nổi lên trên mạng xã hội.

1.3 Giới thiệu đề tài

Tiến hành thu thập dữ liệu từ một mạng xã hội thực tế và đưa chúng vào cơ sở dữ liệu (database) sử dụng MySql. Cơ sở dữ liệu này là đầu vào cho hệ thống phát hiện xu hướng nổi lên. Toàn bộ hệ thống sẽ được xây dựng dựa trên ngôn ngữ Java. Các kết quả đầu ra sẽ được lưu trữ vào cơ sở dữ liệu.

Phân đoạn dữ liệu theo thời gian: dữ liệu được thu thập sẽ chia thành nhiều phân đoạn theo thời gian. Dựa trên những phân đoạn dữ liệu này sau khi tìm được xu hướng nổi lên hệ thống cũng xác định được những xu hướng này nổi lên trong khoảng thời gian nào.

Tiến hành tiền xử lý dữ liệu và tìm từ khóa quan trọng trong từng phân đoạn.

Kết hợp hai phương pháp gom cụm k-means và HAC (Hierarchical Agglomerative Clustering) để gom nhóm các từ khóa quan trọng. Trước tiên, áp dụng k-means gom nhóm các từ khóa quan trọng. Sau đó, áp dụng phương pháp gom cụm HAC trên từng cụm kết quả của phương pháp k-means. Việc kết hợp này nhằm mục đích làm giảm độ phức tạp tính toán khi gom cụm. Phương pháp gom cụm HAC có độ chính xác cao tuy nhiên tốc độ tính toán khá chậm vì có độ phức tạp tăng theo cấp số mũ của số phần tử nên áp dụng k-means ở bước đầu tiên để giảm số cụm.

Ví dụ: Hệ thống thu thập được các thông tin trên mạng xã hội từ ngày 24/3/2015 đến ngày 31/3/2015. Hệ thống sẽ tiến hành chia dữ liệu này theo từng phân đoạn theo thời gian. Mỗi phân đoạn là 6 ngày và có độ phủ lên nhau là 5 ngày. Giả sử hệ thống cho ra kết quả bên dưới:

Bảng 1.3.1: Ví dụ minh họa phân đoạn dữ liệu

Thời gian	Nội dung tài liệu
2015-03-24 - 2015-03-29	máy bay Airbus A320 chở 148 người rơi tại Pháp , Pháp tiết lộ lời cuối cùng từ Airbus A320 trước khi

	rơi , Airbus A320 rơi : 16 học sinh Đức lẽ ra đã thoát nạn , sập giàn giáo công trình Formosa , 12 người tử vong , rớt nước mắt đón thi thể nạn nhân vụ sập giàn giáo Formosa , trải lòng của người thoát chết vụ sập giàn giáo Formosa
2015-03-25 - 2015-03-30	Airbus A320 rơi : 16 học sinh Đức lẽ ra đã thoát nạn , sập giàn giáo công trình Formosa , 12 người tử vong , rớt nước mắt đón thi thể nạn nhân vụ sập giàn giáo Formosa , trải lòng của người thoát chết vụ sập giàn giáo Formosa
2015-03-26 - 2015-03-31	sập giàn giáo công trình Formosa , 12 người tử vong , rớt nước mắt đón thi thể nạn nhân vụ sập giàn giáo Formosa , trải lòng của người thoát chết vụ sập giàn giáo Formosa , U23 VN - U23 MaCau : 2 bàn thắng sớm , U23 Malaysia - U23 VN : ngỡ ngàng Công Phượng , U23 VN chính thức vào VCK U23 Châu Á , HLV Miura: U23 VN không thua 10 bàn là thành công , HLV Miura bị sa thải? , lý do HLV Miura sợ báo chí viết nhiều về Công Phượng

Sau khi tiến hành tiền xử lý văn bản sẽ được giới thiệu ở Chương 3. Tiếp tục áp dụng phương pháp tính trọng số tf*idf để rút trích được các từ khóa quan trọng trong từng phân đoạn. Số lượng từ khóa phụ thuộc vào ngưỡng chặn của giá trị

tf*idf, giá trị này có thể điều chỉnh khi chạy hệ thống trên các tập dữ liệu. Bảng dưới đây thể hiện các từ khóa quan trọng của tập dữ liệu trên:

Bảng 1.3.1: Ví dụ minh họa kết quả tìm từ khóa quan trọng

Thời gian	Nội dung tài liệu
2015-03-24 - 2015-03-29	cơ phó, sập, giàn giáo, người, A320, Formosa, rơi, Pháp, máy bay, Airbus
2015-03-25 - 2015-03-30	sập, giàn giáo, Formosa,
2015-03-26 - 2015-03-31	sập, giàn giáo, Formosa, VN, Miura, U23

Sau khi rút trích được các từ khóa quan trọng trong từng phân đoạn thời gian bước tiếp theo cần tạo ra tập dữ liệu đầu vào cho bước gom cụm mà cụ thể là gom cụm bằng *k*-means ở bước đầu tiên. Tiến hành biểu diễn mỗi từ khóa thành một cấu trúc dữ liệu và đặt tên là DataPoint, mỗi DataPoint gồm 2 thành phần (thuộc tính) quan trọng là tên của từ khóa và vector sức mạnh liên kết (*link-strength*) với mỗi thành phần của vector thể hiện độ liên kết của một từ khóa đến một từ khóa khác trong tập từ khóa quan trọng được rút trích ở bước trước đó. Như vậy dữ liệu đầu vào cho bước gom cụm là một vector các DataPoint với mỗi DataPoint được đại diện cho một từ khóa trong tập các từ khóa [cơ phó, sập, giàn giáo, người, A320, Formosa, rơi, Pháp, máy bay, Airbus, VN, Miura, U23].

Nếu tại bước gom cụm *k*-means chúng ta chọn hệ số k = 2, khi đó dữ liệu sẽ được chia thành 2 cụm với mỗi cụm chứa một vector các DataPoint:

Cụm 1: vector DataPoint [cơ phó, sập, giàn giáo, người, A320, Formosa, rơi, Pháp, máy bay, Airbus].

Cụm 2: vector DataPoint [VN, Miura, U23].

Sử dụng các cụm kết quả tại bước gom cụm *k*-means áp dụng thuật toán HAC hệ thống thu được 3 cụm chủ đề như sau:

Cụm 1: vector DataPoint [cơ phó, A320, rơi, Pháp, máy bay, Airbus].

Cụm 2: vector DataPoint [sập, giàn giáo, người, Formosa].

Cụm 3: vector DataPoint [VN, Miura, U23].

1.4 Cấu trúc của luận văn

Trong chương này chúng tôi đã giới thiệu tổng quan về đề tài. Phần còn lại của luận văn sẽ được tổ chức như sau:

Chương 2: Trình bày các nghiên cứu liên quan, giới thiệu phương pháp và hướng tiếp cận của đề tài.

Chương 3: Trình bày cơ sở lý thuyết được sử dụng để xây dựng hệ thống phát hiện xu hướng nổi lên trên mạng xã hội.

Chương 4: Trình bày chi tiết về hệ thống phát hiện xu hướng nổi lên trên mạng xã hội.

Chương 5: Trình bày kết quả thực nghiệm và đưa ra đánh giá về hệ thống.

Chương 6: Đưa ra kết luận.

Chương 2: CÁC NGHIÊN CỨU LIÊN QUAN

2.1 Giới thiệu:

Sự phát triển lớn mạnh của truyền thông xã hội (*Social Media*) trong thời gian gần đây hẳn chúng ta ai cũng có thể nhận ra. Các doanh nghiệp thì sử dụng mạng xã hội cho việc quản bá tên tuổi, sản phẩm công ty, tạo mạng lưới khách hàng. Các cá nhân thì dùng mạng xã hội cho việc tạo lập tên tuổi cho bản thân, hay chỉ đơn giản là chia sẻ những cảm xúc, kết nối bạn bè.

Với sự phát triển bùng nổ như vậy kéo theo khối lượng dữ liệu trực tuyến, thông tin chia sẽ trên mạng xã hội ngày càng trở nên khổng lồ, khó kiểm soát và sàng lọc. Chính những nhu cầu đó đã thúc đẩy sự phát triển mạnh mẽ của các nghiên cứu có liên quan đến khai phá dữ liệu trong các mạng xã hội (*Social Media Mining*) như:

- Phát hiện khả năng mở rộng của các chủ đề đang nổi trong dòng văn bảng bằng cách băm các ngưỡng quan trọng của : Erich Schubert, Michael Weiler và Hans-Peter Kriegel [2].
- Hệ thống phát hiện xu hướng của các chủ đề dựa trên dòng dữ liệu của các tài khoản Twitter nhất định của: Duc T. Nguyen và Jai E. Jung [3].
- Phát hiện chủ đề nổi lên trên mạng xã hội Twitter của: James Benhardus và Jugal Kalita [4]. Yavuz Selim Yilmaz, Muhammed Fatih Bulut, Cuneyt Gurcan Akcora, Murat Ali Bayir và Murat Demirbas [5]. Mario Cataldi, Luigi Di Caro và Claudio Schifanella [9].
- Khai phá dữ liệu trong các miền web xã hội (social web) khác nhau bao gồm cả những trang nhật ký (*blogs*) và thư điện tử (*email*) của: Matthew A. Russell [6].
- Phương pháp phát hiện xu hướng trên mạng xã hội dựa vào phân tích xu hướng có cấu trúc của : Ceren Budak, Divyakant Agrawal và Amr El Abbadi [7]. Tác giả đưa ra hai định nghĩa mới cho xu hướng có cấu trúc là

xu hướng có liên kết và không liên kết (*coordinated and uncoordinated trends*). Ý nghĩa chính của cách tiếp cận này là sẽ cho điểm số cao đối với các chủ đề được thảo luận nhiều trong một cụm các nút mạng có liên kết chặt chẽ với nhau và những chủ đề được thảo luận nhiều nhưng có ít các nút mạng ngoài cụm của nó liên kết đến nó.

- Phương pháp phát hiện xu hướng nổi lên bằng "dictionary learning" của: Shiva Prasad Kasiviswanathan, Prem Melville, Arindam Banerjee và Vikas Sindhwani [8].

2.2 Các phương pháp phát hiện xu hướng:

Có hai loại kỹ thuật chính khác nhau được áp dụng để phát hiện xu hướng [10]. Chúng là phương pháp bán tự động (*semi-automatic*) và tự động (*automatic*). Các phương pháp này được áp dụng để phát hiện xu hướng từ các miền khác nhau như dữ liệu văn bản, dữ liệu đa truyền thông, và các công bố khoa học.

2.2.1 Phương pháp bán tự động (*semi-automatic*)

Phương pháp đầu tiên được đề cập đến là phương pháp bán tự động, hướng tiếp cận của phương pháp này chỉ dựa trên thông tin thống kê và công sức từ con người để xác định các xu hướng dựa trên các thông tin thống kê đó. Để hỗ trợ con người trong việc phát hiện xu hướng, một giao diện người dùng thường được phát triển để hiển thị các thông tin thống kê một cách có tổ chức.

Hệ thống phân tích thời cơ công nghệ (*Technology Opportunities Analysis System - TOAS*) [11] cung cấp những thông tin đo lường của tài liệu như số lượng từ, trích dẫn, ngày và thông tin nhà xuất bản. TOAS cung cấp thông tin đo lường thu được từ những truy vấn của người dùng trên các miền nghiên cứu.

Trong một hệ thống khác, Envision [12] người dùng có thể khám phá ra được thư viện kỹ thuật số đa phương tiện dưới dạng biểu diễn đồ họa để xác định các xu hướng. Trong Envision những công bố khoa học từ lĩnh vực khoa học máy tính được chứa trong những định dạng khác nhau như text, video và hiệu ứng.

Envision hỗ trợ tìm kiếm theo đoạn trích dẫn hoặc đầy đủ nội dung. Kết quả tìm kiếm được hiển thị dưới dạng đồ họa như ma trận của các biểu tượng màu. Từ sự hiển thị kết quả đó, người dùng có thể quan sát các quan hệ động liên quan giữa các công bố. Như vậy, người dùng có thể xác định được những chủ đề nổi lên và xu hướng trong các miền nghiên cứu.

CIMEL (*Constructive, Collaborative Inquiry-based Multimedia E-learning*) [13] là một hệ thống cộng tác có thể xác định các xu hướng nổi lên từ vùng được lựa chọn bởi người dùng. Để làm được điều đó CIMEL thu thập thông tin từ các hội nghị và workshop có liên quan đến vùng được lựa chọn. Sau đó các thông tin thống kê sẽ được tạo ra. Công sức của con người là cần thiết để tạo ra các ứng viên xu hướng nổi lên từ các thông tin được tạo ra.

2.2.2 Phương pháp tự động (*automatic*)

Phương pháp tiếp theo là phương pháp tự động (*automatic*), áp dụng kỹ thuật khai phá dữ liệu thông minh (*intelligent data mining*) để tìm ra các xu hướng một cách tự động, phương pháp này gồm có hai thành phần chính là khai phá tài liệu và khai phá thời gian. Thành phần khai phá tài liệu tập trung xác định xu hướng dựa vào tài liệu trong khi thành phần khai phá thời gian tập trung trên thông tin thời gian như ngày đăng tải của tài liệu để xác định xu hướng.

Khai phá tài liệu:

Trong thành phần khai phá tài liệu, thông tin của các xu hướng được khai phá từ nội dung của tài liệu. Các tài liệu được nhóm vào các cụm bằng một kỹ thuật gom cụm. Sau đó, thông tin trong cụm sẽ được phân tích để xác định xu hướng.

Hệ thống HDDI [14] áp dụng phương pháp khai phá tài liệu để phát hiện xu hướng từ dữ liệu văn bản. Đầu tiên hệ thống sử dụng kỹ thuật lựa chọn điểm đặc trưng để rút trích những từ khóa quan trọng từ tài liệu. Sau đó, dựa trên những điểm đặc trưng được rút trích hệ thống tính toán độ tương tự của tài liệu và gom nhóm những tài liệu phù hợp. Thông tin trên số lượng của cụm, tần suất và sự kết hợp

những điểm đặc trưng chính trong cụm được sử dụng để xác định xu hướng. Trong HDDI, một kỹ thuật dựa trên mạng neural cũng được đề xuất [15] cho việc phát hiện xu hướng nổi lên.

Các đoạn trích dẫn cung cấp thông tin liên quan giữa các công bố khoa học, nó rất hữu ích cho phát hiện xu hướng nghiên cứu. Các tác giả bài báo [16] đề xuất một kỹ thuật dựa trên các đoạn trích dẫn cho việc phát hiện xu hướng trong CiteSeer sử dụng cơ sở dữ liệu trích dẫn (*citation database*) . Đầu tiên gom cụm các tài liệu trong các citation database dựa trên thông tin kết hợp của chúng. Mỗi một cụm được tạo ra được xem như là một xu hướng. Thông tin thời gian của tài liệu (ví dụ như ngày công bố) trong mỗi cụm được sử dụng để cung cấp thông tin thống kê trên mỗi xu hướng. Kỹ thuật này hỗ trợ tìm xu hướng dựa trên từ khóa người dùng nhập. Tuy nhiên, kỹ thuật này không xem xét thông tin thời gian của tài liệu trong suốt quá trình gom cụm. Do đó, mặc dù các cụm có thể được tìm thấy tương ứng với các truy vấn trên miền nghiên cứu, nhưng nó khó có thể được sử dụng cho việc xác định xu hướng hiện tại trong miền nghiên cứu.

Khai phá thời gian:

Trong thành phần khai phá thời gian, thông tin của các xu hướng được khai phá từ thông tin thời gian của tài liệu. Tài liệu trong một cơ sở dữ liệu văn bản được gom cụm dựa trên ngày công bố. Sau đó, xu hướng có thể được phát hiện dựa trên các mẫu tuần tự của những tài liệu có liên quan thông qua thời gian (hoặc ngày).

Khai phá thông tin thời gian từ một tập dữ liệu văn bản được đánh nhãn được đề cập đến trong tài liệu [17]. Đầu tiên, nó sử dụng phương pháp thống kê để rút trích những điểm đặc trưng quan trọng liên quan đến những thời kỳ nhất định của thời gian. Sau đó, với mỗi thời kỳ tiến hành nhóm các điểm đặc trưng được rút trích thành các chủ đề. Như vậy, TimeMines có thể xác định đúng các chủ đề hoặc sự kiện quan trọng được chứa trong các thời kỳ nhất định của thời gian. Thông tin thống kê của các sự kiện trong mỗi thời kỳ thời gian cũng được cung cấp.

Tương tự với TimeMines, ThemeRiver [18] cũng khai phá thông tin thời gian từ một tập dữ liệu văn bản lớn để xác định các chủ đề. Đầu tiên, ThemeRiver phân loại tài liệu vào các nhóm dựa vào ngày công bố của chúng. Mỗi nhóm được thể hiện bởi một tập các từ khóa được xem như một chủ đề. Do đó, ThemeRiver có thể xác định các chủ đề cho mỗi thời kỳ của thời gian. Bước tiếp theo, ThemeRiver tính toán mức độ tương tự của các chủ đề chứa trong từng thời kỳ thời gian khác nhau, kết hợp lại như một dòng chảy. Khi giao diện đồ họa được hiển thị, một dòng chảy có thể hỗ trợ người sử dụng quan sát sự thay đổi trong các chủ đề theo thời gian một cách trực quan.

Trong dự án TDT [19], một hệ thống "Event Tracking" được phát triển để lấy một tập dữ liệu văn bản như là đầu vào. Dữ liệu trong tập văn bản chứa một vài mẩu chuyện tin tức. Những mẩu chuyện này được sắp xếp theo thời gian. Sau đó, trong mỗi mẩu chuyện, hệ thống TDT rút trích một tập các từ khóa quan trọng. Thông qua việc so sánh tập từ khóa vừa được rút trích từ một mẩu chuyện với những mẩu chuyện khác trong quá khứ, hệ thống TDT có thể phán đoán xem mẩu chuyện này có giống với những mẩu chuyện trong quá khứ hay không. Nếu mẩu chuyện này không giống với những mẩu chuyện trong quá khứ, hệ thống sẽ ghi lại một sự kiện mới được phát hiện. Mỗi một sự kiện được xem là một xu hướng mới được tìm thấy tại thời gian đó.

2.2.3 Phương pháp phân tích cụm dựa trên ngữ cảnh

Các tác giả trong bài báo [20] đã đề xuất một phương pháp phát hiện xu hướng trong miền nghiên cứu dựa trên kỹ thuật phân tích cụm thông qua ngữ cảnh (*Context-based Cluster Analysis - CCA*). Phương pháp này có thể phát hiện các xu hướng trong nghiên cứu dựa trên một cơ sở dữ liệu trích dẫn một cách hoàn toàn tự động. CCA gồm hai quá trình chính: phát sinh quan hệ và ngữ cảnh xuyên qua các cụm.

Chương 3: CƠ SỞ LÝ THUYẾT

Trong chương này tôi sẽ đi sâu vào việc phân tích và diễn giải các cơ sở lý thuyết được chọn để thực hiện hệ thống phát hiện xu hướng nổi lên.

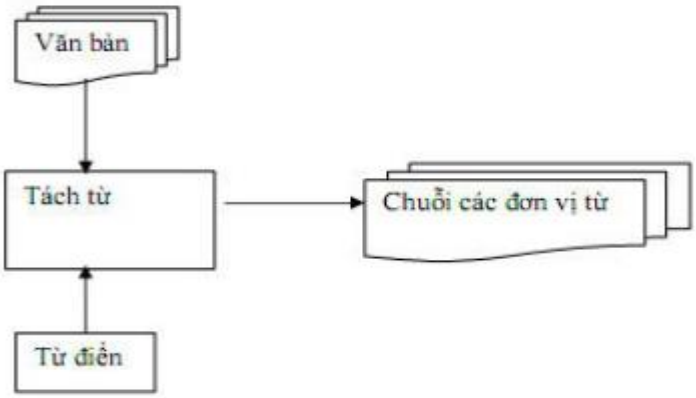
3.1 Tiền xử lý văn bản

Đây là bước rất quan trọng vì trong bước này sẽ làm giảm số từ có trong văn bản qua đó sẽ làm giảm kích thước dữ liệu trong biểu diễn văn bản và tăng độ chính xác khi tiến hành tìm từ các từ khóa quan trọng khi tính toán tf*idf.

Tiền xử lý văn bản sẽ tiến hành các bước sau:

- Lấy từ ghép:

Để lấy được từ ghép trong văn bản tôi sử dụng công cụ tách từ tiếng Việt vnTokenizer[21]. Dựa trên phương pháp so khớp tối đa (Maximum Matching) với tập dữ liệu sử dụng là bảng âm tiết tiếng Việt và từ điển từ vựng tiếng Việt. Mục đích của việc lấy từ ghép là để làm giảm số lượng từ mà vẫn không ảnh hưởng đến nội dung tài liệu. Dưới đây là quy trình thực hiện tách từ theo phương pháp so khớp tối đa:



Hình 3.1.1: Quy trình tách từ

Đầu vào của công cụ tách từ vnTokenizer là một câu hoặc một văn bản.

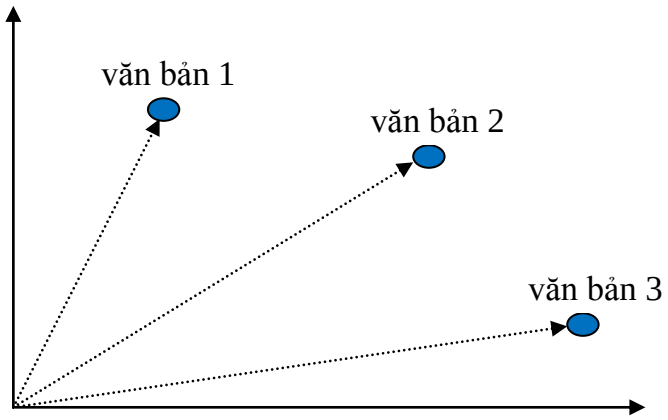
Đầu ra là một chuỗi các đơn vị từ được tách.

- Loại bỏ từ dừng (Stop-words):

  Từ dừng (Stop-words) dùng để chỉ những từ xuất hiện quá nhiều trong các văn bản của tập kết quả, thường thì không giúp ích gì trong việc phân biệt nội dung của tài liệu. Ví dụ như những từ ”và”, ”do”, ”rồi”, ”những”... Vì là những từ không quan trọng nên việc loại bỏ các từ dừng ra ngoài văn bản sẽ không ảnh hưởng đến việc biểu diễn văn bản sau này.

3.2 Vector trọng số tf-idf

Để hiểu rõ phương pháp vector trọng số tf-idf (term frequency – inverse document frequency) trước tiên chúng ta phân tích mô hình không gian vector (Vector Space Model). Mô hình không gian vector là mô hình đại số thể hiện thông tin văn bản như là một vector. Mỗi thành phần của vector là một từ khóa riêng biệt (term) trong tập văn bản gốc và được gán một giá trị là hàm f chỉ mật độ xuất hiện của từ khóa trong văn bản.



Hình 3.2.1: Các vector văn bản được biểu diễn trong không gian 2 chiều

Giả sử ta có một văn bản và nó được biểu diễn bởi vector $V(v_1,v_2,v_3,...,v_n)$. Trong đó v_i là số lần xuất hiện của từ khóa i trong văn bản. Ta xét 3 văn bản sau:

- Văn bản 1 (VB1): Tôi thích bóng đá và tenis
- Văn bản 2 (VB2): Tôi chơi guitar
- Văn bản 3 (VB3): Bóng đá và bóng đá

Sau khi qua bước tiền xử lý văn bản ta biểu diễn chúng dưới dạng vector như sau:

Bảng 3.2.1: Biểu diễn các vector văn bản

Từ	thích	bóng đá	tenis	chơi	guitar
Vector_VB1	1	1	1	0	0
Vector_VB2	0	0	0	1	1
Vector_VB3	0	2	0	0	0

Trong các cơ sở dữ liệu văn bản, mô hình vector là mô hình biểu diễn văn bản được sử dụng phổ biến nhất hiện nay. Mối quan hệ giữa các trang văn bản được thực hiện thông qua việc tính toán trên các vector vì vậy được thi hành khá hiệu quả. Một phương pháp nổi tiếng nhất trong mô hình không gian vector dùng để xác định giá trị các cụm từ trong vector đặc trưng là phương pháp trọng số tf-idf.

Tf-idf là tích của hai số liệu thống kê tần suất xuất hiện và nghịch đảo tần suất xuất hiện.

- Tf (tần suất xuất hiện) được tính theo công thức:

$$tf(t,d) = \frac{f\ (t,d)}{\max\{\ f(w,d) : w\ \in d\ \}}$$

$tf(t,d)$: là tần suất xuất hiện từ t trong văn bản d.

$f(t,d)$: là số lần xuất hiện từ t trong văn bản d.

$\max\{\ f(w,d) : w\ \in d\ \}$: số lần xuất hiện nhiều nhất của một từ bất kỳ trong văn bản d.

- Idf (nghịch đảo tần suất xuất hiện) được tính theo công thức:

$$idf(t) = \log 2(\frac{N}{n(t)})$$

$idf(t)$: là nghịch đảo tần suất xuất hiện từ t.

N: là tổng số văn bản trong tập tài liệu D

$n(t)$: số lượng tài liệu chứa từ t.

Trọng số tf-idf:

$$w_{td} = tf(t,d) * idf(t)$$

Những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao.

3.3 Thuật toán k-means

Gom cụm dữ liệu hay còn gọi là phân cụm, phân vùng dữ liệu là phương pháp gom một tập các ứng viên vào các cụm. Đây là phương pháp học không giám sát. Gom cụm dữ liệu được ứng dụng trong rất nhiều lĩnh vực khác nhau như là khai phá dữ liệu, học máy, xử lý ảnh,…

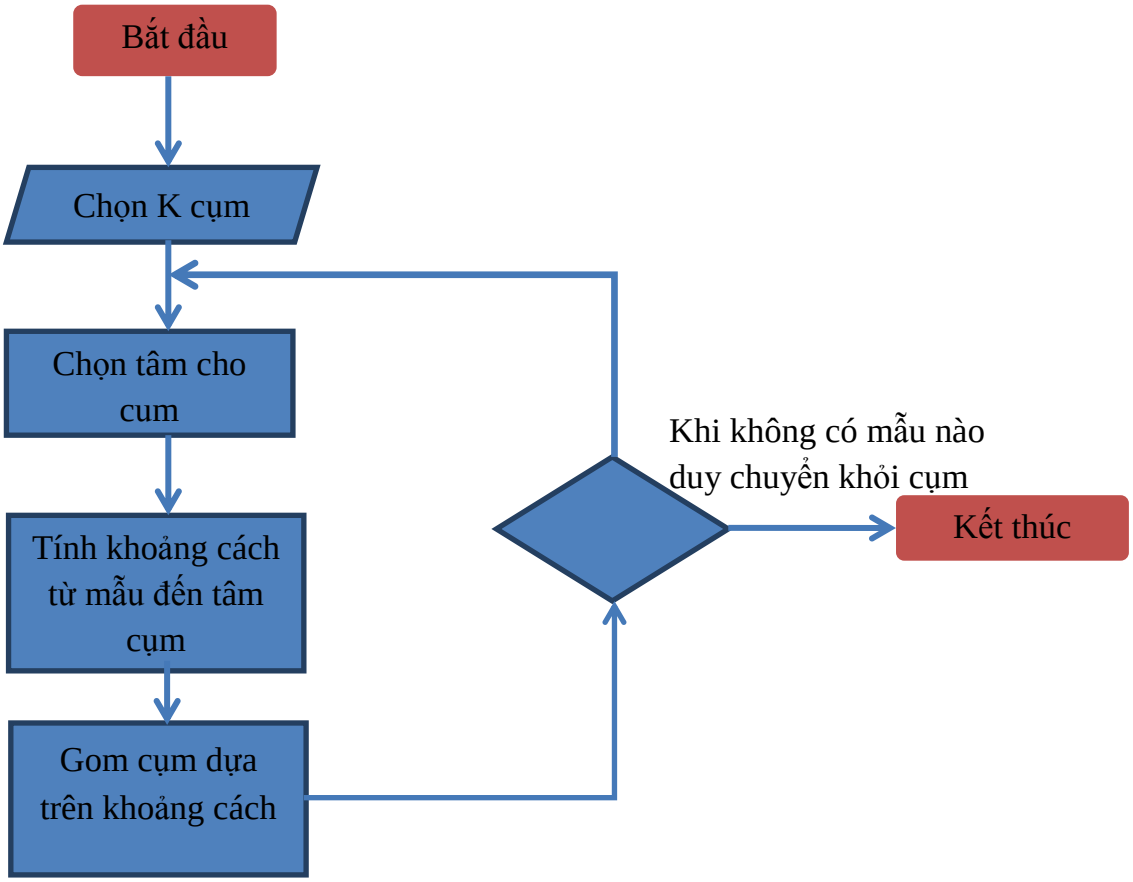
K-Means là thuật toán được sử dụng phổ biến trong trong hướng tiếp cận phân cụm phân hoạch (*Partitional clustering*). Ý tưởng chính của phương pháp phân cụm phân hoạch là phân một tập dữ liệu có n phần tử cho trước thành k nhóm dữ liệu sao cho mỗi phần tử dữ liệu chỉ thuộc về một nhóm dữ liệu và mỗi nhóm dữ liệu có tối thiểu một phần tử dữ liệu.

Thuật toán *k*-means có độ phức tạp thấp O(tkn) với t là số lần lặp, k là số cụm, n là số đối tượng sẽ gom cụm. Thuật toán này có nhiều biến thể khác nhau nhưng được đưa ra đầu tiên bởi J.B MacQueen vào năm 1967. Đầu vào của thuật toán này là một tập gồm n mẫu và một số nguyên k. Cần phân n đối tượng này thành k cụm sao cho sự giống nhau giữa các mẫu trong cùng cụm là cao hơn là giữa các đối tượng khác cụm.

Tư tưởng chính của thuật toán này là đầu tiên chọn ngẫu nhiên k mẫu, mỗi mẫu này coi như biểu diễn một cụm, như vậy lúc này trong mỗi cụm thì mẫu đó cũng là

tâm của cụm. Các mẫu còn lại được gán vào một cụm nào đó trong k cụm đã có sao cho tổng khoảng cách từ mẫu đó đến tâm của nhóm là nhỏ nhất. Sau đó tính lại tâm cho các nhóm và lặp lại quá trình đó cho đến khi hàm tiêu chuẩn hội tụ. Hàm tiêu chuẩn được dùng phổ biến nhất là hàm tiêu chuẩn bình phương sai.

Thuật toán *k*-means được mô tả cụ thể như sau:



Hình 3.3.1: Lưu đồ mô tả thuật toán K-means

– Bước 1: Chọn ngẫu nhiên k mẫu vào k cụm. Coi tâm của cụm là chính là mẫu có trong cụm.

– Bước 2: Tính khoảng cách giữa các mẫu còn lại đến k tâm

– Bước 3: Gán các mẫu vào cụm sao cho khoảng cách từ mẫu đến tâm cụm là nhỏ nhất

– Bước 4:Nếu các cụm không có sự thay đổi nào sau khi thực hiện bước 3 thì chuyển sang bước 5, ngược lại thì quay lại bước 2.

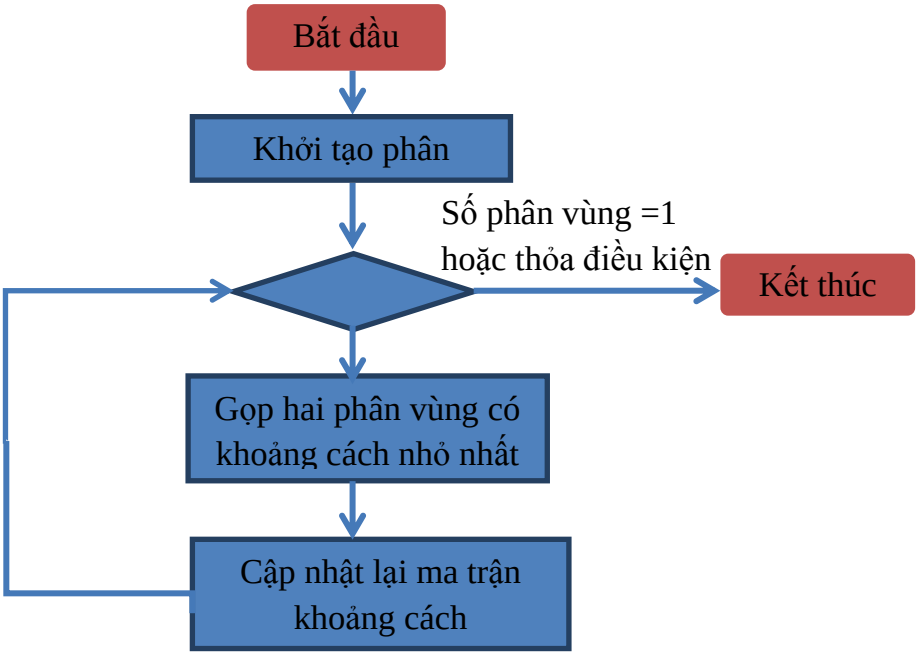
– Bước 5: thuật toán kết thúc.

3.4 Thuật toán HAC

HAC (Hierarchical Agglomerative Clustering) là thuật toán phân cụm không giám sát (không cần biết trước số cụm cần phân vào) nhưng phải cung cấp điều kiện dừng. có 2 phương pháp chính đó là:

- Phương pháp kết hợp cụm từ dưới lên (Agglomerative Aapproach): Ban đầu, chúng ta xem mỗi đối tượng là 1 nhóm (cluster) và nhóm 2 đối tượng gần nhất thành 1 cluster. Quá trình này lặp lại cho đến khi tất cả các đối tượng được nhóm vào 1 cluster cuối cùng hoặc thỏa điều kiện cho trước.
- Phương pháp phân chia cụm từ trên xuống (*Divisive Approach*) : Quá trình ngược lại với Agglomerative Approach, ban đầu chúng ta xem tất cả các đối tượng thuộc cùng 1 cluster, sau đó tiến hành phân thành 2 nhóm con (thường dựa vào khoảng cách lớn nhất). Quá trình này được thực hiện cho đến khi mỗi nhóm chỉ còn 1 đối tượng.

Với hướng tiếp cận của đề tài là gom nhóm các từ khóa quan trọng trên từng cụm kết quả của bước gom cụm bằng thuật toán *k*-means để hình thành nên các chủ đề, chúng tôi chọn phương pháp kết hợp cụm từ dưới lên. Thuật toán HAC được mô tả cụ thể như sau:



Hình 3.4.1: Lưu đồ mô tả thuật toán HAC

Giả sử có N phần tử và ma trận khoảng cách N*N

– Bước 1: Bắt đầu cho mỗi phần tử vào phân vùng của nó. Nếu có N phần tử thì có N phân vùng khởi tạo.

– Bước 2: Tìm cặp phân vùng có khoảng cách nhỏ nhất và hợp lại thành một phân vùng. Lúc này số phân vùng đã giảm đi một.

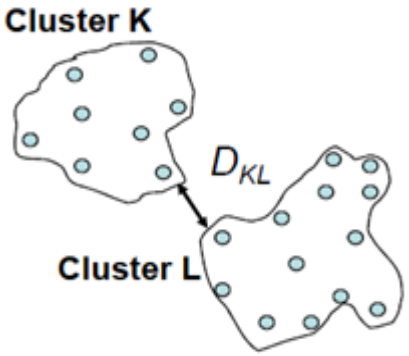
– Bước 3: Tính khoảng cách giữa phân vùng mới với các phân vùng cũ còn lại.

– Bước 4: Lặp lại bước 2,3 cho đến khi chỉ còn lại một phân vùng hoặc thỏa mãn điều kiện dừng nào đó.

Để tính được khoảng cách giữa các cặp phân vùng, chúng ta tìm hiểu một số phương pháp phổ biến để tính khoảng cách giữa hai phân vùng cho thuật toán HAC như bên dưới:

Giả sử có 2 cụm dữ liệu K và L với kí hiệu là C_K và C_L, x_i và x_j lần lượt là các phần tử thuộc về cụm C_K và C_L, $d(x_i, x_j)$ là khoảng cách giữa 2 phần tử x_i và x_j.

- Single Linkage: Khoảng cách ngắn nhất giữa 2 phần tử của mỗi cụm

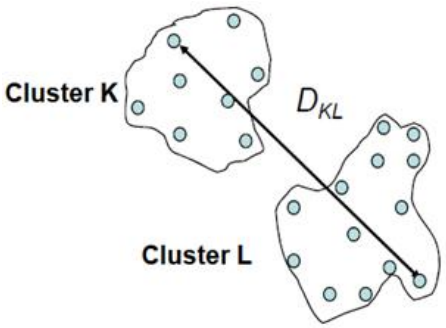


$$D_{KL} = \min_{\substack{i \in C_K \\ j \in C_L}} d(x_i, x_j)$$

Hình 3.4.2: Single Linkage

Với phương pháp này khoảng cách giữa 2 cụm là D_{KL} được tính bằng khoảng cách ngắn nhất min $d(x_i, x_j)$ của x_i và x_j.

- Complete linkage: Khoảng cách lớn nhất giữa 2 phần tử của mỗi cụm

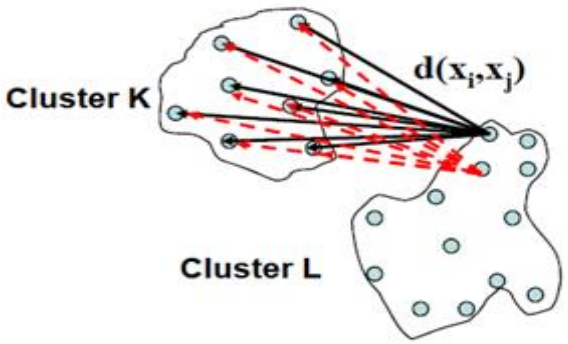


$$D_{KL} = \max_{\substack{i \in C_K \\ j \in C_L}} d(x_i, x_j)$$

Hình 3.4.3: Complete Linkage

Với phương pháp này khoảng cách giữa 2 cụm là D_{KL} được tính bằng khoảng cách lớn nhất max $d(x_i, x_j)$ của x_i và x_j.

- Average Linkage: Trung bình khoảng cách của tất cả các cặp phần tử của hai cụm

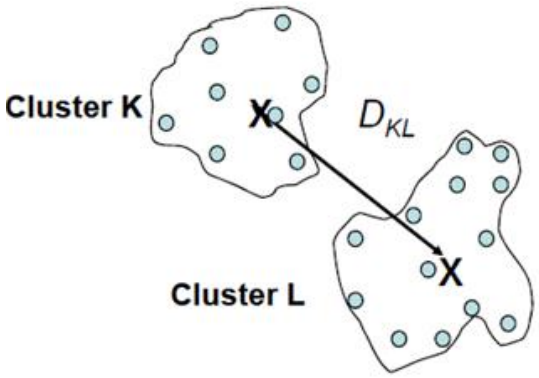


$$D_{KL} = \frac{1}{n_K n_L} \sum_{i \in C_K} \sum_{j \in C_L} d(x_i, x_j)$$

Hình 3.4.3: Average Linkage

Với phương pháp này khoảng cách giữa 2 cụm là D_{KL} được tính bằng trung bình cộng khoảng cách của tất cả các cặp x_i và x_j trong 2 cụm C_K và C_L .

- Centroid linkage: Khoảng cách giữa hai phần tử trung tâm của mỗi cụm



$$D_{KL} = \|\overline{x}_K - \overline{x}_L\|^2$$

Hình 3.4.3: Centroid Linkage

Với phương pháp này khoảng cách giữa 2 cụm là D_{KL} được tính bằng khoảng cách giữa 2 phần tử trung tâm (Centroid) của mỗi cụm C_K và C_L .

Xét một ví dụ sau: Gom nhóm các từ khóa quan trọng sử dụng single linkage:

Bảng 3.4.1: Ma trận khoảng cách khi khởi tạo

	HLV	**Miura**	**cầu thủ**	**U23**	**Việt Nam**
HLV	0.00	0.00	1.23	0.44	1.22
Miura	0.00	0.00	1.23	0.44	1.22
cầu thủ	1.23	1.23	0.00	0.43	1.29
U23	0.44	0.44	0.43	0.00	0.42

Việt Nam	1.22	1.22	1.29	0.42	0.00

- Ta thấy khoảng cách giữa thành phố HLV và Miura bằng 0.00 là gần nhất. Gom nhóm hai từ khóa này lại gọi là ” HLV/Miura”. Tính lại khoảng cách từ ” HLV/Miura” đến các từ khóa khác. Ta được ma trận khoảng cách như sau:

Bảng 3.4.2: Ma trận khoảng cách sau khi gom cụm ”HLV” và ”Miura”

	HLV/Miura	**cầu thủ**	**U23**	**Việt Nam**
HLV/Miura	0.00	1.23	0.44	1.22
cầu thủ	1.23	0.00	0.43	1.29
U23	0.44	0.43	0.00	0.42
Việt Nam	1.22	1.29	0.42	0.00

- Tiếp tục gọp hai từ khóa có khoảng cách là gần nhất: ta thấy U23 và Việt Nam có khoảng cách gần nhất là 0.42 gom nhóm hai từ khóa này lại cập nhật ma trận khoảng cách như sau:

Bảng 3.4.2: Ma trận khoảng cách sau khi gom cụm ”U23” và ”Việt Nam”

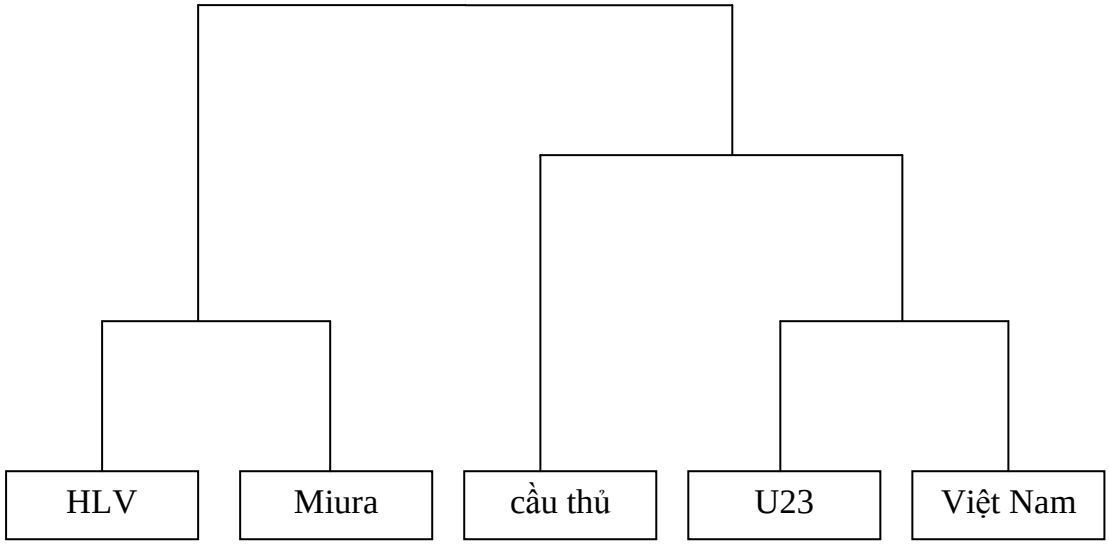
	HLV/Miura	**cầu thủ**	**U23/ Việt Nam**
HLV/Miura	0.00	1.23	0.44
cầu thủ	1.23	0.00	0.43
U23/ Việt Nam	0.44	0.43	0.00

- Lặp lại quá trình gom nhóm ta gọp cụm ”U23/Việt Nam” và ”cầu thủ” do khoảng cách giữa chúng bằng 0.43 là nhỏ nhất cập nhật ma trận khoảng cách:

Bảng 3.4.2: Ma trận khoảng cách sau khi gom cụm ”U23/Việt Nam ” và ”cầu thủ”

	HLV/Miura	**U23/ Việt Nam/cầu thủ**
HLV/Miura	0.00	0.44
U23/ Việt Nam/cầu thủ	0.44	0.00

Cuối cùng ta gọp hai cụm còn lại. Quá trình trên được thể hiện qua cây dendrogram bên dưới:



Hình 3.4.3: Cây dendrogram biểu diễn quá trình gom cụm HAC

3.6 Phương pháp tính khoảng cách khi gom cụm

3.6.1 Giới thiệu về link-strength và correlation

Đối với dữ liệu hệ thống làm việc là dữ liệu text do đó khi tiến hành gom cụm các từ khóa quan trọng để xác định chủ đề chúng ta không thể áp dụng giá trị khoảng cách toán học (*euclidean distance*) để tính toán giá trị khoảng cách giữa các từ khóa để hình thành nên cụm chủ đề. Trong ngữ cảnh là mạng xã hội, một số từ khóa quan trọng thường được người dùng nhắc đến thường xuyên sẽ đi cùng với nhau, chính sự xuất hiện cùng nhau của những từ khóa này mà chúng ta có thể dự đoán được những chủ đề mà người dùng trong mạng xã hội bàn luận đến trong một khoảng thời gian nào đó. Tức là nếu hai từ khóa A và B liên quan đến nhau thì khi nhắc đến A thì thường nhắc đến B. Từ những phân tích đó chúng tôi kết hợp hai kỹ thuật về sức mạnh liên kết và hệ số tương quan để suy ra được cách tính khoảng cách giữa các từ khóa quan trọng trong các interval tại bước gom cụm.

Sức mạnh liên kết (*link-strength*) giữa hai đối tượng A và B được định nghĩa bởi công thức:

$$LS_{AB} = Link - Strength(A, B) = X/(Y - X)$$

Trong đó X là tổng số lần cùng xuất hiện của A và B, Y là tổng số lần xuất hiện của riêng A, riêng B và cùng xuất hiện của A và B.

Ánh xạ vào hệ thống phát hiện xu hướng trên mạng xã hội thì X biểu diễn cho tổng số interval mà cả hai từ khóa A và B cùng xuất hiện trong các interval đó. Y biểu diễn cho tổng số interval mà những interval này có thể chứa A hoặc B hoặc chứa cả A và B. Đối với hệ thống phát hiện xu hướng, kết quả tính link-strength được biểu diễn dưới dạng một ma trận (*matrix*) N*N với mỗi hàng của ma trận là một vector mà tại mỗi giá trị của nó thể hiện độ liên kết giữa một từ khóa và các từ khóa còn lại. Những vector này sẽ được sử dụng để tính độ tương quan giữa các từ khóa khi tiến hành gom cụm.

Trong lý thuyết xác suất và thống kê, hệ số tương quan cho biết độ mạnh của mối tương quan tuyến tính giữa hai biến số ngẫu nhiên. Trong hệ thống phát hiện xu hướng theo phương pháp tiếp cận của chúng tôi hệ số tương quan giữa hai từ khóa

trong mỗi interval sẽ thể hiện giá trị khoảng cách giữa chúng khi tiến hành gom cụm. Hệ số tương quan giữa hai từ khóa càng lớn thì khoảng cách giữa chúng càng nhỏ, tức khả năng chúng thuộc về một cụm càng cao. Hệ số tương quan có công thức tính như sau:

$$r = \frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[N(\sum X^2) - (\sum X)^2][N(\sum Y^2) - (\sum Y)^2]}}$$

Trong đó X, Y lần lượt là các vector biểu diễn độ liên kết giữa các từ khóa, N là số chiều của vector X hoặc Y.

3.6.2 Kết hợp link-strength và correlation để tính khoảng cách

Ví dụ bên dưới trình bày các bước kết hợp giữa link-strength và correlation để tính khoảng cách khi gom cụm. Giả sử ta có 6 interval và các trend words trong mỗi interval như bên dưới:

Bảng 3.5.2.1: Ví dụ về các trend word trong interval.

Interval	Trend words
1	A, B
2	A, B, C
3	A, C
4	A, B
5	D
6	D

- Tính Link-strength ta được ma trận kết quả như bên dưới

Bảng 3.5.2.2: Ma trận tính link-strength.

	A	B	C	D
A	0	3	1	0
B	3	0	0.33	0
C	1	0.33	0	0
D	0	0	0	0

Từ ma trận kết quả link-strength ta được các vector như sau:

Vec_A (0, 3, 1, 0), Vec_B (3, 0, 0.33, 0), Vec_C (1, 0.33, 0, 0), Vec_D (0, 0, 0, 0).

▪ Tính correlation

– Correlation giữa A và B:

$$r_{AB} = \frac{4(1 * 0.33) - (3 + 1) * (3 + 0.33)}{\sqrt{4(3^2 + 1^2) - (3 + 1)^2} * \sqrt{4(3^2 + 0.33^2) - (3 + 0.33)^2}} = 0.48$$

– Correlation giữa A và C:

$$r_{AC} = \frac{4(3 * 0.33) - (3 + 1) * (1 + 0.33)}{\sqrt{4(3^2 + 1^2) - (3 + 1)^2} * \sqrt{4(1^2 + 0.33^2) - (1 + 0.33)^2}} = 0.17$$

– Correlation giữa A và D:

$$r_{AD} = 0$$

Tương tự cho việc tính các hệ số tương quan còn lại.

▪ Tính distance khi clustering

Ta nhận thấy correlation giữa X và Y càng lớn thì distance giữa chúng càng nhỏ và hệ thống chỉ xét hệ số tương quan tuyến tính dương, do đó correlation thuộc [0,1] như vậy có thể suy ra công thức tính khoảng cách như sau:

Distance = 1 - Correlation

Ngoài ra thông thường chúng ta hay dùng độ nghịch đảo để tính đại lượng nghịch vì nó không phụ thuộc vào giá trị lớn nhất và để giảm đi khoảng cách tuyệt đối tương đối lớn có thể ảnh hưởng nhiều đến kết quả tính toán, người ta hay sử dụng giá trị log để khoảng cách được scale nhỏ lại nên trong các ứng dụng thực tế người ta hay tính khoảng cách dựa vào độ tương quan bằng công thức sau:

Distance = log(1/Correlation)

Đối với ví dụ ở trên nếu ta chọn hệ số k = 2 cho thuật toán *k*-means và tâm của cụm là A và D thì ta có thể tìm ra được 2 cụm là {A,B,C} và {D}.

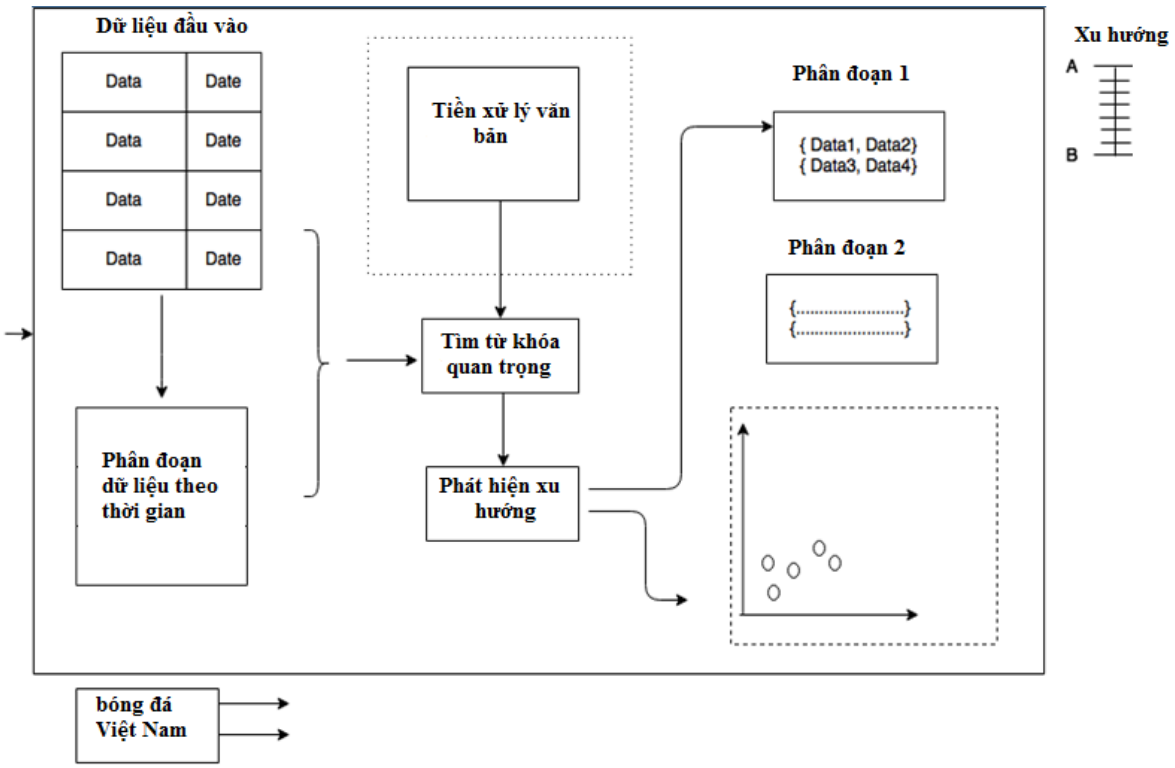
Trong chương này chúng tôi đã đi sâu vào việc phân tích và diễn giải các cơ sở lý thuyết được sử dụng để hiện thực hệ thống phát hiện xu hướng. Phương pháp vector trọng số tf-idf được chọn để biểu diễn dữ liệu văn bản trên mạng xã hội. Tiếp đến giới thiệu hai phương pháp gom cụm *k*-means và HAC cùng với các phương pháp tính khoảng cách giữa hai cụm. Cuối cùng chúng tôi giới thiệu hai kỹ thuật về sức mạnh liên kết và hệ số tương quan để tính giá trị khoảng cách khi gom cụm.

Chương 4: MÔ HÌNH PHÁT HIỆN XU HƯỚNG ĐƯỢC ĐỀ XUẤT

Trong chương này chúng tôi sẽ trình bày chi tiết về kiến trúc của hệ thống và sự tương tác giữa các thành phần chính trong toàn hệ thống phát hiện xu hướng nổi lên trên mạng xã hội.

4.1 Kiến trúc của hệ thống

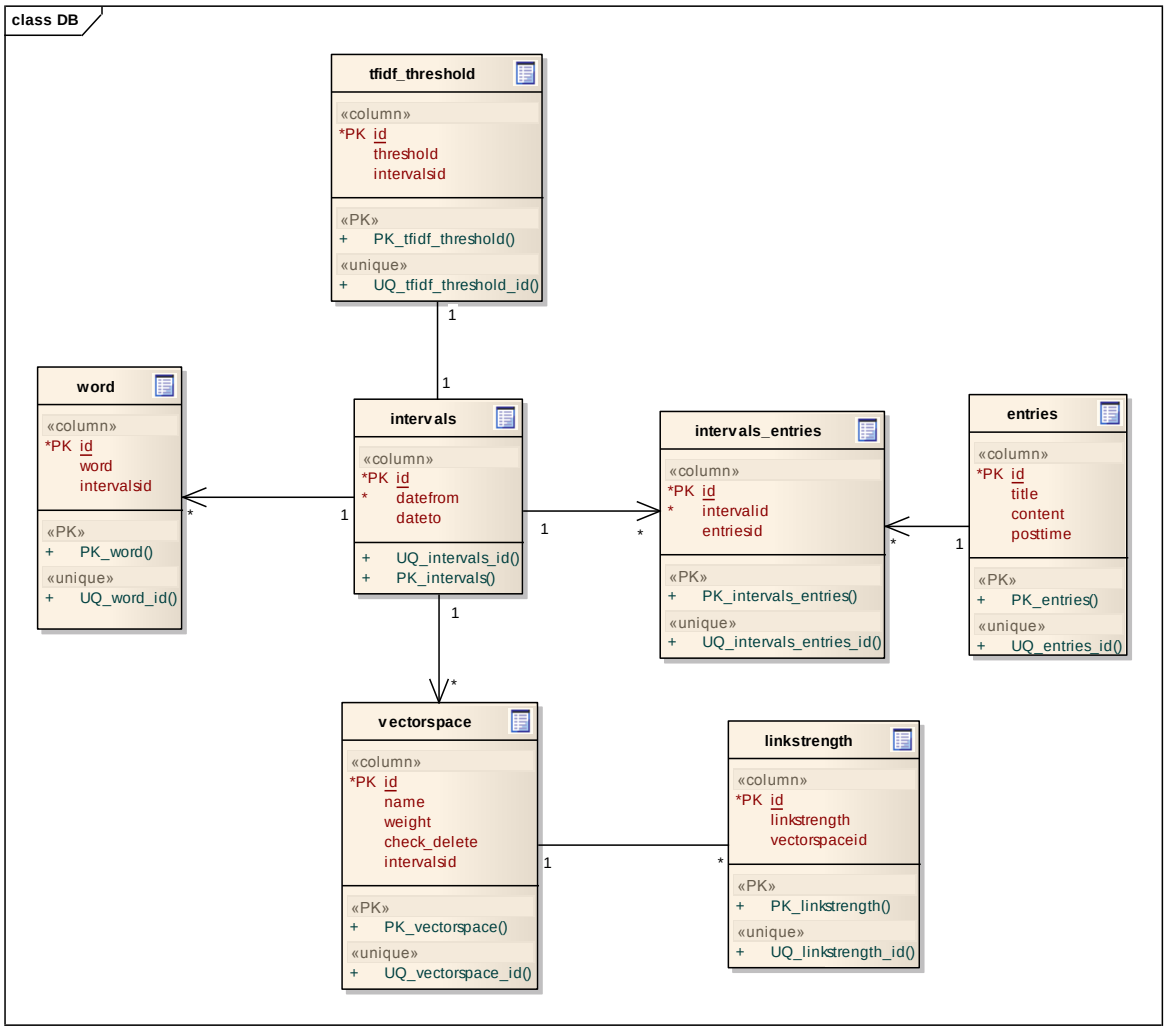
Các thành phần chính của hệ thống được thể hiện theo mô hình sau. Đây là mô hình được đề xuất bởi giáo viên hướng dẫn và được tôi nắm vững và phát triển:



Hình 4.1.1: Mô hình hệ thống phát hiện xu hướng nổi trên mạng xã hội

4.1.1 Dữ liệu đầu vào:

Lưu trữ dữ liệu đầu vào của hệ thống và các giá trị sau khi được tín toán. Sơ đồ quan hệ sau mô tả cấu trúc của cơ sở dữ liệu:



Hình 4.1.2: Sơ đồ cơ sở dữ liệu quan hệ của hệ thống

Bảng entries: lưu trữ dữ liệu đầu vào của hệ thống là các bài viết được người dùng đăng.

Bảng intervals: lưu trữ các phân đoạn thời gian sau khi tính toán từ các entries.

Bảng intervals_entries: cho biết intervals gồm có những bài viết nào.

Bảng tfidf_threshold: sau khi tính toán các giá trị tf-idf của từng intervals. Ngưỡng tf-idf của chúng sẽ được lưu trữ trong bảng này để có thể thay đổi ngưỡng cho phù hợp.

Bảng word: lưu trữ các từ được cắt ra sau quá trình xử lý tf-idf.

Bảng vectorspace: lưu trữ các từ và trọng số của nó sau quá trình xử lý tf-idf.

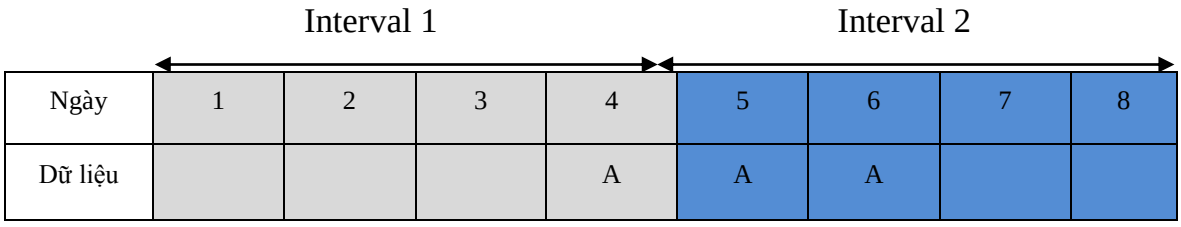
Bảng linkstrength: lưu trữ các giá trị trong ma trận linkstrength của vectorspace.

4.1.2 Phân đoạn dữ liệu theo thời gian

Hướng tiếp cận của đề tài là hệ thống sẽ phân đoạn dữ liệu của mạng xã hội ra thành nhiều phân đoạn (*interval*) theo thời gian, số lượng phân đoạn thời gian phụ thuộc vào độ lớn thời gian của từng phân đoạn, vì dữ liệu trên mạng xã hội được cập nhật thường xuyên nên chúng tôi chọn độ lớn thời gian cho từng phân đoạn là 7 ngày và giá trị này có thể được điều chỉnh khi chạy thực nghiệm hệ thống. Tuy nhiên vấn đề gặp phải khi phân đoạn dữ liệu mạng xã hội thành từng phân đoạn theo thời gian là độ lớn của mỗi phân đoạn (độ lớn thời gian) chỉ mang giá trị tương đối, dẫn đến khả năng bỏ sót một số từ khóa quan trọng khi nó bị chia cắt ở hai hay nhiều phân đoạn liên tiếp. Ta xét một ví dụ đơn giản như bên dưới:

Giả sử hệ thống thu thập được dữ liệu của một mạng xã hội trong 8 ngày (từ ngày 1 đến ngày 8), một từ khóa "A" xuất hiện trong 3 ngày liên tiếp là ngày 4, ngày 5 và ngày 6. Giả định hệ thống phân đoạn dữ liệu ra thành 2 phân đoạn, tức mỗi phân đoạn là 4 ngày và quy định một từ xuất hiện nhiều hơn 2 lần ở một phân đoạn sẽ được coi là một từ khóa quan trọng (trend word).

Bảng 4.1.1: Phân đoạn dữ liệu trên mạng xã hội

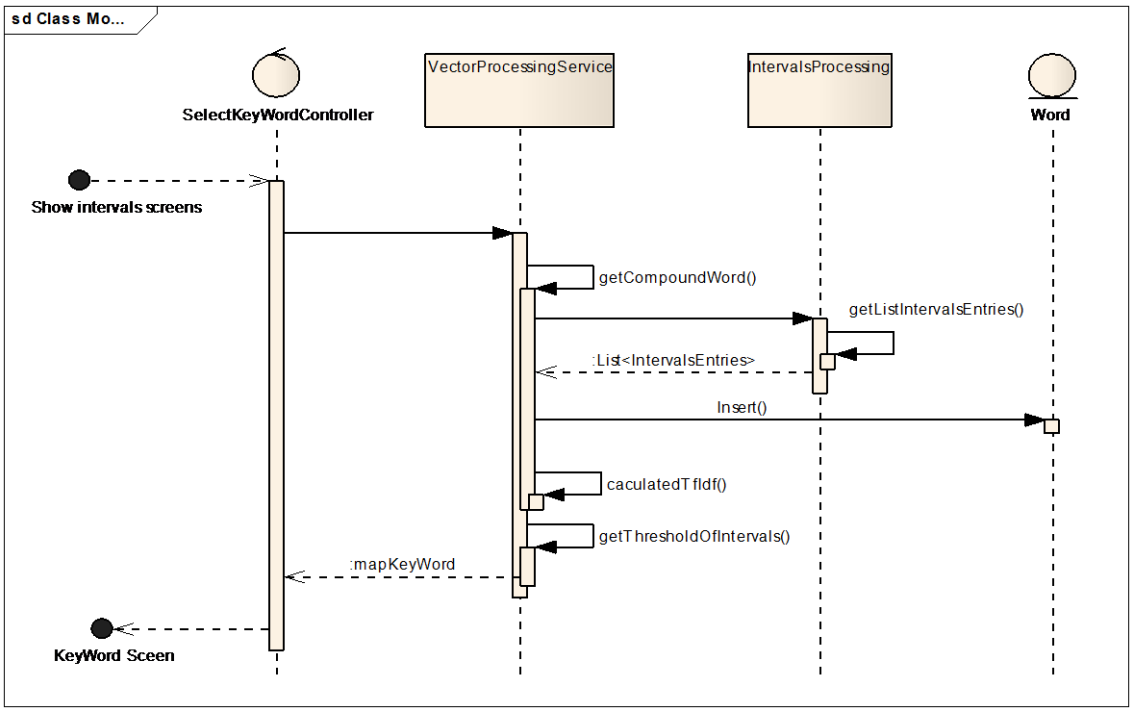


Ngày	1	2	3	4	5	6	7	8
Dữ liệu				A	A	A		

Từ khóa “A” xuất hiện trong phân đoạn thứ nhất với tần suất là 1 và trong phân đoạn thứ hai với tần suất là 2, rõ ràng với cách phân đoạn như vậy thì hệ thống sẽ không tìm được trend word “A” cho dù “A” là một trend word theo như quy định ở trên. Để khắc phục được vấn đề này chúng tôi đưa ra ý tưởng là phân đoạn mạng xã hội ra nhiều phân đoạn theo thời gian nhưng các phân đoạn này phải phủ lên nhau (*overlap*) một khoảng thời gian. Tiếp tục xét ví dụ ở trên nhưng lần này hệ thống phân đoạn mạng xã hội với phân đoạn là 4 ngày và mỗi phân đoạn có ngày bắt đầu phủ lên nhau một ngày, khi đó hệ thống sẽ có được các phân đoạn là [1→4], [2→5], [3→6], [4→7], … Với kết quả phân phân đoạn này thì từ khóa “A” xuất hiện với tần suất 3 lần trong phân đoạn [4→7] do đó hệ thống sẽ xác định một trend word “A” trong phân đoạn từ ngày 4 đến ngày 7.

4.1.3 Tiền xử lý văn bản và Tìm từ khóa quan trọng

Các chức năng tiền xử lý văn bản và tìm từ khóa quan trọng được mô tả trong sơ đồ sau:

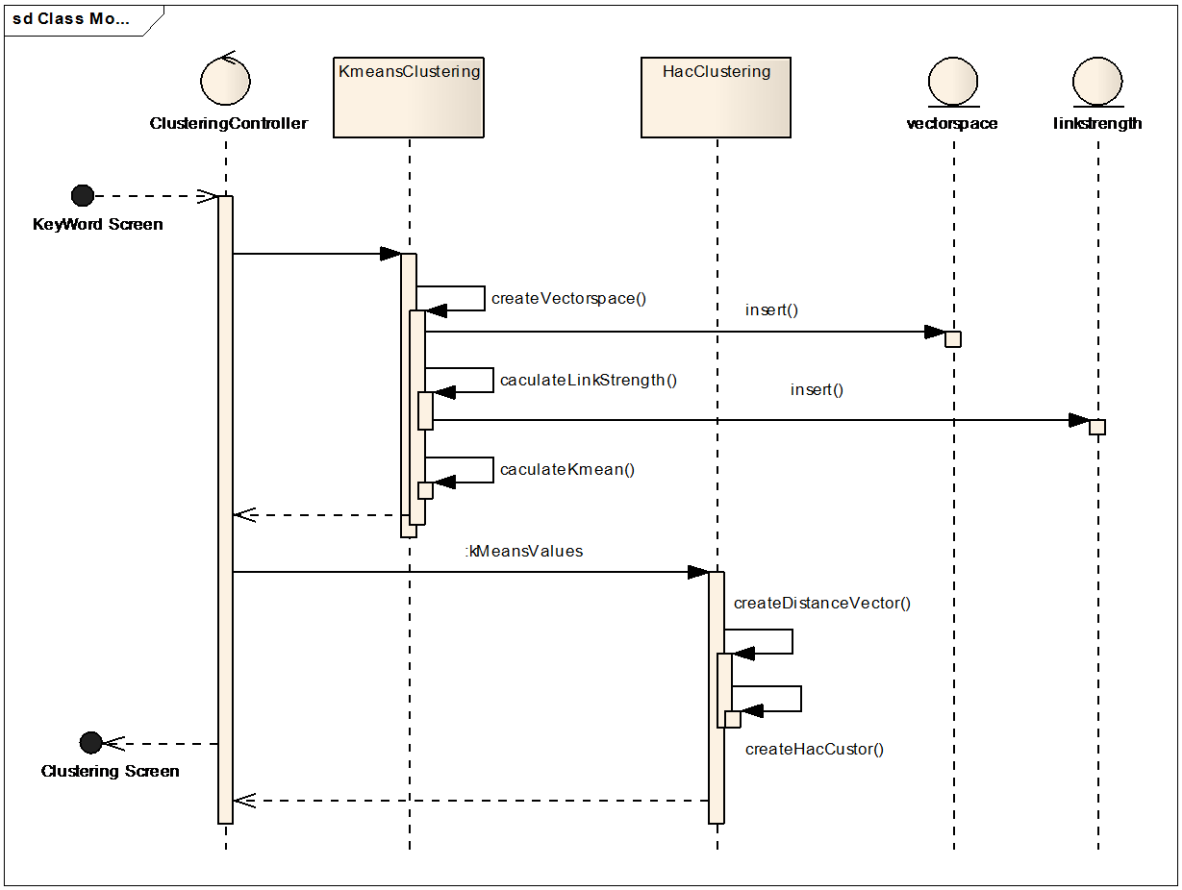


Hình 4.1.3: Sơ đồ sơ đồ mô tả chức năng của similarity module và scoring module

Từ màn hình phân đoạn dữ liệu theo thời gian. Khi click vào nút tìm từ khóa quan trọng hệ thống sẽ gọi controller SelectKeyWordController. Controller này sẽ gọi VectorProcessingService để tiến hành xử lý. Trước tiên hệ thống tách các từ trong phân đoạn, lấy từ ghép, loại bỏ stop words và thêm vào cơ sở dữ liệu. Sau đó tiến hành tính trọng số tf-idf cho từ trong từng intervals và trả kết quả về controller. Controller sẽ gọi trang KeyWord để hiển thị kết quả ra màn hình.

4.1.4 Phát hiện xu hướng:

Dựa trên các từ khóa quan trọng đã tìm được hệ thống bắt đầu thực hiện gom nhóm các từ khóa liên quan. Sơ đồ sau trình bài sự tương tác của hệ thống trong quá trình gom cụm:



Hình 4.1.4: Sơ đồ sơ đồ mô tả chức năng của Trend detection

Từ màn hình các từ khóa quan trọng nhấn vào nút gom cụm dữ liệu hệ thống sẽ gọi controller ClusteringController controller này sẽ gọi KmeanClustering để tiến hành xử lý. Sau khi tạo không gian vector và thêm vào cơ sở dữ liệu hệ thống sẽ tiến tính toán linkstrength. Kế tiếp, hệ thống tiến hành gom nhóm sử dụng thuật toán k-means và trả kết quả về cho controller. Controller sử dụng kết quả của kmeans để tạo ma trận khoảng cách trong HacClustering và tiến hành gom nhóm bằng thuật toán HAC. Cuối cùng, kết quả sẽ được hiển thị trên màn hình Clustering.

Dưới đây sẽ trình bày giải thuật kết hợp thuật toán k-means và HAC:

Input: Danh sách các từ khóa quan trọng $L_p = \{p_1,\ldots, p_n\}$ với p_i là một từ khóa

Output: danh sách cụm HAC $L_{hac\_cluster}$ với mỗi hac_cluster chứa một danh sách các từ khóa.

Process:

1: **begin**

2: $L_{hac\_cluster} \leftarrow Ø$

3: $L_{k\text{-means\_cluster}} \leftarrow Ø$

4: Apply *k*-means algorithm on L_p

5: $L_{k\text{-means\_cluster}} \leftarrow$ result of *k*-means algorithm on L_p

6: **for** each *k*-means_cluster $c_i \in \{c_1,\ldots, c_n\}$ in $L_{k\text{-means\_cluster}}$ **do**

7: $L_{i\_hac\_cluster} \leftarrow$ Apply HAC algorithm on c_i

8: Merge $L_{i\_hac\_cluster}$ to $L_{hac\_cluster}$

9: **end for**

10: **end**

Kết quả của hệ thống bị ảnh hưởng bởi cách chọn hệ số k và chọn tâm phù hợp trong bước gom cụm bằng thuật toán *k*-means. Hiện nay vẫn chưa có giải pháp nào được xem là tốt về tính khoa học để chọn hệ số k này. Thông thường để chọn hệ số k phù hợp với từng hệ thống, trong thực tế người ta hay sử dụng các phương pháp sau:

- Thử hệ thống với các giá trị của k, từ đó chọn k cho kết quả phân cụm tốt nhất. Hệ thống phát hiện xu hướng trên mạng xã hội của chúng tôi chọn hệ số k theo phương pháp này.

- Tham khảo ý kiến của các chuyên gia. Thông thường các chuyên gia trong một lĩnh vực nào đó sẽ có cái nhìn (ban đầu) về dữ liệu cần phân cụm và đề xuất giá trị cho hệ số k.

Chương 5: THỰC NGHIỆM

Trong chương này chúng tôi sẽ trình bày về cách tạo tập dữ liệu thí nghiệm cho hệ thống phát hiện xu hướng, tổng hợp các kết quả từ hệ thống. Cuối cùng tiến hành đánh giá độ chính xác và tốc độ của hệ thống.

5.1 Kết quả thí nghiệm

5.1.1 Cách xây dựng tập dữ liệu thí nghiệm

Tập dữ liệu được thu thập từ cộng đồng tin tức 24h trên mạng xã hội facebook. chúng tôi chọn và tạo ra các tập dữ liệu con nhỏ hơn với mỗi tập dữ liệu có độ lớn về thời gian là 1 tháng.

Để đánh giá độ chính xác của hệ thống chúng tôi xác định thủ công trước các chủ đề được người dùng bàn luận nhiều trong mỗi tập dữ liệu. Sau đó tiến hành chạy hệ thống trên từng tập dữ liệu đã chọn và so sánh kết quả các chủ đề nổi lên mà hệ thống phát hiện được so với các chủ đề đã được xác định trước. Đồng thời so sánh kết quả chạy của hai phương pháp.

Để đánh giá về tốc độ chúng tôi tiến hành chạy độc lập và đo tốc độ của hai phương pháp:

- Phương pháp 1: Chỉ chạy độc lập phương pháp gom cụm HAC cho bước gom nhóm các chủ đề.
- Phương pháp 2: Kết hợp hai phương pháp gom cụm *k*-means và HAC cho bước gom nhóm các chủ đề.

5.1.2 Kết quả thí nghiệm

Sau khi chạy hệ thống trên 4 tập dữ liệu thí nghiệm, chúng tôi tổng hợp được các kết quả như sau:

Bảng 5.1.2.1 So sánh kết quả về thời gian chạy giữa hai phương pháp gom cụm

Độ lớn	**HAC - *Kmeans***	**HAC**
214 từ	2 mili giây	10 mili giây
460 từ	3 mili giây	7 mili giây

740 từ	22 mili giây	31 mili giây
992 từ	81 mili giây	110 mili giây

Tập Dữ liệu 1:

Đầu vào: 214 từ và độ rộng về thời gian là 1 tháng

Kết quả:

Bảng 5.1.2.1 So sánh về kết quả chạy giữa hai phương pháp gom cụm với tập dữ liệu 1

Kmeans-HAC	HAC
[U23 - VN - HLV - Miura - công phượng]	[U23 - VN - HLV - Miura - công phượng]
[Pháp - Airbus - A320 - rơi - máy bay - đâm - cơ phó]	[Pháp - Airbus - A320 - rơi - máy bay - đâm - cơ phó]
[nạn nhân - Formosa - tử vong - sập - giàn giáo]	[nạn nhân - Formosa - tử vong - sập - giàn giáo]

Tập Dữ liệu 2:

Đầu vào: 460 từ và độ rộng về thời gian là 1 tháng

Kết quả:

Bảng 5.1.2.2 So sánh về kết quả chạy giữa hai phương pháp gom cụm với tập dữ liệu 2

Kmeans-HAC	HAC
[chị ve chai - giấy tờ - bà - triệu - Yên - Ngọt]	[Yên - giấy tờ - bà - chị ve chai - triệu - Ngọt]
[xúc động - qua đời - vợ - người - duy nhân]	[duy nhân - xúc động - vợ - qua đời]
[Nepal - VN - động đất]	[người - VN - Nepal - động đất]
[thiếu nữ - sàm sỡ - công viên nước]	[sàm sỡ - công viên nước - thiếu nữ]

Tập Dữ liệu 3:

Đầu vào: 740 từ và độ rộng về thời gian là 1 tháng

Kết quả:

Bảng 5.1.2.3 So sánh về kết quả chạy giữa hai phương pháp gom cụm với tập dữ liệu 3

Kmeans-HAC	HAC
[scandal - MC - nguy kịch]	[nguy kịch - MC]
[hé lộ - gây án - người ở - nghi phạm - khai - 10 - lên tiếng - ai - thảm sát - ngôi nhà - gia đình - bình phước - nghệ an - hung thủ - người chết]	[hung thủ - nghệ an - thảm sát - bình phước- gây án - người ở - người - hé lộ - nghi phạm – khai - ai- 10 - lên tiếng - ngôi nhà - gia đình]
[lịch sử - cơn - lũ - quảng ninh - nhấn chìm]	[quảng ninh - cơn - lũ - lịch sử]

Tập Dữ liệu 4:

Đầu vào: 992 từ và độ rộng về thời gian là 1 tháng

Kết quả:

Bảng 5.1.2.4 So sánh về kết quả chạy giữa hai phương pháp gom cụm với tập dữ liệu 4

Kmeans-HAC	HAC
[ánh viên - giành - Singapore - kình ngư - HCV]	[giành - HCV- ánh viên - kình ngư - Singapore -]
[dài - 2.000 - phim - bá đạo - tập - cô dâu]	[phim - cô dâu - 2.000- dài - bá đạo - tập]
[U23 - indonesia - sea games - trực tiếp - việt nam]	[sea games - trực tiếp - U23 - việt nam - indonesia]
[tức tưởi - bạn trai - chết - mạng - tung - 15 - nữ sinh - sex]	[nữ sinh - chết - tức tưởi - bạn trai - mạng - tung - 15 - sex]

[Mỹ - hợp pháp hóa - hôn nhân - đồng giới - chính thức - cộng đồng - LGBT]	[Mỹ - hợp pháp hóa - hôn nhân - đồng giới - chính thức - cộng đồng - LGBT]
[đầu - trọc - thái lan - tuấn hưng - cạo - mr đàm - tin nhắn - quang lê]	[thái lan - tuấn hưng - đầu - trọc - cạo - mr đàm - tin nhắn - quang lê]

5.2 Đánh giá

Dựa vào kết quả thu thập được sau khi chạy hệ thống trên các bộ dữ liệu thí nghiệm, chúng tôi tiến hành đánh giá hệ thống về 2 tiêu chí là độ chính xác và tốc độ:

- Về độ chính xác của hệ thống

Vậy nếu đánh giá tổng thể trên toàn bộ các tập thí nghiêm kết quả chạy giữa hai phương pháp gom cụm giống nhau trong khoảng 90 -95%. .

- Tốc độ

So sánh tốc độ giữa phương pháp 1 - chạy độc lập thuật toán gom cụm HAC và phương pháp 2 - kết hợp giữa *k*-means và HAC đối với các bộ thí nghiệm được trình bày ở trên, tốc độ của phương pháp 2 (phương pháp kết hợp) nhanh gấp khoảng 2.2 lần phương pháp 1.

Kết luận

Trong khuôn khổ của đề tài những kết quả đánh giá về độ chính xác và tốc độ ở trên chỉ mang tính chất tương đối vì độ chính xác của hệ thống phụ thuộc vào nhiều giá trị khởi tạo đầu vào như cách chọn hệ số k khi tiến hành thuật toán *k*-means và điều kiện dừng của thuật toán HAC. Đối với tốc độ, vì thuật toán *k*-means có độ phức tạp thấp là O(t*k*n) với t và k rất nhỏ so với n và thuật toán HAC trong trường hợp tổng quát có độ phức tạp lớn là $O(n^3)$ do đó nếu tập dữ liệu càng lớn, tức giá trị n càng lớn thì sự tối ưu về tốc độ của phương pháp kết hợp càng được thể hiện rõ nét hơn.

Chương 6: KẾT LUẬN

6.1 Tổng kết

Các công việc về phát hiện xu hướng và thông tin nổi lên trên mạng xã hội đang thu hút nhiều sự quan tâm và nghiên cứu. Kết quả của những nghiên cứu này có ý nghĩa thật sự quan trọng trong việc giúp chúng ta có thể hiểu tốt hơn những mối quan tâm của xã hội và giúp các công ty có những chiến lược quảng cáo hiệu quả nhất.

Trong hướng tiếp cận nghiên cứu này chúng tôi đưa ra một phương pháp mới, sử dụng phương pháp gom cụm (*clustering*) trong khai phá dữ liệu (*data mining*) kết hợp với thông tin thời gian (*temporal information*) để phát hiện những xu hướng nổi lên trên mạng xã hội. Những mục tiêu chính đạt được trong nghiên cứu này được tóm tắt như sau:

- Xây dựng thành phần tương tác với cơ sở dữ liệu của mạng xã hội, thành phần này chia khối lượng dữ liệu rất lớn của mạng xã hội thành nhiều phân đoạn theo thời gian. Với cách phân đoạn này sẽ giúp cho việc thao tác và tính toán trên tập dữ liệu của mạng xã hội được cải tiến rất lớn về mặt tốc độ.

Xây dựng thành phần tiền xử lý văn bản, đảm nhiệm việc tiền xử lí văn bản như loại bỏ từ dừng (*stop-words*) và lấy từ ghép trong tiếng Việt.

- Xây dựng thành phần phát hiện những từ khóa quan trọng dùng kỹ thuật vector trọng số tf.

- Xây dựng thành phần gom cụm các từ khóa quan trọng để hình thành nên các cụm chủ đề nổi lên. Kết hợp hai phương pháp gom cụm *k*-means và HAC để gom nhóm các từ khóa quan trọng tìm được ở bước áp dụng vector trọng số tf. Áp dụng thuật toán *k*-means để làm giảm không gian bài toán ở bước đầu tiên, sau đó tiếp tục áp dụng phương pháp HAC trên từng cụm kết quả của bước *k*-means.

- Xây dựng thành phần xuất kết quả các xu hướng được phát hiện bởi hệ thống.

6.2 Hướng phát triển

Với khối lượng dữ liệu ngày càng bùng nổ trong các mạng xã hội và những đặc thù của loại hình mạng tương tác trực tuyến này, những đề xuất về phương pháp tiếp cận và kỹ thuật được sử dụng để xây dựng hệ thống phát hiện xu hướng nổi lên trong khuôn khổ nghiên cứu của chúng tôi là những bước nền tảng ban đầu. Để có được những kết quả thu được tốt nhất cho các hệ thống phát hiện xu hướng xây dựng trên mạng xã hội, chúng ta cần cảm nhận thực tế, phân tích và đưa ra nhiều kỹ thuật khác nhau phù hợp cho từng mạng xã hội, sau đó tiến hành so sánh những kết quả thực nghiệm để có được phương pháp tốt nhất. Dựa trên tinh thần đó chúng tôi đề xuất hướng phát triển mở rộng của đề tài như sau:

- Nghiên cứu mở rộng và đánh giá các kết quả thí nghiệm để chọn ra các kỹ thuật tốt nhất làm tăng hiệu quả của 2 thành phần phát hiện những từ khóa quan trọng trong từng phân đoạn dữ liệu mạng xã hội và thành phần tính toán khoảng cách khi gom cụm.

TÀI LIỆU THAM KHẢO

[11] A. Porter and M. Detampel (1995), “Technology opportunities analysis” , Technological Forecasting and Social Change, vol. 49, pp. 237-255.

[16] A. Popescul, G. Flake, S. L. S., L. Ungar, and C. Giles (2000), “Clustering and identifying temporal trends in document databases”, IEEE Advances in Digital Libraries, pp. 173-182.

[7] Ceren Budak, Divyakant Agrawal and Amr El Abbadi (2011), “Structural Trend Analysis for Online Social Networks”, Proceedings of the VLDB Endowment,Vol. 4, (No. 10), Pages 646-656.

[5] Cuneyt Gurcan Akcora, Murat Ali Bayir and Murat Demirbas. Trend sensing via Twitter. International Journal of Ad Hoc and Ubiquitous Computing, List of Issues, Volume 14, Issue 1, 2013, pages 16 - 26.

[17] D. J. R. Swan (2003), “TimeMines: Constructing timelines with statistical models of word usage”, Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (Boston, MA, USA).

[3] Duc T. Nguyen and Jai E. Jung. Privacy-Preserving Discovery of Topic-Based Events from Social Sensor Signals: An Experimental Study on Twitter. The Scientific World Journal Volume 2014 (2014), Article ID 204785, 5 pages.

[13] G. Blank, W. Pottenger, G. Kessler, M. Herr, H. Jaffe, S. Roy, D. Gevry,

and Q. Wang (2001), “CIMEL: Constructive, collaborative inquiry-based multimedia e-learning”, Proceedings of the 6th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE), (United Kingdom) page 179.

[2] Erich Schubert, Michael Weiler and Hans-Peter Kriegel. SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds. KDD '14 Proceedings of the 20th ACM SIGKDD international

conference on Knowledge discovery and data mining. Pages: 871-880. Publisher: ACM New York, NY, USA ©2014.

[4] James Benhardus and Jugal Kalita Streaming trend detection in Twitter. International Journal of Web Based Communities, List of Issues, Volume 9, Issue 1, 2013, pages 122 - 139.

[19] J. Allan, R. Papka, and V. Lavrenko (1998), "On-line new event detection and tracking", Proceedings of ACM SIGIR, pp. 37-45.

[12] L. Nowell, R. France, D. H. an L.S. Heath, and E. A. Fox (1996), "Visualizing search results: Some alternatives to query-document similarity", Proceedings of SI-GIR'96, (Zurich, Switzeland) pages 67-75.

[9] Mario Cataldi, Luigi Di Caro and Claudio Schifanella (2010), "Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation", ACM New York, NY, USA.

[6] Matthew A. Russell (2011), *Mining the Social Web,* O'Reilly Media, 1005 Gravenstein Highway North, Sebastopol.

[20] Q.T. Tho, A.C.M. Fong, S.C. Hui, (2007) "A scholarly semantic web system for advanced search functions", Online Information Review, Vol. 31 No.3, pp.353 - 364.

[14] R. Bader, M. Callahan, D. Grim, J. Krause, N. Miller, and W. Pottenger (2001), "The role of the HDDI collection builder in hierarchical distributed dynamic indexing", Proceedings of the Textmine'01 Workshop, First SIAM International Conference on Data Ming.

[18] S. Havre, E. Hetzler, P. Whitney, and L. Nowell (2002), "Themeriver: Visualizing the-matic changes in large document collection", IEEE Transactions on Visualization and Computer Graphics, vol. 8, no. 1 pp 9 - 20.

[8] Shiva Prasad Kasiviswanathan, Prem Melville, Arindam Banerjee and Vikas Sindhwani (2011), "Emerging topic detection using dictionary learning" ACM New York, NY, USA pages 745-754.

[10] W. P. S. R. D. P. A. Kontostathis, L. Galitsky (2003), "A Survey of Emerging Trend Detection in Textual Data Mining", A Comprehensive Survey of Text Mining, Springer-Verlag pp 185-224.

[15] W. Pottenger and T. Yang (2001), "Detecting Emerging Concepts in Textual Data Mining", Computational Information Retrieval, Philadelphia, USA: SIAM pages 89-105.

[1] http://wearesocial.net/tag/vietnam/

[21] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer