HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN ĐỨC KHÔI

KỸ THUẬT LỌC CỘNG TÁC TRONG TƯ VẤN NGƯỜI DÙNG TWITTER

Chuyên ngành: Khoa học máy tính

Mã số: 60.48.01.01

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2013

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: …PGS. TS Từ Minh Phương

Phản biện 1: ……

Phản biện 2: ……

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

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

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỤC LỤC

MỞ ĐẦU

Ngày nay, mạng xã hội ngày càng phát triển và đi sâu vào cuộc sống của con người. Trên thế giới có hàng trăm mạng mạng xã hội khác nhau, trong đó một trong mạng xã hội phát triển nhanh nhất và thành công nhất mặc dù có mặt khá muộn, đó là Twitter.

Với số lượng người sử dụng lên đến trên 500 triệu người, lượng tweet được người dùng đăng lên hàng ngày rất lớn, lên đến 340 triệu tweets mỗi ngày, kèm với đó là một lượng thông tin khổng lồ được chia sẻ và cập nhật mới nhất. Mỗi người dùng có thể lựa chọn theo dõi một cá nhân hoặc tổ chức nào đó, mà người đó quan tâm, và ngược lại cũng có thể được theo dõi bởi các người dùng khác, Twitter sẽ hiển thị những tweet mới nhất được đăng tải bởi các cá nhân hoặc tổ chức mà người dùng đang theo dõi, theo thứ tự thời gian đăng tweet đó.

Một vấn đề đặt ra là khi số lượng tweet tăng lên nhanh như vậy, do người dùng theo dõi quá nhiều cá nhân hoặc tổ chức khác thì vấn đề lớn mà họ gặp phải chính là sự quá tải thông tin. Rất nhiều thông tin hữu ích có thể sẽ bị mất đi do các tweet khác mới hơn được cập nhật và làm đẩy lùi các tweet trước đó, trong khi những tweet đó không phải là những thông

tin thực sự cần thiết mà người dùng quan tâm. Đồng thời, một bài toán khác cũng được quan tâm là rất nhiều người dùng muốn có được những thông tin hữu ích nằm ngoài luồng thông tin mà họ nhận được bởi những người mà họ chủ động theo dõi, những thông tin đó có thể được đăng bởi những người bạn của bạn, hoặc từ những Blog được theo dõi bởi những người bạn của người dùng đó.

Chính vì vậy, việc nghiên cứu xây dựng hệ tư vấn nhằm tư vấn cho người dùng những tweet hữu dụng là một vấn đề quan trọng và có ý nghĩa thực tiễn. Hệ thống sẽ khuyến nghị cho mỗi người dùng Twitter một danh sách các tweet mà nhiều khả năng người đó sẽ quan tâm.Danh sách này được cá nhân hóa, tức là dựa trên mối quan tâm của từng người dùng.

CHƯƠNG 1. MẠNG XÃ HỘI TWITTER VÀ CÁC ĐẶC TRƯNG

Twitter đã phát triển rất nhanh để trở thành mạng xã hội phổ biến trong những năm gần đây và cung cấp một số lượng lớn người dùng sử dụng để đăng các bản tin, hoặc có thể được gọi là các tweet. Các tweet đó được Twitter hiển thị cho người dùng theo thứ tự về thời gian và được gọi là Timeline, người dùng sẽ dựa vào timeline để theo dõi những thông tin mà họ có thể sẽ quan tâm. Tuy nhiên, vấn đề quá tải thông tin đã gây khó khăn cho người sử dụng, đặc biệt khi người dùng đó theo dõi nhiều người dùng khác và có hàng ngàn tweet đến với họ mỗi ngày. Luận văn này sẽ tập trung vào việc đưa ra những tweet hữu ích mà người dùng thực sự quan tâm thông qua các phương pháp tư vấn, giúp người dùng giảm công sức bỏ ra để tìm kiếm những thông tin đó.

1.1 Giới thiệu mạng xã hội Twitter

Twitter là dịch vụ mạng xã hội miễn phí cho phép người dùng sử dụng đọc, nhắn và cập nhật các mẩu tin nhỏ gọi là tweet, đây là một dạng tiểu blog. Những mẩu tweet được giới hạn tối đa 140 ký tự và được lan truyền nhanh chóng trong phạm vi nhóm bạn của người nhắn hoặc có thể được trưng rộng rãi cho mọi người. Thành lập từ năm 2006, Twitter đã trở thành

một hiện tượng phổ biến toàn cầu, những tweet có thể chỉ là dòng tin cá nhân cho đến những cập nhật mang tính thời sự tại chỗ kịp thời và nhanh chóng hơn cả truyền thông chính thông.

1.2 Các đặc trưng thông tin của Twitter

Người dùng Twitter cập nhật các bản tin ngắn bị giới hạn trong 140 ký tự được gọi là các $tweet$, và thuật ngữ để chỉ việc đăng các bản tin đó gọi là $tweeting$. Người dùng Twitter có mối quan hệ trực tiếp với nhau, nếu người dùng A theo dõi người dùng B nhưng B không theo dõi A, A sẽ thấy tất cả các tweet của B nhưng ngược lại, B không thấy tweet của A.

Thuật ngữ mà Twitter đề xuất cho những mối quan hệ giữa người dùng Twitter với nhau gồm có $follower$ và $followee$, $follower$ là những người đang theo dõi một người dùng nào đó, và $followee$ là chỉ những người đang được người dùng theo dõi. Ví dụ trong hình 1.1, A đang theo dõi B, vì thế A sẽ là $follower$ của B, và B là $followee$ của A. Mỗi người dùng sẽ có một danh sách hiển thị những tweet mới được cập nhật, danh sách đó được gọi là Twitter stream theo thứ tự thời gian. Các tweet hiển thị trong danh sách này chính là những tweet được đăng bởi các followee. Trong ví dụ ở hình 1.1, nếu A đang follow B, tất cả các tweet của B sẽ được hiển thị trong danh sách các tweet của A, nhưng nếu B không follow

A thì những tweet của A sẽ không hiển thị trong danh sách tweet của B, B phải lựa chọn ‘follow’ A để có thể thấy các tweet này trong danh sách tweet của mình hoặc truy cập vào trang cá nhân của A để thấy được tất cả các tweet mà A đã đăng.

Người dùng Twitter ngoài việc có thể chia sẻ các tweet dưới dạng một bản tin văn bản ngắn, Twitter còn chọ phép họ cung cấp thêm nhiều thông tin hữu ích trong bản tin đó, một trong những đặc trưng mà Twitter cung cấp giúp người dùng bổ sung thêm những thông tin hữu ích trong tweet của mình là hashtag, mention và retweet.

Tất cả các đặc trưng mà Twitter cung cấp đều góp phần thể hiện một phần quan điểm, sở thích cá nhân của người dùng, những hành động của người dùng cũng sẽ được lưu trữ trong hồ sơ người dùng và có thể trích xuất thông qua giao diện lập trình ứng dụng (API) mà Twitter cung cấp.

1.3 Mối quan hệ người dùng trọng mạng Twitter

Tính năng chính của Twitter là cho phép người dùng gửi tin nhắn văn bản ngắn gọi là tweet. Người dùng có thể theo dõi người sử dụng khác để tự động nhận được tất cả các tweets của họ và có thể thấy chúng đang được hiện thị trên trang chủ

của họ. Những người sử dụng mà một người nào đó theo dõi họ thì là bạn bè của họ, trong khi những người dùng mà đang theo dõi người đó thì sẽ được gọi là những người đi theo – *followers*. Hành động tham chiếu tới một người dùng nào đó trong một tweet của mình thì được gọi là đề cập đến – *mentions*. Mentions là các thông điệp trực tiếp gửi đến một hoặc nhiều người thông qua cơ chế đề cập và là một hình thức đặc biệt của truyền thông trực tiếp giữa những người sử dụng. Twitter cho phép người dùng trả lời – *reply* trực tiếp cho bất kỳ tweet nào tự động thêm một mention để phản hồi lại. Trả lời thường liên quan đến hai hướng trong giao tiếp, vì người dùng thường trả lời để phản hồi lại các thông tin mà họ được đề cập. Twitter cho phép việc trao đổi tin nhắn riêng như một cơ chế bổ sung cho thông tin liên lạc trực tiếp. Mặc dù vậy, nội dung của những tin nhắn này là cá nhân và không thể được truy cập mà không có sự cho phép. Hơn nữa, tin nhắn riêng chỉ chiếm một phần nhỏ của tất cả các tin nhắn trao đổi trên Twitter và do đó nếu chỉ sử dụng chúng để xác định thông tin liên lạc trực tiếp giữa những người sử dụng có thể dẫn đến một hình ảnh không đầy đủ. Bên cạnh truyền thông trực tiếp, tất cả các tweet sẽ được tự động quảng bá đến tất cả các người sử dụng đang theo dõi. Các Tweet có thể được *retweeted* hay nói cách khác, các tweets có thể được chuyển tiếp bởi người sử dụng cho tất

cả các followers của họ. *Retweeting* là một cơ chế truyền thông thực sự hiệu quả, nó giúp truyền bá thông tin trên mạng nhanh chóng hơn. Các thẻ đặc biệt được sử dụng để gán một hoặc nhiều chủ đề của một tweet được gọi là hashtags, các thẻ này được đặc trưng bởi sự hiện diện của ký tự "#" trước tên của chủ đề, như là một phần của văn bản của các tweet. Hashtags được sử dụng bởi Twitter để phân loại các tweet và nhóm chúng thành các loại, có thể xem bởi người sử dụng.

1.4 Các hệ tư vấn người dùng Twitter

Các hệ tư vấn mạng xã hội tư vấn các sản phẩm dựa trên sở thích của bạn bè của người dùng hay các thông tin phương tiện truyền thông xã hội khác, chẳng hạn như các bình luận. Các sản phẩm được tư vấn không nhất thiết là các thành phần của mạng xã hội. Ví dụ, trong trường hợp của Twitter, người ta có thể tư vấn các thông tin tạo được sự chú ý từ người dùng Twitter. Do đó, các tư vấn như vậy có thể được dùng để nhắm tới những người dùng bên ngoài của Twitter.

Các phương pháp tư vấn hiện tại trong mạng xã hội phải đáp ứng được các đặc tính duy nhất trong Twitter. Ví dụ, các phương pháp tư vấn kết nối bạn bè làm việc tốt trong các trang mạng xã hội như Facebook có thể không phát huy tác dụng trong tư vấn liên kết của Twitter.

Một số hệ thống tư vấn đã được đề xuất để giúp người dùng Twitter thực hiện chia sẻ thông tin và tương tác xã hội dễ dàng hơn. Phần nội dung dưới đây sẽ trình bày về phân loại các phương pháp tư vấn theo một số phạm trù được xác định dựa trên các kiểu chức năng người dùng trong Twitter.

Với sự hiểu biết của chúng tôi , đây là lần đầu tiên một phân loại được sử dụng để phân loại các phương pháp tư vấn trong Twitter.

1.4.1 Tư vấn followee

1.4.2 Tư vấn follower

1.4.3 Tư vấn Hashtag

1.4.4 Tư vấn tweet

1.4.5 Tư vấn retweet

1.4.6 Tư vấn tin tức

CHƯƠNG 2. LỌC CỘNG TÁC VÀ ÁP DỤNG TRONG HỆ TƯ VẤN NGƯỜI DÙNG TWITTER

2.1 Giới thiệu hệ tư vấn và kỹ thuật lọc cộng tác

2.1.1 Bài toán tư vấn

Một cách hình thức, bài toán tư vấn được tác giả Adomavicius và Tuzhilin mô tả như sau:

Gọi $U = (u_1, u_2, \dots, u_M)$ là tập hợp tất cả người dùng trong hệ tư vấn, $I = (i_1, i_2, \dots, i_N)$ là tập tất cả các sản phẩm có thể tư vấn.

Một hàm $g = U \times I \to R$ trong đó R là một tập hợp có thứ tự, được dùng để đo mức độ phù hợp của sản phẩm i_n đối với người dung u_m.

Như vậy, với mỗi người dùng u_m thuộc vào U, hệ tư vấn cần chọn ra các sản phẩm $i^{max,u_m} \in I$, chưa biết với mỗi người dùng u_m sao cho hàm g đạt giá trị lớn nhất

$$\forall u_m \in U, i^{max,u_m} = \arg\max g(u_m, i_n) \quad (2.1)$$

Trong các hệ thống tư vấn, mức độ phù hợp của sản phẩm thường được biểu diễn theo đánh giá thang điểm, phụ thuộc vào từng ứng dụng, các đánh giá này có thể được thực

hiện trực tiếp bởi người dùng, ví dụ người dùng tự đánh giá mức độ quan tâm đối với một sản phẩm nào đó hoặc được tính toán bởi hệ thống.

Mỗi người dùng thuộc không gian người dùng U có một hồ sơ riêng (user profile) bao gồm những thông tin về người dùng đó như tên, tuổi, giới tính, quốc tịch đồng thời những thông tin có liên quan giữa người dùng và hệ thống như lịch sử truy cập, số lượng kết nối giữa các người dùng, tham gia các nhóm trong hệ thống …

2.1.2 Các kỹ thuật tư vấn

Có rất nhiều phương pháp được đưa ra nhằm mục đích xây dựng hệ tư vấn cho người dùng, các hệ thống tư vấn hiện tại thường dựa trên 3 cách chính:

- Dựa trên nội dung (content – based): Người dùng sẽ được tư vấn những sản phẩm tương tự như các sản phẩm mà trước đó họ đã đưa ra đánh giá tích cực về sản phẩm đó.
- Cộng tác (collaborative): Hệ thống dựa trên những người dùng có sở thích tương đồng với người dùng hiện tại để tiến hành đưa ra tư vấn.

- Kết hợp (hybrid): Hệ thống kết hợp cả 2 phương pháp nội dung và cộng tác để đưa ra tư vấn

2.2 Áp dụng kỹ thuật lọc cộng tác tư vấn người dùng Twitter

Như đã trình bày ở phần trước, để tiến hành tư vấn cho người dùng thì yêu cầu đặt ra là phải thu thập được càng nhiều dữ liệu phản hồi từ người dùng càng tốt. Nhưng cũng chính vì vậy, một trong những vấn đề mà các hệ thống tư vấn lựa chọn phải đối mặt, đó là, sự đa dạng của dữ liệu phản hồi từ người dùng. Có thể chia dữ liệu này thành hai loại chính: đánh giá tường minh (explicit ratings) và đánh giá không tường minh (implicit ratings).

Với quan điểm là sở thích của người dùng đối với một sản phẩm có thể được thể hiện một cách trực quan thông qua hệ thống đánh giá mức độ yêu thích sản phẩm, hay được sử dụng nhất là việc cho điểm dựa vào số lượng sao, thì hầu hết các hệ thống thương mại điện tử như amazon hoặc MovieLens đều dựa trên loại dữ liệu là các đánh giá rõ ràng để thực hiện tư vấn. Khác với quan điểm này, Twitter và một số mạng xã hội khác, việc người dùng thể hiện sự quan tâm đối với một bản tin nào đó không thông qua hệ thống đánh giá giống như vậy. Chính vì thế, chúng ta cần tìm cách để xác định sở thích của

người dùng dựa trên các tác nhân ẩn mà hệ thống có thể cung cấp cho ứng dụng.

2.2.1 Xây dựng hồ sơ người dùng Twitter

Mỗi người dùng Twitter có các thông tin cá nhân như tên, tuổi, quốc tịch, sở thích… do họ tự thiết lập nhưng đó chỉ là thông tin chung chung mà người dùng có thể thiết lập hoặc không, và cũng có thể chúng không phải là những thông tin chính xác về người dùng. Vậy bằng cách nào chúng ta có thể xây dựng bộ hồ sơ hoàn chỉnh về một người dùng Twitter và thể hiện được họ là ai? Với các đặc trưng mà Twitter cung cấp, chúng ta hoàn toàn có thể tính toán được các hoạt động của người dùng trong mạng xã hội, các mối quan hệ của họ, họ đang theo dõi những ai và những người nào đang theo dõi họ, cũng như những chủ đề mà họ thấy ưa thích, quan tâm [11].

Để làm được những công việc đó, chúng ta cần phải thu thập đầy đủ nhất các thông tin về người dùng Twitter và thao tác trên nguồn dữ liệu thô đó, việc thu thập dữ liệu sẽ được trình bày chi tiết trong chương 3, trong chương này chúng ta sẽ quan tâm đến việc xử lý dữ liệu và xây dựng hồ sơ để đưa ra được tư vấn tốt nhất cho người dùng Twitter.

Trước tiên, chúng ta sẽ xem xét nguồn thông tin đơn giản nhất mà người dùng có, đó là các tweet gần đây của người dùng, trong biểu thức 2.25, cho một người dùng đích U_T nào đó, sẽ có $tweets(U_T)$ là tập các tweet gần đây nhất của người dùng này, mục đích của luận văn là tư vấn người dùng dựa trên sở thích của họ, mà sở thích của con người không bao giờ cố định trong một thời gian dài, vì thế những thông tin của người dùng đưa ra cũng phải mang tính chất thời sự. Tập $tweets(U_T)$ mà chúng ta xây dựng sẽ là 100 tweet gần nhất của người dùng.

$$tweets(U_T) = \{t_1, t_2, \dots., t_k\} \quad (2.2)$$

Mỗi người dùng Twitter sẽ follow một tập các người dùng khác được gọi là followees, và tương tự, mỗi người dùng cũng sẽ được follow bởi một tập các người dùng được gọi là followers. Tập tweet của tất cả những người dùng (followees và followers) cũng có thể sẽ cung cấp một số thông tin về sở thích của người dùng hiện tại.

$$followees(U_T) = \{f_1, f_2, \dots, f_m\} \quad (2.3)$$

$$followers(U_T) = \{g_1, g_2, \dots, g_n\} \quad (2.4)$$

Như chúng ta đã biết, Twitter cho phép người dùng follow một người dùng nào đó, và tất cả các tweet được đăng bởi người dùng đó sẽ được hiển thị trên danh sách tweet của anh ta. Như vậy, một người nào đó khi quyết định sẽ follow

người dùng khác, thể hiện được rằng, họ đã đăng những tweet hoặc hồ sơ của người đó có những thông tin mà anh ta thấy hứng thú, chính vì thế, chúng ta có thể coi các tweet của họ có vai trò tương tự với các tweet của người dùng đang được xem xét tư vấn và xây dựng hồ sơ. Ta có $followeeteets(U_T)$ là tập các tweet được đăng bởi followee của U_T.

$$followeetweets(U_T) = \bigcup_{\forall f_i \in followees(U_T)} \left(tweets(f_i)\right) \quad (2.5)$$

$$followertweets(U_T) = \bigcup_{\forall f_i \in followers(U_T)} \left(tweets(g_i)\right) \quad (2.6)$$

Như vậy, để xây dựng hồ sơ người dùng, chúng ta sẽ dựa vào 5 chiến lược: sử dụng tweet được đăng bởi chính người dùng đó $tweets(U_T)$, bởi tweet của các followee của họ $(followeetweets(U_T))$, tweet của các follower $(followertweets(U_T))$, thông tin của followee $(followees(U_T))$ hoặc thông tin của các followers $(followers(U_T))$.

2.2.2 Phương pháp tư vấn người dùng Twitter

Giả sử $U = \{u_1, u_2, \dots, u_N\}$ là tập các người dùng, và $T = \{t_1, t_2, \dots, t_M\}$ là tập các item, trong bài toán của chúng ta hiện tại, item chính là các tweet. Ta sẽ có $R(u_i) = \{t_1^i, t_2^i, \dots, t_R^i\}$ là tập các tweet tư vấn và được chấp nhận bởi

người dùng u_i. Với $A(u_i)$ và $A(t_i)$ là tập hợp các tweet đưa ra tư vấn được chấp nhận bởi người dùng u_i và tập các người dùng chấp nhận t_i khi tiến hành tư vấn.

Dựa vào hồ sơ người dùng đã được xây dựng ở phần trước, chúng ta sẽ đưa ra các chiến thuật trong việc tư vấn người dùng Twitter những tweet phù hợp với sở thích cá nhân của từng người dùng.

Mỗi chiến thuật đưa ra tư vấn người dùng Twitter được trình bày dưới đây sẽ tiến hành xếp hạng từng tweet trong danh sách những tweet sẽ được sử dụng để khuyến nghị người dùng, giá trị xếp hạng của tweet thể hiện mức độ quan tâm của người dùng đối với tweet đó, giá trị càng cao nghĩa là người dùng quan tâm nhiều đến tweet và sẽ tư vấn những tweet đó cho người dùng. Trong phạm vi khóa luận, em sẽ tiến hành tư vấn 20 tweet có xếp hạng cao nhất đến mỗi người dùng trong quá trình tiến hành kiểm thử. 5 chiến thuật tư vấn trình bày dưới đây sẽ được sử dụng để xếp hạng các tweet, bao gồm: phương pháp phân loại văn bản, tính phổ biến của tweet, mức độ chấp nhận của các followee, tính ngữ nghĩa của các từ khóa và sự tương tác của người dùng. Chúng ta sẽ tìm hiểu chi tiết thuật toán của từng chiến lược trong phần dưới đây.

2.2.2.1 Phân loại văn bản

Trong phần này, chúng ta sẽ dựa trên nội dung của tweet và tiến hành phân loại văn bản để thực hiện đánh giá và xếp hạng cho các tweet, dựa vào kết quả xếp hạng này, những tweet có thứ hạng cao sẽ được sử dụng để tư vấn cho người dùng cụ thể.

Các tweet sẽ được tổ chức thành các hạng mục khác nhau, mỗi hạng mục lại có thể nằm trong một hạng mục khác và tạo thành một cây phân cấp các hạng mục. Xếp hạng tweet là một chuỗi "A.B.C.D" với các hạng mục trong phân cấp được chỉ thị bởi dấu ".", với hạng mục dạng "A.B.C.D" thì A là hạng mục cha của B, B là cha của C…. Ví dụ người dùng Twitter có tên người dùng là phuongtu được xếp trong hạng mục "computer.machine-learning.mobile", người dùng nào đó follow phuongtu có thể sẽ quan tâm đến một trong những hạng mục mà phuongtu được xếp vào. Vì thế, chúng ta có thể sử dụng hạng mục của người dùng để đưa ra tư vấn những tweet mà người dùng đó quan tâm.

Giả sử $f(t_i)$ là một hàm ánh xạ trả về hạng mục của t_i, và có $C = \{c_1, c_2, \dots, c_P\}$ là tập các hạng mục tweet. Chúng ta sử dụng vector $N_i = \left[n_1^i, n_2^i, \dots, n_p^i\right]^T$ để thể hiện số lần mỗi

hạng mục tweet xảy ra trong các tư vấn được chấp nhận trong tập huấn luyện. N_k^i được định nghĩa như sau:

$$n_k^i = |\{t_j | t_j \in A(u_i) \text{ và } f(t_j) = c_k\}| \quad (2.7)$$

Với tất cả các tweet trong danh sách để tư vấn cho người dùng, chúng ta sẽ tiến hành xếp hạng chúng bằng số lần tương ứng mà hạng mục được chấp nhận bởi người dùng, thuật toán tiến hành xếp hạng tweet như sau.

Bảng 2.1 Thuật toán phân loại văn bản

Huấn luyện:

for all $u_i \in U$ **do**

    tính giá trị n_i^k theo biểu thức (3)

end for

Kiểm thử

for all $u_i \in U$ **do**

    for all $t_j \in R(u_i)$ **do**

        if $f(t_j) = c_k$ **then set** $s_j = n_k^i$

        end if

    end for

    Xếp hạng t_j thông qua s_j theo thứ tự giảm dần

```
    end for
```

Với thuật toán xếp hạng này, mỗi tweet sẽ được gán một giá trị, chính là giá trị xếp hạng của tweet t_j đối với người dùng u_i, giá trị s_j càng cao, mức độ quan tâm của u_i đối với t_j càng lớn, hay nói cách khác t_j sẽ dễ được chấp nhận bởi u_i hơn

2.2.2.2 Tính phổ biến của tweet

Tất cả chúng ta đều biết, mức độ nổi tiếng có sức thu hút đáng chú ý đối với hầu hết mọi người dùng. Qua đánh giá trực quan, những sản phẩm càng phổ biến sẽ càng được chấp nhận bởi người dùng hơn. Để đánh giá mức độ phổ biến của một tweet p_j, chúng ta sử dụng số lần mà tweet đó được chấp nhận bởi người dùng trong tập huấn luyện như sau:

$$p_j = \left|A(t_j)\right| \quad (2.8)$$

Chúng ta sẽ tiến hành xếp hạng các tweet trong danh sách sử dụng để tư vấn thông qua mức độ phổ biến của tweet đó, thông qua thuật toán sau:

Bảng 2.2 Thuật toán xếp hạng tweet dựa trên tính phổ biến của tweet

```
Huấn luyện:
```

```
for all t_i ∈ T do
    tính p_j = |A(t_j)| thông qua biểu thức (4)
end for
Kiểm thử:
for all u_i ∈ U do
    for all t_j ∈ R(u_i) do
        set s_j = p_j
     end for
     xếp hạng t_j thông qua s_j theo thứ tự giảm dần
 end for.
```

2.2.2.3 Mức độ chấp nhận của các followee

Khi chúng ta xem xét vào khả năng chấp nhận của người dùng Twitter đối với một tweet bị ảnh hưởng bởi những người mà anh ta đang follow, theo lý thuyết thông tin, càng nhiều người dùng là hàng xóm lân cận của người đó quan tâm đến một tweet, thì khả năng tweet đó được chấp nhận sẽ càng cao. Trong Twitter, nếu càng nhiều người dùng mà anh ta đang follow có quan tâm đến một tweet, nghĩa là item đó có thể anh ta sẽ quan tâm hơn so với các tweet khác. Chúng ta sử dụng $F(u_i)$ để biểu diễn cho tập các followee của người dùng u_i, số

lượng các followee của u_i chấp nhận một tweet t_j có thể được tính thông qua công thức 2.32

$$a_{ij} = |\{u_j | u_j \in F(u_i) \text{ và } u_j \in A(t_j)\}| \quad (2.9)$$

Biểu diễn dưới dạng thuật toán như sau

Bảng 2.3 Thuật toán xếp hạng tweet theo mức độ chấp nhận của các followee

```
for all u_i ∈ U do
    for all t_j ∈ R(u_i) do
        set s_j = a_ij được tính theo công thức (5)
    end for
    xếp hạng t_j thông qua s_j theo thứ tự giảm dần
end for.
```

Với mỗi người dùng u_i, các tweet t_j sẽ được sắp xếp dựa vào thuật toán xếp hạng tweet, những tweet có giá trị xếp hạng s_j cao sẽ được sử dụng để tư vấn cho người dùng.

2.2.2.4 Tính ngữ nghĩa của các từ khóa

Dữ liệu chứa các từ khóa được trích xuất từ tweet, retweet hoặc comment bởi mỗi người dùng trong tập huấn luyện. Từ khóa có dạng "kw1:wight1; kw2:wight2;….;

kwN:weightN". Nếu trọng số càng lớn, mức độ quan tâm của người dùng đối với những từ khóa đó càng nhiều. Mỗi từ khóa được mã hóa dưới dạng một số nguyên duy nhất, và các từ khóa của người dùng từ cùng một bộ từ vựng có dạng bản tin- từ khóa. Đặc biệt, bản tin- từ khóa chứa các từ khóa được trích xuất tương ứng từ hồ sơ người dùng, tổ chức hoặc nhóm của người dùng Twitter. Định dạng là một chuỗi "id1; id2; …..; idN". Sau đó, chúng ta sẽ cố gắng xác định các từ khóa ngữ nghĩa được trích xuất từ một người dùng và một tweet cho việc tư vấn.

Giả sử $W(k_m^i)$ thể hiện trọng số của từ khóa $k_m^i \in K(u_i) = \{k_1^i, k_2^i, \dots, k_W^i\}$ được trích xuất từ u_i và gọi $K(t_j)$ là tập các từ khóa được trích xuất từ u_i và t_j tương ứng. Thuật toán xếp hạng sẽ được tiến hành như sau:

Bảng 2.4 Xếp hạng tweet theo tính ngữ nghĩa của các từ khóa

```
for all u_i ∈ U do
    for all t_j ∈ R(u_i) do
        set s_j = max{W(k_m^i) | k_m^i ∈ K(t_j)}
    end for
```

```
      xếp hạng t_j theo s_j theo thứ tự giảm dần
  end for
```

Chiến thuật sẽ tiến hành xếp hạng các tweet dựa vào s_j theo thứ tự giảm dần, sử dụng những tweet có thứ hạng cao để tư vấn cho người dùng.

2.2.2.5 Sự tương tác người dùng

Dữ liệu chứa các thông tin về hoạt động của người dùng, ví dụ người dùng A retweet người dùng B 5 lần, mention đến B 3 lần và comment B 6 lần, chúng ta sẽ thể hiện dữ liệu dưới dạng "A B 3 5 6" trong dữ liệu về các tương tác của người dùng. Chúng ta có thể xây dựng một đồ thị quan hệ dựa trên đồ thị hai phía của người dùng và tweet đã tồn tại trọng số được xác định như sau

$$w_{ij} = \alpha r_{ij} + \beta m_{ij} + \gamma c_{ij} \quad (2.10)$$

Với r_{ij}, m_{ij}, c_{ij} là số lần người dùng retweet, mention và comment từ u_i đối với t_j và $\theta = [\alpha, \beta, \gamma]^T$ là vector trọng số được thiết lập trong ứng dụng

Thuật toán được mô tả như sau

Bảng 2.5 Xếp hạng tweet dựa trên sự tương tác người dùng

```
Huấn luyện:
for all u_i ∈ U do
    for all t_j ∈ T do
        tính w_ij theo công thức (6)
    end for
 end for
 Kiểm thử:
 for all u_i ∈ U do
     for all t_j in R(u_i) do
         set s_j = ∑_k w_kj , với u_k ∈ F(u_i)
     end for
     Xếp hạng t_j thông qua s_j theo thứ tự giảm dần.
 end for
```

Dựa vào tương tác người dùng, chiến thuật sẽ tiến hành xếp hạng các tweet theo thứ tự giảm dần, giá trị xếp hạng của mỗi tweet t_j chính là giá trị s_j tính được ở trên. Những tweet có thứ hạng cao sẽ được tư vấn cho người dùng.

CHƯƠNG 3. THỬ NGHIỆM VÀ ĐÁNH GIÁ

Sử dụng các thư viện có sẵn do Twitter cung cấp, thực hiện xây dựng bộ dữ liệu từ dữ liệu thực tế hiện có của Twitter cho một số lượng người dùng, sử dụng kỹ thuật xếp hạng cộng tác đã đề xuất ở chương 2 để đưa ra tư vấn về những tweet hữu ích cho một số người dùng nào đó.

3.1 Thu thập dữ liệu thử nghiệm

Để xây dựng hệ tư vấn người dùng Twitter, bước đầu tiên chúng ta cần thu thập dữ liệu, sau đó chia thành hai tập: tập huấn luyện và tập kiểm thử. Do sở thích người dùng có thể thay đổi theo thời gian, và luôn luôn biến động, thêm nữa, với số lượng người dùng vô cùng lớn, các tweet liên tục được cập nhật nên sẽ khó để có một tập dữ liệu có sẵn nào có thể đáp ứng được yêu cầu cho hệ tư vấn Twitter.

Twitter cung cấp một loạt các giao diện lập trình ứng dụng cho phép truy vấn các thông tin về người dùng sau khi được cấp quyền, các truy vấn này có dữ liệu trả về dưới dạng JSON [4]

Thông qua các API được cung cấp bởi Twitter, chúng ta sẽ tiến hành thu thập một dữ liệu đủ lớn để xây dựng tập huấn luyện, với mục đích thử nghiệm các thuật toán, chúng ta sẽ lấy

thông tin của 1000 người dùng trực tiếp từ Twitter API. Để có được thông tin của 10000 người dùng này, ban đầu chúng ta mở rộng từ 10 người dùng là những người dùng trong danh sách bạn bè. Sau đó mở rộng tập người dùng thông qua những người đang theo dõi và đang được theo dõi bởi những người dùng đã biết.

Toàn bộ dữ liệu sẽ được chia thành 2 tập là tập huấn luyện và tập kiểm thử, tập lớn hơn sẽ là tập huấn luyện, bao gồm 9000 người dùng, và tập nhỏ hơn sẽ là tập kiểm thử bao gồm 1000 người dùng. Bảng dưới đây là thông tin về số lượng người dùng và số lượng tweet, số lượng follower và followee trung bình trong 2 tập huấn luyện và kiểm thử.

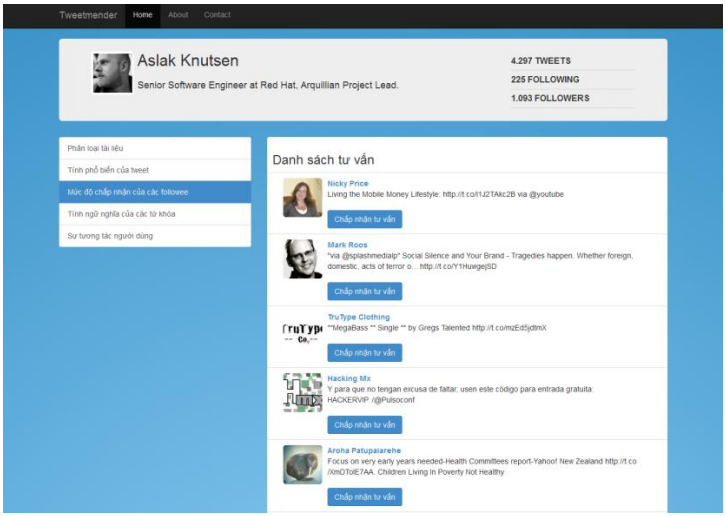
Bảng 3.1 Phân chia tập huấn luyện và tập kiểm thử

	Người dùng	**Tweet**	**Followers**	**Followee**
Huấn luyện	9000	72	124	225
Kiểm thử	1000	57	98	176

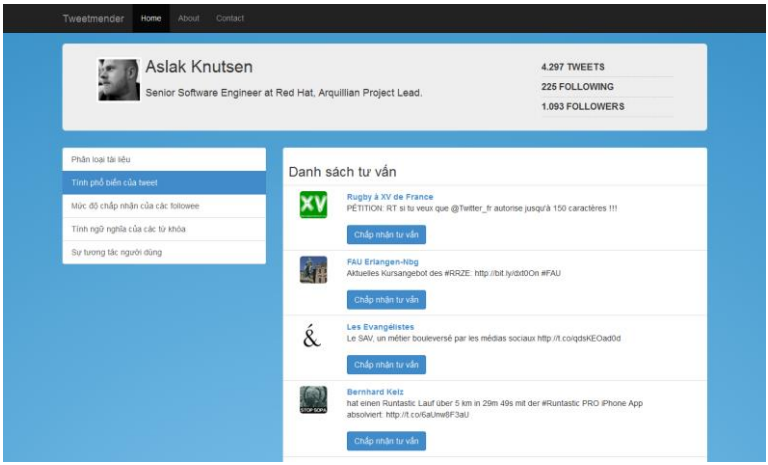
3.2 Ứng dụng mô phỏng thuật toán

Dựa vào dữ liệu đã thu thập được thông qua các API do Twitter cung cấp, chia dữ liệu thành hai tập huấn luyện và kiểm thử với lượng dữ liệu tương ứng là 9000 người dùng và 1000 người dùng. Sử dụng dữ liệu ngẫu nhiên trong tập kiểm thử để tiến hành đánh giá từng phương pháp tư vấn đã được đưa ra ở trên.

Với mỗi người dùng trong tập kiểm thử, lựa chọn một trong các thuật toán trên để đưa ra danh sách 20 tweet có xếp hạng cao nhất



Hình 3.1 Ứng dụng tư vấn người dùng Twitter dựa trên tập huấn luyện
với thuật toán mức độ chấp nhận Followee



Hình 3.2 Kết quả tư vấn dựa trên tính phổ biến của tweet

Ngoài ra, ứng dụng cho phép người dùng mới có thể được tư vấn trực tuyến bằng cách đăng nhập vào tài khoản Twitter và cấp quyền truy cập thông tin người dùng thông qua OAuth API, ứng dụng được triển khai và có khả năng truy cập tại địa chỉ http://tweetmender.herokuapp.com/. Với mỗi người dùng mới đăng nhập hệ thống, toàn bộ thông tin người dùng sẽ được sử dụng như một phần dữ liệu huấn luyện, người dùng sau khi được tư vấn sẽ có thể đưa ra đánh giá những tweet nào trong danh sách có đáp ứng đúng sở thích của người dùng hoặc không.

3.3 Đánh giá các phương pháp tư vấn

Dựa trên 5 thuật toán áp dụng cho việc tư vấn Twitter đã được để cập ở trên, việc đánh giá các phương pháp sẽ sử dụng tập kiểm thử bao gồm có 1000 người dùng đã thu thập được.

Để đánh giá mức độ chính xác của kết quả tư vấn cho người dùng, chúng ta sẽ sử dụng độ đo MAP (Mean Average Precision), tạm gọi là độ chính xác trung bình toàn cục. Giả sử chúng ta tư vấn m tweet trong danh sách xếp hạng cho một người dùng Twitter nào đó, có thể người dùng đó sẽ lựa chọn việc retweet hoặc đánh dấu thích tweet đó, độ chính xác trung bình có thể được sử dụng để thực hiện việc đánh giá này.

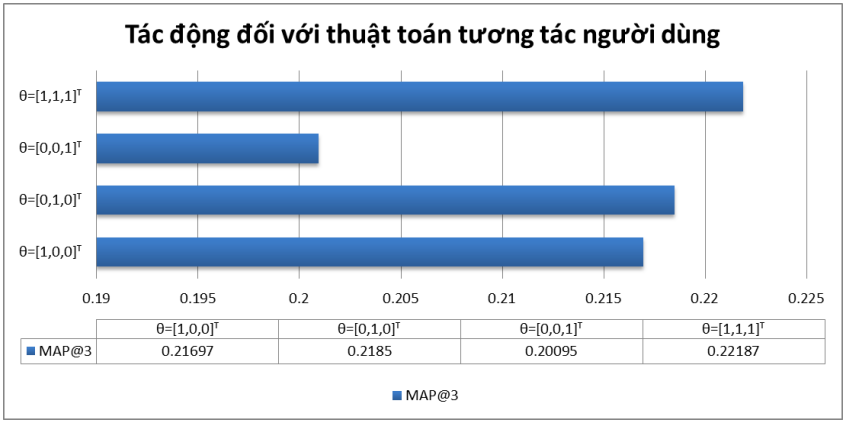
Giá trị độ chính xác trung bình (AP) ứng với truy vấn được xác định bởi công thức sau:

$$AP@n = \sum_{k=1}^{n} \frac{p(k) \times rel(k)}{c(m)} \qquad (3.1)$$

Trong đó, k là ngưỡng, $p(k)$ là hàm trả về độ chính xác tại k, $rel(k)$ là hàm nhị phân cho biết đây có phải là một kết quả đúng hay không, giá trị của $rel(k)$ thể hiện rằng tweet k có được quan tâm bởi người dùng hay không. $C(m)$ là số lượng

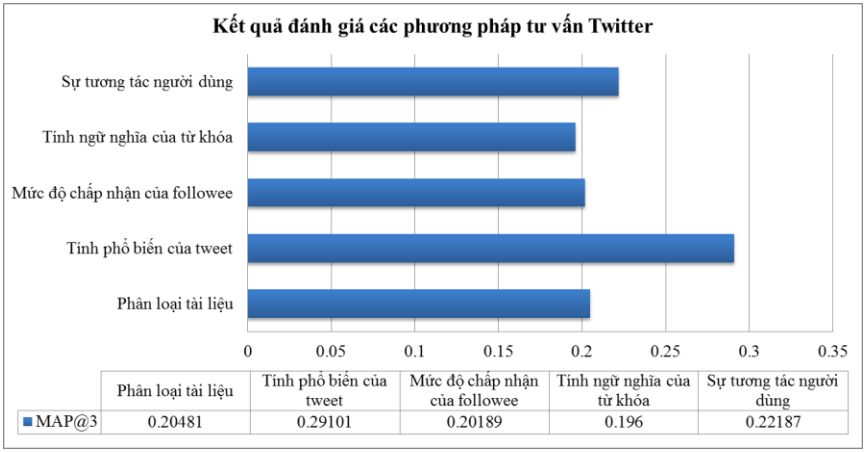
tweet mà người dùng có thể sẽ quan tâm trong danh sách m item đã được xếp hạng.

Sử dụng tập huấn luyện và kiểm thử có được trong phần 3.1, chúng ta tiến hành đánh giá các thuật toán tư vấn. Với thuật toán tư vấn dựa vào tổng hợp các tương tác người dùng, chúng ta cũng lần lượt sử dụng các giá trị $\theta = [\alpha, \beta, \gamma]^T$ là 4 trường hợp $\theta = [1,0,0]^T$, $\theta = [0,1,0]^T$, $\theta = [0,0,1]^T$ và $\theta = [1,1,1]^T$, trong đó hiệu năng với $\theta = [1,1,1]^T$ đạt cao nhất



Hình 3.3 Các tùy chọn đầu vào đối với thuật toán tương tác người dùng

Kết quả kiểm thử của các phương pháp tư vấn Twitter được cho trong bảng dưới đây



Hình 3.4 Kết quả đánh giá các phương pháp tư vấn Twitter.

Với kết quả này, chúng ta có thể thấy cả ba hành động: retweet, mention và comment đều có tác động tích cực đối với việc đưa ra các tweet mà người dùng quan tâm, tuy nhiên hành động comment không có ý nghĩa nhiều như retweet và mention. Trong cả năm phương pháp tư vấn này, phương pháp xếp hạng dựa trên tính phổ biến của tweet đạt kết quả cao nhất, điều đó giải thích rằng những tweet được đăng bởi người nổi tiếng có ý nghĩa và ảnh hưởng lớn đến hầu hết các người dùng Twitter khác. Phương pháp dựa trên ngữ nghĩa của từ khóa không đạt được kết quả cao, điều này hoàn toàn phù hợp với những mạng xã hội như Twitter do sự giới hạn 140 ký tự cho mỗi tweet được đăng.

KẾT LUẬN VÀ KIẾN NGHỊ

Bài toán tư vấn cho người dùng mạng xã hội càng ngày càng trở lên quan trọng do lượng thông tin được cung cấp thông qua mạng xã hội là vô cùng lớn, bản thân hầu hết các mạng xã hội phổ biến hiện nay đều đã tự đưa ra những giải pháp tư vấn, tuy nhiên chưa thực sự phát huy hiệu quả. Với những phương pháp đã được nghiên cứu, cài đặt, kiểm thử và đánh giá trong luận văn này cho thấy độ chính xác trong tư vấn đạt được khá tốt, do những phương pháp này dựa trên sự kết hợp cả nội dung lẫn tương tác người dùng trong mạng xã hội, và điển hình là Twitter.

Các phương pháp được tìm hiểu trong luận văn chủ yếu dựa trên quan điểm của kỹ thuật lọc cộng tác, những người dùng có sở thích gần giống nhau sẽ có những lựa chọn tương đồng. Để xác định sở thích của từng người dùng trong mạng xã hội Twitter, cần tiến hành thu thập dữ liệu và xây dựng hồ sơ người dùng, khác với các hệ tư vấn khác như Amazon hay MovieLens, đánh giá của người dùng trên từng sản phẩm là tường minh, kết quả thu thập được đối với người dùng Twitter là không tường minh, nghĩa là không có đánh giá cụ thể của người dùng trên từng tweet để xác định quan điểm người dùng đối với tweet đó, vì thế bước xây dựng hồ sơ người dùng nhằm xác định sở thích mỗi người dùng dựa trên toàn bộ dữ liệu liên quan đến người dùng đó bao gồm các tweet người dùng đã đăng, đang theo dõi ai và đang có những ai theo dõi, ngoài ra còn cần thông tin về sự tương tác của người dùng trong mạng

xã hội như phản hồi, đề cập đến người nào đó hoặc đánh dấu một tweet là ưa thích.

Từ những kết quả đạt được của các phương pháp tư vấn người dùng Twitter đã trình bày ở trên, có thể thấy mỗi phương pháp có những ưu nhược điểm khác nhau, trong tương lai cần tìm hiểu và thử nghiệm, đánh giá phương pháp kết hợp tất cả các phương pháp đã trình bày, nhằm đưa ra kết quả tư vấn gần với sở thích người dùng nhất. Đồng thời, do đặc thù của mạng xã hội, sở thích của người dùng có thể thay đổi theo thời gian, nên việc đánh giá chỉ dựa vào độ chính xác trung bình toàn cục không cho thấy hết được những tư vấn đưa ra có đúng với sở thích người dùng hay không, cần cho phép bản thân người dùng được tư vấn tự đánh giá trực tuyến, nếu một tư vấn là chính xác hoặc không. Kết quả đánh giá trực tuyến đó sẽ mang tính chính xác hơn đối với từng phương pháp tư vấn.