MỘT ƯỚC LƯỢNG TƯƠNG QUAN GIỮA HÀNH VI VÀ QUAN TÂM CỦA NGƯỜI DÙNG TRÊN MẠNG XÃ HỘI

Nguyễn Thị Hội[*], Trần Đình Quế[+]
[*] Trường Đại học Thương Mại
[+]Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt:
Phát hiện quan tâm của người dùng trên các mạng xã hội là một trong những chủ đề thu hút nhiều quan tâm nghiên cứu và áp dụng trong nhiều ứng dụng như các hệ tư vấn người dùng, các chiến lược quảng cáo, phân loại người dùng, ... Trong bài báo này, chúng tôi đề xuất một mô hình phân tích một số hành vi của người dùng trên các mạng xã hội để phát hiện và so sánh tương quan về quan tâm của họ, sau đó, đánh giá bằng thực nghiệm với dữ liệu thực. Kết quả thực nghiệm cho thấy nếu hai người dùng có nhiều hành vi giống nhau thì sẽ có quan tâm tương tự nhau.

Từ khóa: Mạng xã hội, hành vi người dùng, quan tâm của người dùng, độ đo tương tự

I. MỞ ĐẦU

Theo từ điển Tiếng Việt [18] thì quan tâm là sự chú ý và để tâm một cách thường xuyên đến chủ đề nào đó, các chủ đề quan tâm của người dùng trên các mạng xã hội thường rất đa dạng và không dễ dàng để xếp vào một lĩnh vực nào đó: Chẳng hạn như một người dùng thường xuyên chia sẻ các bài viết về lịch trình các trận tennis, tham gia vào nhóm cổ động viên của Man U, thường xuyên thích hoặc bình luận hình ảnh các vận động viên …có thể xem người dùng đó quan tâm đến chủ đề thể thao, hoặc một người dùng thường xuyên chú ý đến các bộ phim bom tấn, tham gia nhóm hâm mộ một ca sĩ, diễn viên nào đó, thường xuyên theo dõi lịch chiếu và các sự kiện bên lề của các liên hoan phim, …có thể xem người dùng đó quan tâm đến chủ đề phim ảnh, giải trí …Như vậy, có thể nói rằng, quan tâm của người dùng trên các mạng xã hội là sự để tâm và chú ý thường xuyên đến một hoặc một số chủ đề nào đó trên các mạng xã hội.

Cũng theo từ điển Tiếng Việt [18], hành vi là toàn bộ những phản ứng, cách cư xử ra bên ngoài của một người trong một hoàn cảnh cụ thể. Còn theo từ điển Wikipedia [19] thì hành vi là một chuỗi các hành động lặp đi lặp lại. Hành động là toàn thể những hoạt động (phản ứng, cách ứng xử) của cơ thể, có mục đích cụ thể là nhằm đáp ứng lại kích thích ngoại giới, là hành động hoặc phản ứng của đối tượng (khách thể) hoặc sinh vật, thường sử dụng trong sự tác động đến môi trường, xã hội. Hành vi có thể thuộc về ý thức, tiềm thức, công khai hay bí mật, có thể tự giác hoặc không tự giác. Hành vi là một giá trị có thể thay đổi qua thời gian.

Hiện nay, với sự lớn mạnh và ảnh hưởng sâu rộng của các mạng xã hội, các nghiên cứu về quan tâm của người dùng trên các mạng xã hội không những được rất nhiều cá nhân, tổ chức chú ý mà chúng còn có rất nhiều ứng dụng trong các dịch vụ trực tuyến như các hệ thống khuyến nghị người dùng, các chiến lược quảng cáo sản phẩm, các chương trình giới thiệu dịch vụ cho người dùng…Quan tâm của người dùng trên các mạng xã hội là một hướng được rất nhiều nhà nghiên cứu phân tích và đưa ra nhiều cách thức để thu được các kết quả nghiên cứu khác nhau. Theo khảo sát của chúng tôi, có một số cách phát hiện quan tâm người dùng phổ biến dùng trên các phương tiện truyền thông như: trích xuất thông tin từ thông tin cá nhân người dùng (profile) [2, 8, 17]; trích xuất từ các liên kết của người dùng đến các người dùng khác (follows, link) [2, 7, 12]; trích xuất hành vi đánh dấu, đăng bài (tag, post)…của người dùng [9, 10, 12, 13]…

Tuy nhiên, hiện nay các thông tin cá nhân của người dùng trên các mạng xã hội rất khó thu thập do yêu cầu bảo mật người dùng hoặc người dùng cũng thường xuyên không cung cấp đầy đủ thông tin. Thêm nữa, các thông tin của người dùng thường quá ít, quá rời rạc cũng là một trở ngại trong nghiên cứu về quan tâm của người dùng trên các mạng xã hội, vì vậy, các nghiên cứu về quan tâm của người dùng trên các mạng xã hội những năm gần đây thường đi theo hai hướng tiếp cận chính: một là phân tích về các kết nối, quan hệ bạn bè, những danh sách những người được theo dõi, các đánh dấu… của người dùng trên các mạng xã hội như [2, 7, 8]; hai là phân tích các bài đăng (status) và các thuộc tính liên quan đến các bài đăng của người

Tác giả liên hệ: Nguyễn Thị Hội
Email: hoint2002@gmail.com
Đến toàn soạn: 5/2018 , chỉnh sửa: 7/2018 , chấp nhận đăng: 8/2018

dùng trên các mạng xã hội [7, 9, 11, 12]. Các nghiên cứu này chủ yếu đi sâu vào vấn đề xác định hoặc phát hiện quan tâm của từng cá nhân người dùng, chưa chú ý nghiên cứu nhiều về mối liên quan giữa những người dùng trên các mạng xã hội

Bài báo của chúng tôi dựa trên kết quả nghiên cứu đã có về mô hình hành vi của người dùng để ước lượng quan tâm tương tự của các người dùng trên mạng xã hội

Phần còn lại của bài báo được tổ chức như sau: Phần 2 là giới thiệu về mô hình bài viết và ước lượng độ tương tự giữa các bài viết; Phần 3 giới thiệu về hành vi và ước lượng độ tương tự hành vi; Phần 4 là ước lượng quan tâm tương tự của người dùng và phần 5 là phần thực nghiệm và đánh giá;

II. MÔ HÌNH BÀI VIẾT VÀ ĐỘ TƯƠNG TỰ GIỮA CÁC BÀI VIẾT TRÊN MẠNG XÃ HỘI

A. Mô hình bài viết của người dùng

Trên một mạng xã hội, có một tập những người dùng, mỗi người dùng có thể có một hoặc một số bài viết, 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.

Các bài viết trên mạng xã hội thường chia thành hai nhóm: Nhóm thứ nhất là bài viết của người dùng tự viết sau đó đăng lện tường của mình, có thể đánh dấu vị trí, và đánh dấu những người liên quan, đánh dấu cảm xúc … Nhóm thứ hai là bài viết bao gồm nội dung của người viết tự viết và một nội dung được chia sẻ có thể bài viết của chính họ hoặc của người dùng khác, có thể chia sẻ từ mạng xã hội hiện tại hoặc từ một mạng xã hội khác, hoặc chia sẻ từ một phương tiện truyền thông xã hội khác nữa.

Bài báo chỉ quan tâm đến phần chứa văn bản (text) và đánh dấu (tag) hoặc biểu tượng cảm xúc (emotion icon) của bài viết còn các hình ảnh, các video, các âm thaanh sẽ không được xem xét trong bài báo này. Vì vậy trong bài báo, một bài viết được mô tả bởi các đặc tính của chúng, bao gồm: tiêu đề *(caption*), thể loại (*category*), các đánh dấu *(tags),* nội dung *(content),* cảm xúc (*emotion*), quan điểm (*sentiment*), …

B. Độ tương tự các bài viết trên các mạng xã hội

Khi đăng một bài viết trên mạng xã hội người dùng phần nào đã thể hiện thái độ và sự chú ý của mình về một chủ đề nào đó thông qua bài viết, vì vậy, để ước lượng độ tương tự các bài viết đã đăng của người dùng, bài báo xem xét độ tương tự giữa các thành phần đã đăng của người dùng và xây dựng bộ từ khóa tương ứng. Cách thức xây dựng bộ từ khóa dựa trên nghiên cứu trong [10] của nhóm tác giả, sau đó được tính toán và ước lượng dựa trên TF-IDF của các từ khóa của mỗi bài viết

Khoảng cách cosine được sử dụng để tính độ tương tự giữa hai đối tượng, bài báo cũng sử dụng kỹ thuật N-gram được giới thiệu bởi W.B. Cavnar và J.M. Trenkle [16] để xây dựng các tập từ khóa và kế thừa và mở rộng thuật toán đề xuất bởi S.A.Takale và S.S Nandgaonkar [14] cho từng từ Tiếng Anh để xây dựng và phân tích các N-gram áp dụng cho ngôn ngữ Tiếng Việt. Sau khi phân tích, bài báo sử dụng TF-IDF để xây dựng vectơ chứa giá trị của các thành phần trong bộ hành vi của người dùng. TF-IDF (Term Frequency – Inverse Document Frequency) là trọng số của một từ trong tài liệu của người dùng được tính dựa trên thống kê mức độ quan trọng hay số lần xuất hiện của từ này trong tài liệu đó, cách tính như sau:

Gọi n_v là số lần từ khóa k xuất hiện trong vectơ v của bài viết e, N_v là tổng số từ khóa của bài viết e được biểu diễn bởi vectơ v, N_E là tổng số các bài viết của người dùng u, N_k là tổng số các bài viết của người dùng u có chứa từ khóa k. Khi đó:

Tần suất của từ khóa k xuất hiện trong vec tơ v của bài viết e là TF được tính theo công thức (1) như sau:

$$tf(k,v) = \frac{n_v}{N_v}, \quad (1)$$

Tần suất nghịch đảo của từ khóa k xuất hiện trong vec tơ v của bài viết e là IDF được tính theo công thức (2) như sau:

$$idf(k,N_E) = log(\frac{N_E}{N_k}), \quad (2)$$

Trọng số của từ khóa k xuất hiện trong vec tơ v của bài viết e là TF-IDF được tính theo công thức (3) như sau:

và $$tf - idf(k,v) = tf(k,v) * idf(k,N_e) \quad (3)$$

Như vậy dựa trên các công thức (1), (2) và (3) bài báo tính toán các giá trị cho vec tơ thuộc tính của các bài viết của người dùng trên các mạng xã hội như sau: Giả sử U là một tập người dùng trên một mạng xã hội và mỗi $u_i \in U$ có một tập bài viết đã đăng E_i^{post}, với mỗi bài viết được biểu diễn bởi 5 thành phần, ký hiệu tương ứng như sau: nội dung là *cont*, đánh dấu là *tags*, nhóm bài viết là *cate*, quan điểm là *sent* và cuối cùng cảm xúc ký hiệu là *emot.*

Gọi $e_i^k \in E_i^{post}$, $e_j^l \in E_j^{post}$ là hai bài viết tương ứng của $u_i, u_j \in U$, mỗi tập từ khóa của mỗi bài viết $e_i^k \in E_i^{post}$ được biểu diễn bằng một vectơ v_i^k tương ứng.

Sau khi tính TF-IDF của các từ khóa trong hai vectơ biểu diễn hai bài đăng, bài viết thu được các vectơ chứa trọng số của hai bài viết tương ứng v_k^w, v_l^w. Khi đó, độ tương tự của của hai bài viết e_i^k, e_j^l được tính theo công thức (4) như sau:

$$sim_{entry}(e_i^k, e_j^l) = D_{cosine}(v_k^w, v_l^w) \quad (4)$$

Trong đó, v_k^w, v_l^w là các vectơ chứa trọng số tính theo TF-IDF của hai bài viết e_i^k, e_j^l tương ứng

III. MÔ HÌNH HÀNH VI VÀ ƯỚC LƯỢNG TƯƠNG TỰ HÀNH VI CỦA NGƯỜI DÙNG

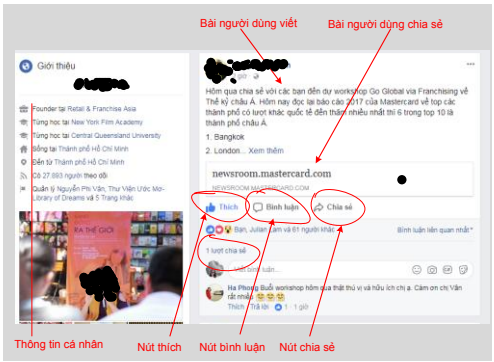
A. Mô hình hành vi người dùng trên mạng xã hội

Trong một mạng xã hội, có nhiều người dùng kết nối với nhau theo các kiểu quan hệ như quan hệ bạn

bè, quan hệ gia đình, quan hệ đồng nghiệp... Trong đó, mỗi người dùng có một không gian cá nhân riêng và người dùng có toàn quyền làm những việc họ muốn, chẳng hạn như đăng các bài viết mô tả trạng thái cá nhân; thích và chia sẻ niềm vui vì hoàn thành công việc hay đạt được một thành tựu nào đó; trích dẫn và chia sẻ lại những bài viết, bài báo, bức ảnh, đoạn phim mà bản thân thấy thú vị...

Những hành động như đăng bài viết, bài chia sẻ, thích hoặc bình luận trong một bài viết hoặc bài chia sẻ, tham gia một nhóm trên mạng xã hội… gọi chung là các *hành vi* của người dùng trên các mạng xã hội. Các hành vi trên mạng xã hội phản ánh một phần cách ứng xử của người dùng đó với các sự kiện hoặc hiện tượng xảy ra trên mạng xã hội

Ví dụ với một người dùng trên mạng xã hội Facebook như Hình 1 thì có các thông tin như giới thiệu về bản thân, đăng nội dung bài viết, chia sẻ nội dung từ phương tiện truyền thông xã hội khác, nhóm tham gia, thích, bình luận, trả lời bình luận, …



Hình 1. Ví dụ về người dùng trên MXH Facebook

Các hành vi trên mạng xã hội có rất nhiều kiểu khác nhau như: đăng bài viết, chia sẻ bài viết, bình luận trong một bài viết, kết bạn, theo dõi một người dùng khác hoặc một trang khác, … nhưng trong bài báo chỉ tập trung nghiên cứu và xem xét các hành vi phổ biến nhất bao gồm: đăng một bài viết *(post)*, chia sẻ bài viết *(share)*, thích bài viết *(like)*, bình luận trong bài viết *(comment)*

Tuy nhiên, hành vi chia sẻ bài viết trên mạng xã hội của người dùng được bao hàm trong hành vi đăng bài nên bài báo xem hành vi chia sẻ đồng nhất với hành vi đăng bài. Trên một số mạng xã hội như Facebook.com, Twitter.com hành vi bình luận về một bình luận trong một bài viết của người dùng được xem như là bình luận trong bài viết để tránh phức tạp và nhập nhằng khi thống kê các bình luận và thống nhất về khái niệm sử dụng.

Như vậy, mỗi người dùng trên mạng xã hội được bài báo tập trung xem xét với các hành vi sau: đăng bài viết, thích bài viết, bình luận trong bài viết

B. *Ước lượng độ tương tự hành vi người dùng*

Giả sử U là một tập người dùng trên một mạng xã hội, khi đó, mỗi $u_i \in U$ có một tập các hành vi B_i, với mỗi $b_i^j \in B_i$ có thể là một trong ba hành vi được ký hiệu tương ứng như: đăng/chia sẻ bài là *post*, thích bài viết là *like*, và bình luận trong bài viết là *comm*

Khi đó, việc ước lượng độ tương tự giữa hai tập hành vi người dùng B_i *của* $u_i \in U$ và B_j của $u_j \in U$ được tính toán bằng cách tích hợp các độ tương tự của ba tập con các thuộc tính *post, like* và *comm*. Để tính độ tương tự giữa các thuộc tính hành vi của bộ hành vi thì bài báo tính toán như sau:

- *Độ tương tự hành vi đăng bài viết (post)*

Gọi $u_i, u_j \in U$ là hai người dùng, mỗi người dùng có tập các bài viết đã được đăng $E_i^{post}, E_j^{post} \in E$ và mỗi người dùng sẽ có một vectơ trọng số biểu diễn các bài viết của họ tương ứng là $\vec{u}_i^w, \vec{u}_j^w$. Với mỗi cặp người dùng $u_i, u_j \in U$ thì mỗi thành phần u_i^k của vecto $\vec{u}_i^w$ được tính như sau:

Mỗi $e_i^k \in E_i^{post}$ của u_i tính độ tương tự của e_i^k với tất cả các bài viết $e_j^l \in E_j^{post}$ của $u_j \in U$. Mỗi thành phần u_i^k được tính theo công thức:

$$u_i^k = max(sim_{entry}(e_i^k, e_j^1), \dots, sim_{entry}(e_i^k, e_j^m)) \quad (5)$$

Trong đó, $e_i^k \in E_i^{post}$ và m là số bài viết của u_j và $sim_{entry}(e_i, e_j)$ là độ tương tự của hai bài viết e_i, e_j tương ứng

Mỗi thành phần u_j^k của vecto $\vec{u}_j^w$ cũng được tính tương tự, khi đó, độ tương tự của hai người dùng $u_i, u_j \in U$ dựa trên hành vi đăng bài viết được tính bằng:

$$sim_{user-entry}(u_i, u_j) = D_{cosine}(\vec{u}_i^w, \vec{u}_j^w) \quad (6)$$

Trong đó $\vec{u}_i^w, \vec{u}_j^w$ là các vec tơ chứa trọng số các bài viết đã được đăng của hai người dùng u_i, u_j tương ứng, có thể thấy rằng $sim_{user-entry}(u_i, u_j)$ nằm trong khoảng [0,1]

- *Độ tương tự hành vi thích/ quan tâm bài viết*

Mỗi người dùng trên mạng xã hội có thể thích hay quan tâm (người dùng thể hiện các thái độ yêu, mỉm cười, lo âu, buồn, giận dữ…trong bài báo đều được coi là có quan tâm đến bài viết) hoặc bỏ qua một bài viết trên mạng xã hội, để ước lượng độ tương tự hành vi thích/quan tâm của hai người dùng về bài viết thì bài báo xem xét và tính toán như sau:

Gọi E_i^{like} và E_j^{like} lần lượt là tập của các bài viết đã được thích/quan tâm của người dùng $u_i, u_j \in U$, khi đó độ tương tự về hành vi thích/quan tâm bài viết của hai người dùng $u_i, u_j \in$ U được tính bằng độ tương tự giữa hai tập bài viết đã được thích/quan tâm của hai người dùng dựa trên công thức (5) và (6) như sau:

$$sim_{user-like}(u_i, u_j) = D_{cosine}(\vec{u}_i^w, \vec{u}_j^w) \quad (7)$$

Trong đó $\vec{u}_i^w, \vec{u}_j^w$ là các vectơ chứa trọng số các bài viết đã được thích của hai người dùng u_i, u_j tương ứng, có thể thấy rằng $sim_{user-like}(u_i, u_j)$ nằm trong khoảng [0,1]

- *Độ tương tự hành vi bình luận trong bài viết*

Mỗi người dùng có thể bình luận hoặc thích một vài bình luận mà các người dùng đã bình luận trong

một bài viết, để ước lượng độ tương tự về hành vi bình luận của hai người dùng, bài báo xem xét các bình luận của hai người dùng và bài viết mà họ đã bình luận trên mạng xã hội. Các bình luận của người dùng đó cùng với bài viết mà người dùng đã bình luận được xây dựng bộ từ khóa dựa trên nghiên cứu của [16, 14] và tính toán TF-IDF tập từ khóa như cách ước lượng trọng số của bài viết đã đăng của người dùng theo các công thức (1), (2) và (3). Khi đó, độ tương tự hành vi bình luận của hai người dùng được bài báo tính toán như sau:

Gọi $u_i, u_j \in U$ là hai người dùng, mỗi người dùng có tập các bình luận cùng các bài viết mà họ đã bình luận trong đó, $E_i^{comm}, E_j^{comm} \in E$ và mỗi người dùng sẽ có một vectơ trọng số biểu diễn các bình luận và bài viết mà họ đã bình luận tương ứng là $\vec{u}_i^w, \vec{u}_j^w$.

Mỗi thành phần của $\vec{u}_i^w, \vec{u}_j^w$ được tính như công thức (5), trong đó các bài viết được kết hợp thêm các bình luận của người dùng trong thuộc tính nội dung *cont* để tính toán và ước lượng

Khi đó, độ tương tự về hành vi bình luận của hai người dùng $u_i, u_j \in U$ được tính dựa trên công thức (5) và (6) như sau:

$$sim_{user-comm}(u_i, u_j) = D_{cosine}(\vec{u}_i^w, \vec{u}_j^w) \quad (8)$$

Trong đó $\vec{u}_i^w, \vec{u}_j^w$ là các vec tơ chứa trọng số các bình luận và bài viết đã được bình luận của hai người dùng u_i, u_j tương ứng, có thể thấy rằng $sim_{user-comm}(u_i, u_j)$ nằm trong khoảng [0,1]

C. Độ tương tự của người dùng theo hành vi

Sau khi ước lượng độ tương tự trên từng tập hành vi của hai người dùng thì độ tương tự của hai người dùng dựa trên các hành vi được tính như sau:

Gọi $u_i, u_j \in U$ là hai người dùng, mỗi người dùng có tập các bộ hành vi $B_i, B_j \in B$ và mỗi người dùng sẽ có một vectơ trọng số biểu diễn các hành vi của họ tương ứng là $\vec{u}_i^w, \vec{u}_j^w$ được tính bằng:

$$\vec{u}_i^w = \begin{cases} sim_{user-entry}(u_i, u_j), \\ sim_{user-like}(u_i, u_j), \\ sim_{user-comm}(u_i, u_j), \end{cases} \quad (9)$$

và

$$\vec{u}_j^w = \begin{cases} sim_{user-entry}(u_j, u_i), \\ sim_{user-like}(u_j, u_i), \\ sim_{user-comm}(u_j, u_i), \end{cases} \quad (10)$$

Khi đó, độ tương tự của hai người dùng $u_i, u_j \in U$ dựa trên các hành vi được tính bằng:

$$sim_{user-behavior}(u_i, u_j) = D_{cosine}(\vec{u}_i^w, \vec{u}_j^w) \quad (11)$$

Trong đó $\vec{u}_i^w, \vec{u}_j^w$ là các vec tơ chứa trọng số các bộ hành vi đã thực hiện trên mạng xã hội của hai người dùng u_i, u_j tương ứng, có thể thấy rằng $sim_{user-behavior}(u_i, u_j)$ nằm trong khoảng [0,1]

IV. ƯỚC LƯỢNG QUAN TÂM TƯƠNG TỰ NGƯỜI DÙNG

A. Xác định các chủ đề trên mạng xã hội

Phát hiện các chủ đề và các quan tâm đến các chủ đề của người dùng đã được rất nhiều nghiên cứu đưa ra như các nghiên cứu của Bhattacharya et al [2], Diana et al [7], Li Xin et al [9], Sheng Bin et al [13]. Bài báo dựa trên các kết quả nghiên cứu trước đó của chính nhóm tác giả [11] để áp dụng cho bài toán phân loại các bài viết của người dùng theo các chủ đề, nhóm nghiên cứu sau khi phân tích đã thu được một danh sách gồm 21 chủ đề chính và 81 chủ đề con được sử dụng phổ biến trên mạng xã hội. Bài báo kế thừa kết quả nghiên cứu đó để áp dụng cho ước lượng và phân loại các bài viết của người dùng trên mạng xã hội vào các chủ đề. Ví dụ một số chủ đề được minh họa trong Bảng I. như sau:

Bảng I. Ví dụ về chủ đề cùng từ khóa của chủ đề

Chủ đề	Danh sách từ khóa
Giáo dục	Giáo dục, tiếng Anh, học tập, kiến thức, thói quen, thế hệ, giảng dạy, đào tạo, nghiên cứu, trải nghiệm, giáo dục, tiểu học, trung học, từ nguyên, từ đồng, tiếng Việt, toàn cầu, Quốc tế, Kinh tế, Xã hội, Văn hóa, Quốc công, cha mẹ, trực tuyến, Liên Hiệp Quốc, học trực tuyến, giáo dục tiểu học, …
Môi trường	Môi trường, tổ hợp, tự nhiên, xã hội, hệ thống, tập hợp, tương tác, định nghĩa, con người, không khí, độ ẩm, sinh vật, loài người, môi trường, vật chất, đối tượng, tập hợp con, …

Mỗi chủ đề sau khi xác định danh sách từ khóa được biểu diễn bằng một vectơ trọng số t_k^w được tính toán theo công thức (3), trong đó, chỉ số *k* là chủ đề thứ *k* trong danh sách các chủ đề và *w* là ký hiệu vec tơ chứa trọng số các từ khóa của chủ đề thứ *k*.

B. Xác định quan tâm theo các chủ đề

Với mỗi người dùng $u_i \in U$, bài báo xác định mức độ quan tâm của các hành vi $(post, like, comm) \in B_i$ theo chủ đề $t_j \in T$ như sau:

Gọi $E_i^{post}, E_i^{like}, E_i^{comm}$ lần lượt là tập các bài viết đã đăng, đã thích, đã bình luận, bài báo ước lượng độ tương tự của mỗi bài viết $e_i^k \in E_i^{post}$ đã đăng của người dùng $u_i \in U$ với mỗi chủ đề $t_j \in T$ được tính bằng công thức:

$$sim_{entry-topic}(e_i^k, t_j) = D_{cosine}(v_i^w, t_j^w) \quad (11)$$

Trong đó, v_i^w là vectơ trọng số của bài viết $e_i^k \in E_i^{post}$ của $u_i \in U$ và t_j^w là vectơ trọng số của chủ đề $t_j \in T$. Nghĩa là độ quan tâm của bài viết theo chủ đề dựa trên độ tương tự của các từ khóa của bài viết và từ khóa của chủ đề đang xem xét. Khi đó:

Độ quan tâm dựa trên hành vi đăng bài viết của người dùng $u_i \in U$ theo chủ đề $t_j \in T$ được tính bằng:

$$interest_{post-topic}(u_i, \mathrm{t_j}) = \max\begin{Bmatrix} 0, \\ (sim_{entry-topic}(e_i^1, t_j), \\ \dots \\ (sim_{entry-topic}(e_i^n, t_j) \end{Bmatrix} \quad (12)$$

Trong đó, *n* là số bài viết đã đăng của người dùng $u_i \in U$ và $t_j \in T$ là chủ đề thứ j trong danh sách các chủ đề đang xem xét

Độ quan tâm dựa trên hành vi thích/quan tâm bài viết của người dùng $u_i \in U$ theo chủ đề $t_j \in T$ được tính bằng:

$$interest_{like-topic}(u_i, \mathrm{t_j}) = \max\begin{Bmatrix} 0, \\ (sim_{entry-topic}(e_i^1, t_j), \\ \dots \\ (sim_{entry-topic}(e_i^m, t_j) \end{Bmatrix} \quad (13)$$

Trong đó, *m* là số bài viết đã thích/quan tâm của người dùng $u_i \in U$ và $t_j \in T$ là chủ đề thứ j trong danh sách các chủ đề đang xem xét

Độ quan tâm dựa trên hành vi bình luận bài viết của người dùng $u_i \in U$ theo chủ đề $t_j \in T$ được tính bằng:

$$interest_{comm-topic}(u_i, t_j) = \max\begin{Bmatrix} 0, \\ (sim_{entry-topic}(e_i^1, t_j), \\ \dots \\ (sim_{entry-topic}(e_i^p, t_j) \end{Bmatrix} \quad (14)$$

Trong đó, *p* là số bài viết đã bình luận của người dùng $u_i \in U$ và $t_j \in T$ là chủ đề thứ *j* trong danh sách các chủ đề đang xem xét

Như vậy, mức độ quan tâm của người dùng $u_i \in U$ với chủ đề $\mathrm{t_j} \in T$ được tính dựa trên các công thức (12), (13), (14)

$$interest_{behavior-topic}(u_i, \mathrm{t_j}) = \max\begin{Bmatrix} 0, \\ interest_{post-topic}(u_i, \mathrm{t_j}), \\ interest_{like-topic}(u_i, \mathrm{t_j}), \\ interest_{comm-topic}(u_i, \mathrm{t_j}), \end{Bmatrix} \quad (15)$$

C. Độ quan tâm tương tự của người dùng theo chủ đề dựa trên hành vi

Với mỗi $u_i, u_j \in U$ trên mạng xã hội cùng tập các hành vi $B_i, B_j \in B$, độ quan tâm của người dùng $u_i \in U$ với chủ đề $t_k \in$ T được biểu diễn bằng vectơ q_i^k (gọi là vectơ độ quan tâm của người dùng u_i đến chủ đề t_j trên mạng xã hội) như sau:

$$\mathrm{interest}_{\mathrm{user-topic}}(u_i, t_k) = \overrightarrow{q_i^k} \text{ trong đó}$$

$$\overrightarrow{q_i^k} = \begin{Bmatrix} interest_{poste-topic}(u_i, t_k), \\ interest_{like-topic}(u_i, t_k), \\ interest_{comm-topic}(u_i, t_k) \end{Bmatrix} \quad (16)$$

Và độ quan tâm của người dùng $u_j \in U$ với chủ đề $t_k \in T$ được biểu diễn bằng vectơ q_j^k như sau:

$$\mathrm{interest}_{\mathrm{user-topic}}(u_j, t_k) = \overrightarrow{q_j^k}$$

$$\overrightarrow{q_j^k} = \begin{Bmatrix} interest_{post-topic}(u_j, t_k), \\ interest_{like-topic}(u_j, t_k), \\ interest_{comm-topic}(u_j, t_k), \end{Bmatrix} \quad (17)$$

Trong đó, các thành phần của hai vec tơ $\overrightarrow{q_i^k}$ và $\overrightarrow{q_j^k}$ được tính theo các công thức (12), (13), (14) và (15)

Khi đó, độ tương tự quan tâm của hai người dùng $u_i, u_j \in U$ với chủ đề $t_j \in T$ dựa trên hành vi được tính bằng:

$$sim_{user-topic}(u_i, u_j, t_k) = D_{cosine}(\overrightarrow{q_i^k}, \overrightarrow{q_j^k}) \quad (18)$$

Có thể thấy rằng $sim_{user-topic}(u_i, u_j, t_k)$ nằm trong khoảng [0,1].

Sau khi đề xuất hướng tiếp cận ước lượng độ tương tự giữa hai người dùng dựa trên các hành vi và độ quan tâm tương tự của người dùng theo chủ đề, câu hỏi đặt ra là: *Nếu hai người dùng tương tự nhau dựa trên các hành vi thì họ có quan tâm đến một số chủ đề tương tự nhau hay không? và ngược lại?* Để trả lời cho câu hỏi này, phần tiếp theo bài báo trình bày thực nghiệm dựa trên dữ liệu thực để kiểm nghiệm và đưa ra câu trả lời cho câu hỏi này!

V. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Như bài báo đã trình bày ở mục IV, mục đích của thực nghiệm là để tìm câu trả lời cho câu hỏi: "*Nếu hai người dùng tương tự nhau dựa trên các hành vi thì họ có quan tâm đến một số chủ đề tương tự nhau hay không? và ngược lại?".*

A. Thu thập dữ liệu và xây dựng tập mẫu

Chúng tôi thực hiện việc thu thập dữ liệu từ trang mạng Facebook.com. Mỗi người dùng được chọn 10 hành vi gần thời điểm lấy dữ liệu nhất bao gồm: 10 bài đăng (*post)*, 10 bài viết đã thích (*like*), 10 bài viết đã bình luận (*comment).* Trong mô hình đề xuất, bài báo chỉ xem xét các bài viết, các bình luận, các bài viết được bình luận, các bài viết đã thích, các bài viết đã chia sẻ chứa văn bản, còn các đối tượng không chứa văn bản bị loại bỏ khỏi tập dữ liệu. Việc loại bỏ các đặc tính không phải văn bản được thực hiện tự động thông qua chương trình tiền xử lý dữ liệu

Sau khi đã xử lý, bài báo thu được 40 người dùng thực, bài viết thực hiện mã hóa tên người dùng thành danh sách từ U001 đến U040 thành các cặp so sánh

Sau khi phân tích và loại bỏ các bớt các cặp trùng lặp, ví dụ C1,2 và C2,1, bài viết loại bỏ C2,1 chỉ để C1,2. Các cặp C1,1 vẫn giữ nguyên. Bài viết thu được tổng cộng số lượng mẫu thử nghiệm ở Bảng II

Bảng II. Bộ dữ liệu mẫu thu được

	Số lượng
Người dùng	40
Số cặp so sánh	840
Số bài đăng	400
Số bài đã thích	400
Số bài đã bình luận	400

B. Thực nghiệm

Kịch bản thực nghiệm được thực hiện theo năm bước như sau:

1: Tách từ khóa và tính TF_IDF của bài viết

2: Ước lượng độ tương tự các bài viết

3: Ước lượng độ tương tự của người dùng

4: Tách từ khóa và TF –IDF của các chủ đề

5: Ước lượng độ tương tự quan tâm

Thực hiện lần lượt với tất cả các bài viết của trong bộ dữ liệu thử nghiệm, các cặp người dùng và các chủ đề đã xác định, chi tiết các bước thử nghiệm như sau:

Bước 1: Phân tích các bài viết thành các từ khóa, loại bỏ từ dừng, lấy định nghĩa các từ khóa theo từ điển, phân tích lại và tính TF-IDF của mỗi từ khóa được minh họa trong Bảng III theo công thức (2)

Bảng III. Phân tích một bài viết và tính TF-IDF

<table>
<tr><td>Một bài viết của U011</td><td>Chắc chết! Thành phố xanh - Blue city của Morocco. Quẹt vào lịch cái thành phố này rồi nhé. Nhớ những người bạn nói tiếng Ả rập lai Pháp. Nhớ bị chặn lại tại sân bay hỏi cô đi với ai. Nhớ bữa ăn tối ...</td></tr>
<tr><td>Từ khóa và TF-IDF tương ứng</td><td>
<table>
<tr><th>Keyword</th><th>Tf-idf</th><th>Keyword</th><th>Tf-idf</th></tr>
<tr><td>thành phố</td><td>0.561</td><td>bữa ăn</td><td>0.223</td></tr>
<tr><td>- blue</td><td>0.281</td><td>kéo dài</td><td>0.281</td></tr>
<tr><td>thành phố</td><td>0.561</td><td>bản địa</td><td>0.281</td></tr>
<tr><td>nói tiếng</td><td>0.281</td><td>nói chuyện</td><td>0.189</td></tr>
<tr><td>tại sân</td><td>0.281</td><td>thơ ca</td><td>0.281</td></tr>
<tr><td>sân bay</td><td>0.281</td><td>triết học</td><td>0.223</td></tr>
</table>
</td></tr>
</table>

Bước 2: Tính độ tương tự giữa các bài viết dựa trên TF-IDF. Ở bước thứ hai các cặp bài viết của các người dùng được ước lượng độ tương tự bằng cosine của hai vec tơ chứa TF-IDF tương ứng của chúng theo công thức (3)

Bảng IV. Độ tương tự hai bài viết theo TF-IDF

Bài viết 1	Chẳng biết đường nào mà lần, nước nào mà lo... ? Thưa bà Phan Hà Thủy, Tổng giám đốc Vinschool Trong buổi họp với Ban phụ huynh sáng và chiều qua tại Vinschool, bà đã có những phát ngôn, và cách ...
Bài viết 2	DON'T TAKE IT PERSONAL! Rất nhiền bạn inbox kể cho tôi nghe bản thân cảm thấy bị tổn thương thế nào vì lời nói của người khác. Các bạn trách sao người ta không nhạy cảm, thờ ơ, thiếu trí tuệ cảm
Sim (e1, e2)	***0.02792***

Bước 3: Ước lượng độ tương tự của người dùng dựa trên các hành vi theo các công thức (5), (6), (7) và (8)

Bước 4: Phân tích định nghĩa của các chủ đề thành các từ khóa theo N-gram, loại bỏ từ dừng và tính TF – IDF của chúng theo công thức (3)

Bước 5: Ước lượng độ tương tự của các bài viết của mỗi người dùng theo các chủ đề theo công thức, để xác định độ quan tâm của họ với mỗi chủ đề kết quả được minh họa trong Bảng V.

Bảng V. Độ quan tâm của người dùng theo chủ đề

	Môi trường	Sức khỏe	Công nghệ	Du lịch	Giáo dục	Hôn nhân
U001	0.0159	0.0133	0.0400	0.0293	0.0135	0.0482
U003	0.0357	0.0259	0.0242	0.0319	0.0338	0.0244
U006	0.0357	0.0167	0.0264	0.0095	0.0281	0.0
U007	0.0349	0.0218	0.0298	0.0247	0.0269	0.0229
U008	0.0366	0.0318	0.0210	0.0170	0.0268	0.1213
U010	0.0429	0.0262	0.0239	0.0282	0.0	0.0274

Độ quan tâm của người dùng đối với các chủ đề phổ biến trên các mạng xã hội được tính theo công thức (15). Nhìn vào Bảng V có thể thấy rằng các ô có giá trị 0.0 là không có bài viết nào tương tự với các chủ đề được xây dựng, hay nói cách khác là người dùng không quan tâm đến chủ đề đó trong thời điểm hiện tại.

Dựa trên Bảng V và công thức (16) để ước lượng độ tương tự quan tâm của người dùng theo các chủ đề dựa trên các hành vi. Để xác định hai người dùng có độ quan tâm tương tự nhau, bài báo lựa chọn ngưỡng $sim_{user-topic}(u_i, u_j, t_k)$ >=0.55. Những cặp nào không thỏa mãn được ngưỡng này được coi là quan tâm ít tương tự nhau theo các chủ đề trên mạng xã hội

Bảng VI. Độ quan tâm tương tự dựa trên hành vi

	U001	U002	U003	…	U039	U040
U001	1.0					
U002	0.633	1.0				
U003	0.510					
…				…		…
U039	0.543	0.116	0.844		1.0	
U040	0.135	0.722	0.507	…	0.644	1.0

C. Đánh giá

Để đánh giá độ tương quan của công thức (11) và công thức (18), bài báo sử dụng giá trị trung bình độ lệch tuyệt đối và giá trị trung bình độ lệch tương đối để đánh giá như sau:

Đánh giá theo trung bình độ lệch tuyệt đối:

TB độ lệch tuyệt đối = TB của các

$$|sim_{user-behavior}(u_i, u_j) - sim_{user-topic}(u_i, u_j, t_k)| \quad (19)$$

Với kết quả từ thực nghiệm trong từ bộ mẫu dữ liệu thì mô hình đề xuất có trung bình độ lệch tuyệt đối là 11.8%, khi đó, độ chính xác của mô hình đề xuất là:

*CR= (1- TB độ lệch tuyệt đối)*100%* *(20)*

Và độ chính xác bằng 88.2%

Đánh giá theo trung bình độ lệch tương đối:

TB độ lệch tương đối = TB của các

$$\frac{|sim_{user-behavior}(u_i,u_j) - sim_{user-topic}(u_i,u_j,t_k)|}{MAX(sim_{user-behavior}(u_i,u_j), sim_{user-topic}(u_i,u_j,t_k))} \quad (21)$$

Với kết quả từ thực nghiệm trong bộ mẫu dữ liệu thì mô hình đề xuất có trung bình độ lệch tương đối sẽ là 14.8%, khi đó, độ chính xác của mô hình đề xuất là:

$$CR = (1 - \text{TB độ lệch tương đối}) * 100\% \quad (22)$$

Và độ chính xác bằng 85.2%

Bảng VII. Đánh giá mô hình và sự tương quan

	TB độ lệch tuyệt đối	TB độ lệch tương đối	Độ chính xác theo độ lệch tuyệt đối	Độ chính xác theo độ lệch tương đối
Facebook	0.118	0.148	88.2%	85.2%

VI. KẾT LUẬN

Bài báo đã đề xuất mô hình ước lượng độ tương tự quan tâm của người dùng dựa trên các hành vi đăng bài viết, thích bài viết và bình luận trong bài viết. Mô hình đề xuất có thể áp dụng trong việc phân loại người dùng trên các mạng xã hội hoặc xác định quan tâm của người dùng theo các chủ đề ứng dụng trong các chương trình quảng cáo, các hệ thống khuyến nghị người dùng…

TÀI LIỆU THAM KHẢO

[1]. Attacharya Parantapa, Zafar Muhammad Bilal, Ganguly Niloy, Ghosh Saptarshi, Gummadi Krishna P. *Inferring User Interests in the Twitter Social Network* Proceedings of the 8th ACM Conference on Recommender Systems, RecSys '14 pgs 357-360, ACM, New York, NY, USA

[2]. Bruno Ohana and Brendan Tierney. *Sentiment classification of reviews using sentiwordnet*. 2009.

[3]. Chihli Hung and Hao-Kai Lin. *Using objective words in sentiwordnet to improve word-of-mouth sentiment classification*. IEEE Intelligent Systems, 28(2):47–54, 2013.

[4]. D. Manning, Prabhakar Raghavan, Hinrich Schutze, 2008, *Introduction to Information Retrieval*, 27 Oct 2013

[5]. Dekang Lin. *An information-theoretic definition of similarity*. In Proc. 15th International Conf. on Machine Learning, pages 296–304. Morgan Kaufmann, San Francisco, CA, 1998

[6]. Diana Palsetia, Md. Mostofa, Ali Patwary, Kunpeng Zhang , Kathy Lee, Christopher Moran, Yves Xie, Daniel Honbo, Ankit Agrawal, Wei-keng Liao, Alok Choudhary. *User-Interest based Community Extraction in Social Networks* ACM, NY, USA, 2012

[7]. Elie Raad, Richard Chbeir, and Albert Dipanda. *User profile matching in social networks*. In Proceedings of the 2010 13th International Conference on NetworkBased Information Systems, NBIS '10, pages 297–304, Washington, DC, USA, 2010. IEEE Computer Society.

[8]. Li Xin, Guo Lei, Zhao Yihong Eric *Tag-based Social Interest Discovery* Proceedings of the 17th International Conference on World Wide Web Beijing, China,pages 675-684, ACM, New York, NY, USA

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

[10]. Nguyễn Thị Hội, Đàm Gia Mạnh, Trần Đình Quế, *Độ tương đồng ngữ nghĩa các bài viết trên mạng xã hội dựa trên Wikipedia.* Hội nghị Khoa học Quốc gia: Nghiên cứu cơ bản và ứng dụng CNTT lần 10 - FAIR'10. Thg8/2017

[11]. Pavan Kapanipathi, Prateek Jain , Chitra Venkataramani, Amit Sheth, *User Interests Identification on Twitter Using a Hierarchical Knowledge Base* 11th ESWC 2014 (ESWC2014) , 2014 , May

[12]. Sheng Bin, Gengxin Sun, Peijian Zhang and Yixin Zhou *Tag-Based Interest-Matching Users Discovery Approach in Online Social Network* International Journal of Hybrid Information Technology Vol. 9, No. 5 (2016), pp. 61-70

[13]. Sheetal A Takale, Sushma S Nandgaonkar, *Measuring semantic similarity between words using web documents* International Journal of Advanced Computer Science and Applications (IJACSA) Volume 1, Issue 4. 2010

[14]. Nguyen T.H., Tran D.Q., Dam G.M., Nguyen M.H. (2018) *Integrated Sentiment and Emotion into Estimating the Similarity Among Entries on Social Network*. In: Chen Y., Duong T. (eds) Industrial Networks and Intelligent Systems. INISCOM 2017. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 221. Springer, Cham

[15]. W.B. Cavnar and J.M. Trenkle. *N-gram-based text categorization.* Ann Arbor MI, 48113(2):161–175, 1994.

[16]. Zhao Zhe, Cheng Zhiyuan, Hong Lichan, Hsin Chi Ed Huai *Improving User Topic Interest Profiles by Behavior Factorization* 2015, Pages 1406-1416, ACM, New York, NY, USA

[17]. Perelman L.C., Paradis J., Barrett E. Mayfield Handbook of Technical and Scientific Writing, Mayfield, Mountain View, California (1997).

[18]. Hoàng Phê (2018), Từ điển Tiếng việt, Viện ngôn ngữ học, NXB Hồng Đức

[19]. Từ điển Wikipedia, https://www.wikipedia.org/

ESTIMATING USER'S INTEREST ON SOCIAL NETWORKS BASED ON BEHAVIORS

Abstract: Discovering interests of users on social networks is one of the issues attracting many researches and being applied to various fields, such as user recommendations, personalized ads, or categorizing users into groups. In this paper, we propose an approach based on the analysis of user's behaviors on social networks to detect and compare the correlations of interest of two users on the network. Our proposal is also empirically evaluated with the real data. The evaluation shows that the more same behaviors two users have, the more similar interests they have. And vice versa, if two users have similar interests, their entries are the same.

Nguyễn Thị Hội, Nhận học vị Thạc sỹ năm 2006. Hiện công tác tại Đại học Thương mại. Lĩnh vực nghiên cứu: Hệ thống thông tin, khai phá dữ liệu, tính toán xã hội. Đang là NCS tại Học viện Công nghệ Bưu chính Viễn thông

Trần Đình Quế, Nhận học vị Tiến sỹ năm 2000. Hiện công tác tại Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Tính toán thông minh và phân tán, Tính toán xã hội và Khai phá dữ liệu.