ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

HỒ TRUNG THÀNH

PHÂN TÍCH MẠNG XÃ HỘI DỰA THEO MÔ HÌNH CHỦ ĐỀ VÀ ỨNG DỤNG

Chuyên ngành Khoa học máy tính
Mã số: 62.48.01.01

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

TP. HỒ CHÍ MINH - NĂM 2017

Công trình được hoàn thành tại **Trường Đại học Công nghệ Thông tin, Đại học Quốc gia TP.HCM.**

Người hướng dẫn khoa học: PGS. TS. Đỗ Phúc

Phản biện 1:
Phản biện 2:
Phản biện 3:

Luận án sẽ được bảo vệ trước
Hội đồng chấm luận án cấp Trường tại:
……
……

Vào lúc …… giờ …….. ngày …….. tháng ……..năm …….

Có thể tìm luận án tại:
- Thư viện Quốc gia Việt Nam.
- Thư viện Trường Đại học Công nghệ Thông tin, ĐHQG-HCM.

TỔNG QUAN VỀ LUẬN ÁN

1. Động cơ nghiên cứu

Mục tiêu phân tích mạng xã hội (MXH) là phân tích sự tương tác giữa con người, tổ chức với nhau và khám phá những thông tin, tri thức tiềm ẩn thông qua sự tương tác đó [27][28][41][59]. Xu hướng gần đây, các nghiên cứu tập trung vào khai thác và phân tích MXH. MXH đã phát triển nhanh chóng vì cho phép cá nhân, tổ chức tương tác dễ dàng. Chính MXH đã tạo nên sự không lệ thuộc vào không gian và thời gian khi giao tiếp của cá nhân và cộng đồng. Mỗi cá nhân trên MXH đều có thể kết bạn và trò chuyện với bất kỳ một cá nhân khác trên cùng MXH đó. Một số MXH trực tuyến điển hình như Facebook, LinkedIn, MySpace, Twitter. Các MXH này mang lại lượng lớn dữ liệu là thông điệp trao đổi của cá nhân thông qua các liên kết xã hội. Có thể biểu diễn dữ liệu này bằng cấu trúc đồ thị của MXH và nội dung dữ liệu là thông tin trao đổi giữa các thành viên trên MXH trong đó bao gồm dữ liệu thông điệp, dữ liệu đa phương tiện,... Đây chính là nguồn dữ liệu để phân tích MXH tìm ra những thông tin, tri thức tiềm ẩn được chứa đựng trong dữ liệu trên MXH.

Thông điệp được cá nhân trao đổi trên MXH, diễn đàn hay hệ thống e-mail có sự pha trộn nhiều chủ đề. Chủ đề trong thông điệp được cá nhân quan tâm trao đổi và chia sẻ tạo nên sự lan truyền thông tin từ cá nhân này đến cá nhân khác hình thành cộng đồng MXH cùng quan tâm đến các chủ đề. Khai thác chủ đề quan tâm của cá nhân cũng như phân tích mối liên kết xã hội giữa các cá nhân qua những thông điệp, dữ liệu trao đổi là một công việc nhiều thách thức, đặc biệt chủ đề thường xuyên được thay đổi theo thời gian hoặc đôi khi một chủ đề có thể được trao đổi thường xuyên, liên tục trong một khoảng thời gian nào đó. Bên cạnh đó, chủ đề của thông điệp được thảo luận có thể là khác nhau tuỳ theo sở thích, hành vi, mức độ quan tâm, trao đổi của từng cá nhân theo từng giai đoạn thời gian. Khám phá chủ đề quan tâm và phân tích vai trò của cá nhân trên MXH là một thách thức đặt ra cho bài toán với mục tiêu trả lời được các câu hỏi "cá nhân đã trao đổi chủ đề gì trên MXH theo thời gian?", "mức độ quan tâm của cá nhân đến chủ đề cụ thể như thế nào?", "có bao nhiêu cá nhân quan tâm đến chủ đề?", "chủ đề nào được quan tâm nhiều nhất?" và "có thay đổi gì về sự quan tâm đến các chủ đề của cá nhân theo từng giai đoạn thời gian?".

Bên cạnh việc khám phá vai trò cá nhân trên MXH, một thách thức khác đặt ra là phân tích MXH để khám phá nhóm cá nhân (cộng đồng) cùng quan tâm chủ

đề theo từng giai đoạn thời gian. Khám phá nhóm cá nhân hay khám phá cộng đồng là cách để nhận biết nhóm các cá nhân có mối liên kết xã hội với nhau trên MXH và cùng chủ đề quan, đồng thời giúp hiểu được sự quan tâm của từng cá nhân trong cộng đồng MXH theo từng chủ đề cụ thể. Những thay đổi xảy ra trong cộng đồng thường liên quan đến các đặc trưng của cộng đồng như: chủ đề quan tâm, số cá nhân tham gia cộng đồng, mức độ quan tâm chủ đề của cộng đồng tại từng thời điểm khác nhau, và sự thay đổi chủ đề quan tâm trong cộng đồng dẫn đến thay đổi hành vi, sự quan tâm và trao đổi chủ đề của các cá nhân trong cộng đồng. Vấn đề đặt ra là làm thế nào để có thể khám phá cộng đồng cá nhân cùng quan tâm đến một hay một nhóm chủ đề thông qua những nội dung thông điệp được trao đổi của tập cá nhân trên MXH? Với một hay nhóm chủ đề cụ thể có những cộng đồng nào trên MXH quan tâm trao đổi? Sự biến thiên chủ đề quan tâm và cá nhân tham gia cộng đồng? Tìm giải pháp cho các câu hỏi này rõ ràng là việc không đơn giản nhưng kết quả nghiên cứu sẽ giúp cho việc phân tích và khám phá chủ đề được cá nhân quan tâm hay tìm ra những cá nhân có ảnh hưởng trong cộng đồng để phục vụ cho những chiến lược phát triển như quản lý cộng đồng cá nhân của công ty, tổ chức hay của một quốc gia; hiểu cá nhân để thực hiện chiến lược tiếp thị hiệu quả, phát triển loại hình đào tạo trực tuyến trong trường đại học,...

2. Mục tiêu nghiên cứu

Từ những động cơ nghiên cứu trên, luận án xây dựng hai mục tiêu chính và các nhiệm vụ nghiên cứu cụ thể. Trong đó, luận án xây dựng các mô hình và phương pháp trong phân tích MXH dựa theo mô hình chủ đề (Topic model) để khám phá chủ đề quan tâm, vai trò của cá nhân và cộng đồng trên MXH theo từng giai đoạn thời gian. Cụ thể hai mục tiêu chính sau:

(i) Xây dựng mô hình khám phá chủ đề quan tâm của cá nhân trên MXH dựa theo mô hình chủ đề có yếu tố thời gian và phân tích sự biến thiên chủ đề quan tâm của cá nhân.

Nhiệm vụ nghiên cứu của mục tiêu (i) bao gồm:

- Xây dựng phương pháp gán nhãn chủ đề quan tâm của cá nhân theo thời gian dựa trên cây phân cấp chủ đề (Topic Taxonomy).
- Xây dựng mô hình Temporal-Author-Recipient-Topic (TART) dựa theo mô hình chủ đề để khám phá chủ đề quan tâm và phân tích vai trò của cá nhân trên MXH đối với từng chủ đề quan tâm cụ thể theo từng giai đoạn thời gian.

(ii) Xây dựng phương pháp khám phá cộng đồng (gom cụm cá nhân có cùng đặc trưng: chủ đề, mức độ và thời gian quan tâm chủ đề) trên MXH dựa theo mô hình chủ đề có yếu tố thời gian và phân tích sự biến thiên những đặc trưng trong cộng đồng MXH.

Nhiệm vụ nghiên cứu của mục tiêu (ii) bao gồm:

- Xây dựng phương pháp khám phá cộng đồng trên MXH có cùng các chủ đề quan tâm theo từng giai đoạn thời gian.
- Xây dựng phương pháp khảo sát sự biến thiên các đặc trưng của cộng đồng. Trong đó, luận án tập trung vào hai đặc trưng là chủ đề quan tâm và cá nhân tham gia cộng đồng.

Các đối tượng trọng tâm trong nghiên cứu của luận án:

- Mô hình chủ đề Latent Dirichlet Allocation (LDA).
- Các phương pháp, mô hình phân tích MXH dựa theo mô hình chủ đề.
- Các liên kết xã hội: chủ đề và thông điệp được cá nhân trao đổi trên MXH.
- Vai trò (cá nhân và cộng đồng): cá nhân là người gửi, người nhận, chủ đề và cộng đồng là nhóm những cá nhân có cùng sự quan tâm trao đổi các chủ đề.
- Thời gian cá nhân quan tâm đến chủ đề thông qua liên kết xã hội là thông điệp.

Từ hai mục tiêu chính và các nhiệm vụ nghiên cứu, hai bài toán chính được đặt ra trong phạm vi luận án, bao gồm:

Bài toán 1. Khám phá chủ đề quan tâm của cá nhân dựa theo mô hình chủ đề có yếu tố thời gian.

Xây dựng mô hình TART dựa theo mô hình chủ đề để khám phá chủ đề quan tâm và phân tích vai trò của cá nhân trên MXH theo từng giai đoạn thời gian và xây dựng phương pháp gán nhãn chủ đề ẩn dựa trên cây phân cấp chủ đề.

Bài toán 1 được chia làm hai bài toán nhỏ: (i) Bài toán 1.1. Khám phá và gán nhãn chủ đề ẩn từ thông điện trên MXH; (ii) Bài toán 1.2. Khám phá chủ đề quan tâm của cá nhân trên MXH có yếu tố thời gian.

Nội dung thực hiện của bài toán 1.1 bao gồm:

- Nghiên cứu cho trường hợp dữ liệu là thông điệp tiếng Việt trên MXH. Thông điệp trên MXH chứa đựng nhiều từ viết tắt, từ không rõ nghĩa, các ký hiệu. Trước khi phân tích thông điệp, luận án phải tiến hành tiền xử lý dữ liệu bằng cách lọc đi những hư từ (stopwords) và hệ thống các từ viết

tắt và ký hiệu được ánh xạ sang từ rõ nghĩa, từ đó hiểu được nội dung thông điệp để phân tích.

- Các nghiên cứu truyền thống xem mỗi thông điệp chỉ thuộc về duy nhất một chủ đề. Tuy nhiên, theo tiếp cận mô hình chủ đề, mỗi thông điệp tiềm ẩn nhiều chủ đề và mỗi chủ đề được đặc trưng bởi tập từ đồng hiện trong thông điệp. Như vậy, vấn đề đặt ra là làm thế nào để khám phá chủ đề ẩn trong thông điệp?
- Chủ đề ẩn được khám phá từ thông điệp chưa được gán nhãn (tên của chủ đề). Như vậy, để gán nhãn và chỉ rõ được chủ đề trao đổi, bài toán 1.1 xây dựng phương pháp xây dựng cây phân cấp chủ đề và phương pháp học máy SVM để gán nhãn chủ đề.

  Bài toán 1.1 được trình bày chi tiết trong chương 2.

Nội dung thực hiện của Bài toán 1.2 bao gồm:

- Xây dựng mô hình Khám phá chủ đề, phân tích mức độ quan tâm chủ đề của cá nhân.
- Phân tích vai trò của cá nhân quan tâm chủ đề trên MXH theo từng giai đoạn thời gian.
- Dùng yếu tố thời gian để chia nhỏ các yếu tố trong mô hình ART như tập cá nhân gửi, tập cá nhân nhận, tập chủ đề và tìm ra được sự thay đổi chủ đề quan tâm của cá nhân trong tập thông điệp theo từng khoảng thời gian so với chủ đề quan tâm trong kho ngữ liệu thông điệp.
- Khảo sát sự biến thiên chủ đề quan tâm của từng cá nhân để chỉ ra trong từng giai đoạn thời gian từng cá nhân quan tâm đến chủ đề gì. Tìm ra chủ đề được cá nhân quan tâm nhiều nhất trên MXH.

  Bài toán 1.2 được trình bày chi tiết trong chương 2.

Bài toán 2. Khám phá chủ đề quan tâm của cộng đồng dựa theo mô hình chủ đề có yếu tố thời gian.

Xây dựng phương pháp khám phá cộng đồng trên MXH có cùng các chủ đề quan tâm theo từng giai đoạn thời gian và phương pháp khảo sát sự biến thiên các đặc trưng của cộng đồng [CB01][CB06][CB10].

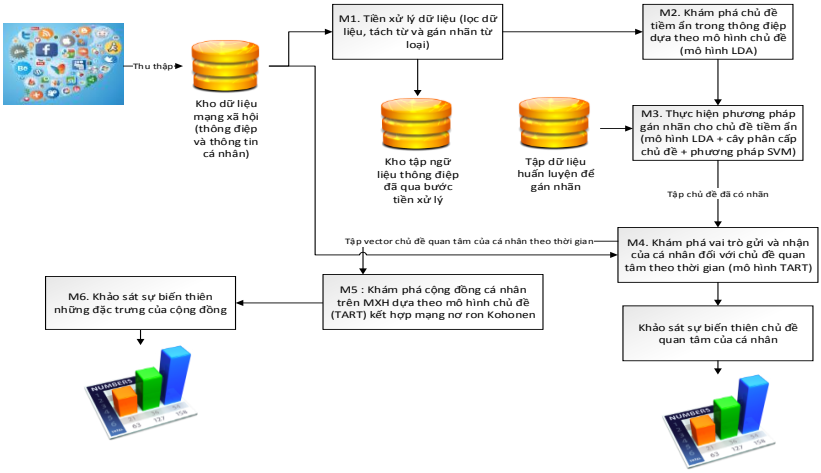
Nội dung thực hiện bài toán 2 bao gồm:

- Một cộng đồng quan tâm nhiều chủ đề và một chủ đề có nhiều cộng đồng quan tâm. Theo tính chất của MXH, nhiều người trao đổi với nhiều người khác về một hay một số chủ đề mà không chịu sự giới hạn của không gian và thời gian. Mục tiêu tìm ra đặc trưng của cộng đồng bao gồm: chủ đề quan tâm, số cá nhân và mức độ quan tâm đến từng chủ đề cụ thể.

- Các đặc trưng của cộng đồng sẽ thay đổi theo từng giai đoạn thời gian. Vì vậy, khảo sát sự biến thiên của đặc trưng chủ đề như: số chủ đề quan tâm, số cá nhân tham gia cộng đồng trong từng thời gian để tìm ra xu thế quan tâm chủ đề của cá nhân và cộng đồng trên MXH.
- Trực quan hoá kết quả khám phá cộng đồng cũng là vấn đề được xem xét trong bài toán 2.
- Xây dựng phương pháp phân tích sự biến thiên các đặc trưng của cộng đồng.

Bài toán 2 được trình bày chi tiết trong chương 3.

3. Sơ đồ nghiên cứu tổng thể luận án



Hình 2. Sơ đồ nghiên cứu tổng thể của luận án

Hình 2 trình bày quy trình nghiên cứu của luận án. Sơ đồ được chia làm 6 mô-đun.

4. Đóng góp của luận án

- Áp dụng mô hình chủ đề vào phân tích MXH để khám phá chủ đề từ nội dung thông điệp trên MXH. Kết quả thể hiện trong các công bố [CB07][CB08][CB09]. Luận án xây dựng phương pháp kết hợp khám phá và gán nhãn chủ đề ẩn từ mối liên kết xã hội là thông điệp được cá nhân trao đổi trên MXH và gán nhãn chủ đề dựa trên cây phân cấp chủ đề. Phương pháp này còn làm nền tảng cho những nghiên cứu tiếp theo về việc khám phá chủ đề, phân tích nội dung và gán nhãn chủ đề nhằm tìm ra những tri thức mới từ các mối liên kết xã hội. Kết quả này được thể hiện trong các công bố [CB03][CB04].

- Xây dựng mô hình TART để khám phá vai trò của cá nhân quan tâm chủ đề dựa theo mô hình chủ đề có yếu tố thời gian. Mô hình này đóng vai trò quan trọng trong việc tìm ra các liên kết xã hội của cá nhân trên MXH dựa theo mô hình chủ đề thông qua việc phân tích chủ đề của thông điệp. Mô hình TART độc lập với ngôn ngữ. Kết quả của đóng góp này được thể hiện trong các công bố [CB02][CB05].
- Xây dựng phương pháp khám phá cộng đồng cá nhân dựa theo mô hình chủ đề. Phương pháp khám phá cộng đồng là sự kết hợp giữa mô hình TART và phương pháp mạng nơ ron Kohonen để khám phá ra các cộng đồng những cá nhân có cùng chủ đề quan tâm. Xây dựng phương pháp phân tích sự biến thiên đặc trưng của cộng đồng trên MXH theo từng giai đoạn thời gian. Kết quả này được thể hiện trong các công bố [CB01][CB06][CB10].
- Để tiến hành thử nghiệm, luận án đã xây dựng một hệ thống phần mềm phân tích MXH thực hiện đầy đủ sáu mô-đun trên sơ đồ nghiên cứu tổng thể của luận án (hình 2 phần tổng quan) từ mô-đun thu thập, tiền xử lý dữ liệu, thực nghiệm khám phá và gán nhãn chủ đề ẩn, thực nghiệm mô hình TART và phương pháp khám phá cộng đồng.

5. Bố cục của luận án

Luận án được cấu trúc thành 4 chương như sau: Giới thiệu tổng quan luận án; Chương 1 trình bàu về phân tích MXH và các nghiên cứu liên quan, nhận định chung và động lực nghiên cứu; Chương 2 trình bày chi tiết về mô hình LDA, kỹ thuật lấy mẫu Gibbs cho mô hình LDA, đề xuất phương pháp gán nhãn chủ đề; Chương 3 trình bày việc phát triển mô hình khám phá chủ đề quan tâm, phân tích vai trò của cá nhân trên MXH có yếu tố thời gian (Temporal ART - TART) và đề xuất phương pháp phân tích sự biến thiên chủ đề quan tâm của cá nhân trên MXH; Chương 4 trình bày chi tiết về đề xuất phương pháp khám phá cộng đồng dựa trên mô hình chủ đề có yếu tố thời gian. Trong đó, luận án khai thác mô hình TART và kết hợp với mạng nơ ron Kohonen để đề xuất phương pháp gom cụm cá nhân (khám phá cộng đồng) dựa trên các đặc trưng của cá nhân trên MXH như chủ đề quan tâm, xác suất và thời gian quan tâm; Cuối cùng là phần kết luận, những đóng góp của luận án và hướng phát triển. Phần cuối là phục lục.

CHƯƠNG 1. PHÂN TÍCH MẠNG XÃ HỘI VÀ CÁC NGHIÊN CỨU LIÊN QUAN

1.1 Giới thiệu chương

Mục tiêu của phân tích MXH là khám phá thông tin và tri thức tiềm ẩn từ những liên kết xã hội của cá nhân, cộng đồng. Phân tích MXH giúp các nhà nghiên cứu, nhà quản lý hiểu rõ mối quan hệ giữa các đối tượng, khám phá tri thức và tìm ra các đặc trưng, hành vi và các nguy cơ trong MXH từ những liên kết xã hội để phục vụ cho công tác nghiên cứu và quản lý. Ban đầu, phương pháp phân tích MXH thường tập trung vào việc tìm hiểu sự tương tác giữa các cá nhân trong MXH mà chưa quan tâm tới nội dung thông tin được chia sẻ. Tuy nhiên, do nhu cầu thực tế mà việc phân tích MXH theo hướng nội dung ngày càng được nhiều nghiên cứu quan tâm. Phân tích MXH để hiểu nội dung thông điệp được trao đổi trên MXH của từng cá nhân, xác định được các cộng đồng MXH, phân tích sự lan truyền thông tin trên MXH, ứng dụng MXH đồng tác giả để phân tích tìm ra lĩnh vực nghiên cứu của các nhà khoa học được đăng tải trên các bài báo khoa học và tìm kiếm chủ đề yêu thích, khai thác thái độ, suy nghĩ và hành vi của cá nhân thông qua những nội dung thảo luận trên MXH, ứng dụng phân tích những vấn đề chính trị trên MXH trong quân đội, phân tích vấn đề về hạt nhân.

1.2 Khái niệm mạng xã hội

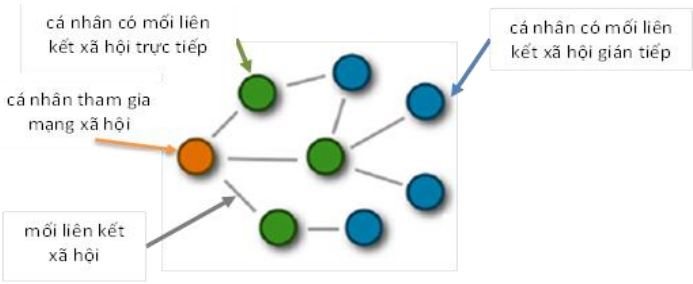
MXH là một cấu trúc xã hội của con người, có sự liên hệ trực tiếp hay gián tiếp với nhau thông qua những liên kết xã hội hoặc thông qua việc cùng quan tâm một vấn đề nào đó trong xã hội. Theo Stanley Wasserman và Katherine Faust, 1994, MXH là sự phản ánh mối quan hệ giữa các cá nhân của một xã hội trong thế giới thực vào trong máy tính được được biểu diễn ở dạng đồ thị.

MXH được mô hình hóa bằng đồ thị $G = (V,E)$ với V là tập các cá nhân (actor), E là tập các liên kết xã hội (social link) giữa các cá nhân:

- Từng cá nhân $v \in V$ có các đặc trưng, vai trò giống hay khác nhau.
- Từng liên kết $e \in E$ cùng có các loại liên kết khác nhau như: liên kết trao đổi thông tin, kết bạn, thích, chia sẻ,…
- MXH cung cấp dữ liệu với lượng lớn thông qua các liên kết xã hội.
- MXH ứng dụng trong nhiều lĩnh vực như kinh tế, giáo dục, chính trị, xã hội,…

Trong MXH, các cá nhân được liên thông qua các liên kết xã hội hay còn gọi là mối liên kết xã hội. Mối liên kết xã hội được chia làm hai loại: mối liên kết xã hội trực tiếp và mối liên kết xã hội gián tiếp. Mối liên kết xã hội trực tiếp

thông qua việc kết bạn trực tiếp hay gửi nhận thông điệp trực tiếp. Hình 1.1 biểu diễn mô hình MXH.



Hình 1.1 Mô hình MXH

Đối với mối liên kết xã hội gián tiếp là thông qua trung gian là một hay nhiều bạn nào đó để kết bạn. Để xây dựng mối quan hệ giữa các cá nhân trong một MXH cụ thể, trước tiên cần phải có phương pháp biểu diễn dữ liệu phù hợp. Trong thực tế, biểu diễn MXH thường được biểu diễn ở dạng đồ thị, phương pháp này có ưu điểm là biểu diễn mọi dạng hình thái của MXH

1.3 Phương pháp phân tích mạng xã hội

1.3.1 Khái niệm về phân tích mạng xã hội

Phân tích MXH (Social Network Analysis - SNA) là phương pháp phân tích những mối liên kết xã hội giữa người với người hay giữa người và tổ chức. Quay trở lại các nghiên cứu trước đây, SNA được thực hiện bằng phương pháp lý thuyết đồ thị và được ứng dụng trong nhiều lĩnh vực như phân tích tâm lý tổ chức, xã hội học và nhân học. SNA tập trung vào bốn mục tiêu: (i) trực quan hoá sự giao tiếp và những mối quan hệ khác nhau giữa người với người hay giữa người với tổ chức bằng các biểu đồ. Trực quan hoá MXH có truyền thống lâu đời và được một khảo sát đưa ra trong; (ii) nghiên cứu các yếu tố ảnh hưởng đến các mối quan hệ như tuổi tác, nền tảng đào tạo liên quan,...) và nghiên cứu mối tương quan giữa các mối quan hệ đó. Điều này thực hiện bằng các kỹ thuật thống kê truyền thống như phân tích mối tương quan, phương sai, phân tích các yếu tố; (iii) rút trích thông tin và khám phá tri thức trong dữ liệu là thông điệp được trao đổi trên MXH; (iv) mục tiêu thứ tư của SNA là tạo ra các khuyến nghị để cải thiện sự giao tiếp của con người và quy trình làm việc trong tổ chức.

1.3.2 Phân tích MXH theo hướng phân tích nội dung

1.3.3 Phân tích MXH dựa theo mô hình chủ đề

1.3.3.1 Khái niệm chủ đề (topic)

Một số thuật ngữ và khái niệm liên quan đến mô hình chủ đề:

- *Từ*: một từ được ký hiệu *w* là một đơn vị cơ bản của dữ liệu rời rạc, từ được định nghĩa là một phần tử của tập từ vựng được đánh chỉ mục bởi {*1, 2, . . . , V*}.
- *Thông điệp*: một thông điệp được ký hiệu *d* là tập hợp được biểu diễn bằng một dãy *N* từ $V=(w_1, w_2, ..., w_N)$ trong đó w_i là từ thứ *i* của dãy trong *d*.
- *Kho ngữ liệu:* kho ngữ liệu là tập hợp *M* thông điệp được ký hiệu là $\mathcal{D} = (d_1, d_2, ..., d_M)$ trong đó d_i là dãy từ biểu diễn cho thông điệp thứ *i* của kho ngữ liệu $\mathcal{D}$. Mỗi thông điệp $d_i \in \mathcal{D}$ chứa một tập từ *W*.
- *Chủ đề* (*theo R. Swan cùng cộng sự, 2000 và theo W.M. Pottenger cùng cộng sự, 2001*) *là*:
  - Đại diện bởi mô hình n-grams cho biết tần suất xuất hiện của từ liên tiếp nhau có trong dữ liệu của kho ngữ liệu và sự đồng hiện của từ *w*.
  - Tập các từ *w* có quan hệ ngữ nghĩa với nhau.
- *Chủ đề* (*theo mô hình chủ đề David Blei cùng cộng sự, 2003*) *là*:
  - Một phân bố của nhiều từ *w*. Những từ được phân bố trong cùng chủ đề có sự đồng hiện với nhau trong thông điệp *d*. Chủ đề trong mô hình chủ đề được ký hiệu là *z*.

Trong nghiên cứu của luận án, khái niệm chủ đề của David Blei cùng cộng sự được luận án áp dụng để xây dựng các mô hình và phương pháp.

1.3.3.2 Mô hình chủ đề trong phân tích MXH

Mô hình chủ đề cho phép kiểm tra và khai thác tập thông điệp dựa trên việc tìm kiếm và thống kê các từ có liên quan đến chủ đề trong mỗi thông điệp, và khám phá ra những chủ đề ẩn trong thông điệp đó. Mục đích của mô hình chủ đề sẽ tìm ra một mô tả từ một văn bản có nhiều chiều thành một văn bản có số chiều ít hơn. Một số tiếp cận hiện nay trong việc mô hình nội dung thông điệp bằng chủ đề dựa trên ý tưởng là tính phân bố xác suất của mỗi từ đặc trưng trong thông điệp. Phân bố này xem mỗi thông điệp là hỗn hợp nhiều chủ đề, mỗi chủ đề là sự kết hợp của nhiều từ kèm phân bố xác suất riêng cho từng từ trong chủ đề.

1.3.3.3 Một số mô hình chủ đề

i. Mô hình Latent Semantic Indexing (LSI)

ii. Mô hình Probabilistic Latent Semantic Indexing (PLSI)

iii. Mô hình chủ đề Latent Dirichlet Allocation (LDA)

Những hạn chế của mô hình PLSI được David Blei cùng cộng sự đề xuất cải tiến trong mô hình chủ đề LDA. Mô hình LDA là một mô hình sinh xác suất cho kho ngữ liệu rời rạc. Về bản chất, LDA là một mô hình mạng Bayes theo

ba cấp, trong đó mỗi thông điệp được mô tả dưới dạng kết hợp ngẫu nhiên của một tập các chủ đề. Mỗi chủ đề là một phân bố rời rạc của một tập từ.

Theo tiếp cận truyền thống xem xét một thông điệp chỉ thuộc về một chủ đề. Tiếp cận theo mô hình chủ đề chỉ ra rằng, mỗi thông điệp được biểu diễn bằng nhiều chủ đề mà thông điệp đó đề cập đến, mỗi chủ đề được biểu diễn bằng tập từ đặc trưng.

1.4 Lý thuyết mạng Bayes và các phân bố xác suất

Tiếp cận phân tích MXH dựa theo mô hình chủ đề, luận án dựa trên nền tảng lý thuyết mạng xác suất Bayes và kỹ thuật Gibbs để xây dựng mô hình và giải quyết các bài toán đặt ra.

1.4.1 Lý thuyết mạng Bayes

Theo định lý Bayes, xác suất xảy ra X khi biết Y được ký hiệu là $P(X|Y)$ phụ thuộc vào ba yếu tố. Khi biết ba yếu tố trên, xác suất của X khi biết Y được cho bởi công thức (1.3):

$$\overbrace{P(X|Y)}^{Xác\ suất\ hậu\ nghiệm} = \frac{\overbrace{P(Y|X)}^{Khả\ năng} \quad \overbrace{P(X)}^{Xác\ suất\ tiên\ nghiệm}}{\underbrace{P(Y)}_{Hằng\ số\ chuẩn\ hoá}} \tag{1.3}$$

Bằng việc tiếp cận mô hình thống kê Bayes để phân tích dữ liệu, cho một tập dữ liệu bao gồm nhiều điểm dữ liệu $\mathcal{D} = \{x_1, x_2, ..., x_N\}$ được giả định được tạo ra từ những phân bố xác suất có tham số là θ. Giả định phân bố xác suất đó được biểu diễn bởi hàm khả năng $P(\mathcal{D}|\theta)$. Trong đó, mặc dù θ chưa biết, nhưng cho một số tri thức tiên nghiệm đến mô hình được tạo ra bởi phân bố $P(\theta|\alpha)$, trong đó α là giá trị biết trước gọi là tham số Dirichlet. Đây là một ý tưởng cơ sở của tiếp cận thống kê Bayes được so sánh với những tiếp cận thống kê truyền thống mà trong đó tham số θ được giả định có một giá trị cố định. Phân bố xác suất liên hợp của kho ngữ liệu quan sát được và những tham số được định nghĩa sau:

$$P(\mathcal{D}, \theta|\alpha) = P(\mathcal{D}|\theta)P(\theta|\alpha) \tag{1.4}$$

Theo thống kê Bayes, cả kho ngữ liệu $\mathcal{D}$ và tham số θ được xem là những biến ngẫu nhiên. Do đó, ứng dụng lý thuyết mạng Bayes để tính phân bố hậu nghiệm của tham số θ như sau:

$$P(\theta|\mathcal{D}; \alpha) = \frac{P(\mathcal{D}|\theta)P(\theta|\alpha)}{P(\mathcal{D}|\alpha)} \tag{1.5}$$

Tích phân hai vế của (1.5) theo θ để tính phân phối biên $P(\mathcal{D}|\alpha)$ của tập dữ liệu $\mathcal{D}$. Kết quả $P(\mathcal{D}|\alpha)$ được được tính dựa theo hàm khả năng $P(\mathcal{D}|\theta)$ và phân bố tiên nghiệm $P(\theta|\alpha)$ như sau:

$$P(\mathcal{D}|\alpha) = \int_{\theta} P(\mathcal{D}|\theta)\ P(\theta|\alpha) d\theta \quad (1.6)$$

Mô hình chủ đề LDA được xây dựng dựa theo mô hình mạng Bayes. Vì vậy, các yếu tố và thành phần trong mạng Bayes luôn được xem xét trong mô hình chủ đề LDA. Luận án kế thừa những ưu điểm của mạng Bayes và tiếp cận dựa theo mô hình chủ đề.

1.4.2 Phân bố Dirichlet – hàm Gamma – hàm Beta

1.4.3 Phân bố hậu nghiệm

Theo công thức (2.5), phân bố hậu nghiệm của mô hình xác suất như sau:

$$\mathrm{P}(\theta|\mathcal{D};\ \alpha\ ) = \frac{P(\mathcal{D}|\theta)P(\theta|\alpha)}{\int_{\theta} P(\mathcal{D},\theta)P(\theta|\alpha)\ d\theta} \quad (1.11)$$

1.4.4 Lý thuyết về phương pháp lấy mẫu Gibbs

1.4.4.1 Lý thuyết xích Markov

1.4.4.2 Kỹ thuật lấy mẫu Gibbs

1.5 Các nghiên cứu liên quan phân tích MXH dựa theo mô hình chủ đề

Mô hình chủ đề được các nghiên cứu liên quan đến phân tích MXH áp dụng. Trong đó, mô hình chủ đề LDA được quan tâm áp dụng nhiều nhất. Luận án khảo sát các nghiên cứu có liên quan về mô hình khám phá chủ đề quan tâm của cá nhân và khám phá chủ đề quan tâm của cộng đồng (khám phá cộng đồng) dựa theo mô hình chủ đề.

1.5.1 Khám phá chủ đề quan tâm của cá nhân trên MXH

1.5.1.1 Khái niệm chủ đề quan tâm của cá nhân

Chủ đề quan tâm là chủ đề có trong thông điệp được cá nhân quan tâm trao đổi. Mỗi cá nhân quan tâm nhiều chủ đề khác nhau và ngược lại mỗi chủ đề có nhiều cá nhân quan tâm. Chủ đề quan tâm được ký hiệu là *z*. Đối với một chủ đề cụ thể, cả cá nhân nhận và cá nhân gửi đều có mức độ quan tâm khác nhau.

1.5.1.2 Mô hình Author và mô hình Author-Topic

1.5.1.3 Mô hình khám phá chủ đề ART

Mô hình Author-Recipient-Topic (ART) là mô hình Tác giả –Người nhận– Chủ đề [11] tiếp cận theo mạng Bayes và là mạng Bayes ba lớp. Theo tiếp cận của mô hình ART, một liên kết xã hội giữa tác giả và người nhận bằng cách tính các phân bố xác suất độc lập giữa tác giả và người nhận cho một thông điệp.

1.5.1.4 Mô hình Dynamic Topic Model

1.5.1.5 Mô hình Topic-Over-Time

1.5.1.6 Mô hình Author - Topic - Time

1.5.2 Khám phá chủ đề quan tâm của cộng đồng trên MXH

1.5.2.1 Mô hình khám phá nhóm GT

1.5.2.2 Mô hình khám phá cộng đồng CUT

1.5.2.3 Mô hình khám phá cộng đồng CART

1.5.2.4 Mô hình Author-Topic-Community

1.6 Nhận định chung và động lực nghiên cứu

Những hoạt động trên MXH luôn có sự thay đổi theo thời gian, vai trò của cá nhân tham gia trên MXH là quan trọng quyết định nên sự luôn vận động và thay đổi các hoạt động trên MXH đó. Trong phân tích MXH, nghiên cứu về mô hình khám phá chủ đề quan tâm của cá nhân và khám phá cộng đồng dựa theo mô hình chủ đề đã có nhiều công trình công bố. Tuy nhiên, đối với nghiên cứu khám chủ đề quan tâm của cá nhân, các mô hình chủ đề như mô hình LDA [24], PLSI [33] hay LSI [57] đều chưa quan tâm đến yếu tố cá nhân gửi và nhận thông điệp cũng như chưa phân tích sự biến thiên chủ đề và tập từ đặc trưng của chủ đề theo từng giai đoạn thời gian.

Bên cạnh đó, trên MXH thông điệp được gởi bởi rất nhiều cá nhân mà không theo một chủ đề nào được xác định trước, vấn đề được đặt ra làm sao hiểu được cá nhân trên MXH quan tâm đến những chủ đề gì và tìm ra chuyên gia theo từng chủ đề, những vấn đề này bước đầu mô hình ART [11] đã giải quyết được. Tuy nhiên, như đã trình bày và phân tích trong mục 1.5.1.3, mô hình ART không quan tâm đến yếu tố thời gian trong phân tích mà chỉ phân tích trên cơ sở chủ đề quan tâm của cá nhân và vai trò của cá nhân trên MXH theo chủ đề cụ thể. Trên thực tế, từng chủ đề quan tâm của cá nhân theo từng chủ đề sẽ thay đổi theo thời gian và cá nhân nào có gây ảnh hưởng nhiều nhất trong cộng đồng mạng. Chính vì vậy, cả mô hình Tác giả, AT [47] và ART chưa phù hợp cho lĩnh vực phân tích MXH với từng chủ đề gắn với yếu tố thời gian và cá nhân gửi và nhận chủ đề chủ đề.

Bên cạnh đó, qua khảo sát trên nhận thấy rằng: đối với mô hình DTM [23], ATT [38], TOT [76] và các mô hình trong [5][6] tiếp cận mô hình chủ đề có yếu tố thời gian, chủ đề được trao đổi trên MXH và sự thay đổi chủ đề quan tâm của cá nhân trên MXH thường xuyên thay đổi theo thời gian. Các mô hình trên đã giải quyết được vấn đề từng chủ đề được gán nhãn thời gian. Tuy nhiên, các mô hình vẫn chưa phân tích vai trò của cá nhân gửi và cá nhân nhận chủ đề. Bên cạnh đó, cả 3 mô hình DTM, TOT và ATT cũng không quan tâm

đến việc khám phá vai trò và chủ đề quan tâm của cá nhân với nhãn thời gian cũng như chưa quan tâm đến phân tích sự biến thiên chủ đề và thành viên, các yếu tố hình thành cộng đồng MXH theo thời gian. Về thử nghiệm, các mô hình trên tập trung vào thu thập và phân tích trên nguồn dữ liệu tiếng Anh và là kho bài báo khoa học và hệ thống Enron Email. Bên cạnh đó, kết quả từ mô hình LDA, ART và các mô hình được luận án khảo sát chưa quan tâm việc gán nhãn chủ đề được khám phá mà chỉ dừng lại việc đánh chỉ mục cho từng chủ đề hoặc gán nhãn chủ đề bằng tay.

Giải quyết những hạn chế này, luận án xây dựng mô hình TART nhằm mục tiêu khắc phục những hạn chế của những nghiên cứu trước đó và đưa ra mô hình phân tích MXH dựa theo mô hình chủ đề để khám phá chủ đề quan tâm, phân tích vai trò của cá nhân quan tâm chủ đề và phân tích sự biến thiên chủ đề quan tâm của cá nhân có yếu tố thời gian (hình 3.5). Chi tiết nội dung này được luận án trình bày trong chương 3. Bên cạnh đó, kết quả mô hình TART là nền tảng để luận án tiếp tục xây dựng phương pháp khám phá cộng đồng cá nhân dựa theo mô hình chủ đề, nội dung này được trình bày chi tiết trong chương 4.

Đối với phương pháp khám phá cộng đồng cá nhân trên MXH, trong các nghiên cứu trước liên quan đến nghiên cứu của luận án đã được giới thiệu trong phần 1.5.2, luận án đã trình bày khảo sát các nghiên cứu về xây dựng mô hình khám phá nhóm hay cộng đồng cá nhân trên MXH cùng quan tâm đến chủ đề [19][22][30][49]. Bên cạnh đó, luận án cũng đã khảo sát một số nghiên cứu về khám phá cộng đồng MXH [1][4][16][25][34][47][65] dựa theo mô hình chủ đề. Các nghiên cứu trên đã đạt kết quả trong khám phá cộng đồng mạng dựa trên việc phân tích nội dung thông điệp là các bài báo khoa học, nội dung email bằng tiếng Anh. Ưu điểm và những hạn chế của các nghiên cứu trước có liên quan đến khám phá cộng cộng cá nhân trên MXH:

- Ưu điểm:
  - Xây dựng mô hình dựa theo mô hình chủ đề.
  - Dùng ART để tạo vector chủ đề quan tâm và sử dụng làm vector đầu vào cho quá trình gom cụm của mô hình.
  - Các mô hình dùng giải thuật gom cụm (K-Means hoặc K-Medoids,…) để khám phá cộng đồng MXH theo vector chủ đề quan tâm.
- Hạn chế:
  - Chưa gom cụm được cộng đồng theo thời gian vì vector đầu vào của ART không có yếu tố thời gian.

- Chưa biểu diễn trực quan kết quả gom cụm cộng đồng theo thời gian và phân tích sự biến thiên đặc trưng của cộng đồng.
- Số cộng đồng MXH là rất lớn, các nghiên cứu dùng giải thuật K-Means hoặc K-Medoids nên khó tính toán trước hệ số K để gom cụm cộng đồng. Nghĩa là khó xác định số cộng đồng.

Mặt khác, vấn đề phân tích sự phân bố chủ đề trong cộng đồng theo thời gian, phân bố chủ đề được quan tâm trong cộng đồng, với một chủ đề thì sự quan tâm của nhiều cá nhân thay đổi ra sao, điều này cũng chưa được các nghiên cứu quan tâm. Hơn thế nữa, các nghiên cứu trên chủ yếu tập trung khám phá cộng đồng dựa trên tập ngữ liệu thông điệp tiếng Anh. Trong luận án nghiên cứu và thử nghiệm trên tập ngữ liệu thông điệp tiếng Việt được thu thập từ MXH. Bên cạnh đó, luận án xây dựng phương pháp khám phá cộng đồng dựa trên mô hình TART kết hợp mạng nơ ron Kohonen để khám phá cộng đồng theo thời gian và trực quan hoá kết quả khám phá cộng đồng dựa trên lớp ra Kohonen. Mạng nơ ron Kohonen gom cụm dữ liệu mà không cần chỉ định trước số cộng đồng. Áp dụng mạng nơ ron Kohonen để gom cụm những cá nhân cùng quan tâm đến chủ đề cụ thể nhưng mức độ quan tâm là khác nhau, vì thế kết quả gom nhóm từ phương pháp đề xuất của luận án đáp ứng tốt tiêu chí trong phương pháp gom cụm.

CHƯƠNG 2. KHÁM PHÁ VÀ GÁN NHÃN CHỦ ĐỀ ẨN TỪ THÔNG ĐIỆP TRÊN MẠNG XÃ HỘI

2.1 Giới thiệu chương

Mỗi thông điệp trên MXH tiềm ẩn nhiều chủ đề được cá nhân quan tâm trao đổi. Theo từng giai đoạn thời gian khác nhau, cá nhân có thể quan tâm đến chủ đề khác nhau. Đây là những yếu tố cơ bản để giúp phân biệt một thông điệp thông thường và một thông điệp trên MXH. Chính vì vậy, việc khám phá chủ đề ẩn trong thông điệp trên MXH cũng khác với phân tích một thông điệp thông thường. Mô hình chủ đề LDA được luận án lựa chọn để làm cơ sở giải quyết bài toán khám phá chủ đề ẩn từ thông điệp trên MXH.

2.2 Khám phá chủ đề ẩn trên MXH áp dụng mô hình chủ đề

Mục tiêu của mô hình khám phá chủ đề ẩn là tìm ra tập vector chủ đề và từ ($Z \times W$) và tập vector thông điệp và chủ đề ($\mathcal{D} \times Z$). Các khái niệm liên quan đến vector chủ đề và vector thông điệp được trình bày

2.2.1 Khái niệm vector chủ đề

2.2.2 Khái niệm vector thông điệp

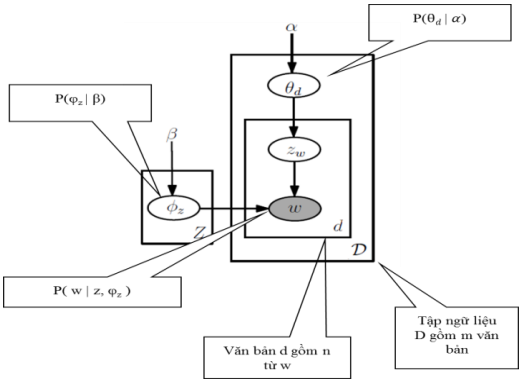
2.2.3 Phát biểu bài toán khám phá chủ đề ẩn từ thông điệp trên MXH

Bài toán khám phá chủ đề ẩn áp dụng mô hình chủ đề LDA được phát biểu:

Cho: $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ tập thông điệp trong kho ngữ liệu, $W = \{w_1, w_2, \ldots, w_N\}$ tập các từ trong kho ngữ liệu $\mathcal{D}$ - mỗi thông điệp $d_i \in \mathcal{D}$ chứa một tập từ của W, K số chủ đề ẩn.

Tìm: Vector chủ đề của k chủ đề (Z x W) ($\phi_{z,w}$), Vector thông điệp của các thông điệp ($\mathcal{D}$ x Z) ($\theta_{d,z}$).

2.3 Mô hình chủ đề LDA



Hình 2.2 Mô hình LDA và phân bố xác suất liên hợp.

2.3.1 Phân bố xác suất liên hợp cho mô hình LDA

Với mỗi θ_j là ma trận chứa các chủ đề của thông điệp thứ j, mỗi $z_t \in z$ là chủ đề được gán cho từ w thứ t, mỗi $\emptyset_i$ là ma trận chứa các từ của chủ đề thứ i. Mục đích của mô hình LDA là khám phá các từ đặc trưng thuộc về một chủ đề từ đó suy diễn chủ đề đó là chủ đề gì. Đây là quá trình tạo sinh và phân bố hậu nghiệm cho các biến ẩn là tập từ đặc trưng cho chủ đề. Nói cách khác, nếu cho trước phân bố từ thuộc chủ đề z là $\emptyset_z$và phân bố chủ đề thuộc thông điệp d là θ_d, thì phân bố xác suất mà một từ w trong d thuộc về chủ đề z sẽ là $\theta_{d,z}, \emptyset_{z,w}$:

$$P(w, z, |\emptyset_z, \theta_d) = P(z|\theta_d)P(w|\emptyset_z) = \theta_{d,z}\emptyset_{z,w} \quad (2.1)$$

Giả sử rằng hai biến $\emptyset_z$ và θ_d được sinh ra bởi phân bố xác suất, ký hiệu là $P(\emptyset_z|\beta)$ và $P(\theta_d|\alpha)$, trong đó, α và β là hai tham số Dirichlet, phân bố xác suất liên hợp của từ w và chủ đề z trong thông điệp $d \epsilon \mathcal{D}$ được trình bày sau:

$$P(w, z, \emptyset_z, \theta_d|\alpha, \beta) = P(\emptyset_z|\beta)P(\theta_d|\alpha)P(z|\theta_d)P(w|\emptyset_z) \quad (2.2)$$

và phân bố xác suất liên hợp của tất cả từ w và chủ đề z trong $d \epsilon \mathcal{D}$ trở thành:

$$P(d, z, \emptyset, \theta_d|\alpha, \beta) = \prod_{i=1}^{K} P(\emptyset_i|\beta)\, P(\theta_d|\alpha) \prod_{t=1}^{N} P(\mathrm{z}_t|\theta_d)\, P\left(w_t\middle|\emptyset_{z_t}\right) \quad (2.3)$$

Trong đó, K là số chủ đề trong d, i là chủ đề z thứ i, N là số từ w trong d, t là từ w thứ t, $w \in d$. Mỗi $z_t \in z$ chỉ ra từ w thứ t được gán vào chủ đề $\mathrm{z} \in d$. Và cuối cùng có được phân bố xác suất liên hợp của tất cả các từ w và chủ đề trong kho ngữ liệu $\mathcal{D}$, đây chính là phân bố xác suất liên hợp của mô hình LDA:

$$P(\mathcal{D}, z, \theta, \emptyset|\alpha, \beta) = \prod_{i=1}^{K} P(\emptyset_i|\beta) \prod_{j=1}^{M} P(\theta_j|\alpha) \prod_{t=1}^{N} P\left(\mathrm{z}_{j,t}|\theta_j\right) P\left(w_{j,t}\middle|\emptyset_{z_{j,t}}\right) \quad (2.4)$$

2.3.2 Kỹ thuật lấy mẫu Gibbs cho mô hình LDA

Các biến ẩn trong mô hình LDA đã trình bày gồm chủ đề z, phân bố từ trong chủ đề $\emptyset$, phân bố chủ đề trong thông điệp θ. Phân bố hậu nghiệm của các biến này được phân tích bằng cách sử dụng lý thuyết Bayes được trình bày trong chương 1. Xét theo từng từ w, ta tính tổng xác suất của mô hình dựa trên từng từ w và từ đó suy ra tổng xác suất của mô hình trên cả kho ngữ liệu $\mathcal{D}$. Trong mô hình LDA, các đại lượng biến ẩn này được tính theo công thức sau:

$$P(\theta, \emptyset, z|w; \alpha, \beta) = \frac{P(\theta, \emptyset, z, w|\alpha, \beta)}{P(w|\alpha, \beta)} = \frac{P(\theta, \emptyset, z, w|\alpha, \beta)}{\int_\theta \int_\emptyset \sum_{i=1}^{K} P(w, z, \theta, \emptyset|\alpha, \beta)\, d\emptyset\, d\theta} \quad (2.6)$$

Tuy nhiên, các yếu tố chuẩn hoá $P(w|\alpha, \beta)$ (hay phân phối biên) không thể tính một cách chính xác [67] vì $P(w|\alpha, \beta)$ không đổi cho bất kỳ chủ đề z nào. Việc áp dụng phương pháp lấy mẫu, phân bố hậu nghiệm cho (2.6) được tính xấp xỉ thông qua các mẫu của phân bố xác suất liên hợp trình bày trong (2.7).

$$P(\theta, \emptyset, z|w; \alpha, \beta) = \frac{P(\theta, \emptyset, z, w|\alpha, \beta)}{P(w|\alpha, \beta)} \propto P(\theta, \emptyset, z, w|\alpha, \beta) \quad (2.7)$$

Việc thực hiện lấy mẫu Gibss nên được thực hiện bằng cách kết hợp giữa phân bố Dirichlet và phân bố xác suất nhiều chiều để tính tích phân theo các tham số đa thức θ và $\emptyset$ trong công thức (2.7) và áp dụng giải thuật Collapsed Gibbs sampling để tính xác suất của một chủ đề z đang được gán vào từ w_i dựa theo tất cả các phép gán của chủ đề z khác vào các từ w khác, tức là tính: $P(z_i|z_{-i}, \alpha, \beta, w)$.

$$\theta_{d,z} = \frac{n_z^{(d)} + \alpha_z}{\sum_{z' \epsilon Z} n_{z'}^{(d)} + \alpha_{z'}}, d \in \mathcal{D}, z \in Z \quad (2.22)$$

Và $\emptyset_{z,w}$ được tính bởi công thức:

$$\emptyset_{z,w} = \frac{n_w^{(z)} + \beta_w}{\sum_{w' \epsilon V} n_{w'}^{(z)} + \beta_{w'}}, z \epsilon Z, w \epsilon V \quad (2.23)$$

2.4 Thử nghiệm phương pháp khám phá chủ đề ẩn bằng mô hình LDA

2.4.1 Mô tả dữ liệu thử nghiệm

2.4.2 Tiền xử lý thông điệp tiếng Việt

2.4.3 Thử nghiệm mô hình LDA trên dữ liệu diễn đàn và MXH

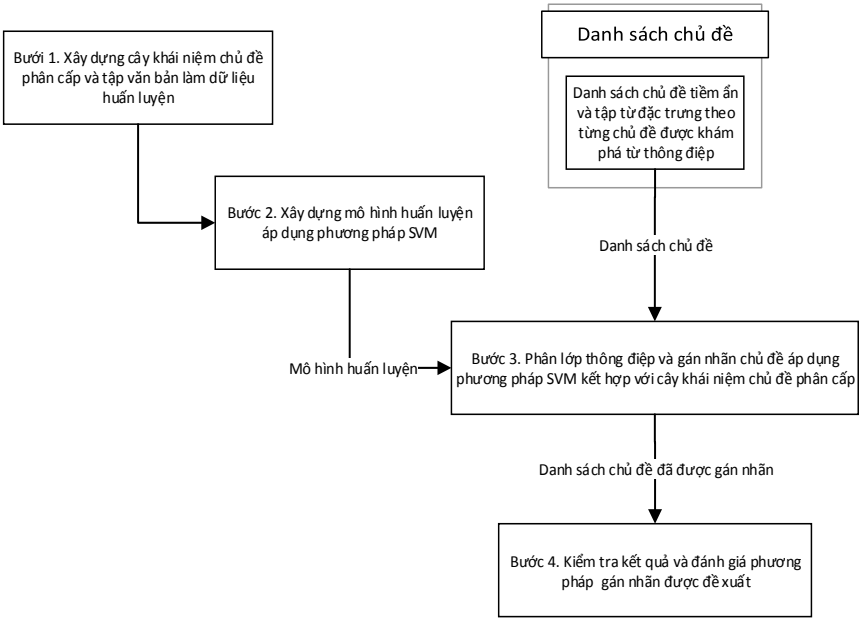
Kết quả từ mô hình LDA là danh sách các chủ đề chưa được gán nhãn. Với tập kết quả này dẫn đến khó nhận biết được cá nhân trên MXH quan tâm đến chủ đề cụ thể nào từ thông điệp được trao đổi

2.4.4 Thử nghiệm mô hình LDA trên dữ liệu của trang VnExpress.net

2.4.5 Hạn chế của mô hình LDA

2.5 Phương pháp gán nhãn chủ đề ẩn

2.5.1 Ý tưởng phương pháp gán nhãn cho chủ đề ẩn



Hình 2.6. Mô hình gán nhãn chủ đề ẩn

2.5.2 Xây dựng cây phân cấp chủ đề

2.5.2.1 Khái niệm cây phân cấp chủ đề

Cây phân cấp chủ đề (Topic Taxonomy) là một cấu trúc phân cấp các thực thể (các lớp hay chủ đề). Các lớp trên cây được sắp xếp dựa trên loại quan hệ cha-con và không có sự ràng buộc trên bất kỳ thuộc tính tại bất kỳ cấp nào trong cấu trúc phân cấp. Mục đích của cây phân cấp chủ đề dùng phân lớp tri thức.

2.5.2.2 Quy trình xây dựng cây phân cấp chủ đề

2.5.3 Thử nghiệm phương pháp gán nhãn và đánh giá kết quả

Bảng 2.8 Trình bày 4 chủ đề đã được gán nhãn (4 vector chủ đề) dựa trên cây phân cấp chủ đề.

Cấp 0		Cấp 1		Cấp 1		Cấp 1	
Chủ đề 1: Hoạt động đoàn		**Chủ đề 2: Câu lạc bộ**		**Chủ đề 3: Hội sinh**		**Chủ đề 4: Đoàn thanh**	
Từ	Xác suất	Từ	Xác	Từ	Xác	Từ	Xác suất
công_tác	0.01197	đại_học	0.01306	hội	0.01339	tình	0.02464
đại_học	0.01051	học_thuật	0.01259	xã_hội	0.01292	hoạt_động	0.02261
tuổi	0.01051	tham_gia	0.01053	tư_tưởng	0.01194	người	0.02002
trường	0.00903	nghiên_cứu	0.00969	phòng	0.01072	thanh_niên	0.01712
năm_học	0.00860	thể_thao	0.00928	olympic	0.01069	sinh_viên	0.01459
học	0.00827	sinh_viên	0.00905	cờ	0.01046	đại_hội	0.01346
sinh_viên	0.00631	tinh_thần	0.00818	hội_thi	0.01003	học_sinh	0.01313
chuyên	0.00616	kỹ_năng	0.00740	chung_kết	0.00844	cống_hiến	0.01274
đoàn	0.00614	đại_học	0.00644	thời_đại	0.00773	chương_trình	0.01188
hoạt_động	0.00579	bóng_đá	0.00633	kỹ_năng	0.00752	kỹ_năng	0.01172
trẻ	0.00543	giao_tiếp	0.00581	sinh_viên	0.00725	đoàn	0.01165
tình_nguyện	0.00524	thi	0.00482	liên	0.00722	trưởng_thành	0.01122
tham_gia	0.00510	ngoại_ngữ	0.00419	thành_tích	0.00614	con	0.01025
phong_trào	0.00417	chương	0.00419	bản_lĩnh	0.00559	nguyện	0.00772
đoàn_viên	0.00373	hoạt_động	0.00414	về_nguồn	0.00515	chiến_dịch	0.00767

Về đánh giá kết quả: luận án áp dụng các hệ số Precision, Recall, độ đo F (F-measure) để đánh giá kết quả gán nhãn chủ đề ẩn.

2.6 Kết luận chương

Trong chương 2, luận án đã xây dựng được mô hình khám phá, phân lớp để gán nhãn chủ đề trong lĩnh vực phân tích MXH và rút trích thông tin dựa theo mô hình chủ đề và thử nghiệm trên kho ngữ liệu thông điệp tiếng Việt được thu thập từ diễn đàn, MXH trong trường đại học. Luận án đã cải tiến bước xử lý và làm sạch dữ liệu đầu vào nhằm cải tiến quá trình tách từ và gán nhãn từ loại tiếng Việt. Đóng góp chính trong chương 2: (1) xây dựng cây phân cấp chủ đề gồm tập khái niệm trong trường đại học và tập từ đặc trưng cho từng chủ đề trên cây phân cấp chủ đề, (2) áp dụng mô hình chủ đề LDA để khám phá chủ đề ẩn từ tập thông điệp trên MXH, nội dung này cũng được luận án áp dụng trong [CB07] và [CB08] về phân tích tầm ảnh hưởng trên MXH, (3) dùng phương pháp học máy SVM dựa trên tập dữ liệu huấn luyện là cây phân cấp chủ đề để phân lớp thông điệp và gán nhãn chủ đề ẩn. Mô hình đã cho kết quả tốt, các mô hình và phương pháp thực hiện trong chương 2 được tích hợp trên hệ thống phần mềm SNA được luận án xây dựng để tự động làm sạch dữ liệu, tự động khám phá và gán nhãn chủ đề ẩn với độ chính xác cao.

CHƯƠNG 3. KHÁM PHÁ CHỦ ĐỀ QUAN TÂM CỦA CÁ NHÂN DỰA THEO MÔ HÌNH CHỦ ĐỀ

3.1 Giới thiệu chương

Trong chương này, luận án tập trung trình bày phương pháp xây dựng mô hình khám phá chủ đề quan tâm của cá nhân có yếu tố thời gian, phân tích những ưu điểm và hạn chế của các mô hình. Dựa trên cơ sở đó, luận án đề xuất phát triển mô hình khám phá chủ đề quan tâm và phân tích vai trò của cá nhân quan tâm đến chủ đề có yếu tố thời gian dựa theo mô hình chủ đề, được gọi là mô hình TART hay mô hình Thời gian-Tác giả-Người nhận-Chủ đề. Kết quả chương 3 được thể hiện trong công bố chính [CB05] về xây dựng mô hình TART tiếp cận theo mô hình chủ đề nhằm mục tiêu phân tích chủ đề quan tâm của cá nhân có yếu tố thời gian, và mô hình khám phá chủ đề được cá nhân quan tâm nhiều nhất trên MXH được thể hiện trong [CB02].

3.2 Khám phá chủ đề quan tâm của cá nhân trên MXH theo thời gian

3.2.1 Khái niệm chủ đề quan tâm của cá nhân theo thời gian

3.2.2 Bài toán khám phá chủ đề quan tâm của cá nhân trên MXH có yếu tố thời gian

Cho: MXH $G = <V, E>$, V là tập cá nhân và E là tập các liên kết xã hội giữa các cá nhân. Gọi $\mathcal{D}$ là tập các thông điệp được cá nhân trao đổi trên MXH, Z là chủ đề quan tâm được cá nhân trao đổi trong các thông điệp thông qua các liên kết xã hội, K là số chủ đề, thời gian trao đổi thông điệp.

Tìm:

(i) Vector chủ đề quan tâm của cá nhân $< f(v_{i1}), f(v_{i2}), \ldots, f(v_{ik}) >$ theo từng giai đoạn thời gian, trong đó thành phần $f(v_{ik})$ phản ánh xác suất quan tâm chủ đề Z_k của actor v_j trong thông điệp d. Mỗi giai đoạn thời gian T_i, actor i có xác suất quan tâm chủ đề Z_k là khác nhau. Ta có, thành phần $f(v_{ik})$ của mỗi actor vector $< f(v_{i1}), f(v_{i2}), \ldots, f(v_{ik}) >$ cũng khác nhau.

Nghĩa là ta phải tìm các phân bố xác suất: phân bố Z (chủ đề) x W (từ): $\emptyset_{zw,}$ phân bố A (tác giả) x Z (chủ đề): Θ_{az}, phân bố R (cá nhân nhận) x Z (chủ đề): Θ_{rz}, phân bố Z (chủ đề) x T (thời gian): ψ_{zt}.

(ii) Sự biến thiên chủ đề quan tâm của cá nhân theo thời gian.

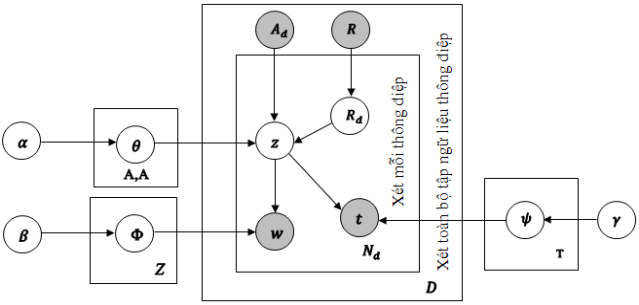
3.3 Mô hình khám phá chủ đề quan tâm cá nhân theo thời gian

3.3.1 Xây dựng mô hình TART

Mô hình TART (Temporal-Author-Recipient-Topic) [CB05] trong hình 3.1 được xây dựng dựa theo mô hình LDA và ART, mô hình TART giải quyết những hạn chế tồn tại đã được trình bày trong phần 1.6.

Nhiệm vụ của mô hình TART (hình 3.1):

- Khám phá chủ đề quan tâm của cá nhân trên MXH có yếu tố thời gian. Nghĩa là tìm tập actor vector có yếu tố thời gian.
- Phân tích vai trò của cá nhân tham gia mạng xã hội dựa theo mô hình chủ đề có yếu tố thời gian.
- Dùng yếu tố thời gian để chia nhỏ các yếu tố trong mô hình ART như tập cá nhân gửi, tập cá nhân nhận, tập chủ đề và tìm ra sự thay đổi chủ đề quan tâm của cá nhân trong tập thông điệp theo từng khoảng thời gian so với chủ đề quan tâm trong kho ngữ liệu thông điệp.
- Khảo sát sự biến thiên chủ đề quan tâm của từng cá nhân.



Hình 3.1. Mô hình TART khám phá chủ đề quan tâm của cá nhân theo thời gian

3.3.2 Phân bố xác suất liên hợp cho mô hình TART

Theo mô hình TART được trình bày trong hình 3.1, cho trước các tham số Dirichlet α, β, γ, cá nhân gửi A_d và tập cá nhân nhận R_d mỗi thông điệp d, ta có phân bố xác suất liện hợp của thông điệp – chủ đề θ_{ij} cho mỗi cặp cá nhân gửi – cá nhân nhận (i, j), chủ đề - từ $\emptyset_z$ cho mỗi chủ đề z, tập cá nhận nhận R, tập chủ đề z và tập từ w trong thông điệp d được cho bởi công thức (3.1). Xét trên từng thông điệp d, ta có công thức phân bố xác suất liên hợp trên không gian d nhiều chiều, các chiều gồm: cá nhân gửi, tập cá nhân nhận, yếu tố thời gian, tập chủ đề và tập từ đặc trưng:

$$P(w, T, z, R_d \mid A_d, R, \alpha, \beta, \gamma) \tag{3.1}$$
$$= P(R_d|R)\ P(z|A_d, R_d, \alpha)\ P(w|z, \beta)\ P(T, \psi|z, \gamma)$$
$$= \prod_{u=1}^{R_d} \prod_{n=1}^{N_d} [P(r_u|R)\ P(z|A_d, r_u, \alpha)\ P(w|z, \beta)\ P(T|z, \gamma)]$$

Trong đó, T là thời gian mà chủ đề z được quan tâm trao đổi bởi cặp cá nhân gửi A_d hay a – cá nhân nhận r_u, N_d là số từ trong thông điệp d, r_u là tập cá nhân nhận thông điệp d, với $r_u \in R_d$. Do các giá trị trong $\theta, \emptyset, \psi$ dựa trên các tham số Dirichlet tương ứng là α, β, γ. Các tham số Dirichlet này không phụ thuộc nhau vì vậy ta phân rã công thức (3.1) để tính tích phân từng phần theo

θ phụ thuộc vào α, $\emptyset$ phụ thuộc β và ψ phụ thuộc vào γ. Từ đó, ta có được các tích phân được phân rã như trong (i), (ii) và (iii) như sau:

(i). Tích phân theo θ để tìm phân bố chủ đề z theo cá nhân gửi A_d và cá nhân nhận R_d dựa trên tham số α:

$$P(z|A_d, r_u, \alpha) = \int P(\theta|\alpha)P(z|\theta, A_d, r_u)d\theta = \int \prod_{i=1}^{A}\prod_{j=1}^{A}[P(\theta_{ij}|\alpha)\, P(z|\theta_{ij})]\, d\theta \quad (3.2)$$

$$= \int \prod_{i=1}^{A}\prod_{j=1}^{A}\left[P(\theta_{ij}|\alpha)\prod_{t=1}^{K}P(z_t|\theta_{ij})\right]d\theta$$

$$= \int \prod_{i=1}^{A}\prod_{j=1}^{A}P(\theta_{ij}|\alpha)\prod_{i=1}^{A}\prod_{j=1}^{A}\prod_{t=1}^{K}\theta_{ijt}\, d\theta$$

(ii). Tính tích phân theo $\emptyset$ để tìm phân bố hậu nghiệm của từ w theo chủ đề z dựa trên tham số β:

$$P(w|z, \beta) = \int P(\emptyset|\beta)\, P(w|\emptyset, z)d\emptyset = \int \prod_{t=1}^{K}[P(\emptyset_t|\beta)\, P(w|\emptyset_t)]\, d\emptyset \quad (3.3)$$

$$= \int \prod_{t=1}^{K}\left[P(\emptyset_t|\beta)\prod_{v=1}^{V}P(w_v|\emptyset_t)\right]d\emptyset = \int \prod_{t=1}^{K}P(\emptyset_t|\beta)\prod_{t=1}^{K}\prod_{v=1}^{V}\emptyset_{tv}\, d\emptyset$$

Trong đó, K là số chủ đề trong thông điệp d, $\emptyset_{tv} \in \emptyset$ là thành phần vector của chủ đề thứ t gán cho từ w thứ v.

(iii). Và tích phân theo ψ để tìm phân bố hậu nghiệm của thời gian T theo chủ đề z dựa trên tham số γ:

$$P(T|Z, \gamma) = \int P(\psi|\gamma)\, P(T|\psi, Z)d\psi \quad (3.4)$$

$$= \int \prod_{t=1}^{K}[P(\psi_t|\gamma)P(T|\psi_t)]$$

$$= \int \prod_{t=1}^{K}\left[P(\psi_t|\gamma)\prod_{y=1}^{T}P(T_y|\psi_t)\right]d\psi$$

$$= \int \prod_{t=1}^{T}P(\psi_t|\gamma)\prod_{t=1}^{K}\prod_{y=1}^{T}\psi_{ty}\, d\psi$$

3.3.3 Kỹ thuật lấy mẫu Gibbs cho mô hình TART

Mục đích của kỹ thuật lấy mẫu Gibbs là tính xấp xỉ phân bố điều kiện của biểu thức $P(w, T, z, R_d|A_d, R, \alpha, \beta, \gamma)$ trong công thức (3.9). Nghĩa là cần đạt được phân bố xác suất điều kiện của một chủ đề z_{di} gán cho từ w_{di} được cho bởi tất cả chủ đề khác, nghĩa là tính $P(z_{di}|z_{-di}, w, T, R_d, A_d, R, \alpha, \beta, \gamma)$ theo công thức (3.10). Dựa theo luật chuỗi (chain rule) trong luật Bayes để tính. Sau quá trình lấy mẫu Gibbs cho mô hình TART, đạt được phân bố xác suất hậu nghiệm cho $\theta, \emptyset$ và ψ được tính bởi công thức (trong quá trình thực hiện mô hình TART, hệ thống thực hiện lưu lại 4 ma trận để phân tích vai trò và chủ đề quan tâm của cá nhân theo thời gian, bao gồm: T (chủ đề) x W (từ), A (tác giả)

x T (chủ đề), R (người nhận) x T (chủ đề) và T (chủ đề) x T (thời gian) như sau:

$$\theta_{az} = \frac{n_{az} + \alpha}{\sum_z (n_{az} + \alpha)} \quad (3.14) \qquad \phi_{zw} = \frac{m_{zw} + \beta}{\sum_w (m_{zw} + \beta)} \quad (3.15)$$

$$\psi_{zt} = \frac{n_{zt} + \gamma}{\sum_t (n_{zt} + \gamma)} \quad (3.16) \qquad \theta_{rz} = \frac{n_{rz} + \alpha}{\sum_z (n_{rz} + \alpha)} \quad (3.17)$$

Giải thuật 3.1 cho mô hình TART dựa trên dựa mô hình chủ đề:

Giải thuật 3.1. Mô hình sinh của TART [CB05]

1	Đầu vào: Mạng xã hội $G = <V, E>$, V là tập cá nhân và E là tập các liên kết xã hội giữa các cá nhân là các thông điệp được trao đổi giữa các cá nhân gửi và nhận, thời gian trao đổi thông điệp.
2	Đầu ra: Vector chủ đề quan tâm của cá nhân $<f_3(v_{i1}),f_3(v_{i2}),\ldots,f_3(v_{ik})>$ theo từng giai đoạn thời gian, trong đó thành phần $f_3(v_{ik})$ phản ánh xác suất quan tâm chủ đề Z_k của actor v_i trong thông điệp d. Nghĩa là tìm các ma trận: Z (chủ đề) x W (từ) là $\emptyset_{zw}$, A (cá nhân gửi) x Z (chủ đề) là Θ_{az}, R (cá nhân nhận) x Z (chủ đề) là Θ_{rz}, Z (chủ đề) x T (thời gian) là ψ_{zt}
3	Khởi tạo các tham số đầu vào
4	For each cá nhân gửi $a = 1,..., A_d$
5	Rút Θ_a từ phân bố Dirichlet (α);
6	For each cá nhân nhận $r = 1,\ldots, R_d$
7	Rút Θ_r từ phân bố Dirichlet (α);
8	For each chủ đề $z = 1,..., K$ của thông điệp d;
9	Rút Θ_z từ phân bố Dirichlet (α);
10	Rút $\emptyset_z$ từ phân bố Dirichlet (β);
11	Rút Ψ_z từ phân bố Dirichlet (γ);
12	For each từ $w = 1,..., N_d$ của thông điệp d
13	Rút một cá nhân gửi a từ tập các cá nhân gửi $A_{d;}$
14	Rút một cá nhân nhận r từ tập các cá nhân nhận R_d;
15	Rút một chủ đề z từ phân bố (Θ_a) điều kiện trên a;
16	Rút một từ w từ phân bố ($\emptyset_z$) điều kiện trên z;
17	Rút thời gian t tương ứng với chủ đề z từ phân bố (ψ_z) điều kiện trên t;
18	Lấy mẫu Gibbs cho mô hình TART.

Độ phức tạp của giải thuật được tính toán dựa trên bốn vòng lặp tại (xét một thông điệp):

- Dòng 4: lặp theo A_d số người gửi chủ đề *z* thuộc thông điệp *d*
- Dòng 6: lặp theo R_d số người nhận chủ đề *z* thuộc thông điệp *d*
- Dòng 8: lặp theo số chủ đề *K* thuộc thông điệp *d*
- Dòng 12: lặp theo N_d từ trong thông điệp *d*

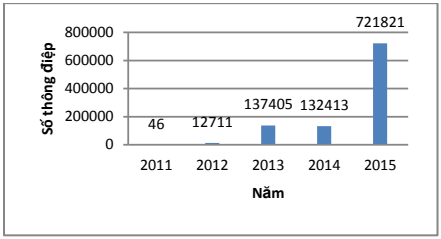
Tổng chi phí thời gian thực hiện của giải thuật cho mô hình TART là: $A_d*R_d*K*N_d$

Từ đó suy ra độ phức tạp của giải thuật cho mô hình TART là: $O(A_d*R_d*K*N_d)$. Xét trên từng thông điệp.

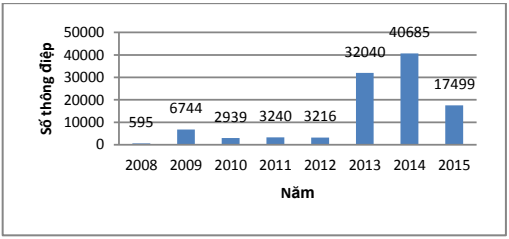
Trong trường hợp, nếu xét trên toàn tập ngữ liệu *D* bao gồm *M* thông điệp, ta có độ phức tạp của giải thuật cho mô hình TART là $O(M*A_d*R_d*K*N_d)$.

3.4 Thử nghiệm mô hình TART và thảo luận kết quả

3.4.1 Mô tả dữ liệu thử nghiệm



Hình 3.5. Lịch sử thông điệp được gửi và nhận trong giai đoạn từ năm 2011 đến năm 2015



Hình 3.2. Lịch sử thông điệp được gửi theo từng năm trong kho ngữ liệu thu thập

3.4.2 Thử nghiệm mô hình TART trên dữ liệu diễn đàn sinh viên và MXH

Bảng 3.5. Kết quả phân tích 4 chủ đề trên cây phân cấp chủ đề trong tháng 08-2014

Cấp 0

Chủ đề 1: Hoạt động đoàn hội	
Từ	Xác suất
công_tác	0.01197
đại_học	0.01051
tuổi	0.01051
trường	0.00903
năm_học	0.0086
học	0.00827
sinh_viên	0.00631
chuyên	0.00616
đoàn	0.00614
hoạt_động	0.00579
trẻ	0.00543
tình_nguyện	0.00524
tham_gia	0.0051
phong_trào	0.00417
đoàn_viên	0.00373
ID Cá nhân gửi	**Xác suất**
97179	0.670330
97568	0.600000
71804	0.560976
97126	0.471698

Cấp 1

Chủ đề 2: Câu lạc bộ	
Từ	Xác suất
đại_học	0.01306
học_thuật	0.01259
tham_gia	0.01053
nghiên_cứu	0.00969
thể_thao	0.00928
sinh_viên	0.00905
tinh_thần	0.00818
kỹ_năng	0.0074
đại_học	0.00644
bóng_đá	0.00633
giao_tiếp	0.00581
thi	0.00482
ngoại_ngữ	0.00419
chương	0.00419
hoạt_động	0.00414
ID Cá nhân gửi	**Xác suất**
78686	0.53982
79249	0.38777

Cấp 1

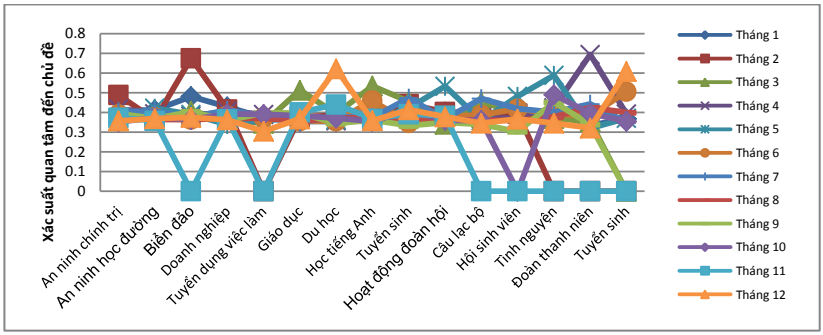
Chủ đề 3: Hội sinh viên	
Từ	Xác suất
hội	0.01339
xã_hội	0.01292
tư_tưởng	0.01194
phòng	0.01072
olympic	0.01069
cờ	0.01046
hội_thi	0.01003
chung_kết	0.00844
thời_đại	0.00773
kỹ_năng	0.00752
sinh_viên	0.00725
liên	0.00722
thành_tích	0.00614
bản_lĩnh	0.00559
về_nguồn	0.00515
ID Cá nhân gửi	**Xác suất**
79554	0.68212
71151	0.39683

Cấp 1

Chủ đề 4: Đoàn thanh niên	
Từ	Xác suất
tình	0.02464
hoạt_động	0.02261
người	0.02002
thanh_niên	0.01712
sinh_viên	0.01459
đại_hội	0.01346
học_sinh	0.01313
cống_hiến	0.01274
chương_trình	0.01188
kỹ_năng	0.01172
đoàn	0.01165
trường_thành	0.01122
con	0.01025
nguyện	0.00772
chiến_dịch	0.00767
ID Cá nhân gửi	**Xác suất**
67484	0.83740
70824	0.77049

74568	0.469388	79096	0.37143	64325	0.39130	68395	0.75439
ID Cá nhân nhận	**Xác suất**	79556	0.36000	72750	0.37931	69361	0.75385
97126	0.670330	69660	0.33333	64374	0.36585	68925	0.74545
77692	0.560976	**ID Cá nhân nhận**	**Xác suất**	**ID Cá nhân nhận**	**Xác suất**	**ID Cá nhân nhận**	**Xác suất**
81027	0.548387	72365	0.44000	90018	0.68212	64595	0.84946
67317	0.538462	90191	0.42222	73490	0.48148	72692	0.83133
76996	0.485714	72597	0.40741	96166	0.45946	71138	0.79221
		71955	0.31034	73376	0.36000	64864	0.74545
		74183	0.31212	76427	0.35484	76590	0.73585

Bảng 3.5 trình bày 4 vector chủ đề quan tâm của tập cá nhân được trình bày giới hạn trong 5 cá nhân gửi và 5 cá nhân nhận và được sắp xếp giảm dần trong mỗi chủ đề. Mỗi vector bao gồm các thành phần như tập các cá nhân quan tâm gửi và nhận chủ đề đó hay còn gọi là vector chủ đề quan tâm.



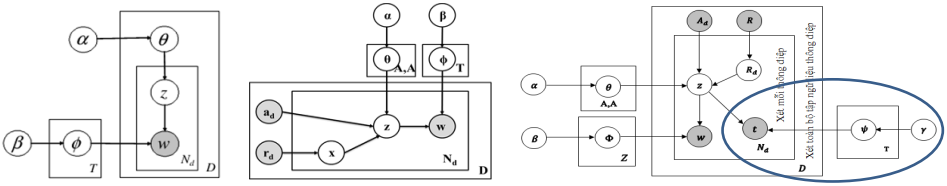
Hình 3.9. Kết quả phân tích trên 15 chủ đề trong thời gian từ tháng 01 đến tháng 12 năm 2015.

3.4.3 Thử nghiệm mô hình TART trên dữ liệu trang báo VnExpress.net

3.4.4 Khảo sát sự thay đổi chủ đề quan tâm của cá nhân theo thời gian

3.4.5 Tổng hợp so sánh mô hình TART với mô hình LDA và ART

3.4.5.1 So sánh tham số mô hình



(1) Mô hình LDA (2) Mô hình ART (3) Mô hình TART

Hình 3.15. So sánh mô hình TART với mô hình LDA và mô hình ART

Bảng 3.9 dưới đây trình bày chi tiết về việc so sánh dựa trên các phương pháp tiếp cận và tham số được xây dựng và tích hợp trong từng mô hình. Trong đó, mô hình TART do luận án đề xuất.

Bảng 3.9. So sánh ba mô hình TART, LDA và ART

Mô hình	Các phương pháp tiếp cận và tham số của các mô hình									
	Phân tích MXH	Mô hình chủ đề	Gán nhãn chủ đề	Lấy mẫu Gibbs	Mạng Bayes, phân bố XS hậu nghiệm	Phân bố XS liên hợp	Chủ đề Z và từ W	Cá nhân gửi chủ đề A	Cá nhân nhận chủ đề R	Yếu tố thời gian T
LDA		x		x	x	x	x			
ART	x	x		x	x	x	x	x	x	
TART	x	x	x	x	x	x	x	x	x	x

3.5 Kết luận chương

Trong chương 3, luận án đã tập trung nghiên cứu mục tiêu chính thứ nhất là khảo sát và nhận định ưu điểm và hạn chế của các mô hình liên quan đến khám phá vai trò của cá nhân quan tâm chủ đề trên MXH. Từ đó, luận án xây dựng mô hình TART. Mô hình TART độc lập với ngôn ngữ và được xây dựng dựa trên sự tích hợp yếu tố thời gian vào để khám phá vai trò của cá nhân dựa theo mô hình chủ đề. Trong đó, luận án đã xây dựng công thức phân bố xác suất liên hợp và áp dụng kỹ thuật lấy mẫu Gibbs cho mô hình TART để tìm ra 4 phân bố xác suất hậu nghiệm của 4 tham số: $\emptyset_{zw}$, Θ_{az}, Θ_{rz} và ψ_{zt}.

CHƯƠNG 4. KHÁM PHÁ CỘNG ĐỒNG DỰA THEO MÔ HÌNH CHỦ ĐỀ

4.1 Giới thiệu chương

Chủ đề quan tâm của cá nhân thường thay đổi dẫn đến chủ đề quan tâm của cộng đồng thay đổi theo. Chủ đề, mức độ và thời gian quan tâm chủ đề cùng với cá nhân tham gia cộng đồng là những đặc trưng của cộng đồng. Sự thay đổi các đặc trưng của cộng đồng thường phụ thuộc vào hai nguyên nhân chính: (i) là thông qua sở thích của từng cá nhân trên mạng cùng kết bạn với nhau hoặc cùng quan tâm đến những chủ đề dựa trên nội dung thông điệp mà cá nhân quan tâm trao đổi; (ii) là hình thành hay thay đổi từ nhóm các bạn bè biết trước và cùng kết bạn trên mạng hoặc thông qua sự giới thiệu bạn bè cùng kết bạn. Thách thức đặt ra trong nghiên cứu này mỗi cộng đồng quan tâm đến nhiều chủ đề và mỗi chủ đề có nhiều cộng đồng quan tâm. Bên cạnh đó, đặc trưng của cộng đồng như: chủ đề quan tâm và thành viên tham gia thường thay đổi theo thời gian. Đây cũng là một thách thức đặt ra cho việc phân tích sự biến thiên đặc trưng của cộng đồng.

Kết quả chương này được thể hiện trong công bố [CB10] về phương pháp khám phá các cụm cá nhân dựa trên đặc trưng của các vector chủ đề của cá

nhân, việc tìm kiếm cụm cá nhân trong công bố này chưa quan tâm đến yếu tố thời gian; trong công bố [CB06] về khám phá cộng đồng cá nhân theo thời gian; trong công bố [CB01] phân tích sự biến thiên cộng đồng MXH dựa trên các đặc trưng của cộng đồng như là cá nhân, chủ đề quan tâm và mức độ mà cộng đồng quan tâm đến từng chủ đề trong từng giai đoạn thời gian.

4.2 Khám phá cộng đồng trên mạng xã hội

4.2.1 Khái niệm cộng đồng mạng xã hội theo chủ đề

Tập hợp các cộng đồng trên mạng được ký hiệu là *C* và một cộng đồng đang xét được ký hiệu là *c*, ta có $c \in C$.

Định nghĩa 5.1: *Cộng đồng* [49]

Cộng đồng là một tập thể cùng sống và làm việc trong cùng một môi trường.

Định nghĩa 5.2: *Cộng đồng MXH* [71]

Cộng đồng MXH là một tập hợp các cá nhân tương tác thông qua các phương tiện truyền thông cụ thể, có khả năng vượt qua những ranh giới địa lý và chính trị để theo đuổi lợi ích hay mục tiêu chung.

Định nghĩa 5.3: *Cộng đồng MXH theo chủ đề* [CB01]

Dựa theo mô hình chủ đề, cộng đồng là tập hợp các cá nhân cùng quan tâm đến các chủ đề. Mỗi cá nhân trong cộng đồng được đặc trưng bằng một vector chủ đề quan tâm và mức độ cùng quan tâm đến chủ đề trong cộng đồng nhiều hơn so với những cộng đồng khác. Cho *c* là một cộng đồng theo chủ đề, $c \in C$, trong đó *C* là tập hợp các cộng đồng. Cộng đồng là một phân hoạch với các đặc tính như cụm, ký hiệu $C = \{C_1, C_2, C_3, C_4,...,C_K\}$ với K là số cộng đồng, mỗi cộng đồng C_i có tập vector chủ đề:

(1) Rời nhau: $C_i \cap C_j = \emptyset$ nếu hai cộng đồng không cùng quan tâm đến một hay nhiều chủ đề cụ thể nào đó (hình 5.2).

(2) Và hợp các cộng đồng $\bigcup_{i=1}^{K} C_i = C$.

4.3 Xây dựng phương pháp khám phá cộng đồng dựa theo mô hình chủ đề

4.3.1. Ý tưởng về khám phá cộng đồng

4.3.2 Phương pháp gom cụm và vấn đề trực quan hóa dữ liệu

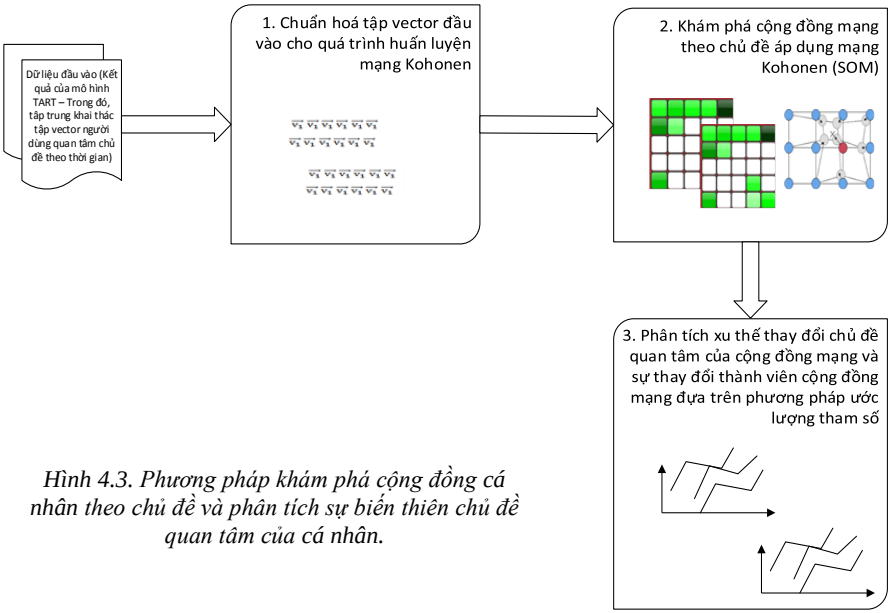
4.3.3 Xây dựng phương pháp khám phá cộng đồng

Phương pháp khám phá cộng đồng cá nhân trên MXH dựa theo mô hình chủ đề để khám phá cộng đồng [CB1][CB6] với 2 nhiệm vụ chính: (i) xây dựng phương pháp khám phá cộng đồng dựa theo mô hình chủ đề có yếu tố thời gian. Trong đó, thông qua kết quả khảo sát, phân tích và đánh giá các mô hình liên quan khám phá cộng đồng, luận án chọn phương pháp huấn luyện

Kohonen; (ii) huấn luyện mạng nơ ron Kohonen kết hợp chuẩn hóa tập dữ liệu đầu vào (kết quả từ mô hình TART [CB05]) là tập các vector chủ đề quan tâm của cá nhân theo thời gian

i. Chuẩn hoá vector nhập

ii. Khám phá cộng đồng sử dụng mạng nơ ron Kohonen



Hình 4.3. Phương pháp khám phá cộng đồng cá nhân theo chủ đề và phân tích sự biến thiên chủ đề quan tâm của cá nhân.

iii. Phân tích sự biến thiên đặc trưng của cộng đồng

4.3.4 Phát biểu bài toán khám phá chủ đề quan tâm của cộng đồng MXH

Cho: tập vector nhập (vector chủ đề quan tâm của cá nhân) $\{v_i\}$ là kết quả từ mô hình TART. Vector v_i có m chiều, $v_i <v_{i1}, v_{i2}, \dots, v_{im}>$, m là số chủ đề quan tâm.

Tìm: danh sách các cộng đồng cá nhân $C = \{C_1, C_2, C_3, C_4, \dots, C_K\}$ quan tâm đến tập chủ đề theo từng giai đoạn thời gian. Với K là số cộng đồng.

Phương pháp: áp dụng phương pháp mạng nơ ron Kohonen kết hợp mô hình chủ đề theo thời gian TART [CB05].

4.4 Thử nghiệm phương pháp khám phá cộng đồng

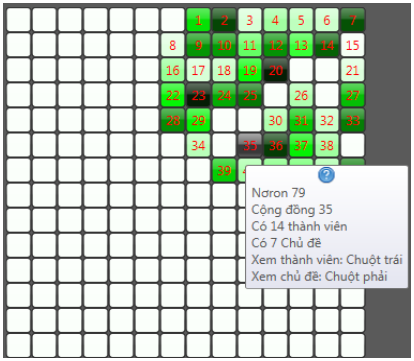
4.4.1 Mô tả dữ liệu thử nghiệm

4.4.2 Chuẩn hoá vector nhập

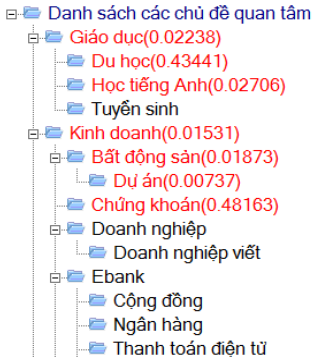
4.4.3 Thử nghiệm phương pháp khám phá cộng đồng

Một kết quả trên hình 4.5, với từng nơron có màu sậm và nhạt tương ứng với số lượng cá nhân nhiều hay ít tham gia vào cộng đồng. Màu sắc trên mỗi

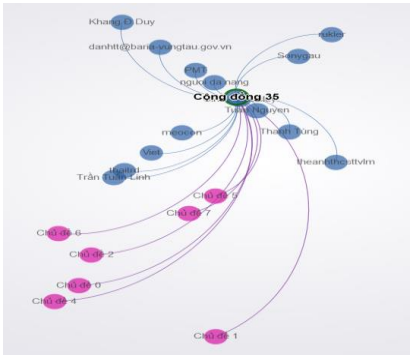
nơron càng đậm đại diện cho số cá nhân trong cộng đồng nhiều hơn những nơron có màu nhạt hơn hoặc cộng đồng không có bất kỳ cá nhân nào (nơron trống không tồn tại cộng đồng).



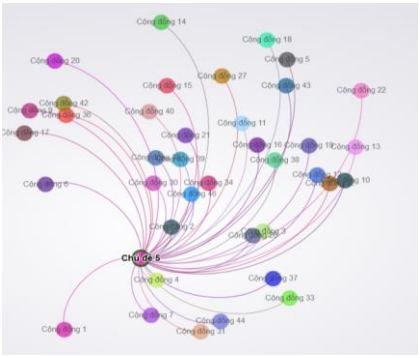
Hình 4.5. Trực quan hóa kết quả khám phá cộng đồng cá nhân trong tháng 01-2015 hiển thị trực quan trên lớp ra Kohonen.



Hình 4.6. Danh sách các chủ đề và xác suất quan tâm của cộng đồng 35 trên lớp ra Kohonen.



Hình 4.7. Trực quan hóa kết quả cộng đồng 35 và các đặc trưng trong cộng đồng.



Hình 4.8. Trực quan hóa kết quả khám phá chủ đề 5 được các cộng đồng quan tâm.

4.4.4 Khảo sát sự biến thiên số cộng đồng dựa trên lớp ra Kohonen

4.5. Phân tích sự biến thiên đặc trưng của cộng đồng theo thời gian

4.5.1 Sự biến thiên đặc trưng của cộng đồng

Sự biến thiên cá nhân tham gia cộng đồng c được biết dựa trên tần suất thay đổi số cá nhân a trong cộng đồng. Ký hiệu là $A(c, t, N_a)$. Trong đó $c \in C$ là cộng đồng, t là thời gian và N_a là số cá nhân tham gia trong cộng đồng c (hay nói cách khác N_a là số cá nhân trong cộng đồng c) theo khoảng thời gian t.

Cá nhân trong cộng đồng cũng là đặc trưng cho cộng đồng đó và việc xác định sự thay đổi số cá nhân trong cộng đồng cũng dựa vào cơ sở này. Sự thay đổi

này thể hiện qua sự khác nhau giữa thành phần của hai tập số cá nhân trong cộng đồng tại thời điểm $t-1$ là $A(c, t\text{-}1, N_a)$ và tại thời điểm t là $A(c, t, N_a)$ mà số cá nhân tham gia cộng đồng. Để đo lường mức độ thay đổi (tính động) số cá nhân a của cộng đồng c tại thời điểm t, luận án đề xuất độ đo $\partial_\theta(c, t-1, t, N_a)$. Độ đo này là tỉ lệ giữa: hiệu số (số cá nhân N_a và phần giao giữa số cá nhân trong cộng đồng tại thời điểm *t-1* với cá nhân trong cộng đồng tại thời điểm t) chia cho cá nhân đề N_a, giá trị của $\partial_\theta(c, t-1, t, N_a)$ nằm trong khoảng [0,1]:

$$\partial_\theta(c, t-1, t, N_a) = \frac{N_a - |A(c, t-1, N_a) \cap A(c, t, N_a)}{N_a} \in [0,1] \quad (4.6)$$

4.5.2 Phương pháp phân tích sự biến thiên đặc trưng của cộng theo thời gian

Cho: tập các cộng đồng được khám phá trên các lớp ra Kohonen theo từng giai đoạn thời gian.

Tìm: sự biết thiên đặc trưng cộng đồng theo từng giai đoạn thời gian.

Phương pháp thực hiện: khảo sát sự liên hệ các cụm trên lớp ra Kohonen. Cụ thể, bài toán tập trung phân tích sự biến thiên chủ đề quan tâm của cộng đồng và cá nhân trên MXH theo từng giai đoạn thời gian.

4.5.3 Kết quả thử nghiệm

4.6 Đánh giá kết quả thử nghiệm phương pháp khám phá cộng đồng

4.6.1 Đánh giá kết quả thông qua khảo sát hệ số Precision, Recall và độ đo F

Áp dụng các hệ số Precision, Recall và độ đo F [66] để đánh giá kết quả gom cụm bằng mạng nơ ron Kohonen. Luận án so sánh kết quả gom cụm vector chủ đề quan tâm của cá nhân theo phương pháp được đề xuất và bằng tay (kết quả gom cụm bằng tay dựa trên dữ liệu là các thông điệp được xây dựng sẵn chủ đề trên trang VnExpress.net) được nhóm theo từng chủ đề bởi cá nhân trên diễn đàn.

4.6.2 Đánh giá kết quả thông qua so sánh với phương pháp gom cụm K-Medoids

Bên cạnh việc áp dụng các hệ số Precision, Recall và độ đo F để đánh giá kết quả thử nghiệm, luận án còn áp dụng giá trị RMSSTD (Root Mean Square Standard Deviation) và giá trị RS (R-Squared) để so sánh kết quả giữa phương pháp gom cụm đề xuất trong luận án và giải thuật K-Medoids).

Sau khi tính giá trị trung bình RMSSTD, phương pháp mạng nơ ron Kohone cho kết quả RMSSTD thấp nhất cho tất cả các lựa chọn số cụm. Điều này cho thấy rằng, phương pháp mạng nơ ron Kohonen có kết quả thực hiện vượt trội hơn so với giải thuật K-Medoids. Bên cạnh đó, kết quả cho thấy rằng thuật toán dựa theo phương pháp mạng nơ ron Kohonen (SOM) mang lại những giá trị RMSSTD là thấp nhất và giá trị RS là cao nhất.

Bảng 4.8. Bảng kết quả giá trị trung bình RMSSTD dựa trên thử nghiệm hai phương pháp gom cụm

Số cụm k	Kohonen	K-Medoids
2	**0.69635**	0.75288
3	**0.58297**	0.65064
4	**0.52873**	0.59444
5	**0.49807**	0.55666
6	**0.47517**	0.52774
7	**0.45634**	0.50502
8	**0.44195**	0.48648

Bảng 4.9. Bảng kết quả giá trị trung bình RS dựa trên thử nghiệnm hai phương pháp gom cụm

Số cụm k	Kohonen	K-Medoids
2	**0.49659**	0.40112
3	**0.63921**	0.55356
4	**0.70391**	0.63431
5	**0.74951**	0.68794
6	**0.78086**	0.72456
7	**0.8034**	0.75273
8	**0.82022**	0.77574

4.7 Kết luận chương

Trong chương 4, luận án tập trung khai thác mô hình TART (được trình bày trong chương 4) kết hợp mạng nơ ron Kohonen. Phương pháp gồm 2 nhiệm vụ chính: (i) khám phá cộng đồng những cá nhân cùng quan tâm đến chủ đề được gọi là cộng đồng MXH theo chủ đề. Phương pháp này dựa trên mô hình chủ đề TART và mạng nơ ron Kohonen; (ii) phân tích sự biến thiên đặc trưng của cộng đồng trên MXH. Kết quả thử nghiệm trên tập vector chủ đề quan tâm của cá nhân có yếu tố thời gian và đánh giá bằng các độ đo Precision, Recall và F, cho thấy phương pháp khám phá cộng đồng được luận án xây dựng, đã giải quyết được yêu cầu đặt ra của bài toán 2 và cho kết quả khả quan.

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận

Những nội dung từ chương 1 đến chương 4, luận án đã bám sát vào mục tiêu, nhiệm vụ và bài toán nghiên cứu được đặt ra, thử nghiệm mô hình đề xuất trên dữ liệu được thu thập từ MXH, kết quả thử nghiệm được thảo luận và đánh giá một cách cẩn thận. Điều này chứng tỏ kết quả đạt được về mặt khoa học và thực tiễn của luận án cũng như giúp xác định những vấn đề cần nghiên cứu trong hướng phát triển luận án. Các đóng góp chính của luận án:

(i) Đóng góp thứ nhất: xây dựng phương pháp kết hợp khám phá chủ đề ẩn từ mối liên kết xã hội là thông điệp được cá nhân trao đổi trên MXH và gán nhãn chủ đề dựa trên cây phân cấp chủ đề.

(ii) Đóng góp thứ hai: xây dựng mô hình TART dựa theo mô hình chủ đề để khám phá chủ đề quan tâm của cá nhân có yếu tố thời gian và phân tích vai trò của cá nhân trên MXH. Mô hình TART độc lập với ngôn ngữ.

(iii) Đóng góp thứ ba: xây dựng phương pháp khám phá cộng đồng cá nhân dựa theo mô hình chủ đề có yếu tố thời gian và phân tích sự biến thiên đặc trưng của cộng đồng.

(iv) Đóng góp thứ tư: Luận án đã xây dựng một hệ thống phần mềm phân tích MXH thực hiện đầy đủ sáu mô-đun trên sơ đồ nghiên cứu tổng thể của luận án (hình .2) từ mô-đun thu thập, tiền xử lý dữ liệu, thực nghiệm khám phá và gán nhãn chủ đề ẩn, thực nghiệm mô hình TART và phương pháp khám phá cộng đồng. Kết quả thực nghiệm đã cho thấy được hướng ứng dụng nghiên cứu của luận án và khả năng khai thác hiệu quả của phần mềm vào ứng dụng thực tế.

5.2 Hạn chế và hướng phát triển

Kết quả nghiên cứu của luận án tập trung vào việc giải quyết các bài toán về khám phá chủ đề ẩn, phân tích chủ đề quan tâm của cá nhân và khám phá cộng đồng cá nhân dựa trên chủ đề được khám phá từ liên kết xã hội là thông điệp mà cá nhân trao đổi trên MXH. Tuy nhiên, luận án còn những hạn chế và đặt ra hướng nghiên cứu tiếp theo:

- Phân tích MXH trên các liên kết xã hội khác như: thích (like), chia sẽ (share), đính kèm (tag),…Trên cơ sở đó, luận án sẽ phát triển và thử nghiệm mô hình LDA với dữ liệu lớn (Big data).
- Phân tích ảnh hưởng lan truyền chủ đề trên MXH. Mục tiêu phân tích ảnh hưởng lan truyền thông điệp trên MXH nhằm xác định "đường đi" và tìm ra nguồn gốc của thông tin.
- Xây dựng hệ thống khoảng thời gian (có tính chất overlap) để phân tích trực tuyến MXH.

Bên cạnh đó, luận án sẽ tiếp tục nghiên cứu và ứng dụng kết quả của luận án trong các lĩnh vực khác như:

- Tìm kiếm chuyên gia.
- Phân tích hành vi khách hàng.

DANH MỤC CÁC CÔNG BỐ CHÍNH

[CB01] **Thanh Ho** and Phuc Do (2015), *Analyzing the Changes in Online Community based on Topic Model and Self-Organizing Map*, International Journal of Advanced Computer Science and Applications (IJACSA), 6(7), 2015, pp. 100-108, ISSN: 2158-107X, DOI: 10.14569/IJACSA.2015.060715, ESCI, Thomson Reuters, 2015.

[CB02] **Thanh Ho**, Duy Doan, Phuc Do (2014), *Discovering Hot Topics On Social Network Based On Improving The Aging Theory*, Advances in Computer Science: an International Journal. Volume 3, Issue 3, pp. 48-53, ISSN: 2322-5157, 2015.

[CB03] **Hồ Trung Thành,** Đỗ Phúc (2014), *Ontology tiếng Việt trong lĩnh vực giáo dục đại học*, Tạp chí Khoa học Công nghệ - Viện Hàn lâm Khoa học Công nghệ Việt Nam, Tập 52, số 1B, pp. 89-100, ISSN: 0866-708x, 2014.

[CB04] **Hồ Trung Thành,** Đỗ Phúc (2014), *Mô hình tích hợp khám phá và gán nhãn chủ đề tiếp cận theo mô hình chủ đề*, Tạp chí Phát triển Khoa học Công nghệ ĐHQG-HCM, số K4, tập 17, pp. 73-85, ISSN: 1859-0128, 2014.

[CB05] **Thanh Ho**, Phuc Do (2014), *Analyzing Users' Interests with the Temporal Factor Based on Topic Modeling*, ACIIDS 03-2015, Indonesia, Springer, pp. 106-115, ISSN: 0302-9743, ISBN: 978-3-319-15704-7, DOI: 10.1007/978-3-319-15705-4_11, Scopus, 2015.

[CB06] **Thanh Ho**, Phuc Do (2015), *Discovering Communities of Users on Social Networks Based on the Topic Model Combined with Kohonen Network*, KSE 10/2015, UIT, Vietnam, 10/2015, INSPEC Accession Number: 15699266, pp. 268-273, DOI:10.1109/KSE.2015.54**,** IEEE, 2015.

DANH MỤC CÁC CÔNG BỐ LIÊN QUAN

[CB07] Nghe Nguyen, **Thanh Ho** and Phuc Do (2015), *Finding the Most Influential User of a Specific Topic on the Social Networks*, Advances in Computer Science : an International Journal, Volume 4, Issue 2, pp. 31-40, ISSN: 2322-5157, 2015.

[CB08] Phan Hồ Viết Trường, **Hồ Trung Thành**, Đỗ Phúc (2013), *Phân tích tầm ảnh hưởng đối tượng theo chủ đề trong mạng xã hội*, Tạp chí Khoa học Công nghệ, Viện Hàn lâm Khoa học Công nghệ Việt Nam, tập 52, số 1B, pp. 101-111, ISSN: 0866-708x, 2013.

[CB09] Muon Nguyen, **Thanh Ho**, Phuc Do (2013), *Social Networks Analysis Based on Topic Modeling*, The 10th IEEE RIVF International Conference on Computing and Communication Technologies, Hanoi, pp. 119-123, ISBN: 978-1-4799-1350-3, 2013.

[CB10] Tran Quang Hoa, Vo Ho Tien Hung, Nguyen Le Hoang, **Ho Trung Thanh**, Do Phuc (2014), *Finding the Cluster of Actors in Social Network based on the Topic of Messages*, ACIIDS 04-2014, ThaiLand, Springer, pp. 183-190, ISBN: 983-3-319-054756-6, Scopus, 2014.

THAM GIA ĐỀ TÀI

Xây dựng hệ thống phân tích mạng xã hội theo chủ đề và ứng dụng vào mạng xã hội trong trường đại học. Mã số đề tài: B2013-26-02. Chủ nhiệm đề tài: PGS.TS Đỗ Phúc. Đề tài cấp ĐHQG-HCM loại B, đã nghiệm thu vào tháng 10/2015, đạt loại tốt.