# KHUYẾN NGHỊ DỰA TRÊN HÀNH VI NGƯỜI DÙNG MẠNG XÃ HỘI

**Nguyễn Mạnh Sơn, Nguyễn Duy Phương**
Khoa Công nghệ thông tin 1
Học viện Công nghệ Bưu chính Viễn thông

***Tóm tắt:*** Sự ra đời của mạng xã hội được xem là sự kiện có ảnh hưởng sâu rộng nhất đối với cộng đồng người dùng Internet hiện nay. Nhiều mạng xã hội trực tuyến như Facebook, Twitter, LinkedIn đã trở nên phổ biến làm thay đổi phương thức trao đổi thông tin truyền thống. Truyền thông giữa các thực thể trong mạng xã hội tạo nên một nguồn tài nguyên phong phú, đa dạng là cơ hội tốt trong phân tích, khai phá và phát triển ứng dụng. Trong bài báo này, chúng tôi đề xuất một phương pháp khuyến nghị bài viếtcho người dùng thông qua các hành vi rate, post, like, comment trong mạng xã hội. Dựa trên dữ liệu các hành vi, chúng tôi đề xuất phương pháp khuyến nghị cho người dùng các các bài viết, hoặc các sản phẩm, dịch vụ họ có thể sử dụng. Kết quả thử nghiệm trên bộ dữ liệu thu thập được trên Facebook cho thấy, phương pháp đề xuất cho lại sai số dự đoán khá tốt so với các phương pháp tiếp cận dựa vào mô hình tin cậy.

***Từ khóa:*** Tư vấn cộng tác (Collaborative Filtering Recommendation), tư vấn theo nội dung (Content-based FilteringRecommendation), hệ tư vấn lai (Hybrid Filtering Recommendation System), hệ tư vấn xã hội (Social Recommender Systems).

## I. GIỚI THIỆU BÀI TOÁN

Hệ thống khuyến nghị (Recommender Systems) là thành phần quan trọng trong các giao dịch trực tuyến hiện nay. Theo kết quả công bố trên 80% phim đã xem trên Netflix [6] và 60% số lần nhấp vào video trên YouTube đến từ hệ thống khuyến nghị [7]. Các hệ thống khuyến nghị được xây dựng từ tập N người dùng U = {u1, u2, …, un} và tập M sản phẩm P = {p1, p2,…, pm}. Trong đó, tập người dùng U được thu thập ngay từ khi người dùng đăng ký tham gia hệ thống, tập sản phẩm P có thể là hàng hóa, phim ảnh, hay dịch vụ được sở hữu bởi người xây dựng các cổng giao dịch thương mại điện tử trực tuyến. Mỗi người dùng i□U đưa ra đánh giá của mình cho một số sản phẩm x□P bằng một số rix. Nhiệm vụ các hệ thống khuyến nghị là điền vào các giá trị đánh giá của người dùng i□U cho các sản phẩm x□P có giá trị phù hợp nhất đối với người dùng này [1, 2]. Dựa vào ma trận đánh giá R = {rix: i=1, 2,.., n; x = 1, 2, .., m}, các phương pháp khuyến nghị truyền thống khai thác những khía cạnh liên quan đến nội dung hoặc thói quen sử dụng của cộng đồng đồng người dùng có cùng chung sở thích để khuyến nghị cho người dùng những sản phẩm mà họ ưa thích. Triết lý chủ đạo của các hệ khuyến nghị truyền thống là những người dùng có sở thích tương tự nhau trong quá khứ thì họ có thể có chung sở thích trong tương lai. Quan điểm của mỗi người dùng trong hệ khuyến nghị là độc lập với người dùng còn lại. Ma trận đánh giá R là đầu vào duy nhất của các phương pháp khuyến nghị truyền thống [1, 3, 5].

Sự ra đời của các mạng xã hội đã làm thay đổi phương thức trao đổi thông tin toàn cầu. Người dùng sử dụng các dịch vụ tư vấn trực tuyến không còn độc lập với những người dùng khác. Mỗi người dùng bị ảnh hưởng hoặc ảnh hưởng đến tập người dùng còn lại thông qua tập hành vi của họ trong mạng xã hội. Hành vi friend cho phép người dùng kết bạn với những người dùng có cùng chung sở thích. Hành vi post cho phép người dùng bày tỏ quan điểm tích cực hoặc tiêu cực của mình đối với các sản phẩm hoặc dịch vụ họ biết hoặc đã từng sử dụng. Hành vi like cho phép người dùng bày tỏ cảm súc của của mình đối với bài post về một sản phẩm cụ thể. Hành vi comment cho phép người dùng bày tỏ quan điểm riêng của mình đối với sản phẩm. Tất cả những hành vi này sẽ có tác động không nhỏ đến thói quen và sở thích của người dùng trong hệ tư vấn. Điều này đã phá vỡ đi những nguyên tắc cơ bản của các phương pháp khuyến nghị truyền thống [4].

Có nhiều đề xuất khác nhau đã được đưa ra để nâng cao chất lượng dự đoán cho các hệ khuyến nghị trong mạng xã hội. Hầu hết các phương pháp đề xuất được thực thi bằng mô hình tin cậy [4, 10]. Phương pháp TidalTrust sử thuật toán tìm kiếm theo chiều rộng tính toán độ tin tưởng giữa các cặp người dùng có đường đi ngắn nhất với trọng số là độ tin cậy [10]. Phương pháp MoleTrust đề xuất giải pháp tương tự như TidalTrust sử dụng thuật toán tìm kiếm theo chiều sâu để xác định đường đi ngắn nhất giữa các cặp người dùng có độ dài không nhỏ hơn d. Giá trị d được xác định thông qua kiểm nghiệm và phụ thuộc vào từng bộ dữ liệu [9]. Phương pháp TrustWalker đề xuất việc sử dụng thuật toán ramdom walk để kết hợp mô hình tin cậy, mô hình dự đoán dựa vào người dùng, mô hình dự đoán dựa vào sản phẩm [8]. Tuy nhiên, các phương pháp kể trên chỉ sử dụng dữ liệu về hành vi friend nên chất lượng khuyến nghị có kết quả chưa cao so với các phương pháp truyền thống [4].

Trong bài báo này chúng tôi đề xuất một phương pháp tiếp cận mới xây dựng mô hình dữ liệu và mô hình dự đoán cho các hệ khuyến nghị trên mạng xã hội. Mô hình dữ liệu được xây dựng bằng cách kế thừa các kết quả nghiên cứu trong xử lý ngôn ngữ tự nhiên để ước lượng quan điểm của người dùng đối với sản phẩm thông qua các hành vi của họ trong mạng xã hội. Bằng cách này ta có thể khai thác được nhiều nguồn dữ liệu vào quá trình huấn luyện và dự đoán quan điểm của người dùng đối với các sản phẩm hoặc dịch

Tác giả liên hệ: Nguyễn Mạnh Sơn,
Email: sonnm@ptit.edu.vn
Đến tòa soạn: 8/2023, chỉnh sửa: 9/2023, chấp nhận đăng: 10/2023.

vụ. Dựa trên nguồn dữ liệu của các hành vi, chúng tôi xây dựng mô hình dự đoán bằng cách kết hợp tất cả các hành vi của người dùng để nâng cao chất lượng khuyến nghị. Để trọng tâm vào những đóng góp mới của bài báo, Mục tiếp theo chúng tôi trình bày phương pháp xây dựng mô hình dữ liệu cho hệ khuyến nghị xã hội. Mục 3 trình bày mô hình dự đoán cho hệ khuyến nghị trong mạng xã hội. Mục 4 trình bày phương pháp xây dựng bộ dữ liệu thử nghiệm và đánh giá. Mục cuối dùng là kết luận và hướng phát triển tiếp theo của bài báo.

## II. MÔ HÌNH DỮ LIỆU CHO HỆ TƯ VẤN KHUYẾN NGHỊ BÀI VIẾT CHO NGƯỜI DÙNG TRONG MẠNG XÃ HỘI

Như đã được trình bày ở trên, các phương pháp tư vấn trong mạng xã hội hướng về mô hình tin cậy [4, 9]. Tập dữ liệu được bổ sung thêm vào mô hình dự đoán là quan hệ kết bạn trong mạng xã hội [15, 16]. Đây là nguyên nhân chính làm cho các phương pháp khuyến nghị trong mạng xã hội có kết quả dự đoán không cao [4]. Trong mục này, chúng tôi trình bày phương pháp mở rộng mô hình dữ liệu cho các hành vi post, like, comment của người dùng.

### *2.1. Mở rộng mô hình dữ liệu biểu diễn hành vi post*

Giả sử ta có mạng xã hội gồm n người dùng U = {u1, u2, .., un}. Mỗi người dùng i∈U đưa ra đánh giá của mình cho một số sản phẩm x∈P bằng một số rix∈Ω. Trong đó, P = {p1, p2, .., pm}, Ω là tập các số thực biểu diễn các mức độ ưa thích khác nhau của người dùng đối với sản phẩm. Ví dụ Ω = {1.0, 0.8, 0.6, 0.4, 0.2} tương ứng với các mức độ {perfect, very good, good, bad, very bad }. Giá trị rix có thể được thu thập trực tiếp hoặc gián tiếp thông qua cơ chế phản hồi của người dùng, rix = 0 được hiểu là người dùng i chưa đánh giá hoặc chưa hề biết đến sản phẩm x. Ma trận R = {rix: i=1, 2, .., n; x = 1, 2, .., m} là đầu vào của các phương pháp khuyến nghị truyền thống được biểu diễn theo công thức (1) [1, 2].

$$r_{ix} = \begin{cases} v: \text{ người dùng } i \text{ đánh giá } x \text{ ở mức độ } v \in \Omega \\ 0 \qquad\qquad otherwise \end{cases} \quad (1)$$

Hành vi post cho phép người dùng đưa thông tin về sản phẩm lên mạng xã hội để cộng đồng đánh giá. Thông tin đưa lên mạng xã hội có thể là bài viết, hình ảnh, video, hoặc tổ hợp các hình thức khác nhau. Nội dung thông tin hành vi post của người dùng có thể được thu thập tự động thông qua API của các mạng xã hội [14]. Hành vi post của người dùng luôn chứa đựng nội dung thông tin phản ánh sản phẩm. Nội dung thông tin có thể biểu diễn ở ba trạng thái khác nhau của người dùng đối với sản phẩm: positive (tích cực), negative (tiêu cực), neutral (trung tính) [11, 12]. Thông tin bài post được đánh giá là positive nếu nội dung bài post chứa đựng thông tin phản ánh tốt về sản phẩm. Thông tin bài post được đánh giá là negative nếu nội dung bài post chứa đựng thông tin phản ánh không tốt về sản phẩm. Thông tin bài post được đánh giá là neutral nếu nội dung bài post không xác định được quan điểm của người dùng là positve hay negative. Vấn đề đặt ra là làm thế nào để xác định tự động quan điểm cá nhân của người dùng đối với sản phẩm thông qua hành vi post.

Để giải quyết vấn đề nêu trên chúng tôi đề xuất việc sử dụng các kết quả nghiên cứu của xử lý ngôn ngữ tự nhiên trong phân tích quan điểm của người dùng (opinion mining, sentiment analysis) [11, 12]. Gọi ContentPost(x) là nội dung bài post của người dùng i∈U đối với sản phẩm x∈P. Nội dung bài post ContentPost(x) có thể dễ dàng trích rút tự động thông qua API của các mạng xã hội [14]. Gọi Sentiment(ContentPost(x)) là hàm ước lượng quan điểm của người dùng i∈U đối với sản phẩm x∈P thông qua bài post ContentPost(x). Hiện tại có nhiều API dùng để ước lượng quan điểm của người dùng dựa vào văn bản. Trong nghiên cứu này, chúng tôi sử dụng API ước lượng quan điểm người dùng cho một bài Post được nhóm nghiên cứu của trường đại học Stanfort đề xuất [13, 14]. Ứng với mỗi bài post của người dùng i đối với sản phẩm x, Sentiment(ContentPost(x)) cho lại một số thực trong khoảng [0, 1] được xác định theo công thức (3). Nếu Sentiment(ContentPost(x)) vượt quá một ngưỡng θ đủ lớn thì ta nói bài post của người dùng i∈U có quan điểm tích cực đối với sản phẩm x∈P. Trong trường hợp khác ta nói bài post của người dùng i∈U có quan điểm tiêu cực hoặc không xác định được quan điểm đối với sản phẩm x∈P. Giá trị θ được xác định thông qua kiểm nghiệm và tùy thuộc vào từng bộ dữ liệu. Trong bài báo này chúng tôi sử dụng ngưỡng θ = 0.85.

$$Post(i,x) = \begin{cases} 1 & \text{nếu } Sentiment(\ ContentPost(x)) > \theta \\ 0 & otherwise \end{cases} \quad (2)$$

$$POST = \{Post(i,x) | i \in U, x \in P\} \quad (3)$$

Hệ khuyến nghị dựa vào hành vi post của người dùng được xác định theo (1) và (3) được biểu diễn thành hai đồ thị hai phía. Đồ thị hai phía thứ nhất biểu diễn đánh giá của người dùng có các cạnh (i, x) nối giữa đỉnh người dùng i∈U và đỉnh sản phẩm x∈P. Trọng số cạnh (i, x) của đồ thị này được đánh là rix theo công thức (1). Đồ thị hai phía thứ hai biểu diễn quan điểm của người dùng i∈U đối với sản phẩm x∈P thông qua hành vi post. Trọng số cạnh (i, x) của đồ thị này được đánh là Post(i, x) theo công thức (2).

### *2.2. Mở rộng mô hình dữ liệu cho hành vi like*

Hành vi like cho phép người dùng bày tỏ cảm súc của mình đối với một sản phẩm thông qua bài post của một người dùng khác. Người dùng cũng có thể like hoặc không like bài post của một người dùng khác dù cho bài post đó có nội dung positive, negative hay neutral. Một bài post có thể nhận được rất nhiều người dùng like. Nếu hầu hết người dùng đều like bài post nói về một chủ đề hoặc sản phẩm thì ta nói đó là "trào lưu" hay "định hướng" người dùng của lĩnh vực phân tích thông tin trong mạng xã hội [4]. Việc tìm ra "trào lưu" hay "định hướng" của người dùng trong mạng xã hội thông qua hành vi like cũng là yếu tố quan trọng để nâng cao chất lượng khuyến nghị. Vấn đề đặt ra là làm thế nào để ước lượng "trào lưu" hay "định hướng" của người dùng thông qua hành vi like trong mạng xã hội.

Để xác định "trào lưu" hay "định hướng" người dùng thông qua hành vi like cho hệ tư vấn chúng tôi đề xuất phương pháp tiến hành như sau:

Gọi UL⊆U là tập người dùng i∈U đã like bài post có nội dung ContentPost(x) chứa đựng quan điểm tích cực đối với sản phẩm x∈P được xác định theo công thức (4). Giá trị Sentiment(ContentPost (x)) phải lớn hơn một ngưỡng θ đủ lớn và nhận được like của người dùng i∈UL. Sentiment(ContentPost(x)). Bằng cách này ta ước lượng được số lượng người dùng gián tiếp có quan điểm tốt đối với sản phẩm x∈P. UL và ContentPost(x) dễ dàng được lấy tự động thông qua API của các mạng xã hội [14].

$$U_L = \{i \in U \text{ like } ContentPost(x) | Sentiment(ContentPost(x)) > \theta\} \quad (4)$$

Gọi UR⊆U là tập người dùng đã có đánh giá cao cho sản phẩm x∈P được xác định theo công thức (5). Giá trị α theo (5) được xác định đủ lớn để xác định người dùng i∈U có đánh giá cao cho sản phẩm x∈P. Tập UR chính là tập người dùng like sản phẩm x∈P không cần dựa vào bài post của bất kỳ người dùng nào thông qua giá trị rix. Đây cũng là điểm khác biệt riêng của hệ tư vấn với lĩnh vực phân tích thông tin [12].

$$U_R = \{i \in U | r_{ix} > \alpha\} \quad (5)$$

Gọi Like(i, x) là "trào lưu" hay "định hướng" của mỗi người dùng $i \in U_L \cup U_R$ đối với sản phẩm x∈P cho bài post ContentPost(x) được xác định theo công thức (6). Trong công thức (6), chúng tôi sử dụng hằng số γ (0≤γ≤1). Nếu $\frac{|U_L \cup U_R|}{|U|}$ gần với 0, điều này có nghĩa rất ít người dùng like bài post ContentPost(x) và cũng rất ít người dùng có đánh giá tốt về sản phẩm x∈P. Khi đó ta nói "trào lưu" hay "định hướng" người dùng không có phản hồi tích cực đối với bài post ContentPost (x ). Nếu $\frac{|U_L \cup U_R|}{|U|}$ gần với 1, điều này có nghĩa hầu hết người dùng đều like bài post ContentPost(x) và hầu hết người dùng có đánh giá tốt về sản phẩm x∈P. Khi đó ta nói "trào lưu" hay "định hướng" người dùng phản ánh tích cực đối với bài post ContentPost (x).

$$Like(i,x) = \begin{cases} 1 & \text{nếu } i \in U_L \cup U_R \text{ và } \frac{|U_L \cup U_R|}{|U|} > \gamma \\ 0 & otherwise \end{cases} \quad (6)$$

$$LIKE = \{Like(i,x) | i \in U, x \in P\} \quad (7)$$

Hệ khuyến nghị dựa vào hành vi like của người dùng được xác định theo (1) và (7) được biểu diễn thành hai đồ thị hai phía. Đồ thị hai phía thứ nhất biểu diễn đánh giá của người dùng có các cạnh (i, x) nối giữa đỉnh người dùng i∈U và đỉnh sản phẩm x∈P. Trọng số cạnh (i, x) của đồ thị này được đánh là rix theo công thức (1). Đồ thị hai phía thứ hai biểu diễn xu hướng của tập người dùng i∈U đã like sản phẩm x∈P gián tiếp thông qua hành vi post được ước lượng theo công thức (6).

### *2.3. Mở rộng mô hình dữ liệu cho hành vi comment*

Song hành cùng hành vi like, post trong mạng xã hội là hành vi comment. Hành vi comment cho phép người dùng bày tỏ quan điểm của mình đối với sản phẩm thông qua bài post của một người dùng khác. Một bài post có thể được nhiều người dùng khác comment. Mỗi comment của người dùng i∈U là một đoạn văn bản phản ánh quan điểm của người dùng này cho nội dung bài post của một người dùng khác nói về sản phẩm x∈P. Gọi CommentContent(x) là nội dung commnent của người dùng i∈U nói về sản phẩm x∈P. CommentContent(x) có thể được lấy tự động thông qua API của các mạng xã hội [14]. Chúng tôi sử dụng phép ước lượng quan điểm của người dùng i∈U cho văn bản CommnentContent(x) được đề xuất trong [13]. Gọi Comment(i, x) là giá trị xác định quan điểm của người dùng i∈U theo nội dung CommentContent(x) được xác định theo công thức (8).

$$Comment(i,x) = \begin{cases} 1 & \text{nếu } Sentiment(CommentContent(x) > \theta \\ 0 & otherwise \end{cases} \quad (8)$$

$$COMMENT = \{Comment(i,x) | i \in U, x \in P\} \quad (9)$$

Hệ khuyến nghị dựa vào hành vi comment của người dùng được xác định theo (1) và (9) được biểu diễn thành hai đồ thị hai phía. Đồ thị hai phía thứ nhất biểu diễn đánh giá của người dùng có các cạnh (i, x) nối giữa đỉnh người dùng i∈U và đỉnh sản phẩm x∈P. Trọng số cạnh (i, x) của đồ thị này được đánh là rix theo công thức (1). Đồ thị hai phía thứ hai biểu diễn quan điểm của người dùng i∈U comment về một sản phẩm x∈P.

## III. MÔ HÌNH DỰ ĐOÁN DỰA VÀO HÀNH VI NGƯỜI DÙNG

Như đã trình bày ở trên, đồ thị biểu diễn đánh giá người dùng cho các sản phẩm được xác định theo (1), đồ thị biểu diễn quan điểm của người dùng đối với sản phẩm thông qua hành vi post được xác định theo (3), đồ thị biểu diễn xu hướng sử dụng sản phẩm của người dùng đối với sản phẩm thông qua hành vi like được xác định theo (7), đồ thị biểu diễn quan điểm của người dùng đối với sản phẩm thông qua hành vi comment được xác định theo (9) đều là những đồ thị hai phía. Ngoài lợi thế về mặt biểu diễn dữ liệu, đồ thị hai phía cho phép ta tính toán được độ tương tự giữa các cặp người dùng hoặc độ tương tự giữa các cặp sản phẩm một cách hiệu quả. Trong mục này, chúng tôi đề xuất phương pháp tư vấn kết hợp giữa đánh giá người dùng và các hành vi post, like, comment dựa vào người dùng trong mạng xã hội. Phương pháp tư vấn Social-UserBased đề xuất được thực hiện tuần tự theo bốn bước như trong Hình 1.

Tại bước 1 của thuật toán chúng tôi tính toán độc lập mức độ tương tự giữa các cặp người dùng dựa vào ma trận đánh giá R, POST, LIKE, COMMENT. Vì đồ thị biểu diễn ma trận đánh giá R, POST, LIKE, COMMENT đều là đồ thị hai phía nên chúng tôi đề xuất việc tính toán độ tương tự giữa các cặp người dùng dựa trên tổng trọng số của tất cả các đường đi từ đỉnh người dùng đến đỉnh người dùng. Trong phương pháp này, việc xác định độ tương tự giữa các cặp người dùng bằng cách tính tổng trọng số các đường đi độ dài L từ đỉnh người dùng i∈U đến đỉnh người dùng j∈U trên đồ thị hai phía. Cặp người dùng i, j∈U có tổng trọng số các đường đi độ dài L lớn nhất sẽ tương tự nhau nhiều nhất. Do đồ thị biểu diễn là đồ thị hai phía vì vậy L luôn là một số chẵn được xác thông qua thực nghiệm. Cụ thể, ký hiệu $R^{(L)}$ là ma trận ước lượng độ tương tự giữa các cặp người dùng dựa vào tổng trọng số các đường đi độ dài L, $R^T$ là ma trận chuyển vị của R, β là một hằng số (0< β <1). Khi đó, mức độ tương tự giữa các cặp người dung theo ma trận đánh giá được xác định theo công thức (11).

$$R^{(L)} = \begin{cases} \beta.R.R^T & \text{nếu } L = 2 \\ \beta.R.R^T.R^{(L-2)} & \text{nếu } L = 4, 6, \ldots \end{cases} \quad (10)$$

Do quan hệ giữa người dùng và sản phẩm thông qua hành vi post, like, comment được xác định theo (3), (7), (9) cũng được biểu diễn thành các đồ thị hai phía vì vậy độ tương tự giữa các cặp người dùng theo các hành vi này cũng được xác định tương tự như đối với ma trận đánh giá người dùng.

Gọi $POST^{(L)}$, $LIKE^{(L)}$, $COMMENT^{(L)}$ là độ tương tự giữa các cặp người dùng dựa vào các hành vi post, like, comment. Gọi $POST^T$, $LIKE^T$, $COMMENT^T$ là ma trận chuyển vị của các ma trận tương ứng. Khi đó, độ tương tự giữa các cặp người dùng theo hành vi post, like, comment được xác định theo công thức (11), (12), (13) theo thứ tự.

**Thuật toán Social-UserBased:**

**Đầu vào** :
- Ma trận đánh giá R được xác định theo công thức (1).
- Ma trận Post(i, x) được xác định theo công thức (3).
- Ma trận Like(i, x) được xác định theo công thức (7).
- Ma trận Comment(i, x) được xác định theo công thức (9).
- Người dùng i∈U là người dùng cần được tư vấn.

**Đầu ra** :
- Danh sách k sản phẩm mới phù hợp nhất đối với người dùng i.

**Các bước tiến hành**:

**Bước 1**. Tính toán mức độ tương tự giữa các cặp người dùng:

1.1. Tính toán mức độ tương tự giữa các cặp người dùng dựa trên ma trận đánh giá R theo công thức (10):

$$R^{(L)} = \begin{cases} \beta.R.R^T & nếu\ \ L = 2 \\ \beta.R.R^T.R^{(L-2)} & nếu\ L = 4,6,\dots \end{cases}$$

1.2. Tính toán mức độ tương tự giữa các cặp người dùng dựa trên hành vi post theo công thức (11):

$$POST^{(L)} = \begin{cases} \beta.POST.POST^T & nếu\ \ L = 2 \\ \beta.POST.POST^T.POST^{(L-2)} & nếu\ L = 4,6,\dots \end{cases}$$

1.3. Tính toán mức độ tương tự giữa các cặp người dùng trên hành vi like theo công thức (12): $LIKE^{(L)} = \begin{cases} \beta.LIKE.LIKE^T & nếu\ \ L = 2 \\ \beta.LIKE.LIKE^T.LIKE^{(L-2)} & nếu\ L = 4,6,\dots \end{cases}$

1.4. Tính toán mức độ tương tự giữa các cặp người dùng dựa trên hành vi comment theo công thức (13):

$$COMMENT^{(L)} = \begin{cases} \beta.COMMENT.COMMENT^T & nếu\ \ L = 2 \\ \beta.COMMENT.COMMENT^T.COMMENT^{(L-2)} & nếu\ L = 4,6,\dots \end{cases}$$

**Bước 2**. Tìm tập láng giềng cho người dùng cần tư vấn i∈U:

**2.1** . $K_R^{(i)} =<$ *Tập láng giềng của dùng j ∈ U dựa vào ma trận đánh giá R* >.

**2.2** . $K_P^{(i)} =<$ *Tập láng giềng của dùng j ∈ U dựa vào hành vi POST*>.

**2.3** . $K_L^{(i)} =<$ *Tập láng giềng của dùng j ∈ U dựa vào hành vi LIKE*>.

**2.4** . $K_C^{(i)} =<$ *Tập láng giềng của dùng j ∈ U dựa vào hành vi COMMNENT* >.

**2.5** . $K^{(i)} = K_R^{(i)} \cap K_P^{(i)} \cap K_L^{(i)} \cap K_C^{(i)}$;

**Bước 3**. Dự đoán quan điểm của người dùng i∈U đối với các sản phẩm mới x∈P [1]:

$$r_{ix} = \frac{1}{|K_i|}\sum_{j \in K_i} r_{jx};$$

**Bước 4**. Tạo nên tư vấn cho người dùng i∈U các sản phẩm mới x∈P:

4.1. Sắp xếp $r_{ix}$ theo thứ tự tăng dần của trọng số.

*Hình 1. Thuật toán Social-UserBased.*

$$POST^{(L)} = \begin{cases} \beta.POST.POST^T & nếu\ \ L = 2 \\ \beta.POST.POST^T.POST^{(L-2)} & nếu\ L = 4,6,\dots \end{cases} \quad (11)$$

$$LIKE^{(L)} = \begin{cases} \beta.LIKE.LIKE^T & nếu\ \ L = 2 \\ \beta.LIKE.LIKE^T.LIKE^{(L-2)} & nếu\ L = 4,6,\dots \end{cases} \quad (12)$$

$$COMMENT^{(L)} = \begin{cases} \beta.COMMENT.COMMENT^T & nếu\ \ L = 2 \\ \beta.COMMENT.COMMENT^T.COMMENT^{(L-2)}\ L = 4,6,\dots \end{cases} \quad (13)$$

Tại bước 2, thuật toán tìm tập láng giềng cho người dùng cần tư vấn i∈U. Gọi $K_R^{(i)}$, $K_P^{(i)}$, $K_L^{(i)}$, $K_C^{(i)}$ là tập láng giềng của người dùng i∈U được xác thông qua độ tương tự giữa các cặp người dùng dựa vào $R^{(L)}$, $POST^{(L)}$, $LIKE^{(L)}$, $COMMENT^{(L)}$ đã được tính toán ở bước 1. Phương pháp tìm $K_R^{(i)}$, $K_P^{(i)}$, $K_L^{(i)}$, $K_C^{(i)}$ được thực hiện đơn giản bằng cách lấy k người dùng j∈U có mức độ tương tự lớn nhất đối với người dùng i∈U làm $K_R^{(i)}$, $K_P^{(i)}$, $K_L^{(i)}$, $K_C^{(i)}$. Tại bước 2.5, chúng tôi tiến hành tìm tập láng giềng cho người dùng i∈U bằng cách tìm $K^{(i)}$ là tập người dùng thuộc tập giao giữa các tập $K_R^{(i)}$, $K_P^{(i)}$, $K_L^{(i)}$, $K_C^{(i)}$. Tập $K^{(i)}$. $K^{(i)}$ chính là tập người dùng vừa tương tự nhau theo đánh giá, Post, Like và Comment. Bước 3 của thuật toán thực hiện dự đoán quan điểm của người dùng i∈U đối với các sản phẩm mới x∈P bằng cách lấy trung bình các đánh giá khác 0 của người dùng đối với sản phẩm trong tập láng giềng theo công thức (14) [1, 3]. Bước 4 của thuật toán thực hiện sinh ra tư vấn cho người dùng i∈U bằng cách chọn k sản phẩm mới có giá trị dự đoán cao nhất gợi ý cho người dùng này.

$$r_{ix} = \frac{1}{|K_i|}\sum_{j \in K_i} r_{jx} \quad (14)$$

## IV. THỬ NGHIỆM VÀ ĐÁNH GIÁ

Như đã trình bày ở trên, do các bộ dữ liệu của lọc cộng tác trong mạng xã hội hiện tại mới chỉ bao gồm dữ liệu đánh giá người dùng đối với sản phẩm và dữ liệu phản về mối quan hệ kết bạn thông qua hành vi friend [15, 16]. Để đánh giá hiệu quả của các phương pháp tư vấn kết hợp với hành vi người dùng trong mạng xã hội đề xuất, chúng tôi tiến xây dựng bộ dữ liệu và thử nghiệm. Phương pháp xây dựng bộ dữ liệu và kết quả thử nghiệm trình bày ở trên được đánh giá và so sánh với các phương pháp khác theo thủ tục mô tả dưới đây.

### 4.1. Dữ liệu thử nghiệm

Thuật toán Social-UserBased đề xuất được thử nghiệm trên tập dữ liệu thu thập bằng tiếng Anh do nhóm nghiên cứu tự xây dựng. Dữ liệu thu thập được bao gồm 6090 người dùng, 1754 khách sạn, 4999 đánh giá người dùng cho mỗi khách sạn, 5710 hành vi add friend, 961 bài post có quan điểm tích cực đối với khách sạn, 4757 hành vị like, 2995 hành vi comment. Đánh giá của người dùng cho mỗi khách sạn được thể hiện theo 5 thang bậc đánh giá Ω = {1.0, 0.8, 0.6, 0.4, 0.2} tương ứng với {Perfect, very good, good, bad, very bad". Dữ liệu về các hành vi người dùng được lấy tự động và xử lý như sau:

- Hành vi Add Friend của tập người dùng được lấy trực tiếp thông qua Facebook API [14]. Dữ liệu của tập này được tiền xử lý và chỉ giữ lại các mối quan hệ kết bạn của 6090 người dùng trong cơ sở dữ liệu.

- Nội dung các bài post của người dùng được lấy tự động thông qua Facebook API [14]. Ứng với mỗi bài post, chúng tôi sử dụng API của xử lý ngôn ngữ tự nhiên để ước lượng quan điểm người dùng thông qua hành vi post [12]. Chọn θ=0.85 để ước lượng quan điểm của người dùng i∈U đối với sản phẩm x∈P. Điều này có nghĩa nếu người dùng i∈U có bài post về khách sạn x∈P với quan điểm tích cực nếu Sentiment(ContentPost(x))>0.85.

- Các hành vi like bài post có quan điểm tích cực đối với khách sạn được lấy tự động thông qua Facebook API [14]. Điều này đã mặc định chọn θ=0.85 để tính toán theo (4). Chọn α=0.4 để xác định UR theo công thức (5). Chọn γ =0. 5 để xác định xu hướng sử dụng của người dùng theo công thức (6).

- Nội dung các comment của người dùng cũng được lấy tự động thông qua Facebook API [14]. Lấy θ=0.85 để ước lượng quan điểm người dùng đối với khách sạn thông qua API của xử lý ngôn ngữ tự nhiên [12, 13]. Lấy β=0.5 để tính toán tổng trọng số các đường đi độ dài L cho các công thức (10), (11), (12), (13).

4.2. Phương pháp thử nghiệm

Trước tiên, toàn bộ dữ liệu thử nghiệm được chia thành hai phần, một phần Utr được sử dụng làm dữ liệu huấn luyện, phần còn lại Ute được sử dụng để kiểm tra. Tập Utr chứa 80% đánh giá và tập Ute chứa 20% đánh giá. Dữ liệu huấn luyện được sử dụng để xây dựng mô hình theo thuật toán mô tả ở trên. Với mỗi người dùng i thuộc tập dữ liệu kiểm tra, các đánh giá (đã có) của người dùng được chia làm hai phần Oi và Pi. Oi được coi là đã biết, trong khi đó Pi là đánh giá cần dự đoán từ dữ liệu huấn luyện và Oi[1, 2, 3].

Sai số dự đoán MAEu với mỗi khách hàng u thuộc tập dữ liệu kiểm tra được tính bằng trung cộng sai số tuyệt đối giữa giá trị dự đoán và giá trị thực đối với tất cả mặt hàng thuộc tập Pu.

$$MAE_u = \frac{1}{|P_u|}\sum_{y \in P_u} |\hat{r}_{uy} - r_{uy}| \qquad (15)$$

Sai số dự đoán trên toàn tập dữ liệu kiểm tra được tính bằng trung bình cộng sai số dự đoán cho mỗi khách hàng thuộc Ute. Giá trị MAE nhỏ thì phương pháp dự đoán có độ chính xác cao [1, 6, 7].

$$MAE = \frac{\sum_{u \in U_{te}} MAE_u}{|U_{te}|} \qquad (16)$$

4.3. So sánh và đánh giá

Phương pháp Social-UserBased đề xuất trong Mục 3 được cài đặt bằng Python. Phương pháp thử nghiệm và so sánh với những phương pháp sau:

Phương pháp k láng giềng gần nhất dựa vào người dùng sử dụng độ tương quan Pearson (ký hiệu là CF-UserBased) [1, 2]. Phương pháp này chỉ sử dụng dữ liệu đánh giá của người dùng đối với sản phẩm [1, 7].

Phương pháp k láng giềng gần nhất dựa vào sản phẩm sử dụng độ tương quan Pearson (ký hiệu là CF-ItemBased) [2]. Phương pháp này chỉ sử dụng dữ liệu đánh giá của người dùng đối với sản phẩm [1].

Phương pháp TidalTrust sử thuật toán tìm kiếm theo chiều rộng tính toán độ tin tưởng giữa các cặp người dùng có đường đi ngắn nhất với trọng số là độ tin cậy [8, 10]. Phương pháp này chỉ sử dụng dữ liệu đánh giá của người dùng đối với sản phẩm cùng với dữ liệu thông qua hành vi kết bạn add friend [8].

Phương pháp MoleTrust sử dụng thuật toán tìm kiếm theo chiều sâu để xác định đường đi ngắn nhất giữa các cặp người dùng có độ dài nhỏ hơn d [10]. Phương pháp này chỉ sử dụng dữ liệu đánh giá của người dùng đối với sản phẩm cùng với dữ liệu thông qua hành vi kết bạn add friend [10].

Phương pháp TrustWalker sử dụng thuật toán ramdom walk để kết hợp mô hình tin cậy, mô hình dự đoán dựa vào người dùng, mô hình dự đoán dựa vào sản phẩm [8]. Phương pháp này chỉ sử dụng dữ liệu đánh giá của người dùng đối với sản phẩm cùng với dữ liệu thông qua hành vi kết bạn add friend [8].

Lấy ngẫu nhiên 4000 người dùng trong tập dữ liệu làm dữ liệu huấn luyện. Chọn ngẫu nhiên 1000 người dùng trong số còn lại để làm tập dữ liệu kiểm tra. Giá trị MAE trong Bảng 1 được lấy trung bình của 10 lần thử nghiệm ngẫu nhiên. Giá trị MAE nhỏ chứng tỏ phương pháp có kết quả dự đoán tốt [1, 3].

Kết quả trong Bảng 1 cho thấy phương pháp khuyến nghị dựa vào người dùng và phương pháp khuyến nghị dựa vào sản phẩm cho lại giá trị MAE lớn nhất. Với số lượng người dùng của tập láng giềng lần lượt là 50, 100, 120, 150, giá trị MAE của các phương pháp này đều lớn hơn 0.35. Kết quả này có thể lý giải cả hai phương pháp thực hiện dự đoán dựa trên duy nhất ma trận đánh giá người dùng có số lượng đánh giá khác 0 rất thưa. Phương pháp TidalTrust, MoleTrust, TrustWalker cải thiện không đáng kể sai số dự đoán MAE. Ứng với số lượng người dùng trong tập láng giềng k=50, giá trị MAE của các phương pháp này đều lớn hơn 0.35. Khi tăng số lượng người dùng trong tập láng giềng k=100, 120, 150 giá trị MAE có giảm đi nhưng rất nhỏ. Điều này có thể khẳng định hành vi friend của người dùng trong mạng xã hội có tác động không đáng kể đến kết quả dự đoán. Kết quả này cũng hoàn toàn phù hợp với những nghiên cứu trước đây [3, 7].

*Bảng 1. Giá trị MAE của các phương pháp*

| Phương pháp | Số lượng người dùng của tập láng giềng | | | |
|---|---|---|---|---|
| | 50 | 100 | 120 | 150 |
| CBF-UserBased | 0.3612 | 0.3522 | 0.3492 | 0.3405 |
| CF-ItemBased | 0.3598 | 0.3573 | 0.3514 | 0.3541 |
| TidalTrust | 0.3558 | 0.3497 | 0.3419 | 0.3412 |
| MoleTrust | 0.3584 | 0.3473 | 0.3397 | 0.3384 |
| TrustWalker | 0.3529 | 0.3315 | 0.3229 | 0.3271 |
| Social-UserBased | 0.1478 | 0.1485 | 0.1461 | 0.1432 |

Giá trị MAE của phương pháp Social-UserBased nhỏ xấp xỉ một nửa so với các phương pháp còn lại. Nghi ngờ có sự nhầm lẫn nào đó trong khi thực nghiệm, chúng tôi tiến hành kiểm tra độc lập việc kết hợp giữa đánh giá người dùng với từng hành vi riêng rẽ. Trước tiên, chúng tôi kiểm tra việc kết hợp giữa tập đánh giá người dùng và hành vi post. Kết quả cho thấy giá trị MAE của phương pháp này đều nhỏ hơn 0.2. Điều này chứng tỏ hành vi post có quan điểm tích cực của người dùng đối với sản phẩm tác động không nhỏ đến chất lượng dự đoán của hệ tư vấn. Kết quả này cũng phù hợp với số liệu của Facebook đưa ra: Có trên 67% các giao dịch điện tử thành công thông qua hành vi post. Tiếp đến, chúng tôi kiểm tra việc kết hợp giữa đánh giá người dùng và hành vi like và comment, kết quả cho thấy các phương pháp kết hợp này cho lại giá trị MAE<0.25. Nhưng khi mở rộng việc kết hợp lên hai hành vi ((post, like), (post, comment)) và ba hành vi (post, like, comment) kết quả tốt dần lên như trong Bảng 1. Điều này chỉ có thể lý giải phương pháp ước lượng quan điểm của người dùng i∈U đối với sản phẩm x∈P theo cách tiếp cận

của xử lý ngôn ngữ tự nhiên là hoàn toàn tin cậy. Phương pháp tính toán độ tương tự giữa các cặp người dùng hoặc sản phẩm bằng cách xác định tổng trọng số các đường đi độ dài L cho lại kết quả tốt hơn so với độ đo tương tự dựa vào trust. Phương pháp xác định tập láng giềng cho người dùng hoặc sản phẩm dựa vào các hành vi rate, post, like, comment chính xác hơn so với phương pháp dựa vào mô hình tin cậy.

## V. KẾT LUẬN

Bài báo đã đề xuất một phương pháp khuyến nghị kết hợp với hành vi người dùng trong mạng xã hội. Trong đó, mô hình dữ liệu của hệ tư vấn được xây dựng bằng cách sử dụng các API của mạng xã hội kết hợp với API của xử lý ngôn ngữ tự nhiên để ước lượng quan điểm của người dùng đối với sản phẩm. Bằng cách này, chúng tôi dịch chuyển đồ thị biểu diễn mạng xã hội tổng quát về mạng thu nhỏ của các hành vi. Mạng xã hội thu nhỏ theo các hành vi được biểu diễn thành các đồ thị hai phía cho phép ta sử dụng các độ đo tương tự trên đồ thị để xây dựng mô hình dự đoán. Mô hình dự đoán được xây dựng bằng cách xác định tập người dùng có cùng chung sở thích theo đánh giá người dùng, có cùng chung sở thích post, có cùng chung sở thích like, cùng chung sở thích comment đối với các sản phẩm làm tập láng giềng. Thực nghiệm trên tập dữ liệu thực do nhóm tự xây dựng cho thấy sử dụng tập láng giềng kết hợp giữa các hành vi cho lại kết quả dự đoán tốt hơn rất nhiều so với các nghiên cứu gần đây dựa trên mô hình tin cậy. Với kết quả thực nghiệm nhận được ta có thể khẳng định một số điểm sau:

Phương pháp ước lượng quan điểm của người dùng đối với hành vi post và comment của người dùng trong mạng xã hội có thể được thực hiện hiệu quả thông qua các kết quả nghiên cứu về khai phá quan điểm người dùng của xử lý ngôn ngữ tự nhiên.

Phương pháp xác định xu hướng sử dụng sản phẩm của người dùng trong mạng xã hội cho hệ tư vấn có thể được xác định bằng cách kết hợp giữa đánh giá người dùng với hành vi like của người dùng trong mạng xã hội.

Độ tương tự giữa các cặp người dùng hoặc sản phẩm khi kết hợp với hành vi người dùng trong mạng xã hội cho lại kết quả chính xác hơn các độ đo tương tự truyền thống. Đặc biệt, kết quả tư vấn kết hợp giữa đánh giá người dùng và các hành vi người dùng trong mạng xã hội có kết quả chính xác cao hơn so với mô hình tin cậy.

Với cách tiếp cận này, bài báo còn có thể mở rộng nghiên cứu cho trường hợp dữ liệu lớn, dữ liệu thưa, người dùng mới và sản phẩm mới của hệ khuyến nghị. Những vấn đề này sẽ được chúng tôi trình bày ở những nghiên cứu tiếp theo của bài báo.

## TÀI LIỆU THAM KHẢO


[1] Su X., Khoshgoftaar T. M., "A Survey of Collaborative Filtering Techniques.,". Advances in Artificial Intelligence ,2009, pp.1-20.

[2] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. User Model. User-Adapt. Interact. 30, 1 (2020), 127–158.

[3] Eva Zangerle and Christine Bauer. Evaluating Recommender Systems: Survey and Framework. ACM Comput. Surv. 55, 8, Article 170 (December 2022).

[4] Jyoti Shokeen, Chhavi Rana. Social Recommender System: Techniques, Domains, Metrics, Datasets and Future Scope. Journal of Intelligent Informaton Systems. Vol: 54, pp: 633-667 (2020).

[5] Guy, I., Carmel, D.: Social recommender systems. In: Proceedings of the 20th international conference companion on World wide web, pp. 283–284. ACM (2011).

[6] C. A. Gomez-Uribe and N. Hunt, "The netflix recommender system: Algorithms, business value, and innovation," ACM Transactions on Management Information Systems (TMIS), vol. 6, no. 4, p. 13, 2016.

[7] J. Davidson et al., "The YouTube video recommendation system," in Proceedings of the fourth ACM conference on Recommender systems, 2010, pp. 293-296: ACM.

[8] Jamali, M., Ester, M.: Trustwalker: a random walk model for combining trust-based and item-based recommendation. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 397–406. ACM (2009).

[9] Guy, I., Jacovi, M., Shahar, E., Meshulam, N., Soroka, V., Farrell, S.: Harvesting with sonar: the value of aggregating social network information. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1017–1026. ACM (2008).

[10] Massa, P., Avesani, P.: Trust-aware recommender systems. In: Proceedings of the 2007 ACM conference on Recommender systems, pp. 17–24. ACM (2007).

[11] Bo Pang, Lillian Lee, "Opinion mining and sentiment analysis". Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1–135.

[12] Bilal Saberi, Saidah Saad. Sentiment Analysis or Opinion Mining: A Review. International Journal on Advanced Science Engineering Information Technology. Vol 17 (2017), pp: 1660-1666.

[13] Christopher D. Manning, Mihai Surdeanu, John Bauer. The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, Baltimore, Maryland USA, June 23-24, 2014.

[14] https://developers.facebook.com/

[15] http://www.epinions.com

[16] http://www.Flixter.com


## RECOMMENDATION BASED ON SOCIAL NETWORKS USER'S BEHAVIOURS


***Abstract:*** The birth of social network is considered to be the most profound event for Internet users at the moment. Some of famous online social networks, such as Facebook, Twitter, LinkedIn have become popular and changed tranditional communication ways. Communication between entities within social networks create a rich and diverse resource that is a good opportunity for analysis, exploration and application development. In this paper, we propose a collaborative filtering method that combine with user behaviours in social networks. The method is conducted by analyzing user opinions expressed through rate, post, like, comment behaviours in social networks. Based on the analysis results of opinion statements, we propose algorithms to recommend suitable items for each user. The experimental results on a real data set that collected on Facebook social network show that the proposed methods achieve superior performance compared to approach methods based on trust models.



***Keywords:*** Collaborative Filtering Recommendation, Content-based Filtering Recommendation, Hybrid

Filtering Recommendation System, Social Recommender Systems.

**Nguyễn Duy Phương** nhận học vị Tiến sỹ năm 2010. Hiện là trưởng khoa Công nghệ thông tin 1 – Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, các hệ thống tư vấn, lý thuyết đồ thị và ứng dụng, các kỹ thuật kiểm thử tự động, các kỹ thuật tối ưu cho hệ thống lập trình trực tuyến.

**Nguyễn Mạnh Sơn** nhận học vị Thạc sỹ năm 2009. Hiện công tác tại Khoa Công nghệ thông tin 1 – Học viện Công nghệ Bưu chinh Viễn thông. Lĩnh vực nghiên cứu: khai phá dữ liệu, các kỹ thuật lọc cộng tác, các kỹ thuật tối ưu hệ thống lập trình trực tuyến.