ƯỚC LƯỢNG QUAN TÂM NGƯỜI DÙNG TRÊN MẠNG XÃ HỘI DỰA TRÊN TƯƠNG TỰ BÀI VIẾT

ESTIMATING USER'S INTEREST ON SOCIAL NETWORKS BASED ON ENTRIES SIMILARITY

Nguyễn Thị Hội[1], Trần Đình Quế[2]

[1]Trường Đại học Thương mại; hoint@tmu.edu.vn
[2]Học viện Công nghệ Bưu chính Viễn thông; tdque@yahoo.com

Tóm tắt - Phát hiện quan tâm của người dùng trên các mạng xã hội là một trong những chủ đề thu hút nhiều nghiên cứu và được áp dụng trong nhiều ứng dụng như khuyến nghị người dùng, các chương trình quảng cáo cá nhân hóa hoặc phân loại người dùng. Trong bài báo này, nhóm tác giả đề xuất một mô hình dựa trên việc phân tích các bài viết của người dùng trên các mạng xã hội để phát hiện và so sánh tương quan về quan tâm của họ. Mô hình đề xuất được đánh giá bằng thực nghiệm với dữ liệu thực. Kết quả thực nghiệm cho thấy rằng nếu hai người dùng có nhiều bài viết giống nhau thì sẽ có quan tâm tương tự nhau và ngược lại, nếu hai người dùng có quan tâm giống nhau thì cũng có nhiều bài viết tương tự nhau.

Từ khóa - quan tâm của người dùng; mạng xã hội; bài viết trên mạng xã hội; độ đo tương tự; người dùng trên mạng xã hội

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 posts 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 similar entries two users have, the more similar interests they have and vice versa. If two users have similar interests, their entries are the same.

Key words - user's interest; social network; entry; similarity measure; users on social networks

1. Đặt vấn đề

Theo từ điển Tiếng Việt thì quan tâm là sự chú ý và để tâm một cách thường xuyên đến chủ đề, sự vật, hiện tượng đang xảy ra trong những hoàn cảnh cụ thể. Trên các mạng xã hội (social network sites) các chủ đề quan tâm của người dùng thường rất đa dạng và không dễ dàng để xếp vào một lĩnh vực cụ thể. Người dùng trên mạng xã hội là những người tham gia vào một trang mạng xã hội bất kỳ, có tài khoản trên trang mạng xã hội đó và sử dụng mạng để trao đổi, tương tác với người dùng khác. 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 phương pháp giáo dục trẻ em, về nội dung các cuốn sách giáo khoa phổ thông, … thì có thể xem người dùng đó quan tâm đến chủ đề giáo dục; hoặc một người dùng thường xuyên chú ý đến các sự kiện thể thao đang diễn ra như các trận bóng đá, các giải thi đấu, …thì có thể xem người dùng đó quan tâm đến chủ đề thể thao … 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.

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 (recommendation system), các chiến lược quảng cáo sản phẩm (product advertising strategy), 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 nhóm tác giả, có một số cách phát hiện quan tâm người dùng phổ biến 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 (link, follow) [2, 7, 12]; trích xuất hành vi 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 cá nhân người dùng thường quá ít cũng là một trở ngại trong phân tích và 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 trong những năm gần đây thường đi theo hai hướng tiếp cận chính: một là phân tích về các kết nối, quan hệ bạn bè, danh sách những người được theo dõi, các đánh dấu, … của người dùng trên các mạng xã hội [2, 7, 8]; hai là phân tích các bài đăng (status) và các thuộc tính liên quan đến các bài đăng của người 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 mà chưa chú ý nghiên cứu nhiều về mối liên quan giữa những người dùng trên các mạng xã hội.

Bài báo của nhóm tác giả đi theo hướng thứ hai, phân tích các bài viết của người dùng trên các mạng xã hội để trả lời cho câu hỏi: Nếu hai người dùng có cùng chủ đề quan tâm trên các mạng xã hội, liệu rằng các bài đăng của họ có nhiều điểm tương tự với nhau hay không? Và ngược lại, nếu hai người dùng có các bài đăng tương tự nhau trên các mạng xã hội, liệu rằng họ có quan tâm đến các chủ đề tương tự nhau hay không?

Trong bài báo này, kỹ thuật N-gram và TF-IDF được sử dụng để phân tích và ước lượng mối tương quan giữa các bài viết và các chủ đề quan tâm của người dùng. Sau đó, mô hình đề xuất được đánh giá và so sánh bằng thực nghiệm.

Phần còn lại của bài báo được tổ chức như sau: Phần 2 là đề xuất cách thức ước lượng mối tương quan giữa quan tâm và bài viết của người dùng; Phần 3 là phần thực nghiệm và đánh giá; Phần 4 là kết luận.

2. Độ tương tự giữa các bài viết và ước lượng quan tâm của người dùng

2.1. Độ tương tự giữa các bài viết trên mạng xã hội

2.1.1. Mô hình và độ tương tự bài viết trên mạng xã hội

Mỗi người dùng trên các mạng xã hội có thể không có, hoặc có ít nhất một hoặc nhiều bài đăng trên tường của họ. Mỗi bài đăng có thể là một câu hoặc một văn bản, một hoặc một số hình ảnh, một video hoặc là một sự kết hợp của các nội dung trên.

Mỗi bài đăng của người dùng trên một mạng xã hội được gọi là một *bài viết* (entry) và được biểu diễn bởi năm thành phần hay đặc trưng, bao gồm: nội dung (content); đánh dấu (tags); thể loại (category); quan điểm (sentiment) và cảm xúc (emotion).

Ví dụ với một bài viết của người dùng có thể được biểu diễn minh họa trong Bảng 1. Giá trị các thành phần được xác định theo phương pháp như trong một nghiên cứu của nhóm tác giả [15] sẽ có các giá trị như sau: *content* bao gồm nội dung phần của bài viết; *tags* là phần được lấy sau dấu # hoặc tên người dùng được đưa vào trong bài viết, như trong ví dụ này là: *#TrangTraiTrungThuc, Mít Tơ Bót*; *category* được xác định dựa trên đề xuất trong nghiên cứu [15] thì có giá trị là "*nông nghiệp, sản phẩm nông nghiệp …*"; *sentiment* và *emotion* sẽ có giá trị là "*tích cực*" và "*biết ơn*".

Bảng 1. *Ví dụ về bài viết và các thành phần phân tích*

Bài viết	Từ khóa tương ứng	
16 tháng qua với #TrangTraiTrungThuc, tôi đã thất bại 5 vụ dưa lưới. Mít Tơ Bót đã chạy vạy khắp nơi để học để tìm ra con đường trồng dưa lưới sạch nhưng chưa một lần thành công! Nhưng ước nguyện của tôi cũng đang dần trở thành sự thực, người làm việc đó là EcoFarm - Bình Phước…	Cont	(tôi đã, thất bại, chạy vạy, khắp nơi, tìm ra, con đường, trở thành, thành sự, sự thực, …)
	Tags	(người làm, làm việc, trang trại, quy trình, …)
	Cate	(Nông nghiệp, sản phẩm, …)
	Sent	(tích cực)
	Emot	(biết ơn)

2.1.2. Ước lượng độ tương tự giữa hai bài viết

Giả sử U là một tập người dùng trên một mạng xã hội. Khi đó, mỗi $u_i \in U$ có một tập bài viết E_i, với mỗi $e_i^j \in E_i$ được biểu diễn bởi 5 thành phần được 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*.

Khi đó, việc ước lượng độ tương tự giữa hai bài viết $e_i^k \in E_i$ *của* $u_i \in U$ *và* $e_j^l \in E_j$ của $u_j \in U$ được tính toán bằng cách tích hợp có trọng số các độ tương tự của 5 thuộc tính của hai bài viết. Trong bài báo này, khoảng cách cosine được sử dụng để tính độ tương tự giữa hai bài viết. Đồng thời, bài báo 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 thuật toán được đề xuất bởi S. A. Takale và S. S Nandgaonkar [14], trong nghiên cứu này S.A.Takale và S.S Nandgaonkar tách các word đơn và tìm từ khóa theo NetWord trên văn bản Tiếng Anh, bài báo này áp dụng và mở rộng trên các N-gram và tìm định nghĩa theo Từ điển Wikipedia, sử dụng cho ngôn ngữ Tiếng Việt trên các bài viết của người dùng trên mạng xã hội. Sau đó, sử dụng TF-IDF để xây dựng véc-tơ chứa giá trị của các thành phần trong bài viết của người dùng. TF-IDF (Term Frequency – Inverse Document Frequency) là trọng số của một từ trong bài viết của người dùng được tính dựa trên thống kê mức độ quan trọng hay số lần xuất hiện của từ này trong một bài viết. Mỗi bài viết $e_i^k \in E_i$ được xét nằm trong một tập hợp các bài viết của người dùng $u_i \in U$. Cách tính TF-IDF trong bài báo được thực hiện dựa trên công trình nghiên cứu [5] như sau:

Mỗi bài viết $e_i^k \in E_i$ được biểu diễn bằng một véc-tơ v_i^k tương ứng. Gọi n_v là số lần từ khóa k xuất hiện trong véc-tơ v của bài viết e, N_v là tổng số từ khóa của véc-tơ 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:

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

và

$$tf - idf(k,v) = tf(k,v) * idf(k,N_e) \quad (2)$$

Sau khi tính TF-IDF của các từ khóa trong hai véc-tơ biểu diễn, ta có véc-tơ các giá trị trọng số của hai bài viết tương ứng $\overrightarrow{v_i^w}, \overrightarrow{v_j^w}$. Độ tương tự của hai bài viết khi đó được tính như sau:

$$sim_{entry}(e_i, e_j) = D_{cosine}(\overrightarrow{v_i^w}, \overrightarrow{v_j^w}) \quad (3)$$

Trong đó, $\overrightarrow{v_i^w}, \overrightarrow{v_j^w}$ là các véc-tơ chứa TF-IDF của hai bài viết e_i, e_j tương ứng.

$D_{cosine}(\overrightarrow{v_i^w}, \overrightarrow{v_j^w}.)$ được tính cho văn bản như sau: Giả sử có véc-tơ biểu diễn cho hai văn bản i và j lần lượt có dạng: $D_i = < w_1^i, w_2^i, \dots w_t^i >$ với w_t^i là trọng số của từ thứ t trong văn bản i. $D_j = < w_1^j, w_2^j, \dots w_t^j >$ với w_t^j là trọng số của từ thứ t trong văn bản j. Độ đo tương tự được tính là Cosine của góc giữa hai véc-tơ biểu diễn cho hai văn bản D_i và D_j. Độ tương tự của chúng được tính theo công thức sau:

$$sim(D_{ij}) = \frac{\sum_{k=1}^{t} w_k^i w_k^j}{\sqrt{\sum_{k=1}^{t}(w_k^i)^2 \sum_{k=1}^{t}(w_k^j)^2}}$$

Dễ dàng thấy rằng giá trị của $sim_{entry}(e_i, e_j)$, nằm trong khoảng [0, 1].

2.1.3. Độ tương tự của người dùng theo bài viết

Ước lượng độ tương tự của hai người dùng dựa trên các bài viết được tính như sau:

Gọi $u_i, u_j \in U$ là hai người dùng, mỗi người dùng có tập các bài viết $E_i, E_j \in E$ và mỗi người dùng sẽ có một véc-tơ trọng số biểu diễn các bài viết của họ tương ứng là $\overrightarrow{u_i^w}, \overrightarrow{u_j^w}$.

Với mỗi cặp người dùng $u_i, u_j \in U$ thì mỗi thành phần u_i^k của véc-tơ $\overrightarrow{u_i^w}$ được tính như sau: Với mỗi $e_i^k \in E_i$ của u_i tính độ tương tự của e_i^k với tất cả các bài viết $e_j^l \in E_j$ của $u_j \in U$. Mỗi thành phần u_i^k được tính theo công thức:

$$u_i^k = \frac{\sum_1^m sim_{entry}(e_k, e_j)}{m} \quad (4)$$

Mỗi thành phần u_j^k của véc-tơ $\overrightarrow{u_j^w}$ cũng được tính tương tự.

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

$$sim_{user-entry}(u_i, u_j) = D_{cosine}(\overrightarrow{u_i^w}, \overrightarrow{u_j^w}.) \quad (5)$$

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

2.2. Ước lượng quan tâm của người dùng theo chủ đề

2.2.1. 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 và cộng sự [2], Diana và cộng sự [7], Li Xin và cộng sự [9], Sheng Bin và cộng sự [13]. Bài báo dựa trên các kết quả nghiên cứu có được từ tiếng Anh, sau đó tiến hành xây dựng và cải tiến danh sách chủ đề phổ biến bằng tiếng Việt trong một nghiên cứu trước đó của nhóm tác giả [11]. Sử dụng kết quả từ nghiên cứu [11], nhóm tác giả có được một danh sách gồm 21 chủ đề chính và 81 chủ đề con được sử dụng phổ biến trên mạng xã hội. Ví dụ một số chủ đề được minh họa trong Bảng 2.

Bảng 2. *Ví dụ về chủ đề và danh sách từ khóa tương ứng*

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 véc-tơ trọng số $\overrightarrow{t_k^w}$ được tính toán theo công thức (2). Trong đó, chỉ số k là chủ đề thứ k trong danh sách các chủ đề và w là ký hiệu véc-tơ chứa trọng số các từ khóa của chủ đề thứ k.

2.2.2. Xác định quan tâm bài viết theo các chủ đề

Với mỗi bài viết $e_i \in E$ của $u_i \in U$ theo chủ đề $t_j \in T$ thì mức độ quan tâm được tính bằng công thức sau đây:

$$sim_{entry-topic}(e_i, t_j) = D_{cosine}(\overrightarrow{v_i^w}, \overrightarrow{t_j^w}) \quad (6)$$

Trong đó, $\overrightarrow{v_i^w}$ là véc-tơ trọng số của bài viết $e_i \in E$ của $u_i \in U$ và $\overrightarrow{t_j^w}$ là véc-tơ 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. Dễ dàng thấy rằng $sim_{entry-topic}(e_i, t_j)$ nằm trong khoảng [0, 1].

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

Bây giờ ta có thể định nghĩa mức độ quan tâm của người dùng theo chủ đề như sau. Với mỗi $u_i \in U$ trên mạng xã hội cùng tập các bài viết $E_i \in E$, độ quan tâm của người dùng $u_i \in U$ với chủ đề $t_j \in T$ được biểu diễn bằng véc-tơ $\overrightarrow{q_i^j}$ (gọi là véc-tơ độ quan tâm của người dùng u_i đến chủ đề t_j trên mạng xã hội) như sau:

$$interest_{user-topic}(u_i, t_j) = \overrightarrow{q_i^j} = (q_{i1}^j, q_{i2}^j, .., q_{in}^j)$$

Trong đó, q_{ik}^j với k = 1 ... n là độ quan tâm của mỗi bài viết $e_i^k \in E_i$ của người dùng u_i với chủ đề t_j tính theo công thức (6).

Gọi $\overrightarrow{q_i^k}$ là véc-tơ quan tâm của người dùng $u_i \in U$ trên mạng xã hội đến chủ đề $t_j \in T$ và $\overrightarrow{q_j^k}$ là véc-tơ quan tâm của người dùng $u_j \in U$ trên mạng xã hội đến chủ đề $t_j \in T$. Khi đó, độ tương tự quan tâm của hai người dùng $u_i, u_j \in U$ với chủ đề $t_j \in T$ được tính bằng:

$$sim_{user-topic}(u_i, u_j, t_k) = D_{cosine}(\overrightarrow{q_i^k}, \overrightarrow{q_j^k}) \quad (7)$$

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 bài viết và độ quan tâm tương tự của người dùng theo chủ đề, bài báo đề xuất giả thuyết rằng: *Nếu hai người dùng tương tự nhau dựa trên các bài viết thì họ sẽ quan tâm đến một số chủ đề tương tự nhau và ngược lại.* Phần 3 bài báo trình bày thực nghiệm dựa trên dữ liệu thực để kiểm nghiệm và đánh giá lại giả thuyết này.

3. Thực nghiệm và đánh giá

Như bài báo đã trình bày cuối mục 2.2.3, mục đích của thực nghiệm là kiểm nghiệm giả thuyết đã nêu đánh giá dựa trên dữ liệu thực.

3.1. Thu thập dữ liệu và xây dựng tập mẫu

Nhóm tác giả thực hiện việc thu thập dữ liệu từ trang mạng xã hội Facebook.com và Twitter.com với 150 người dùng cho mỗi trang. Mỗi người dùng được chọn 10 bài viết gần với thời điểm lấy dữ liệu nhất. Trong mô hình đề xuất, bài báo chỉ xem xét các bài viết chứa văn bản tiếng Việt, còn các bài viết không chứa văn bản, hoặc chứa các ngôn ngữ khác bị loại bỏ khỏi tập dữ liệu. Sau khi đã xử lý, nhóm tác giả thu được 150 người dùng và thực hiện việc xây dựng bộ mẫu dữ liệu thực nghiệm như sau:

Mỗi mẫu là một cặp người dùng với tập 10 bài viết tiếng Việt tương ứng được sinh tự động bằng cách ghép cặp các người dùng, sau đó, tự động loại bỏ các cặp trùng nhau, ví dụ (A, B) và (B, A) sẽ bị loại bỏ đi một, các cặp dạng (A, A) cũng bị loại bỏ khỏi bộ mẫu. Cuối cùng, nhóm tác giả thu được bộ mẫu dữ liệu trong Bảng 3.

Bảng 3. *Bộ mẫu dữ liệu thực nghiệm*

	Facebook.com	Twitter.com
Số lượng người dùng	150	150
Số lượng bài viết	1.500	1.500
Số cặp người dùng	11.100	11.100

3.2. Các bước thực nghiệm

Để tiến hành đánh giá mối tương quan dựa trên thực nghiệm, mỗi mẫu trong bộ dữ liệu lần lượt được thực hiện như sau:

Bước 1: Mỗi bài viết $e_i \in E_i$ của mỗi người dùng $u_i \in U$ được phân tích và ước lượng véc-tơ trọng số theo công thức (2) và lưu lại kết quả.

Bước 2: Ước lượng độ tương tự của hai người dùng dựa trên các bài viết theo công thức (5) và lưu lại kết quả. Minh họa kết quả trình bày trong Bảng 4.

Bước 3: Xây dựng véc-tơ trọng số cho mỗi chủ đề.

Bước 4: Xác định độ quan tâm của người dùng với các chủ đề theo công thức (6). Minh họa kết quả ở Bảng 5.

Bước 5: Ước lượng độ tương tự quan tâm của người dùng theo chủ đề theo công thức (7). Minh họa kết quả trong Bảng 6.

Bước 6: Ước lượng độ tương quan giữa kết quả của Bảng 4 và Bảng 6.

Bước 7: Đánh giá và thảo luận các kết quả.

Bảng 4. *Độ tương tự của người dùng theo bài viết*

	U001	U003	U006	U007	U008	U010
U001	1,0					
U003	0,712	1,0				
U006	0,623	0,804	1,0			
U007	0,644	0,912	0,733	1,0		
U008	0,810	0,941	0,687	0,711	1,0	
U010	0,743	0,894	0,791	0,765	0,824	1,0

Độ tương tự của hai người dùng được tính theo công thức (5) và minh họa trong Bảng 4. Trong bài báo này, hai người dùng được coi là tương tự nhau dựa trên bài viết nếu $sim_{user-entry}(u_i, u_j) \geq 0{,}55$, ngược lại được coi là có nhiều bài viết khác nhau. Từ Bảng 4, có thể thấy rằng nếu hai người dùng càng có nhiều bài viết tương tự nhau thì độ tương tự sẽ gần đến giá trị 1. Ngược lại, nếu có nhiều bài viết không tương tự nhau thì độ tương tự của hai người dùng càng xa giá trị 1.

Bảng 5. *Độ quan tâm của người dùng với các chủ đề*

	Môi trường	**Chính trị**	**Sức khỏe**	**Công nghệ**	**Du lịch**	**Giáo dục**	**Hôn nhân**
U001	0,0159	0,0	0,0133	0,0400	0,0293	0,0135	0,0482
U003	0,0357	0,0242	0,0259	0,0242	0,0319	0,0338	0,0244
U006	0,0357	0,0265	0,0167	0,0264	0,0095	0,0281	0,0
U007	0,0349	0,0326	0,0218	0,0298	0,0247	0,0269	0,0229
U008	0,0366	0,0400	0,0318	0,0210	0,0170	0,0268	0,1213
U010	0,0429	0,0499	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 (6). Nhìn vào Bảng 5 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 vào Bảng 5 và công thức (7) để ước lượng độ tương tự quan tâm của người dùng theo các chủ đề. 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 6. *Độ tương tự quan tâm của người dùng theo chủ đề*

	U001	U003	U006	U007	U008	U010
U001	1,0					
U003	0,633	1,0				
U006	0,590	0,720	1,0			
U007	0,573	0,803	0,733	1,0		
U008	0,643	0,816	0,644	0,679	1,0	
U010	0,674	0,872	0,667	0,654	0,742	1,0

3.3. Đánh giá

Để đánh giá độ tương quan của công thức (5) và công thức (7), 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:

Trung bình độ lệch tuyệt đối được tính bằng giá trị tuyệt đối của trung bình chung hiệu giữa độ đo tương tự của các cặp người dùng theo bài viết và độ đo tương tự của mỗi cặp người dùng theo chủ đề và được tính như sau:

$$\text{TBTĐ}=|sim_{user-entry}(u_i, u_j) - sim_{user-topic}(u_i, u_j, t_k)| \quad (8)$$

Với kết quả từ thực nghiệm trong bộ mẫu dữ liệu thì mô hình đề xuất có trung bình độ lệch tuyệt đối là 0,077. Khi đó, độ chính xác của mô hình đề xuất là:

Độ chính xác = (1 - trung bình độ lệch tuyệt đối)*100% (9)

Và độ chính xác bằng 92,3%.

Đánh giá theo trung bình độ lệch tương đối:

Trung bình độ lệch tương đối được tính bằng thương của trung bình chung của giá trị tuyệt đối của độ tương tự của hai người dùng theo bài viết và độ tương tự của hai người dùng theo chủ đề chia cho giá trị lớn nhất của độ đo tương tự theo bài viết và độ đo tương tự theo chủ đề và được tính theo công thức:

$$\text{TBTgĐ}= \frac{|sim_{user-entry}(u_i,u_j) - sim_{user-topic}(u_i,u_j,t_k)|}{\text{MAX}(sim_{user-entry}(u_i,u_j), sim_{user-topic}(u_i,u_j,t_k))} \quad (10)$$

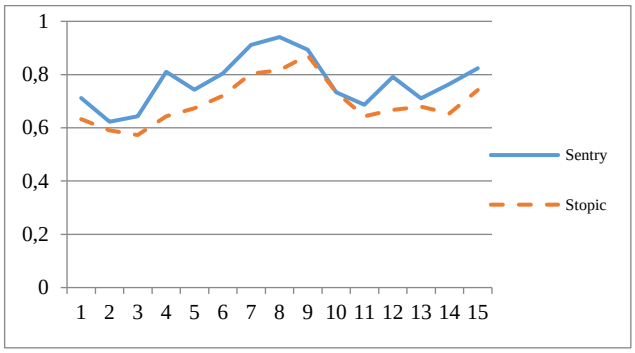
Với kết quả từ thực nghiệm trong bộ mẫu dữ liệu thì mô hình đề xuất có trung bình độ lệch tương đối sẽ là 0,084. Khi đó, độ chính xác của mô hình đề xuất là:

Độ chính xác = (1 - trung bình độ lệch tương đối)*100% (11)

Và độ chính xác bằng 91,2%.

Bảng 7. *Đánh giá mô hình và sự tương quan*

	Trung bình độ lệch tuyệt đối	**Trung bình độ lệch tương đối**	**Độ chính xác theo độ lệch tuyệt đối**	**Độ chính xác theo độ lệch tương đối**
Facebook	0,76	0,84	92,4%	91,6%
Twitter	0,87	0,91	91,3%	90,9%



Hình 1. *Độ tương tự người dùng dựa trên bài viết và các chủ đề*

Biểu diễn ví dụ minh họa với một số cặp người dùng đầu tiên thể hiện trong Hình 1. Hình 1 cho thấy rõ có sự tương quan giữa các bài viết của người dùng và các chủ đề người dùng quan tâm trên các mạng xã hội.

4. Kết luận

Bài báo này đã đề xuất mô hình ước lượng độ tương tự quan tâm của người dùng dựa trên các bài viết và mối tương quan giữa các bài viết và chủ đề quan tâm của người dùng trên các mạng xã hội. 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] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, và Jure Leskovec, *Effects of user similarity in social media*, Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM'12, New York, NY, USA, 2012, ACM, pp. 703-712.

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

[3] Bruno Ohana and Brendan Tierney, *Sentiment Classification of Reviews Using Sentiwordnet*, 2009.

[4] Chihli Hung and Hao-Kai Lin, "Using Objective Words in Sentiwordnet to Improve Word-of-Mouth Sentiment Classification", *IEEE Intelligent Systems*, 28(2), 2013, pp. 47-54.

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

[6] Dekang Lin, *An Information-Theoretic Definition of Similarity*, in Proc. 15th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA, 1998, pp. 296-304.

[7] 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.

[8] Elie Raad, Richard Chbeir, and Albert Dipanda, *User Profile Matching in Social Networks*, in Proceedings of the 2010 13th International Conference on Network Based Information Systems, NBIS'10, Washington, DC, USA, 2010, IEEE Computer Society, pp. 297-304.

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

[10] 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), 2015, pp. 235-239.

[11] 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, 8/2017.

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

[13] 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.

[14] Sheetal A Takale, Sushma S Nandgaonkar, "Measuring Semantic Similarity Between Words Using Web Documents", *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 1, Issue 4, 2010, pp. 78-85.

[15] Nguyen T. H., Tran D. Q., Dam G. M., Nguyen M. H., *Integrated Sentiment and Emotion into Estimating the Similarity Among Entries on Social Network*, International Conference on Industrial Networks and Intelligent Systems, INISCOM 2017: Industrial Networks and Intelligent Systems, Vol. 221, 2018, pp. 242-253.

[16] W. B. Cavnar and J. M. Trenkle, *N-gram-Based Text Categorization*, Environmental Research Institute of Michigan, Ann Arbor MI, 48113(2), 1994, pp. 161-175.

[17] Zhao Zhe, Cheng Zhiyuan, Hong Lichan, Hsin Chi Ed Huai, *Improving User Topic Interest Profiles by Behavior Factorization*, Department of EECS, University of Michigan, ACM, New York, NY, USA, 2015, pp. 1406-1416.

(BBT nhận bài: 01/4/2018, hoàn tất thủ tục phản biện: 03/6/2018)