HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

****

NGUYỄN THỊ HỘI

MÔ HÌNH HÀNH VI VÀ QUAN TÂM CỦA NGƯỜI DÙNG TRÊN CÁC MẠNG XÃ HỘI

Chuyên ngành: Hệ thống thông tin

Mã số : 9.48.01.04

TÓM TẮT LUẬN ÁN TIẾN SĨ KỸ THUẬT

HÀ NỘI – 2021

Công trình hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học:

1. PGS.TS Trần Đình Quế

2. PGS.TS. Đàm Gia Mạnh

Phản biện 1: ……

Phản biện 2: ……

Phản biện 3: ……

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

Học viện Công nghệ Bưu chính Viễn thông

Vào hồi:……giờ, ngày……..tháng…….năm…….

Có thể tìm hiểu luận án tại:

Thư viện Quốc gia Việt Nam

Thư viện Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Tính cấp thiết của luận án

Ảnh hưởng của mạng xã hội đến mọi mặt trong đời sống xã hội đang ngày càng khẳng định rõ vai trò của chúng trong nhiều lĩnh vực từ giáo dục, kinh doanh, sức khỏe, du lịch… đến các vấn đề xã hội như phát hiện gian lận hoặc lừa đảo, phát hiện tâm lý tội phạm, bạo lực xã hội, phát hiện tin tức giả (fake news) được thể hiện trong nhiều công trình nghiên cứu như [30] [37] [38] [73] [81] [93] [137] [146].

Các nghiên cứu về khai phá quan tâm của người dùng (*user interest*) có vai trò quan trọng đối với các tổ chức, doanh nghiệp trong các chiến dịch quảng bá thương hiệu, giới thiệu sản phẩm, gợi ý dịch vụ, đặc biệt có nhiều ứng dụng trong thực tế như [1] [2] [9] [12] [16] [18] [22] [25]: xây dựng hệ thống khuyến nghị người dùng (*user recommendation system*); các ứng dụng của các chương trình hay chiến lược quảng cáo (*advertising campaign*); ứng dụng hệ thống giới thiệu sản phẩm (*product introduction systems*)…

Theo khảo sát của luận án, có một số cách phát hiện mối quan tâm của người dùng phổ biến trên các trang mạng xã hội bao gồm:

- Phát hiện quan tâm của người dùng dựa trên trích xuất thông tin cá nhân (*profile*) [14] [31] [103] [166];
- Phát hiện quan tâm của người dùng dựa trên phân tích các liên kết của người dùng (*follows, link*) [4] [25] [28] [43] [48] [107];
- Phát hiện quan tâm của người dùng dựa trên phân tích hành vi thích, đánh dấu hoặc đăng bài (*like, tags, post*) [50] [63] [76] [77] [87] [108] [121] [144].

Các nghiên cứu về phát hiện quan tâm của người dùng trên các mạng xã hội gần đây thường đi theo hai hướng tiếp cận chính:

- Tập trung phân tích về các liên kết, cấu trúc của mạng xã hội, các kết nối quan hệ bạn bè, danh sách những người được theo dõi... của người dùng trên các mạng xã hội như trong [4] [21] [23] [28] [43] [60] [105] [108] [111]
- Tập trung phân tích các bài đăng, các thẻ đánh dấu, các bài chia sẻ, các bình luận và các đối tượng được tạo ra trong quá trình hoạt động của người dùng trên các mạng xã hội [107], [114] [118] [124] [125] [143] [145] [157] [159], hướng tiếp cận này sẽ loại bỏ được vấn đề về cấu trúc mạng, sự khó khăn trong tiếp cận thông tin cá nhân người dùng cũng như trong thu thập các liên kết bạn bè của người dùng.

Có rất ít nghiên cứu xem xét sự liên quan hay mối tương quan giữa những người dùng có cùng quan tâm với nhau. Ví dụ như: có hai người dùng ***a*** và ***b***, cùng quan tâm đến các trận đấu bóng đá ngoại hạng. Họ thường xuyên đăng, thích, bình luận các bài viết về các trận đấu, về một số cầu thủ, về lịch trình thi đấu của một số câu lạc bộ... Khi đó có thể nói rằng hai người dùng ***a*** và ***b*** có cùng quan tâm đến nội dung bóng đá hoặc rộng hơn là chủ đề thể thao.

Câu hỏi đặt ra là: Khi có một bài viết về một trận đấu bóng đá mà người dùng ***a*** thích và chia sẻ lại thì liệu người dùng ***b*** có thích và chia sẻ lại bài viết đó hay không? Hoặc liệu hai người dùng này có thể cùng tham gia một nhóm có các chủ đề về bóng đá hay không? Hoặc khi có một sự kiện thể thao nào đó xảy ra trên mạng xã hội, nếu người dùng ***b*** chú ý đến và theo dõi sự kiện đó thì liệu người dùng ***a*** có quan tâm và theo dõi sự kiện đó hay không?

Để trả lời các câu hỏi này, ngoài việc xác định được chủ đề quan tâm của từng cá nhân người dùng thì còn cần phải làm rõ ràng hơn *mối tương quan giữa các chủ đề quan tâm của người dùng đó với những người dùng khác trên mạng xã hội*.

Mục tiêu của luận án

- Thứ nhất, mô hình hóa bài viết của người dùng trên các mạng xã hội dựa trên nhiều đặc trưng và phân loại các bài viết đó theo các chủ đề. Các bài viết được luận án đề xuất biểu diễn dựa trên năm đặc trưng gồm: nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc. Dựa trên cách biểu diễn này luận án ước lượng độ tương quan của các bài viết với các chủ đề nhằm phát hiện các quan tâm của người dùng theo các chủ đề đó.
- Thứ hai, mô hình hóa người dùng trên các mạng xã hội theo các hành vi và phân loại họ dựa trên các chủ đề mà họ quan tâm. Luận án đề xuất biểu diễn người dùng trên các mạng xã hội dựa trên các hành vi đăng bài viết, chia sẻ bài viết, thích bài viết, tham gia nhóm trên các mạng xã hội. Dựa trên cách biểu diễn người dùng này, luận án ước lượng độ tương quan giữa các người dùng theo các chủ đề để tìm ra các quan tâm của họ.
- Cuối cùng, ước lượng độ tương tự giữa hai người dùng theo các chủ đề và xem xét mối tương quan giữa những người dùng đó dựa trên các hành vi họ đã thực hiện.

Đối tượng nghiên cứu

Với mục tiêu đã đề ra của luận án, đối tượng nghiên cứu của luận án bao gồm: Các kỹ thuật và phương thức tiền xử lý cho các văn bản ngắn; Các mô hình và phương pháp ước lượng độ tương tự giữa hai đối tượng có nhiều đặc trưng .

Phạm vi nghiên cứu

- Nghiên cứu và phân tích các đối tượng chứa văn bản sinh ra dựa trên hoạt động của người dùng cùng các hành vi của người dùng trên mạng xã hội.
- Nghiên cứu và phân tích các chủ đề trên mạng xã hội cùng các độ đo tương tự giữa các đối tượng trên mạng xã hội.

Các phương pháp nghiên cứu:

- Phân tích, so sánh, tổng hợp, đánh giá trên các kết quả nghiên cứu đã có, từ đó đề xuất hướng giải quyết và cách tiếp cận của luận án
- Kiểm nghiệm các mô hình đề xuất bằng các thực nghiệm và đánh giá

Phương pháp đánh giá

Trong luận án này, việc thực hiện đánh giá hiệu suất hoặc độ chính xác của các mô hình đề xuất được tính toán dựa theo một số phương pháp như sau: Đánh giá dựa trên độ chính xác (*accuracy*), độ nhạy (*recall*) và đánh giá dựa trên độ lệch trung bình như các nghiên cứu [13] [15] [42] [56] [80] [106] [156].

Những đóng góp chính của luận án

- Thứ nhất đề xuất biểu diễn bài viết và các chủ đề bằng véctơ; xây dựng độ đo tương tự giữa hai bài viết và độ tương quan giữa bài viết với các chủ đề.
- Thứ hai đề xuất mô hình biểu diễn bài viết mở rộng dựa trên năm đặc trưng là nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc; xây dựng độ đo tương tự giữa hai bài viết mở rộng và độ tương quan giữa bài viết với các chủ đề.
- Thứ ba đề xuất mô hình biểu diễn người dùng dựa trên các hành vi đăng/chia sẻ bài viết, thích bài viết, bình luận trong bài viết và tham gia các nhóm trên mạng xã hội; xây dựng độ đo tương tự giữa hai người dùng theo các hành vi và độ tương quan giữa hành vi của người dùng với các chủ đề.

Bố cục luận án

Ngoài phần mở đầu, kết luận và hướng phát triển cùng tài liệu tham khảo, luận án được chia thành 4 chương như sau:

Chương 1: Tổng quan về hành vi, quan tâm và mô hình người dùng trên các mạng xã hội.

Chương 2: Mô hình và quan tâm của người dùng theo nội dung bài viết.

Chương 3: Mô hình và quan tâm của người dùng dựa trên bài viết mở rộng nhiều đặc trưng.

Chương 4: Hành vi và quan tâm của người dùng theo các hành vi.

CHƯƠNG 1: TỔNG QUAN VỀ HÀNH VI, QUAN TÂM VÀ MÔ HÌNH NGƯỜI DÙNG TRÊN CÁC MẠNG XÃ HỘI

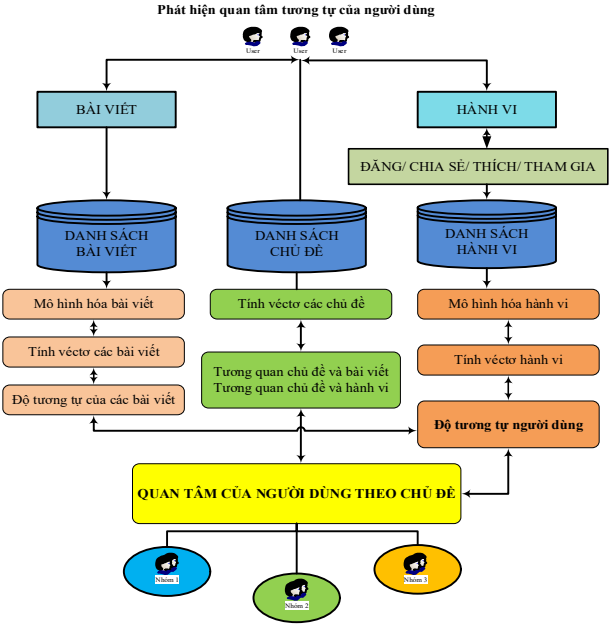
1.1. Mạng xã hội và hành vi của người dùng trên mạng xã hội

Mạng xã hội

Mạng xã hội hay còn gọi là mạng xã hội ảo (social network) là một cấu trúc xã hội được tạo ra bởi cá nhân hoặc các tổ chức (gọi là các "node - nút"). Theo nghiên cứu [41] và [156] thì các mạng xã hội là các dịch vụ dựa trên web cho phép các cá nhân có thể: (1) tạo lập một hồ sơ công khai hoặc bán công khai trong hệ thống có giới hạn, (2) kết nối hoặc chia sẻ với một danh sách người dùng, và (3) cho phép xem, chia sẻ những nội dung thực hiện bởi những người dùng khác trong hệ thống.

Dữ liệu trên mạng xã hội

Theo nghiên cứu [132] [156] thì dữ liệu trên mạng xã hội hay dữ liệu xã hội (social data) là dữ liệu nhận được từ các phương tiện truyền thông xã hội như các trang mạng xã hội, các trang web tìm kiếm, các trang thương mại điện tử, các trang chia sẻ hình ảnh, video ...



Hình 0.2: Những vấn đề nghiên cứu của luận án

Người dùng và cộng đồng người dùng trên các mạng xã hội

Người sử dụng hay người dùng (user) trên các mạng xã hội là những người tham gia vào các mạng xã hội đó, họ thiết lập các kết nối với người dùng khác và có thể trao đổi với nhau, đọc tin tức, chơi trò chơi, tham gia vào các nhóm, tạo ra các thông tin, chia sẻ thông tin, chia sẻ dữ liệu trên các mạng xã hội [8] [9] [23] [35] [41] [51]

Cộng đồng người dùng theo [4] [9] [35] [41] [54] [64] [111] là một tập hợp người dùng trên một mạng xã hội cùng chia sẻ các sở thích, quan tâm chung về một sự kiện, đối tượng hay chủ đề nào đó. Họ có mối liên kết chặt chẽ với nhau theo cùng một mối quan tâm chung hơn so với những người dùng khác.

Mô hình người dùng trên các mạng xã hội

Mô hình người dùng (*user moderling*) là cách thức biểu diễn thông tin cá nhân của người dùng thông qua các đặc trưng mà người dùng thể hiện trên các mạng xã hội. Mô hình người dùng theo các nghiên cứu [8] [9] [135] [18] thường được xây dựng dựa trên các đặc trưng sau của người dùng:

- Đặc điểm cá nhân hoặc nhân khẩu học (*personal characteristics or demographics*)
- Quan tâm và sở thích (*interests and preferences*)
- Nhu cầu và mục tiêu (*needs and goals*)
- Trạng thái tinh thần và thể chất (*mental and physical state*)
- Nền tảng tri thức (*knowledge and background*)
- Hành vi của người dùng (*user behavior*)
- Ngữ cảnh (*context*) là những thông tin mô tả đặc trưng của tình huống mà sự việc xảy ra, trên mạng xã hội
- Đặc điểm tính cách cá nhân (*individual traits*)

Quan tâm của người dùng trên mạng xã hội

Chủ đề trên các trang mạng xã hội

Hành vi của người dùng trên các mạng xã hội

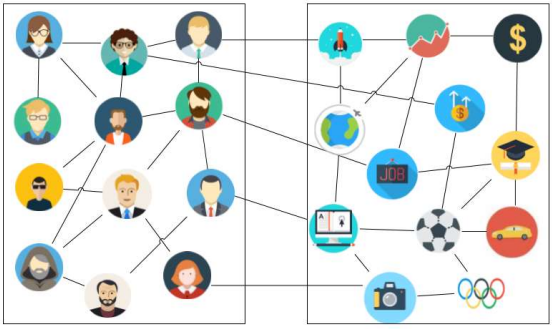
1.2. Phát biểu bài toán và hướng tiếp cận

Phát biểu bài toán và câu hỏi nghiên cứu

Bài toán phát hiện các chủ đề quan tâm của người dùng dựa trên hành vi có thể phát biểu như sau: *Cho một tập các chủ đề trên một mạng xã hội và một tập hợp người dùng cùng các đặc trưng của họ trên mạng xã hội đó, cần đưa ra danh sách các chủ đề mà những người dùng quan tâm, chú ý đến dựa trên việc phân tích các hành vi đặc trưng của những người dùng đó.*

Những câu hỏi cần giải quyết của bài toán bao gồm:

- *Đối tượng nghiên cứu được lựa chọn của bài toán là gì?*
- *Những người dùng trên các mạng xã hội được biểu diễn như thế nào để phân tích và ước lượng nhằm phát hiện các quan tâm của họ?*
- *Các phương pháp hay các kỹ thuật nào sẽ được sử dụng?*
- *Các chủ đề quan tâm được xây dựng và biểu diễn như thế nào?*



Hình 1.1. Minh họa bài toán phát hiện chủ đề quan tâm của người dùng

(Nguồn: Dhelm S.N. et al. [47])

Ứng dụng của phát hiện quan tâm của người dùng trên mạng xã hội

Theo [132] thì các nghiên cứu dữ liệu xã hội chủ yếu dựa trên ba học thuyết: *thuyết tương quan xã hội, thuyết cân bằng và thuyết trạng thái*.

Các nghiên cứu dựa trên các ứng dụng cho người dùng như phát hiện cộng đồng, phân loại các nhóm người dùng và phát hiện người dùng xấu.

Các nghiên cứu dựa trên các mối quan hệ của các người dùng như dự đoán các kết nối của người dùng, dự đoán các kết nối xã hội chặt chẽ và dự đoán các mối quan hệ lâu dài của các nhóm người dùng.

Các nghiên cứu dựa trên nội dung của các đối tượng được sinh ra bởi người dùng như các bài toán khuyến nghị người dùng, các bài toán trích chọn đặc trưng và các bài toán phân tích quan điểm.

Các hướng tiếp cận của bài toán

Theo [10] [54] và [60] thì bài toán phát hiện quan tâm của người dùng trên các mạng xã hội thường được xem xét dựa trên *nguồn thông tin được phân tích, cách thức biểu diễn các chủ đề được so sánh, các kỹ thuật được sử dụng để khai thác các mô hình và các phương pháp để đánh giá*

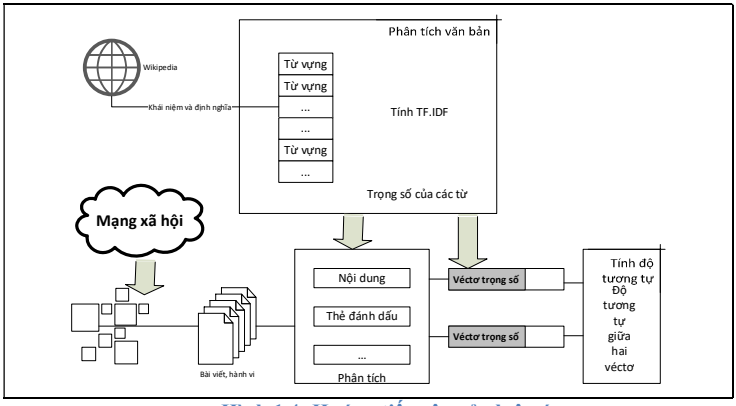
Các bước xây dựng hồ sơ quan tâm của người dùng

Theo [9] và [54] thì quá trình xây dựng hồ sơ quan tâm của người dùng (*user interest profile*) là quá trình thu thập, trích xuất và biểu diễn cho các chủ đề quan tâm của người dùng. Quá trình này thường có ba giai đoạn: *Thu thập dữ liệu, xây dựng đặc trưng và đưa vào các ứng dụng*

Hướng nghiên cứu của luận án

Hình 1.4 mô tả hướng nghiên cứu của luận án với bài toán xây dựng hồ sơ thông tin quan tâm của người dùng gồm hai giai đoạn chính:

- ***Giai đoạn thu thập dữ liệu phân tích***
- ***Giai đoạn xây dựng hồ sơ quan tâm của người dùng***



Hình 1.4: Hướng tiếp cận của luận án

CHƯƠNG 2: MÔ HÌNH VÀ QUAN TÂM CỦA NGƯỜI DÙNG THEO NỘI DUNG BÀI VIẾT

2.1. MÔ HÌNH NGƯỜI DÙNG THEO NỘI DUNG BÀI VIẾT

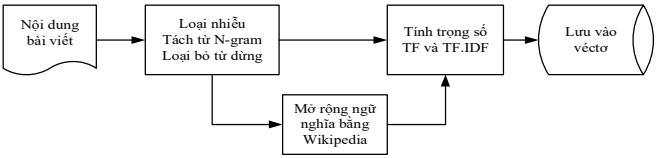
2.1.1. Biểu diễn véctơ bài viết bằng TF.IDF

a. Bài viết trên mạng xã hội

Bài viết của người dùng trên các mạng xã hội là các bài đăng mà người dùng tạo ra hoặc chia sẻ lại từ các nguồn khác trên mạng Internet, một bài viết trên một mạng xã hội có thể là một video clip, một hoặc một số bức ảnh, một văn bản, hoặc một sự kết hợp những thành phần này.

b. Xử lý văn bản ngắn

Theo [33] [53] [80] [119] [130] thì phương pháp xử lý cho dữ liệu văn bản ngắn gồm hai bước chính: Thứ nhất, làm sạch và tách từ theo N-gram; Thứ hai, mở rộng ngữ nghĩa (nếu cần), loại bỏ từ dừng và tính trọng số của từ.



Hình 2.3: Quy trình xử lý nội dung bài viết của luận án

Các bước tiền xử lý dữ liệu văn bản của bài viết được luận án thực hiện qua các bước sau*: làm sạch dữ liệu, tách bài viết thành các từ và thuật ngữ, chuẩn hóa danh sách từ, loại bỏ từ dừng, mở rộng danh sách từ theo Wikipedia*

Quy trình thêm từ vựng bằng mở rộng ngữ nghĩa cho các bài viết được luận án thực hiện theo Thuật toán 2.1 trong Bảng 2.5

Bảng 2.5: Thuật toán 2.1 (Mở rộng ngữ nghĩa theo Wikipedia)

Thuật toán mở rộng từ vựng theo Wikipedia, openWordWiki(x,y)

Input: *Danh sách từ, thuật ngữ của bài viết ngắn x*
Output: *Danh sách từ, thuật ngữ đã mở rộng của bài viết*
Thực hiện:

```
W ← ∅ // Khởi tạo
For i=1 to all(x)
Begin
W[i] ← W[i] ∪ getDefineWiki(x[i]) ;//Lấy định nghĩa
 For j ←2 to 4 do //Tách từ cho định nghĩa
    y ← separateNgram(W[i],j);
 End For
  y ← y ∪ removeStopWord(y);
EndFor
Return
```

c. Biểu diễn văn bản bằng véctơ trọng số

Định nghĩa 2.1:

Cho một tập các văn bản $\mathcal{D} = \{D_1, D_2, \dots, D_p\}$*, mỗi một văn bản được biểu diễn bằng một tập các thuật ngữ* $D_i = \{d_{i1}, d_{i2}, .. d_{ip_i}\}$*. Gọi* $\mathcal{V} = \{v_1, v_2, \dots, v_q\}$*,*

là tập hợp các thuật ngữ khác nhau từng đôi một. Khi đó, trọng số của thuật ngữ $d \in \boldsymbol{\mathcal{V}}$ *đối với* D_i *được tính như sau:*

$$w_d = tf(d, D_i) \times idf(d, \boldsymbol{\mathcal{D}}) \quad (2.1)$$

Trong đó, $tf(d, D_i)$ *là số lần xuất hiện của thuật ngữ* d $trong$ D_i *và* $idf(d, \boldsymbol{\mathcal{D}})$ *được tính bằng*

$$idf(d, \boldsymbol{\mathcal{D}}) = log\ (\frac{\|\boldsymbol{\mathcal{D}}\|}{1+\|\{D_i | d \in D_i\}\|}) \quad (2.2)$$

Để tiện cho việc tính toán, mỗi véctơ được chuẩn hóa về khoảng đơn vị [0,1]. Khi đó có thể định nghĩa văn bản $D_i \in \boldsymbol{\mathcal{D}}$ theo véctơ trọng số như sau:

Định nghĩa 2.2:

Cho một tập các văn bản $\boldsymbol{\mathcal{D}} = \{D_1, D_2, \dots, D_p\}$*, mỗi một văn bản được biểu diễn bằng một tập các thuật ngữ* $D_i = \{d_{i1}, d_{i2}, .. d_{ip_i}\}$*. Gọi q là số các thuật ngữ khác nhau từng đôi một trong không gian* $\boldsymbol{\mathcal{D}}$*. Khi đó, mỗi* D_i *được biểu diễn bởi một véctơ có q chiều:* $\boldsymbol{w_i} = (w_{i1}, w_{i2}, .. w_{iq})$ *trong không gian* $\boldsymbol{\mathcal{D}}$*. Trong đó,* w_{ik} *được tính theo Định nghĩa 2.1.*

d. Biểu diễn nội dung bài viết bằng véctơ trọng số

Định nghĩa 2.3:

Một mạng xã hội $\boldsymbol{\mathcal{N}}$ *là một bộ bốn:* $\boldsymbol{\mathcal{N}} =< U, E, G, B >$*. Trong đó:*

- $U = \{u_i\}$ *là tập những người dùng (user) trên mạng xã hội* $\boldsymbol{\mathcal{N}}$*,* u_i *là kí hiệu người dùng thứ i trong tập U.*
- $E = \{e_i\}$ *là tập các bài đã đăng/đã chia sẻ (entry) trên mạng xã hội* $\boldsymbol{\mathcal{N}}$*,* e_i *là kí hiệu bài đăng thứ i trong tập E.*
- $G = \{g_i\}$ *là tập các nhóm/ cộng đồng người dùng đã tham gia trên mạng xã hội* $\boldsymbol{\mathcal{N}}$*,* g_i *là kí hiệu nhóm thứ i trong tập G.*
- *B là tập các hành vi của người dùng trên mạng xã hội* $\boldsymbol{\mathcal{N}}$**,** *các hành vi được luận án xem xét và phân tích trong chương 4 của luận án*

Bài viết e trên mạng xã hội $\boldsymbol{\mathcal{N}}$ *là một văn bản ngắn được biểu diễn bởi một tập các từ, ký hiệu:* $e = \{w_i\}, i = 1,2, \dots i_q, e \in E$*, với E là tập các bài viết trên mạng xã hội* $\boldsymbol{\mathcal{N}}$*.*

Định nghĩa 2.4:

Cho một tập các bài viết của người dùng $E = \{e_1, e_2, \dots, e_q\}$*, mỗi bài viết được biểu diễn bằng một tập thuật ngữ* $e_i = \{e_{i1}, e_{i2}, .. e_{iq_i}\}$*. Gọi q là số thuật ngữ khác nhau từng đôi một trong không gian E. Khi đó, mỗi* E_i *được biểu diễn bởi một véctơ có q chiều:* $\boldsymbol{w_i} = (w_{i1}, w_{i2}, .. w_{iq})$ *trong không gian E. Trong đó, mỗi* w_{ik} *được tính như trong định nghĩa 2.1.*

d. Các thuật toán tiền xử lý dữ liệu văn bản

Thuật toán 2.2: *Thuật toán phân tách văn bản và xác định từ, thuật ngữ*

Thuật toán 2.3: *Xây dựng véctơ trọng số* cho nội dung các bài viết.

Bảng 2.8: Thuật toán 2.2 (Phân tích văn bản và xác định từ, thuật ngữ)

Thuật toán 2.2: Phân tích bài viết và xây dựng từ, thuật ngữ getTerm(x,y)
Input: Một bài viết trên mạng xã hội
Ouput: Danh sách các từ của văn bản, Term
1: x ←Text; y ←∅; T1 ← ∅; T2 ← ∅; W ← ∅; T3 ← ∅; //Khởi tạo
2: x ← cleanText(x); // Làm sạch văn bản x
3: x ← formatText(x); //Chuẩn hóa các từ vựng trong x
4: For i ←2 to 4 do //Tách từ cho x
T1 ← T1 ∪ separateNgram(x,i) ; // N=2,3,4
End For
5: T2 ← removeStopWord(T1); //Loại bỏ từ dừng
6: If count(T2) ← 10 then //Mở rộng từ vựng nếu cần
Open_word(T2,T3)
Else T3 ← T2;
End If
7: Return T3

Bảng 2.9: Thuật toán 2.3 (Xây dựng các véctơ trọng số cho bài viết)

Thuật toán 2.3: Tính các véctơ trọng số getWeightWord(x)
Input: Danh sách từ, thuật ngữ của bài viết e trên mạng xã hội N
Ouput: Véctơ trọng số TF-IDF của bài viết e
1: w ←∅; wtfidf ←∅; //Khởi tạo
2: For i ← to count(x) do //Đếm tần suất của các từ khóa trong x
w[i] ← count(x[i]) ; N ← tổng số lượng các tài liệu
df_i *← số lượng các tài liệu mà từ* w_i *xuất hiện.*
If w[i]>=1 then
wtfidf[i] ← $(1 + log(f_{ij}))\, log\left(\frac{N}{1+df_i}\right)$
else wtfidf[i] ←0; //Tính TF.IDF
End For
3: Return wtfidf;

2.1.2. Biểu diễn người dùng bằng véctơ

Mỗi người dùng được biểu diễn bởi một véctơ gồm i_{k_i} thành phần, mỗi thành phần là một véctơ được xây dựng theo định nghĩa 2.4. Ký hiệu như sau:

$$u_i = \boldsymbol{u_i} = \left(\boldsymbol{w_{i1}}, \boldsymbol{w_{i2}}, .. \boldsymbol{w_{ik_i}}\right), \boldsymbol{w_{ik}} = (w_{ik1}, w_{ik2}, .. w_{ikq}) | \; k = 1, \ldots, i_{kq}$$

trong không gian E. (2.3)

Cụ thể mỗi người dùng trên mạng xã hội có thể được biểu diễn như sau:

$$u_i = \begin{pmatrix} e_{i1}=\boldsymbol{w}_{i1}=(w_{i11}, w_{i12}, \ldots, w_{i1q}), \\ e_{i2}=\boldsymbol{w}_{i2}=(w_{i21}, w_{i22}, \ldots, w_{i2q}), \\ \ldots \\ e_{ik_i}=\boldsymbol{w}_{ik_i}=(w_{ik1}, w_{ik}\ , \ldots, w_{ikq}) \end{pmatrix} \quad (2.4)$$

Với q là số chiều của không gian E trên mạng xã hội đang xem xét.

2.1.3. Độ đo tương tự và độ tương quan giữa hai đối tượng

Luận án sử dụng độ đo Cosine để tính độ tương tự giữa hai đối tượng theo các véc tơ biểu diễn của hai đối tượng tương ứng như sau: độ tương tự của u và v được tính bằng:

$$sim(u, v) = \frac{<\boldsymbol{u}, \boldsymbol{v}>}{\|\boldsymbol{u}\| * \|\boldsymbol{v}\|} \quad (2.5)$$

Để tính độ tương quan giữa hai đối tượng, luận án sử dụng độ tương quan Pearson theo công thức như sau:

$$cor(\boldsymbol{u},\boldsymbol{v}) = \frac{\sum_i (u_i-\bar{u})(v_i-\bar{v})}{\sqrt{\sum_i (u_i-\bar{u})^2} * \sqrt{\sum_i (v_i-\bar{v})^2}} \quad (2.6)$$

Trong đó, $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$ và $\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$ khi đó, $cor(\boldsymbol{u},\boldsymbol{v})$ là độ tương quan giữa $\boldsymbol{u}$ và $\boldsymbol{v}$.

2.1.4. Độ tương tự giữa hai người dùng theo nội dung bài viết

a. Độ tương tự giữa hai bài viết

Độ tương tự giữa hai bài viết e_{il} và e_{jk} được tính bằng độ tương tự giữa hai véctơ trọng số tương ứng của e_{il} và e_{jk} như sau:

$$sim(\boldsymbol{e_{il}}, \boldsymbol{e_{jk}}) = \frac{<\boldsymbol{e_{il}}, \boldsymbol{e_{jk}}>}{\|\boldsymbol{e_{il}}\| \times \|\boldsymbol{e_{jk}}\|} \quad (2.7)$$

Độ tương tự giữa hai tập bài viết E_i và E_j được tính bằng độ tương tự giữa hai tập các véctơ trọng số tương ứng của u_i và u_j được ký hiệu là:

$$sim(\boldsymbol{E_i}, \boldsymbol{E_j}) = \max_{\text{ik,il}}(sim(\boldsymbol{e_{il}}, \boldsymbol{e_{jk}})) \quad (2.8)$$

b. Độ tương tự giữa hai người dùng theo nội dung bài viết

Định nghĩa 2.5:

Cho hai người dùng u_i và u_j với hai tập bài viết E_i và E_j tương ứng trên mạng xã hội $\boldsymbol{\mathcal{N}}$. Độ tương tự của hai người dùng được tính bằng:

$$sim(u_i, u_j) = sim(\boldsymbol{u_i}, \boldsymbol{u_j}) = sim(\boldsymbol{E_i}, \boldsymbol{E_j}) \quad (2.9)$$

2.2. MÔ HÌNH QUAN TÂM CỦA NGƯỜI DÙNG THEO CHỦ ĐỀ

2.2.1. Biểu diễn véctơ trọng số của chủ đề

Khái niệm về chủ đề như sau: *Cho một tập các chủ đề về các lĩnh vực trên mạng xã hội. Khi đó, mỗi một chủ đề sẽ được biểu diễn bởi một tập hợp từ, thuật ngữ đặc trưng để mô tả và diễn giải về chủ đề đó.*

Giả sử rằng $\boldsymbol{\mathcal{T}} = \{T_1, T_2, \dots, T_p\}$ là tập các chủ đề trên mạng xã hội $\boldsymbol{\mathcal{N}}$, trong đó mỗi chủ đề được biểu diễn bằng một tập các từ $T_i = \{t_{i1}, t_{i2}, \dots, t_{ip_i}\}$.

Định nghĩa 2.6:

Cho một tập các chủ đề $\boldsymbol{\mathcal{T}} = \{T_1, T_2, \dots, T_P\}$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$, khi đó, mỗi chủ đề T_i được biểu diễn bởi một tập các thuật ngữ hoặc các từ: $T_i = \{t_{i1}, t_{i2}, \dots, t_{ip_i}\}$. Gọi $\boldsymbol{\mathcal{V}_T}$ là tập gồm q từ khác nhau từng đôi một trong tất cả các $T_i \in \boldsymbol{\mathcal{T}}$. Khi đó, mỗi T_i tương ứng một véctơ trọng số được ký hiệu như sau:

$$\boldsymbol{t_i} = (w_{i1}, w_{i2}, \dots, w_{iq}) \quad (2.10)$$

Trong đó, mỗi w_{ik} được tính như trong Định nghĩa 2.1

2.2.2. Xây dựng các chủ đề trên mạng xã hội

Luận án thực hiện lựa chọn các chủ đề bằng cách thống kê các chủ đề trên một số trang tin tức điện tử phổ biến ở Việt Nam và trên thế giới, phương pháp này đã được các nghiên cứu [25] [145] [125]. Các chủ đề phổ biến được thống kê từ 10 trang tin tức điện tử của Việt Nam có lượng người dùng truy cập lớn nhất theo thống kê của *https://toplist.vn/top-list/website* cùng với 5 trang tin tức điện tử bằng Tiếng Anh phổ biến trên thế giới của *https://www.similarweb.com/top-websites/category/news-and-media.* Luận án thu được danh sách gồm 21 chủ đề có tần suất xuất hiện nhiều nhất trên 15 trang tin tức như trong Bảng 2.11 và Bảng 2.12

Thuật toán 2.4: *Xây dựng danh sách từ vựng cho chủ đề*

Thuật toán 2.5: *Xây dựng véctơ trọng số cho mỗi chủ đề.*

Bảng 2.13: Thuật toán 2.4 (Xây dựng danh sách từ vựng cho các chủ đề)

Thuật toán 2.4: Xây dựng từ vựng cho các chủ đề, topicWord()
Input: Chủ đề t trên mạng xã hội N
Ouput: Danh sách các từ vựng của chủ đề t
1: x ←∅; tW ← ∅; //Khởi tạo
2: x ← getDefineWiki(t); // Lấy Định nghĩa từ Wikipedia cho t
3: For i←2 to 4 do //Tách từ cho x
tW ← tW ∪ separateNgram(x,i) ; // N=2,3,4
End For
4: tW ← removeStopWord(tW); //Loại bỏ từ dừng
5: Return tW;

Bảng 2.15: Thuật toán 2.5 (Xây dựng véctơ trọng số cho mỗi chủ đề)

Thuật toán 2.5: Xây dựng véctơ trọng số getWeightTopic()
Input: Một danh sách từ vựng của chủ đề t
Ouput: Véctơ trọng số TF-IDF của chủ đề t

1: w←∅; wtfidftp←∅; //Khởi tạo
2: For i ← to count(t) do //Đếm tần suất của các từ khóa trong t
w[i] ← count(tW[i]) ; N ← số lượng các chủ đề trong T
df_i *← số lượng các chủ đề mà từ khóa* w_i *xuất hiện.*
If w[i]>=1 then
wtfidftp[i] ←$(1 + log(f_{ij})) \, log\left(\frac{N}{df_i}\right)$
else wtfidftp[i] ←0; //Tính TF.IDF
End For
3: Return w, wtfidftp;

Sau khi tính toán xong, luận án thu được một tập gồm 21 véctơ tương ứng với 21 chủ đề chứa danh sách từ và véctơ trọng số tương ứng như công thức (2.11).

$$\mathcal{T} = \begin{pmatrix} t_1 = \mathbf{t_1} = (w_{i1}, w_{i2}, \ldots, w_{iq}), \\ t_2 = \mathbf{t_2} = (w_{i1}, w_{i2}, \ldots, w_{iq}), \\ \ldots \\ t_{21} = \mathbf{t_{21}} = (w_{i1}, w_{i2}, \ldots, w_{iq}) \end{pmatrix} \quad (2.11)$$

Trong đó, mỗi w_{ik} *được tính như trong Định nghĩa 2.1*

2.2.3. Biểu diễn véctơ nội dung bài viết theo chủ đề

Định nghĩa 2.7:

Giả sử $e_{ij} \in e_i$ là một bài viết của người dùng u_i trên mạng xã hội $\mathcal{N}$, được mô tả bởi một tập hợp các từ, khi đó, véctơ trọng số của bài viết e_{ij} đối với chủ đề T_k được định nghĩa như sau:

$$\boldsymbol{e_{ij}^k} = \left(e_{ij}^1, e_{ij}^2, \dots, e_{ij}^{t_{kp}}\right) \quad (2.12)$$

Trong đó, $e_{ij}^l = tf(t_{il}, e_{ij}) \times idf(t_{il}, E_i)$ với $t_{il} \in \mathcal{V}_T$

2.2.4. Độ quan tâm của người dùng theo các chủ đề trên mạng xã hội

Mức độ liên quan giữa bài viết e_{ij} của người dùng u_i đối với chủ đề t_k:

$$\alpha_{ij}^k = cor\left(e_{ij}, t_k\right) \quad (2.13)$$

Mức độ liên quan của bài viết e_{ij} đến p chủ đề trong $\mathcal{T}$ ký hiệu là:

$$cor\left(e_{ij}, p\right) = (\alpha_{ij}^1, \alpha_{ij}^2, \dots, \alpha_{ij}^p) \quad (2.14)$$

Có thể thấy rằng:

(1) Khi số lượng các bài viết của một người dùng về cùng một chủ đề tăng lên thì mức độ quan tâm của người dùng đến chủ đề đó cũng tăng lên.

(2) Khi số lượng các người dùng quan tâm đến một chủ đề tăng lên thì mức độ quan tâm của người dùng đến chủ đề đó cũng tăng lên.

Định nghĩa 2.8:

Hàm số: $int: \mathcal{U} \times \mathcal{P}(E) \times \mathcal{T} \to [0,1]$ được gọi là độ đo quan tâm nếu nó thỏa mãn điều kiện sau: $int(u, U, t) \leq int(v, V, t)$, đối với mọi $U, V \in P(E_u)$ với $U \subseteq V$

Để cho đơn giản khi tính toán và biểu diễn, trong luận án này ký hiệu hàm quan tâm của người dùng u_i đến chủ đề t là $int(u_i, t)$. Dễ dàng chứng minh rằng:

Mệnh đề 2.8.1: *Các hàm số sau:*

$$(i) \quad intMax(u_i, t) = max_j\left(cor\left(e_{ij}, t\right)\right) \quad (2.15)$$

$$(ii) \quad intCor(u_i, t) = \frac{\sum_j cor(e_{ij}, t))}{\|E_i\|} \quad (2.16)$$

$$(iii) \quad intSum(u_i, t) = \frac{1}{2}\left(\frac{n_i^t}{\sum_{l \in \mathcal{T}} n_i^l} + \frac{n_i^t}{\sum_{u_k \in \mathcal{U}, l \in \mathcal{T}} n_k^l}\right)_j \quad (2.17)$$

là các độ đo quan tâm của người dùng đối với các chủ đề.

Trong đó, $cor\left(e_{ij}, t\right)$ là mức độ liên quan của bài viết e_{ij} đến chủ đề *t*, n_i^t là số lượng các bài viết liên quan đến chủ đề *t* của người dùng u_i trên mạng xã hội $\mathcal{N}$.

2.2.5. Tương tự quan tâm theo chủ đề của người dùng

Định nghĩa 2.9:

Độ quan tâm của người dùng u_i đến p chủ đề trong $\mathcal{T}$ là một véctơ quan tâm, được biểu diễn như sau: $\boldsymbol{u_i^t} = (u_i^1, u_i^2, \dots, u_i^p)$ (2.18)

Trong đó, mỗi u_i^k là độ quan tâm của u_i đến chủ đề thứ k, k=1, 2, …, p, các u_i^k được tính theo một trong ba công thức của mệnh đề 2.9.1.

Định nghĩa 2.10:

Độ tương tự theo các chủ đề quan tâm của hai người dùng u_i, u_j được tính bằng độ tương tự cosine giữa hai véctơ quan tâm đến tất cả các chủ đề theo công thức: $sim_{int}(u_i, u_j) = sim(\boldsymbol{u_i^t}, \boldsymbol{u_j^t}) = \frac{<\boldsymbol{u_i^t}, \boldsymbol{u_j^t}>}{\|\boldsymbol{u_i^t}\| \times \|\boldsymbol{u_j^t}\|}$ (2.19)

Trong đó, $<\boldsymbol{u_i^t}, \boldsymbol{u_j^t}>$ là tích vô hướng của hai véctơ, $\|\boldsymbol{X}\|$ là độ dài của véctơ. Dễ dàng thấy rằng, $sim_{int}(u_i, u_j)$ nằm trong khoảng [0,1].

CHƯƠNG 3: MÔ HÌNH VÀ QUAN TÂM CỦA NGƯỜI DÙNG DỰA TRÊN BÀI VIẾT MỞ RỘNG

3.1. XÁC ĐỊNH QUAN TÂM CỦA NGƯỜI DÙNG THEO BÀI VIẾT

3.2. MÔ HÌNH BÀI VIẾT MỞ RỘNG

3.2.1. Mô hình bài viết

Định nghĩa 3.1:

Một bài viết $e_i \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$ được biểu diễn bởi năm đặc trưng: $e_i = \{cont_i, cat_i, tag_i, sent_i, emo_i\}$. Trong đó:

- *$cont_i$ là nội dung (content) của bài viết $e_i \in E$,*
- *cat_i là thể loại (category) của bài viết $e_i \in E$,*
- *tag_i là thẻ đánh dấu (tag) của bài viết $e_i \in E$,*
- *$sent_i$ là quan điểm (sentiment) của bài viết $e_i \in E$,*
- *emo_i là cảm xúc (emotion) trong bài viết $e_i \in E$.*

Như vậy, mỗi bài viết $e_i \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$, được biểu diễn bởi năm đặc trưng là nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc. Các đặc trưng của bài viết được mô tả chi tiết như sau:

- *Nội dung (Content) của bài viết e_i ký hiệu là: $cont_i$.*
- *Thể loại (Category) của bài viết e_i ký hiệu là: cat_i*
- *Thẻ đánh dấu (Tag) của bài viết e_i ký hiệu là: tag_i.*
- *Quan điểm (Sentiment) của bài viết e_i ký hiệu là: $sent_i$*
- *Cảm xúc (Emotion) của bài viết e_i ký hiệu là: emo_i.*

Theo định nghĩa 3.1 và dựa trên các đặc trưng đã xem xét thì mỗi bài viết $e_i \in E$ có thể biểu diễn một cách hình thức như công thức (3.1):

$$e_i = (cont_i, cat_i, tag_i, sent_i, emo_i),\ i = 1,..n, \forall e \in E|\boldsymbol{\mathcal{N}} \quad (3.1)$$

3.2.2. Biểu diễn bài viết bằng véctơ

Các thành phần được phân tích như Định nghĩa 2.2.

Ký hiệu $\boldsymbol{E} = \{e_1, e_2, \dots, e_n\}$ là tập tất các các bài viết đang xét trên mạng xã hội $\boldsymbol{\mathcal{N}}$, khi đó theo Định nghĩa 2.2 ở Chương 2, luận án ký hiệu lần lượt:

- E_{cont} là tập tất cả các từ vựng khác nhau từng đôi một của đặc trưng nội dung của tất cả các bài viết trong E
- E_{cat} là tập tất cả các từ vựng khác nhau từng đôi một của đặc trưng thể loại của tất cả các bài viết trong E

- E_{tag} là tập tất cả các từ vựng khác nhau từng đôi một của đặc trưng thẻ đánh dấu của tất cả các bài viết trong E
- E_{sent} là tập tất cả các từ vựng khác nhau từng đôi một của đặc trưng quan điểm của tất cả các bài viết trong E
- E_{emo} là tập tất cả các từ vựng khác nhau từng đôi một của đặc trưng cảm xúc của tất cả các bài viết trong E

Đặc trưng nội dung: $cont_i = \boldsymbol{v}_{cont} = (w_{i1}, w_{i2}, .. w_{iq})$ (3.2)

Đặc trưng thẻ đánh dấu: $tag_i = \boldsymbol{v}_{tag} = (w_{i1}, w_{i2}, .. w_{ip})$ (3.3)

Đặc trưng thể loại: $cat_i = \boldsymbol{v}_{cat} = (w_{i1}, w_{i2}, .. w_{il})$ (3.4)

Đặc trưng cảm xúc: $emo_i = \boldsymbol{v}_{emo} = (w_{i1}, w_{i2}, .. w_{ir})$ (3.5)

Đặc trưng quan điểm: $sent_i = \boldsymbol{v}_{sent} = (w_{i1}, w_{i2}, .. w_{it})$ (3.6)

Mỗi bài viết $e_i \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$, được mô hình hóa bởi năm đặc trưng nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc, được biểu diễn bởi một véctơ có năm thành phần như trong công thức (3.7).

$$e_i = \begin{cases} cont_i = \boldsymbol{v}_{cont} = (w_{i1}, w_{i2}, .. w_{iq}), \\ cat_i = \boldsymbol{v}_{cat} = (w_{i1}, w_{i2}, .. w_{ip}), \\ tag_i = \boldsymbol{v}_{tag} = (w_{i1}, w_{i2}, .. w_{il}), \\ sent_i = \boldsymbol{v}_{emo} = (w_{i1}, w_{i2}, .. w_{ir}), \\ emo_i = \boldsymbol{v}_{sent} = (w_{i1}, w_{i2}, .. w_{it}) \end{cases} \quad (3.7)$$

3.2.3. Độ tương tự giữa hai bài viết mở rộng

a. Mô hình ước lượng tổng quát

Độ tương tự giữa hai bài viết $e_i, e_j \in E$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$ theo định nghĩa 3.1 được tính như sau:

$$\begin{aligned} s_{entry}(e_i, e_j) = {} & w_{cont} * s_{cont}(cont_i, cont_j) + w_{cat} * s_{cat}(cat_i, cat_j) \\ & + w_{tag} * s_{tag}(tag_i, tag_j) + w_{sent} * s_{sent}(sent_i, sent_j) \\ & + w_{emo} * s_{emo}(emo_i, emo_j) \end{aligned} \quad (3.8)$$

Trong đó, $w_{cont}, w_{cat}, w_{tag}, w_{sent}, w_{emo}$ lần lượt là trọng số trên các đặc trưng nội dung, thể loại, thẻ đánh dấu, quan điểm, và cảm xúc của bài viết, thỏa mãn điều kiện: $w_{cont} + w_{cat} + w_{tag} + w_{sent} + w_{emo} = 1$.

Ước lượng độ tương tự trên từng đặc trưng của bài viết

- *Độ tương tự trên đặc trưng nội dung*

$$s_{cont}(cont_i, cont_j) = sim_{cont}\left(\boldsymbol{v}_{cont_i}, \boldsymbol{v}_{cont_j}\right) = \frac{<\boldsymbol{v}_{cont_i}, \boldsymbol{v}_{cont_j}>}{\|\boldsymbol{v}_{cont_i}\| \times \|\boldsymbol{v}_{cont_j}\|} \quad (3.9)$$

- *Độ tương tự trên đặc trưng thể loại:*

$$s_{cat}(cat_i, cat_j) = sim_{cat}\left(\boldsymbol{v}_{cat_i}, \boldsymbol{v}_{cat_j}\right) = \frac{<\boldsymbol{v}_{cat_i}, \boldsymbol{v}_{cat_j}>}{\|\boldsymbol{v}_{cat_i}\| \times \|\boldsymbol{v}_{cat_j}\|} \quad 3.10)$$

- *Độ tương tự trên đặc trưng thẻ đánh dấu:*

$$s_{tag}(tag_i, tag_j) = sim_{tag}\left(\boldsymbol{v}_{tag_i}, \boldsymbol{v}_{tag_j}\right) = \frac{<\boldsymbol{v}_{tag_i}, \boldsymbol{v}_{tag_j}>}{\left\|\boldsymbol{v}_{tag_i}\right\| \times \left\|\boldsymbol{v}_{tag_j}\right\|} \quad (3.11)$$

- *Độ tương tự trên đặc trưng quan điểm:*

$$s_{sent}(sent_i, sent_j) = sim_{sent}\left(\boldsymbol{v}_{sent_i}, \boldsymbol{v}_{sent_j}\right) = \frac{<\boldsymbol{v}_{sent_i}, \boldsymbol{v}_{sent_j}>}{\left\|\boldsymbol{v}_{sent_i}\right\| \times \left\|\boldsymbol{v}_{sent_j}\right\|} \quad (3.12)$$

- *Độ tương tự trên đặc trưng cảm xúc:*

$$s_{emo}(emo_i, emo_j) = sim_{emo}\left(\boldsymbol{v}_{emo_i}, \boldsymbol{v}_{emo_j}\right) = \frac{<\boldsymbol{v}_{emo_i}, \boldsymbol{v}_{emo_j}>}{\left\|\boldsymbol{v}_{emo_i}\right\| \times \left\|\boldsymbol{v}_{emo_j}\right\|} \quad (3.13)$$

3.3. MÔ HÌNH NGƯỜI DÙNG THEO BÀI VIẾT MỞ RỘNG

3.3.1. Biểu diễn người dùng theo bài viết mở rộng

Mỗi người dùng trên mạng xã hội $\boldsymbol{\mathcal{N}}$ được biểu diễn bởi một véctơ gồm m_i thành phần, mỗi thành phần là một véctơ được xây dựng theo công thức 3.7. Ký hiệu như sau: $u_i = \boldsymbol{u_i} = \left(\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{im_i}}\right)$ (3.14)

Cụ thể mỗi người dùng trên mạng xã hội có thể được biểu diễn như sau:

$$u_i = \left(\begin{array}{c} \boldsymbol{e_{i1}} = \begin{cases} cont_{i1} = \boldsymbol{v}_{cont} = (w_{i1}, w_{i2}, .. w_{iq}), \\ cat_{i1} = \boldsymbol{v}_{cat} = (w_{i1}, w_{i2}, .. w_{ip}), \\ tag_{i1} = \boldsymbol{v}_{tag} = (w_{i1}, w_{i2}, .. w_{il}), \\ sent_{i1} = \boldsymbol{v}_{emo} = (w_{i1}, w_{i2}, .. w_{ir}) \ , \\ emo_{i1} = \boldsymbol{v}_{sent} = (w_{i1}, w_{i2}, .. w_{it}) \end{cases} \\ \ldots \ldots \\ \boldsymbol{e_{im_i}} = \begin{cases} cont_{i\boldsymbol{m_i}} = \boldsymbol{v}_{cont} = (w_{i1}, w_{i2}, .. w_{iq}), \\ cat_{i\boldsymbol{m_i}} = \boldsymbol{v}_{cat} = (w_{i1}, w_{i2}, .. w_{ip}), \\ tag_{i\boldsymbol{m_i}} = \boldsymbol{v}_{tag} = (w_{i1}, w_{i2}, .. w_{il}), \\ sent_{i\boldsymbol{m_i}} = \boldsymbol{v}_{emo} = (w_{i1}, w_{i2}, .. w_{ir}) \ , \\ emo_{i\boldsymbol{m_i}} = \boldsymbol{v}_{sent} = (w_{i1}, w_{i2}, .. w_{it}) \end{cases} \end{array} \right)$$

Với *q, p, l, r, t* là số chiều của các không gian $E_{cont}, E_{cat}, E_{tag}, E_{sent}, E_{emo}$ trên mạng xã hội đang xem xét.

3.3.2. Độ tương tự giữa hai người dùng theo mô hình bài viết mở rộng

Độ tương tự giữa hai tập bài viết E_i và E_j được tính bằng độ tương tự giữa hai tập các véctơ trọng số tương ứng của u_i và u_j được tính như sau:

$$sim(\boldsymbol{E_i}, \boldsymbol{E_j}) = \max_{\text{ik,il}}(sim(\boldsymbol{e_{il}}, \boldsymbol{e_{jk}}))$$

Trong đó các $sim(\boldsymbol{e_{il}}, \boldsymbol{e_{jk}})$ được tính theo công thức (3.8). Khi đó độ tương tự của hai người dùng được tính bằng:

$$sim(u_i, u_j) = sim(\boldsymbol{u_i}, \boldsymbol{u_j}) = sim(\boldsymbol{E_i}, \boldsymbol{E_j}) \quad (3.15)$$

3.4. QUAN TÂM CỦA NGƯỜI DÙNG THEO MÔ HÌNH BÀI VIẾT MỞ RỘNG

3.4.1. Biểu diễn bài viết theo chủ đề

Gọi $e_{ij} \in E_i$ là một bài viết của người dùng u_i trên mạng xã hội $\mathcal{N}$, được mô tả bởi năm đặc trưng, mỗi đặc trưng là một tập hợp các từ. Khi đó, véctơ trọng số của bài viết e_{ij} đối với chủ đề T_k được định nghĩa như sau:

$$\boldsymbol{e_{ij}^{k}} = \left(e_{ij}^{1}, e_{ij}^{2}, \dots, e_{ij}^{t_{kp}}\right) \quad (3.16)$$

Trong đó, $e_{ij}^{l} = w_k * tf(t_{il}, e_{ij}) \times idf(t_{il}, E_i)$ với $t_{il} \in \mathcal{V}_T$, $w_k, k = 1,..5$ là trọng số của các đặc trưng tương ứng của bài viết.

3.4.2. Xác định mối tương quan giữa người dùng và các chủ đề

Mức độ liên quan giữa bài viết e_{ij} của người dùng u_i đối với chủ đề t_k:

$$\alpha_{ij}^{k} = cor\left(e_{ij}, t_k\right) \quad (3.17)$$

Khi đó, mức độ liên quan của bài viết e_{ij} đến q chủ đề trong $\boldsymbol{\mathcal{T}}$ ký hiệu:

$$cor\left(e_{ij}, \boldsymbol{\mathcal{T}}\right) = (\alpha_{ij}^{1}, \alpha_{ij}^{2}, \dots, \alpha_{ij}^{q}) \quad (3.18)$$

3.4.3. Độ tương tự quan tâm của người dùng theo chủ đề

Mức độ quan tâm của người dùng theo các chủ đề:

$$\boldsymbol{u_i^t} = (\, u_i^1,\, u_i^2, \dots,\, u_i^p) \quad (3.19)$$

Độ tương tự của hai người dùng theo các chủ đề:

$$sim_{int}\left(u_i, u_j\right) = sim(\, \boldsymbol{u_i^t},\, \boldsymbol{u_j^t}) = \frac{< u_i^t, u_j^t >}{\| u_i^t \| \times \| u_j^t \|} \quad (3.20)$$

CHƯƠNG 4: HÀNH VI VÀ QUAN TÂM CỦA NGƯỜI DÙNG THEO HÀNH VI TRÊN MẠNG XÃ HỘI

4.1. HÀNH VI CỦA NGƯỜI DÙNG TRÊN MẠNG XÃ HỘI

4.1.1. Hành vi và phân loại các hành vi của người dùng trên mạng xã hội

Theo [65] [91] [147] [154] và [104] thì hành vi của người dùng trên các trang mạng xã hội là các cách thức người dùng hoạt động và tương tác với các sự kiện, hiện tượng trên mạng xã hội. Các hành vi này được phân loại theo hành vi cá nhân (*individual behavior*) và hành vi tập thể (*collective behavior*).

Theo thống kê từ [65] [91] [147] [104], và [132] thì trên một mạng xã hội, các hành vi của một người dùng bất kỳ thường bao gồm: *Đăng bài viết (Post) trên trang cá nhân*; *Thích (Like)*; *Bình luận (Comment)*; *Tham gia hay gia nhập nhóm (Join group)*; *Kết bạn (Add friend)*; *Theo dõi (Follow)*; *Tạo/tham gia các sự kiện (Event)*; *Đánh dấu (Tag)*; *Chia sẻ (Share)* …

4.1.2. Phát hiện quan tâm của người dùng dựa trên hành vi

4.1.3. Nhóm hay cộng đồng người dùng trên mạng xã hội

Định nghĩa 4.1:

Một nhóm hay một cộng đồng $g_i \in G$ trên mạng xã hội N, được đặc trưng bởi ba đặc trưng : $g_i = \{name_i, sty_i, des_i\}$. Trong đó:

- $name_i$ *là tên (name) của nhóm* g_i,
- sty_i *là kiểu (style) của nhóm* g_i
- des_i *là mô tả (description) về nhóm* g_i.

4.2. MÔ HÌNH NGƯỜI DÙNG THEO HÀNH VI

4.2.1. Mô hình biểu diễn người dùng

Định nghĩa 4.2:

Trong mạng xã hội $\mathcal{N} =< U, E, G, B >$, tập các hành vi của người dùng B trên mạng xã hội đang xem xét bao gồm:

- $P = \{post_i\}$ *tập hành vi đăng/chia sẻ (post) bài viết trên mạng xã hội N của người dùng,* p_i *là kí hiệu hành vi đăng bài i trong tập P.*
- $L = \{like_i\}$ *tập hành vi thích (like) bài viết trên mạng xã hội N,* l_i *là kí hiệu hành vi thích bài viết i trong tập L.*
- $C = \{comt_i\}$ *tập các bình luận của người dùng trong bài viết trên mạng xã hội đó,* c_i *là kí hiệu bình luận thứ i trong tập C*
- $J = \{join_i\}$ *tập các hành vi gia nhập nhóm hay cộng đồng người dùng trên mạng xã hội đó,* j_i *là kí hiệu hành vi gia nhập nhóm thứ i trong tập J*

Mỗi người dùng u_i khi biểu diễn theo các hành vi sẽ là một bộ bốn như sau: $u_i =< P_i, L_i, C_i, J_i >$

Định nghĩa 4.3:

P là hành vi đăng bài viết (Post an entry). Theo đó, người dùng $u_i \in U$ đăng bài viết $e_j \in E$ trên mạng xã hội $\mathcal{N}$ được xác định bởi một ánh xạ:

$f_{post}: U \times E \rightarrow \{0,1\}$, *xác định như sau:*

$$\begin{cases} f_{post}(u_i, e_j) = 1 \text{ nếu } u_i \text{ đăng bài viết } e_j \in E \\ f_{post}(u_i, e_j) = 0 \text{ nếu } u_i \text{ không đăng bài viết } e_j \in E \end{cases}$$

Định nghĩa 4.4:

L là hành vi thích bài viết (Like an entry). Theo đó, người dùng $u_i \in U$ thích bài viết $e_j \in E$ trên mạng xã hội $\mathcal{N}$ được xác định bởi một ánh xạ:

$f_{like}: U \times E \rightarrow \{0,1\}$, *xác định như sau:*

$$\begin{cases} f_{like}(u_i, e_j) = 1 \text{ nếu } u_i \text{ thích bài viết } e_j \in E \\ f_{like}(u_i, e_j) = 0 \text{ nếu } u_i \text{ không thích bài viết } e_j \in E \end{cases}$$

Định nghĩa 4.5:

Tập các bài viết của người dùng $u_i \in U$ đã đăng/chia sẻ trên mạng xã hội $\mathcal{N}$ được định nghĩa như sau: $E_i^{post} = \{e_j \in E \mid \forall j, f_{post}(u_i, e_j) = 1\}$

Tập các bài viết $e_j \in E$ mà người dùng $u_i \in U$ đã thích trên mạng xã hội $\mathcal{N}$ được định nghĩa như sau: $E_i^{like} = \{e_j \in E \mid \forall j, f_{like}(u_i, e_j) = 1\}$

Định nghĩa 4.6:

C là hành vi bình luận trong bài viết (Comment in an entry). Theo đó, người dùng $u_i \in U$ *bình luận trong bài viết* $e_j \in E$ *trên mạng xã hội* $\boldsymbol{\mathcal{N}}$ *được xác định bởi một ánh xạ:*

$f_{comt}: U \times E \rightarrow \{0,1\}$*, xác định như sau:*

$$\begin{cases} f_{comt}(u_i, e_j) = 1 \text{ nếu } u_i \text{ bình luận trong bài viết } e_j \in E \\ f_{comt}(u_i, e_j) = 0 \text{ nếu } u_i \text{ không bình luận trong bài viết } e_j \in E \end{cases}$$

Định nghĩa 4.7:

J là hành vi tham gia nhóm/cộng đồng (Join a group/page). Theo đó, người dùng u_i *tham gia vào nhóm* g_j *được xác định bởi một ánh xạ:* $f_{join}: U \times G \rightarrow \{0,1\}$*, xác định như sau:*

$$\begin{cases} f_{join}(u_i, g_j) = 1 \text{ nếu } u_i \text{ có tham gia vào nhóm } g_j \in G \\ f_{join}(u_i, g_j) = 0 \text{ nếu } u_i \text{ không tham gia vào nhóm } g_j \in G \end{cases}$$

Định nghĩa 4.8:

Tập các nhóm/cộng đồng mà người dùng $u_i \in U$ *đã tham gia trên mạng xã hội N được Định nghĩa như sau:* $G_i^{join} = \{g_k \in G \mid \forall k, f_{join}(u_i, g_k) = 1\}$

Theo Định nghĩa 4.2, mỗi người dùng được biểu diễn bởi các hành vi là đăng/chia sẻ bài viết, thích bài viết, bình luận và tham gia vào nhóm hoặc cộng đồng trên mạng xã hội.

- *Hành vi đăng (post)* bài viết $e_i \in E$ của một người dùng $u_i \in U$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$, ký hiệu là: $post_i$,
- *Hành vi chia sẻ một bài viết* cũng được được xếp vào hành vi đăng bài viết bởi vì việc chia sẻ chính là hành vi đăng lại một bài viết, một nội dung nào đó từ chính mạng xã hội.
- *Hành vi thích (like)* bài viết $e_i \in E$ của một người dùng $u_i \in U$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$, ký hiệu là: $like_i$
- *Hành vi bình luận trong bài viết (comment):* Nếu người dùng bình luận trong bài viết đã đăng hoặc chia sẻ của người dùng $e_i \in E$ của một người dùng $u_i \in U$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$, ký hiệu là: $comt_i$,
- *Hành vi tham gia hay gia nhập nhóm (join group)* $g_i \in G$ của một người dùng $u_i \in U$ trên mạng xã hội $\boldsymbol{\mathcal{N}}$, ký hiệu là: $join_i$,

Khi đó mỗi người dùng u_i được biểu diễn dựa trên các hành vi:

$$u_i =< P_i, L_i, C_i, J_i > = \{post_i, like_i, comt_i, join_i\} | u_i \in U \qquad (4.1)$$

4.2.2. Biểu diễn mô hình người dùng bằng véc tơ trọng số

a. Tính giá trị cho các hành vi

- Giá trị của hành vi đăng bài viết

$$u_{ipost} = post_i = \boldsymbol{p_i} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{in}}) \qquad (4.2)$$

- Giá trị của hành vi thích bài viết

$$u_{ilike} = like_i = \boldsymbol{l_i} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{im}}) \quad (4.3)$$

- Giá trị của hành vi bình luận trong bài viết

$$u_{icomt} = comt_i = \boldsymbol{c_i} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{ip}}) \quad (4.4)$$

- Giá trị của hành vi gia nhập một nhóm trên mạng xã hội

$$u_{ijoin} = join_i = \boldsymbol{j_i} = (\boldsymbol{g_1}, \boldsymbol{g_2}, \dots, \boldsymbol{g_p}) \quad (4.5)$$

Mỗi người dùng u_i trên mạng xã hội được biểu diễn bằng một véctơ dựa trên các hành vi có các thành phần như sau:

$$u_i = (\boldsymbol{p_i}, \boldsymbol{l_i}, \boldsymbol{c_{i,}}\, \boldsymbol{j_i}) \quad (4.6)$$

Nói cách khác có thể biểu diễn người dùng dựa trên các hành vi như sau:

$$u_i = (post_i, like_i, comt_i, join_i) = \begin{cases} E_i^{post} = \boldsymbol{p_i} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{in}}), \\ E_i^{like} = \boldsymbol{l_i} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{im}}), \\ E_i^{comt} = \boldsymbol{c_i} = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{ik}}), \\ G_i^{join} = \boldsymbol{j_i} = (\boldsymbol{g_{i1}}, \boldsymbol{g_{i2}}, .. \boldsymbol{g_{ip}}) \end{cases} \quad (4.7)$$

4.2.3. Độ tương tự giữa hai người dùng theo hành vi

Mô hình ước lượng tổng quát

Giả sử có hai người dùng u_i và u_k trên mạng xã hội N, độ đo tương tự của hai người dùng theo hành vi:

$$s_{beha}(u_i, u_k) = w_{post} * s_{post}(u_i, u_k) + w_{like} * s_{like}(u_i, u_k) + w_{comt} * s_{comt}(u_i, u_k) + w_{join} * s_{join}(u_i, u_k) \quad (4.8)$$

Trong đó: $w_{post}, w_{like}, w_{comt}, w_{join}$, lần lượt là trọng số của hành vi đăng/ chia sẻ bài viết, hành vi thích bài viết, hành vi bình luận trong bài viết và hành vi tham gia một nhóm trên mạng xã hội, và chúng thỏa mãn điều kiện: $w_{post} + w_{like} + w_{comt} + w_{join} = 1$. $s_x(u_i, u_k)$ là độ tương tự trên từng hành vi của hai người dùng u_i, u_k.

Độ tương tự trên từng hành vi

Độ tương tự dựa trên hành vi đăng bài viết:

$$s_{post}(u_i, u_k) = sim(E_i^{post}, E_k^{post}) = sim(\boldsymbol{p_i}, \boldsymbol{p_k}) \quad (4.9)$$

Độ tương tự dựa trên hành vi thích bài viết:

$$s_{like}(u_i, u_k) = sim(E_i^{like}, E_k^{like}) = sim(\boldsymbol{l_i}, \boldsymbol{l_k}) \quad (4.10)$$

Độ tương tự dựa trên hành vi bình luận trong bài viết:

Đặt $\begin{cases} S1 = sim(E_{posi}^i, E_{posi}^k) + sim(E_{nega}^i, E_{nega}^k) \\ S2 = sim(E_{posi}^i, E_{nega}^k) + sim(E_{nega}^i, E_{posi}^k) \end{cases}$

Độ tương tự về hành vi bình luận của người dùng u^i và u^k được định nghĩa bằng công thức sau:

$$sim_{comt}(u_i, u_k) = min\,(1, max(0, |S1 - S2|)) \quad (4.11)$$

Độ tương tự dựa trên hành vi gia nhập nhóm:

$$sim_{join}(u_i, u_k) = sim\left(G_i^{join}, G_k^{join}\right) = sim(\boldsymbol{j_i}, \boldsymbol{j_k}) \quad (4.12)$$

4.3. QUAN TÂM CỦA NGƯỜI DÙNG THEO MÔ HÌNH HÀNH VI

4.3.1. Biểu diễn mô hình hành vi người dùng theo không gian chủ đề

Mỗi bài viết được xét trong hành vi đăng, hành vi thích, hành vi bình luận và mỗi nhóm người dùng đã tham gia được biểu diễn theo không gian các chủ đề theo công thức (3.16) như vậy mỗi người dùng sẽ được biểu diễn bằng:

$$u_i^t = \begin{cases} E_i^{post} = \boldsymbol{p_i}^t = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{in}}), \\ E_i^{like} = \boldsymbol{l_i}^t = (\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, .. \boldsymbol{e_{im}}), \\ E_i^{comt} = \boldsymbol{c_i}^t = (\boldsymbol{c_{i1}}, \boldsymbol{c_{i2}}, .. \boldsymbol{c_{ik}}) \\ G_i^{join} = \boldsymbol{j_i}^t = \left(\boldsymbol{g_{i1}}, \boldsymbol{g_{i2}}, .. \boldsymbol{g_{ip}}\right) \end{cases} \quad (4.13)$$

Trong đó, $\boldsymbol{e_{ij}^k} = \left(e_{ij}^1, e_{ij}^2, \dots, e_{ij}^{t_{kp}}\right), e_{ij}^k = tf(t_{il}, e_{ij}) \times idf(t_{il}, E_i)$ *với* $t_{il} \in \boldsymbol{\mathcal{T}}$

4.3.2. Xác định chủ đề quan tâm theo hành vi

Giả sử rằng $\boldsymbol{\mathcal{T}} = \{T_1, T_2, \dots, T_p\}$ là một tập các chủ đề trên mạng xã hội N, khi đó, mức độ liên quan của các hành vi đăng bài viết, thích bài viết và gia nhập của người dùng u_i với các chủ đề trong $\boldsymbol{\mathcal{T}}$ được tính bằng mức độ liên quan của các tập bài viết $E_i^{post}, E_i^{like}, G_i^{join}$ với các chủ đề đang xem xét.

Ký hiệu tương ứng là:

$$\boldsymbol{u_{ipost}^t} = (u_i^1, u_i^2, \dots, u_i^p) \quad (4.12)$$

$$\boldsymbol{u_{ilike}^t} = (u_i^1, u_i^2, \dots, u_i^p) \quad (4.13)$$

$$\boldsymbol{u_{icomt}^t} = (u_i^1, u_i^2, \dots, u_i^p) \quad (4.13)$$

$$\boldsymbol{u_{ijoin}^t} = (u_i^1, u_i^2, \dots, u_i^p) \quad (4.14)$$

Khi đó, mức độ quan tâm của người dùng u_i với các chủ đề trong $\boldsymbol{\mathcal{T}}$ được tính theo công thức:

$$\boldsymbol{u_i^t} = w_p * \boldsymbol{u_{ipost}^t} + w_l * \boldsymbol{u_{ilike}^t} + w_j * \boldsymbol{u_{ijoin}^t} \quad (4.15)$$

Trong đó, w_p, w_l, w_j là trọng số của các hành vi thỏa mãn $w_p + w_l + w_j = 1$ và các $\boldsymbol{u_k^t}$ là các độ đo mức quan tâm của người dùng đến các chủ đề trong tập $\boldsymbol{\mathcal{T}}$.

4.3.3. Độ tương tự quan tâm của người dùng theo chủ đề

Khi đó độ quan tâm tương tự của hai người dùng theo hành vi dựa trên chủ đề được tình bằng

$$sim_{int}(u_i, u_j) = sim(\boldsymbol{u_i^t}, \boldsymbol{u_j^t}) \quad (4.16)$$

Trong đó các $\boldsymbol{u_i^t}, \boldsymbol{u_j^t}$ được tính theo công thức (4.15), và $sim(\boldsymbol{u_i^t}, \boldsymbol{u_j^t})$ được tính như công thức (2.16). Từ đó có thể thấy rằng $sim_{int}(u_i, u_j)$ nằm trong khoảng [0,1].

4.5. SO SÁNH VỚI MỘT SỐ MÔ HÌNH KHÁC

4.5.1. Các mô hình so sánh

Luận án thực hiện việc so sánh kết quả thực hiện mô hình với 03 mô hình tính toán dựa trên TF.IDF và dữ liệu là văn bản ngắn gồm: Mô hình ước lượng độ quan tâm dựa trên thẻ đánh dấu của Sheng Bin et al. [125]; Mô hình ước lượng phát hiện các chủ đề quan tâm của người dùng dựa trên các Tweet của Hossen M. F. et al. [63] và mô hình ước lượng chủ đề quan tâm dựa trên hành vi đăng bài (post) và hành vi thích (like) của Kim J. Ko et al. [77].

KẾT LUẬN

Những kết quả nghiên cứu của luận án

- Đề xuất mô hình biểu diễn bài viết của người dùng trên mạng xã hội dựa trên năm đặc trưng là nội dung, thể loại, thẻ đánh dấu, quan điểm và cảm xúc. Mỗi bài viết được tính toán, mở rộng ngữ nghĩa theo Wikipedia và biểu diễn dưới dạng một véctơ có trọng số theo TF.IDF theo các đặc trưng của chúng.
- Đề xuất mô hình biểu diễn hành vi của người dùng dựa trên các hành vi đăng/chia sẻ bài viết, hành vi thích bài viết, bình luận trong bài viết và hành vi gia nhập nhóm/cộng đồng trên mạng xã hội.
- Đề xuất cách xác định các chủ đề quan tâm của người dùng dựa trên ước lượng độ tương quan giữa các bài viết của người dùng với các chủ đề. Độ tương quan giữa tập hợp các bài viết của người dùng với các chủ đề là mức độ quan tâm của người dùng đến các chủ đề đó trên mạng xã hội.
- Đề xuất cách thức ước lượng độ tương tự hai người dùng theo mô hình bài viết và mô hình hành vi. Độ tương tự giữa hai người dùng theo mô hình bài viết được tính dựa trên việc tích hợp có trọng số độ tương tự các đặc trưng của bài viết và giữa hai tập bài viết của người dùng. Độ tương tự giữa hai người dùng theo hành vi cũng được tính dựa trên tích hợp có trọng số độ tương tự giữa các hành vi của người dùng.

Hướng nghiên phát triển của luận án

Thứ nhất là mở rộng dữ liệu nghiên cứu từ dữ liệu kiểu văn bản sang dữ liệu ảnh, dữ liệu video hoặc các liên kết trong các bài viết của người dùng trên mạng xã hội; Thứ hai là tiếp tục khảo sát, nghiên cứu dữ liệu văn bản nhưng áp dụng các thuật toán có hiệu quả hơn để phân tích hoặc xây dựng các bản thể học (ontology) trong phát hiện quan tâm của người dùng trên mạng xã hội; Thứ ba là vấn đề các quan tâm của người dùng luôn thay đổi theo thời gian, trong luận án chưa đề cập đến yếu tố thời gian khi thu thập dữ liệu mặc dầu các thời điểm thu thập dữ liệu thực đều tiến hành lấy các dữ liệu gần thời điểm đó nhất.

DANH MỤC CÁC CÔNG TRÌNH NGHIÊN CỨU

TẠP CHÍ KHOA HỌC

[1]. Manh Hung Nguyen, Thi Hoi Nguyen. *A general model for similarity measurement between objects*. International Journal of Advanced Computer Science and Applications (IJACSA), 6(2):235 - 239, 2015.

[2]. Thi Hoi Nguyen, Dinh Que Tran, Gia Manh Dam, Manh Hung Nguyen, *Estimating the similarity of social network users based on behaviors*, Vietnam Journal of Computer Science (2018) 5:165–175, Springer Opens

[3]. Nguyễn Thị Hội, Trần Đình Quế, *Ước lượng quan tâm người dùng trên mạng xã hội dựa trên tương tự bài viết,* Tạp chí Khoa học và Công nghệ - Đại học Đà Nẵng (JST-UD), Trường Đại học Đà Nẵng, ISSN 1859-1531 – Số 7(128). 2018

[4]. Nguyen Thi Hoi, Tran Dinh Que, *Estimating user's interest on social networks based on behaviors*, Journal of Science and Technology on Information and Communications, Vol 3, CS.01 (2018), 9-15, ISSN 2525 – 2224

[5]. Dinh Que Tran, Thi Hoi Nguyen, Phuong Thanh Pham, *Modeling user's interests, similarity and trustworthiness based on vectors of entries in social networks*, Southeast Asian Journal of Sciences, Vol. 09, No 1 (2020), pp. 01–10

HỘI THẢO KHOA HỌC

[6]. Thi Hoi Nguyen, Dinh Que Tran, Gia Manh Dam, and Manh Hung Nguyen. *Multi-feature Based Similarity Among Entries on Media Portals*, Advances in Information and Communication Technology, Proceedings of the International Conference, ICTA 12 - 2016, Advances in Intelligent Systems and Computing, ISBN 978-3-319-49072-4, Springer International Publishing. Advances in Intelligent Systems and Computing, 538 AISC, pp. 373-382, (2017).

[7]. Nguyen, Thi Hoi; Tran, Dinh Que; Dam, Gia Manh; Nguyen, Manh Hung. *Integrated Sentiment and Emotion into Estimating the Similarity among Entries on Social Network*, 3rd EAI Sep 4, 2017, Springer International Publishing. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, 221, pp. 242-253, (2018).

[8]. Nguyễn Thị Hội, Đàm Gia Mạnh, Trần Đình Quế. *Độ tương đồng ngữ nghĩa các bài viết trên mạng xã hội dựa trên Wikipedia*, Kỷ yếu Hội thảo Fundamental and Applied IT Research - FAIR'10, Đà Nẵng 08/2017, NXB Khoa học Tự nhiên và Công nghệ.

[9]. Nguyễn Thị Hội, Trần Đình Quế. *Ước lượng tương tự quan tâm người dùng trên mạng xã hội dựa vào các nhóm tham gia*, Kỷ yếu Hội thảo Fundamental and Applied IT Research - FAIR'11, Hà Nội 08/2018, NXB KHTN và CN