**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN 1**

****

**ĐỒ ÁN**

**TỐT NGHIỆP ĐẠI HỌC**

***Đề tài:***

**ƯỚC LƯỢNG ĐỘ TƯƠNG ĐỒNG CỦA NGƯỜI DÙNG DỰA VÀO HÀNH VI TRÊN MẠNG XÃ HỘI**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn** | **: TS. NGUYỄN MẠNH HÙNG** |
| **Sinh viên thực hiện** | **: TRẦN XUÂN TIẾN** |
| **Lớp** | **: D12CNPM4** |
| **Khoá** | **: 2012-2017** |
| **Hệ** | **: ĐẠI HỌC CHÍNH QUY** |

**Hà Nội - 2016**

**LỜI CẢM ƠN**

**MỞ ĐẦU**

Trong sự bùng nổ của công nghệ thông tin ngày nay, Internet và *mạng xã hội* đã và đang ngày càng phát triển và là xu hướng của thế giới. Internet đã và đang trở nên không thể thiếu đối với con người hiện đại, bởi nó có ảnh hưởng và mang lại rất nhiều tiện ích hữu dụng ở mọi lĩnh vực trong đời sống. Mạng Internet là môi trường mở cho phép người sử dụng được tự do cung cấp, tìm kiếm và sử dụng thông tin, có tính truyền tải nhanh, diện tham chiếu rộng, thông tin gần như tức thì và dễ tạo hiệu ứng xã hội theo chiều rộng, đó là tiền đề hình thành nên *mạng xã hội*. *Mạng xã hội* xuất hiện, nhanh chóng khẳng định vị trí phổ biến trên dịch vụ web và trở thành hoạt động được quan tâm nhất trên Internet hiện nay. Tính ưu trội của *mạng xã hội* tạo ra khả năng giao lưu, chia sẻ, kết nối cộng đồng thuận lợi nên thu hút được số lượng vô cùng lớn người sử dụng. Con người đã sử dụng *mạng xã hội* để chia sẻ và tìm kiếm thông tin, tìm kiếm bạn bè, hay tập hợp nhiều người thành nhóm có cùng chung mục đích, sở thích hoặc cùng giải quyết, phát triển vấn đề nào đó hoặc để tâm sự, giãi bày nhiều điều trong cuộc sống. Những dữ liệu phong phú, đa dạng được chia sẻ từ số lượng người sử dụng vô cùng lớn đó làm cho *mạng xã hội* trở thành một kho dữ liệu khổng lồ, cung cấp thông tin vô cùng giá trị cho việc nghiên cứu và phát triển. Chính vì vậy, việc trích xuất dữ liệu từ *mạng xã hội* để sử dụng cho việc phân tích, đánh giá về một vấn đề hay con người, một cách tự động đang ngày càng được quan tâm. Một loạt các bài toàn toán về phân tích hành vi, phân tích quan điểm, so sánh, ước lượng độ tương đồng của người dùng dựa trên nguồn tài nguyên đó như: tìm kiếm bạn bè có chung mục đích, sở thích, tìm một nửa còn lại phù hợp với mình, gợi ý xem phim, tư vấn mua sắm thông qua hành vi của những người có chung quan điểm, dự đoán tương lai thông qua hành vi, … được ra đời và phát triển.

Do đó, đồ án chọn tập trung nghiên cứu về đề tài “*ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội*”.

Phương hướng giải quyết của đồ án: Để so sánh được hai người dùng dựa trên hành vi trên *mạng xã hội* thì trước tiên đồ án sẽ thu thập khối dữ liệu của hai người dùng đó trên *mạng xã hội* họ sử dụng, sau đó sẽ thực hiện phân tích hành vi của họ dựa trên dữ liệu thu thập được và thực hiện so sánh. Việc phân tích sẽ tập trung vào các câu hỏi như: dữ liệu đó đề cập tới chủ đề gì? Hai khối dữ liệu có cùng đề cập tới một chủ đề hay không? Quan điểm được thể hiện trong dữ liệu đó như thế nào? Hai khối dữ liệu có chung quan điểm về cùng một vấn đề hay không? Đồ án sẽ tập trung tìm hiểu phương pháp so sánh nội dung văn bản, so sánh các cặp thuộc tính với nhau và đề xuất thành các công thức cụ thể.

Bố cục đồ án bao gồm 3 chương:

* Chương 1: Bài toán ước lượng độ tương đồng của người dùng dựa vào hành vi trên *mạng xã hội*
  + Giới thiệu *mạng xã hội*
  + Giới thiệu bài toán “ước lượng độ tương đồng của người dùng dựa vào hành vi trên *mạng xã hội*
  + Phương pháp tiếp cận bài toán của đồ án
* Chương 2: Thuật toán ước lượng độ tương đồng của người dùng dựa vào hành vi trên *mạng xã hội*
  + Mô hình hoá bài đăng, hành vi của người dùng trên *mạng xã hội*
  + Ước lượng độ tương đồng bài đăng, hành vi, người dùng trên *mạng xã hội*
* Chương 3: Đánh giá thuật toán ước lượng độ tương đồng của người dùng dựa vào hành vi trên *mạng xã hội*
  + Đánh giá: bộ dữ liệu thử nghiệm, bộ trọng số, kịch bản, kết quả đánh giá
  + Ứng dụng: mô tả, kiến trúc tổng qua, đánh giá kết quả

**CHƯƠNG 1: BÀI TOÁN ƯỚC LƯỢNG ĐỘ TƯƠNG ĐỒNG CỦA NGƯỜI DÙNG DỰA VÀO HÀNH VI TRÊN MẠNG XÃ HỘI**

Chương 1 sẽ trình bày các nội dung sau:

* Giới thiệu về *mạng xã hội*
* Giới thiệu bài toán “ước lượng độ tương đồng của người dùng dựa vào hành vi trên *mạng xã hội*
* Phương pháp tiếp cận bài toán của đồ án
  1. **Giới thiệu về mạng xã hội**
* **Mạng xã hội:**



Hình 1.1: Mạng xã hội (Ảnh minh hoạ)

*Mạng xã hội (social network)* hay gọi là *mạng xã hội ảo*, tên đầy đủ là *mạng xã hội trực tuyến (online social network)*.

Một *mạng xã hội* là một xã hội ảo hình thành trên nền mạng Internet, được tạo ra bởi nhiều cá nhân (hoặc tổ chức) liên kết lại với nhau bở một hoặc nhiều loại mối quan hệ, mục đích, giao tiếp như: bạn bè, họ hàng, lợi ích chung, trao đổi tài chính, kiến thức, mua bán, …

* **Lịch sử phát triền**

*Mạng xã hội* xuất hiện lần đầu tiên năng 1995 với sự ra đời của trang Classmate với mục đích kết nối bạn học. Tiếp theo là sự xuất hiện của SixDegrees vào năm 1997 với mục đích giao lưu kết bạn dựa theo sở thích. Năm 2002, Friendster trở thành một trào lưu mới tại Hoa Kỳ với hàng triệu thành viên ghi danh. Năm 2004, MySpace ra đời với các tính năng như phim ảnh và nhanh chóng thu hút hàng chục ngàn thành viên mới mỗi ngày, các thành viên cũ của Friendster cũng chuyển qua MySpace và trong vòng một năm, MySpace trở thành mạng xã hội đầu tiên có nhiều lượt xem hơn vả Google và được tập đoàn News Corpration mua lại với giá 580 triệu USD. Năm 2006, sự ra đời của Facebook đánh dấu bước ngoặt mới cho hệ thống mạng xã hội trực tuyến với nền tảng lập trình (Facebook Platform) cho phép thành viên tạo ra những ứng dụng (applications) mới cho cá nhân mình cũng như các thành viên khác dùng.

* **Cấu trúc chung của mạng xã hội**



Hình 1.2: Mô hình cấu trúc mạng xã hội (Ảnh minh hoạ)

*Mạng xã hội* có cấu trúc như một đồ thị dạng lưới, bao gồm các nút - chính là các cá nhân (hoặc tổ chức) – gọi là thành viên, và các mối liên kết giữa các nút – thành viên đó. Việc tương tác giữa một nút – thành viên với một hoặc nhiều nút – thành viên trong mạng lưới là thông qua mối liên kết giữa các nút – thành viên đó. Một *mạng xã hội* tạo ra thì tự nó sẽ nhân rộng trong cộng đồng, mở rộng số lượng thành viên và mối liên kết, thông qua các tương tác của thành viên trong chính cộng đồng đó.

Một trang *mạng xã hội* hoạt động như một tâm điểm kết nối, giúp các thành viên trong mạng thiết lập mối quan hệ với một hoặc nhiều thành viên khác trong mạng. Một trang *mạng xã hội* cho phép thành viên:

+ Tạo hồ sơ cá nhân trực tuyến và quản lý thông tin cá nhân, thiết lập quyền kiểm soát

+ Tạo danh sách bạn bè, kết nối bạn bè

+ Nhận được các thông báo, tin tức từ các kết nối của họ

+ Thực hiện các hành động như: đăng bài, thích bài đăng, bình luận bài đăng, tham gia các nhóm yêu thích, email, hội thoại, xem ảnh, video, chia sẻ file, …

Trong *mạng xã hội* tồn tại hai kiểu liên kết:

+ *Liên kết trực tiếp*: mối liên kết trực tiếp giữa hai thành viên với nhau, tức là hai thành viên này có mối quan hệ như: là bạn, đồng nghiệp, có cùng chung sở thích, cùng mối quan tâm, lợi ích, …

+ *Liên kết gián tiếp*: mối liên kết giữa hai thành viên thông qua một hoặc nhiều thành viên trung gian khác, ví dụ như mối quan hệ bắc cầu: *X* liên kết trực tiếp với *Y*, *Y* liên kết trực tiếp với *Z*, *X* và *Z* không có liên kết trực tiếp nhưng có liên kết gián tiếp thông qua *Y* và có ít nhất một điểm chung là *Y*.

Như vậy, một thành viên trong *mạng xã hội* có kiểu liên kết nào cũng sẽ tồn tại ít nhất một đặc điểm chung nào đó với các thành viên khác cùng nằm trong liên kết của họ trong mạng lưới.

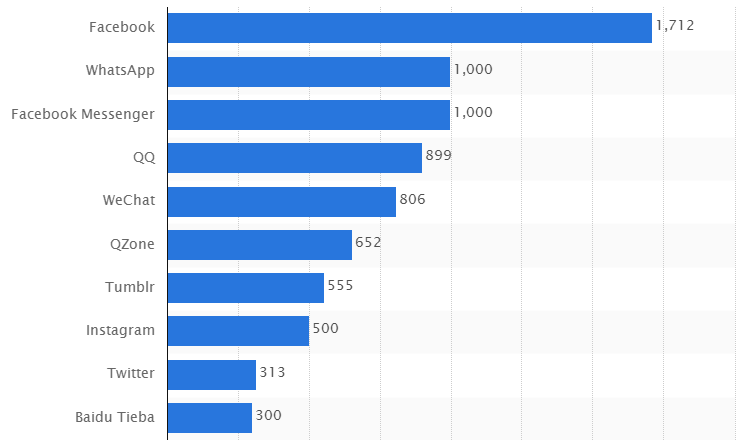
* **Đặc trưng của mạng xã hội**

Điểm nổi bật của *mạng xã hội* là tính kết nối và khả năng chia sẻ rất mạnh mẽ qua mạng Internet không phân biệt không gian và thời gian, phá vỡ ngăn cách về địa lý, ngôn ngữ, giới tính lẫn quốc gia.

*Mạng xã hội* nổi bật trong vai trò truyền thông trong môi trường Internet với khả năng cập nhật thông tin nhanh và lan truyền trên diện rộng.

* **Một số mạng xã hội phổ biến**

Trên thế giới hiện nay, số lượng các trang *mạng xã hội* là không xác định bởi sức mạnh và sự ảnh hưởng của nó. Dự báo số lượng người dùng *mạng xã hội* sẽ tiếp tục tăng, *mạng xã hội* ngày càng trở nên thân thiện và phổ cập với mọi người dùng. Dưới đây là thông kê số lượng người dùng của 10 trang *mạng xã hội* phổ biến nhất bao gồm: Facebook, WhatsApp, QQ, Facebook Messenger, Qzone, WeChat, Tumblr, Instagram, Twitter, Baidu Tieba.



Hình 1.3: Biểu đồ thống kê người dùng của top 10 mạng xã hội phổ biến tính đến 9/2016   
(Đơn vị: Triệu người dùng)

* **Mạng xã hội Twitter:**

Twitter là *mạng xã hội* cho phép người dùng có thể tải hình ảnh lên, đọc, viết và cập nhật các *mẩu tin* (được gọi là *tweet*) có độ dài giới hạn 140 ký tự.

Hình 1.4: Giao diện Twitter

Twitter được sáng lập năm 2006 bởi đồng sáng lập Jack Dorsey, Evan Willams và Biz Stone, được quản lý bởi công ty Twitter Inc., có website chính thức là <https://twitter.com/>

Tính năng của người dùng trên Twitter: để sử dụng được, người dùng phải đăng ký một tài khoản thành viên bằng email cá nhân. Người dùng có thể:

+ Quản lý thông tin cá nhân

+ Viết hay đăng trạng thái trên trang cá nhân

+ Theo dõi tài khoản thành viên khác

+ Chia sẻ bài đăng của thành viên khác, có thể kèm theo lời bình

+ Thích, nhận xét/bình luận bài đăng của các thành viên

+ Gửi tin nhắn tới những thành viên đang theo dõi

Các cách để theo dấu thông tin trên Twitter:

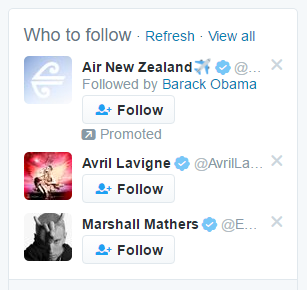
+ Theo dõi theo “*hashtag*” – một dạng như từ khoá. Ví dụ: *#Rio2016*

+ Theo dõi theo Url cụ thể

+ Theo dõi theo các *retweet* – *bài đăng được chia sẻ*

Để trích xuất được dữ liệu, Twitter cũng cung cấp REST API cho tương tác, đọc ghi sử dụng cơ chế xác nhận bảo mật OAuth.

* 1. **Giới thiệu bài toán**

*Mạng xã hội* đã và đang ngày càng phát triển, thể hiện bởi các tính năng của các trang *mạng xã hội* và số lượng thành viên tham gia. Là *mạng xã hội* được coi là lớn nhất thời điểm hiện tại, Facebook có tới 1712 triệu thành viên, theo sau là WhatsApp với 1000 triệu thành viên, số lượng thành viên thấp hơn khoảng 3 lần là Twitter với 313 triệu thành viên. Để đáp ứng được nhu cầu sử dụng, tạo điều kiện thuận lợi khi sử dụng của số lượng các thành viên lớn như vậy, các hệ thống *mạng xã hội* phải luôn cập nhật cũng như đưa ra các tính năng mới đem lại hiệu quả, lợi ích tốt hơn, dựa trên chính kho dữ liệu khổng lồ tạo bởi số lượng lớn người sử dụng đó. Có thể thấy một số tính năng như: gợi ý kết bạn, gợi ý nhóm – cộng đồng, quảng cáo, tư vấn mua sắm online, … Các tính năng này đều liên quan tới sự tương đồng giữa các người dùng trong các mối liên kết của họ trên mạng xã hội.

Hình 1.5: Tính năng gợi ý tham gia nhóm – cộng đồng trên Facebook

Hình 1.6: Tính năng gợi ý theo dõi trên Twitter

* **Sự tương đồng giữa người dùng trên mạng xã hội**

*Mạng xã hội* cung cấp nhiều tính năng mới nhằm nâng cao trải nghiệm của người sử dụng. Khi tham gia *mạng xã hội*, để sử dụng các tính năng này, người dùng phải là thành viên của các trang *mạng xã hội* bằng cách tạo tài khoản cá nhân. Người dùng sẽ chấp nhận chia sẻ thông tin cá nhân ở các mức độ giới hạn khác nhau tuỳ theo chính sách của từng trang *mạng xã hội*. Các thông tin cá nhân như: tên, tuổi, giới tính, nơi ở, học tập, làm việc, sở thích, album ảnh, video, danh sách bạn bè, … Khi đã là thành viên, người dùng có thể thực hiện rất nhiều các hành động trên trang cá nhân, trang bạn bè, trang hội nhóm – cộng đồng đã tham gia như đăng bài, chia sẻ bài đăng, bình luận, tham gia nhóm – cộng đồng, theo dõi, kết bạn, …, sử dụng các tính năng *mạng xã hội* cung cấp như xem phim, ảnh, gửi tin nhắn, trò chuyện hội nhóm, … Như vậy có thể rút ra được, một người dùng trên *mạng xã hội* sẽ có hai đặc trưng chính là: các thuộc tính cá nhân và các hành vi. Việc nghiên cứu, đánh giá, nhận xét, so sánh giữa người dùng với nhau sẽ dựa trên các thuộc tính cá nhân và hành vi của họ. Đó chính là cơ sở hình thành của các loại bài toán về sự tương đồng giữa người dùng trên *mạng xã hội* như:

+ Loại 1: Sự tương đồng của người dùng dựa vào thuộc tính cá nhân

+ Loại 2: Sự tương đồng của người dùng dựa vào hành vi.

Trong phạm vi đồ án này sẽ tập trung nghiên cứu bài toán thuộc loại 2, bài toán *ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội*.

* **Bài toán ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội**

Các hành vi của người dùng trên *mạng xã hội* là các hoạt động, tương tác hàng ngày theo các mối liên kết của người dùng như:

+ *Đăng bài – “Post”* lên trang cá nhân, lên các hội nhóm – cộng đồng đã tham gia, bài đăng có thể là hình ảnh, video hay bất kỳ trạng thái nào của người dùng.

+ *Chia sẻ bài đăng – “Share”* lên trang cá nhân, lên các hội nhóm – cộng đồng đã tham gia.

+ *Nhận xét/Bình luận – “Comment”* về một bài đăng nào đó hay đưa ra quan điểm, nhận xét về một bài đăng nào đó.

+ *Thích – “Like”*, *“Favorite”* một bài đăng hoặc một bình luận nào đó hay đưa ra phản hồi tích cực về bài đăng hoặc bình luận nào đó.

+ *Tham gia – “Join”* một hội nhóm – cộng đồng nào đó.

+ *Theo dõi – “Follow”* một trang cá nhân của thành viên nào đó.

+ *Kết bạn – “Add Friend”* với một thành viên nào đó.

+ *Tạo* hoặc *tham gia* các *sự kiện* – *“Event”* nào đó.

Tất cả các tương tác trên được gọi là hành vi của người dùng trên *mạng xã hội*. Do đó, bài toán ước lượng độ tương đồng của người dùng dựa vào hành vi trên *mạng xã hội* chính là ước lượng độ tương đồng dựa trên từng hành vi của người dùng. Độ tương đồng của hai người dùng dựa vào hành vi được ước lượng dựa theo độ tương đồng của các bài đăng đã *post*, các bài đăng đã *like*, các bài đăng đã *comment/like*, các bài đăng đã *share*, các hội nhóm – cộng đồng đã *join*, các người dùng cùng *follow*, các sự kiện cùng *tham gia*. Càng nhiều các hành vi tương đồng, độ tương đồng theo từng hành vi càng cao thì độ tương đồng của người dùng dựa trên hành vi càng cao và ngược lại, độ tương đồng theo từng hành vi càng thấp thì độ tương đồng của người dùng dựa trên hành vi càng thấp. Dựa trên cơ sở này, đồ án sẽ đưa ra phương pháp tiếp cận bài toán.

* 1. **Phương pháp tiếp cận bài toán của đồ án**

Ước lượng độ tương đồng của người dùng dựa vào hành vi trên *mạng xã hội* là đánh giá độ tương đồng giữa các khối dữ liệu tạo ra bởi các hành vi của người dùng trên *mạng xã hội* họ tham gia. Tiêu chí tiếp cận đánh giá độ tương đồng của đồ án sẽ sử dụng 4 hành vi phổ biến trong số các hành vi của người dùng trên mạng xã hội bao gồm: *đăng bài – post*, *thích bài đăng – like*, *nhận xét hay bình luận bài đăng – comment*, *tham gia hội nhóm – join* để mô hình hoá tổng quát bài toán. Với mỗi bài đăng tạo ra bởi các hành vi của người dùng như: *đăng*, *thích/không thích*, *nhận xét/bình luận* đều thể hiện ít nhất một vấn đề, quan điểm về vấn đề nào đó, nên bài đăng sẽ được mô hình hoá và ước lượng độ tương đồng theo các yếu tố chủ đề và quan điểm và có trọng số phụ thuộc dữ liệu, tương tác thực tế. Việc ước lượng độ tương đồng của các bài đăng như vậy của hai người dùng là một trong quá trình ước lượng độ tương đồng giữa hai người dùng đó. Với mỗi hội nhóm – cộng đồng mà người dùng tham gia đều có: tên, mô tả, chủ đề nào đó, nên hội nhóm – cộng đồng sẽ được mô hình hoá và ước lượng độ tương đồng theo các yếu tố tên, mô tả, chủ đề và có trọng số phụ thuộc dữ liệu, tương tác thực tế. Việc ước lượng độ tương đồng của các nhóm như vậy của hai người dùng cũng là một trong qua trình ước lượng độ tương đồng giữa hai người dùng đó.

Ví dụ: quá trình ước lượng độ tương đồng giữa hai người dùng *X* và *Y* cùng tham gia một *mạng xã hội* dựa trên các hành vi của họ lần lượt theo thứ tự: *đăng bài*, *thích/không thích bài đăng*, *nhận xét/bình luận bài đăng*, *tham gia nhóm* là *PX*, *LX*, *CX*, *JX* và *PY*, *LY*, *CY*, *JY* là quá trình đánh giá độ tương đồng giữa các tập bài đăng tương tác bởi *X* và *Y* là: *PX* và *PY*, *LX* và *LY*, *CX* và *CY*, *JX* và *JY*.

Các vấn đề sẽ được trình bày chi tiết hơn trong chương 2 của đồ án.

* 1. **Kết luận**

Trong chương 1, đồ án đã giới thiệu về *mạng xã hội*, bài toán ước lượng độ tương đồng của người dùng dựa vào hành vi trên *mạng xã hội*, phương pháp tiếp cận bài toán sử dụng trong đồ án.

Trong chương 2, đồ án sẽ trình bày *thuật toán ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội*.

**CHƯƠNG 2: THUẬT TOÁN ƯỚC LƯỢNG ĐỘ TƯƠNG ĐỒNG CỦA NGƯỜI DÙNG DỰA VÀO HÀNH VI TRÊN MẠNG XÃ HỘI**

Chương 2 sẽ trình bày các nội dung sau:

* Mô hình hoá bài đăng, hành vi của người dùng trên *mạng xã hội*
* Ước lượng độ tương đồng bài đăng, hành vi, người dùng trên *mạng xã hội*
  1. **Mô hình hoá bài đăng, hành vi của người dùng trên mạng xã hội**

Một *mạng xã hội* bao gồm tập các người dùng và tập các nội dung được chia sẻ thông qua các mối liên kết. Các nội dung được tạo ra và nhận sự tương tác từ nhiều hành vi của người dùng. Mô hình hoá các nội dung, hành vi của người dùng sẽ dựa trên các đặc trưng của các nội dung, hành vi của người dùng đó. Để cụ thể hoá, xây dựng các công thức của thuật toán, đồ án định nghĩa và giả định dưới đây.

* + 1. **Bài đăng**

Một người dùng trên mạng xã hội gọi là một *user*. Một *user* có thể đăng các trạng thái, ảnh, video lên trang cá nhân hoặc đăng trên các hội nhóm đã tham gia. Trạng thái, ảnh hay video đều là các nội dung được người dùng đăng lên, gọi là *bài đăng (entry)*.

Một *bài đăng* có thể được xem bởi nhiều người dùng và một người dùng cũng có thể xem nhiều *bài đăng* khác nhau.

Một *bài đăng* có thể thể hiện một hoặc nhiều vấn đề mà người dùng đang quan tâm như thị trường chứng khoán, tình hình biển Đông, … Một *bài đăng* có thể thể hiện tâm trạng vui sướng khi gặp gỡ bạn bè hay sự thất vọng trước tỷ số của một trận đấu bóng đá, … Một *bài đăng* có thể thể hiện sở thích, quan điểm nào đó của người dùng hoặc cũng có thể không thể hiện một vấn đề hay quan điểm nào cả. Như vậy, một *bài đăng* sẽ được mô tả bởi *chủ đề* và *quan điểm* của *bài đăng*, hay một *bài đăng* được đặc trưng bởi hai thuộc tính là *chủ đề* và *quan điểm* của *bài đăng* đó. Trong mô hình của đồ án sẽ xem xét một *bài đăng* có hai thuộc tính này.

* + 1. **Hành vi của người dùng**

Một trong các hành vi cơ bản của một người dùng trên mạng xã hội là *đăng bài (post)* lên trang cá nhân, lên các trang, hội nhóm.

Đối với một *bài đăng*, một người dùng có thể thực hiện các hành vi như:

+ *Thích (like) bài đăng*

+ Đưa ra *nhận xét/bình luận* *(comment)* cho *bài đăng*

+ *Chia sẻ (share)* *bài đăng* đó

Mỗi người dùng có thể *thích*, *nhận xét* mỗi *bài đăng*, *thích một tập các nhận xét* của một *bài đăng*.

Một người dùng có thể *thích các trang (like page)*, *tham gia các hội nhóm (join group)*, có thể *đăng bài đăng*, *thích hay nhận xét các bài đăng* trong các hội nhóm đã tham gia. Khi đó, người dùng được gọi là thành viên của *cộng đồng trên mạng xã hội (community)*.

Một người dùng có thể tạo danh sách bạn bè. Danh sách bạn bè là tập hợp các *user* khác có quan hệ bạn bè trên *mạng xã hội*.

Trong mô hình của đồ án sẽ xem xét trên 4 hành vi phổ biến là: *đăng bài*, *thích bài*, *nhận xét/thích nhận xét* và *tham gia hội nhóm/cộng đồng*. Tiếp theo đồ án sẽ trình bài tới ước lượng độ tương đồng *bài đăng*, hành vi, người dùng trên *mạng xã hội*.

* 1. **Ước lượng độ tương đồng bài đăng, hành vi, người dùng trên mạng xã hội**
     1. **Ước lượng độ tương đồng giữa các bài đăng**

Một *bài đăng* được mô tả bởi các thuộc tính xác định nên việc ước lượng độ tương đồng giữa các *bài đăng* được xem như là ước lượng độ tương đồng giữa các đối tượng dựa trên độ tương đồng giữa các thuộc tính của các đối tượng đó. Trong mô hình của đồ án sẽ xem xét bài đăng có 2 thuộc tính sau:

+ *Chủ đề của bài đăng (category)*: Một *bài đăng* có thể có: không, một, hoặc nhiều hơn một chủ đề. Với mỗi chủ đề là một tập các hợp ngữ. Vì vậy, ước lượng độ tương đồng về chủ đề của *bài đăng* chính là ước lượng độ tương đồng giữa hai tập các hợp ngữ. Trong mô hình này chỉ xem xét tới các *bài đăng* có ít nhất một chủ đề phù hợp, những *bài đăng* không thuộc chủ đề nào sẽ xoá bỏ.

+ *Quan điểm của một bài đăng (sentiment)*: Qua việc đăng bài, người dùng có thể cho thấy tình cảm, cảm xúc của họ thuộc 3 dạng: tích cực, tiêu cực hoặc trung tính. Do vậy, ước lượng độ tương đồng về *quan điểm* của *bài đăng* sẽ dựa vào các giá trị *quan điểm* đó. Trong mô hình của đồ án sẽ xem xét *quan điểm* của *bài đăng* thuộc *tập các quan điểm* gồm có: joy (vui sướng), sadness (buồn bã), happyfor (vui cho/vì ai đó), sorryfor (có lỗi với ai đó), sorry (có lỗi), hope (hi vọng), fear (sợ hãi), confused (bối rối), love (yêu), disgust (thù ghét), regret (hối tiếc), disappointed (thất vọng), pride (tự hào), anger (phẫn nộ), gratitude (biết ơn), admiration (khâm phục), và ước lượng độ tương đồng về *quan điểm* của *bài đăng* dựa vào các giá trị *quan điểm* này.

Mỗi thuộc tính *chủ đề* hay *quan điểm* của *bài đăng* đều được mô tả bởi một tập các hợp ngữ. Vì vậy, việc ước lượng độ tương đồng giữa hai *bài đăng* chính là ước lượng độ tương đồng giữa hai tập hợp ngữ.

* **Ước lượng độ tương đồng giữa hai tập các hợp ngữ**

Giả sử *A1 = {a11, a12, …, a1m}*, *A2 = {a21, a22, …, a2n}* là hai tập các hợp ngữ. Với *m*, *n* lần lượt là kích thước của *A1*, *A2*. Gọi k là kích thước của giao hai tập hợp *A1*, *A2*. Theo phương pháp tính toán độ tương đồng giữa các đối tượng được đề xuất bởi H. M. Nguyen và H. T. Nguyen, độ tương đồng của *A1*, *A2* được ước lượng theo công thức sau:

Ví dụ: Xét hai tập các hợp ngữ như sau:

*A1 = {Sport, Music, History}*, kích thước *m* = 3.

*A2 = {Culinary, Sport, Literary}*, kích thước *n* = 3.

Tập giao của hai tập các hợp ngữ trên là *{Sport}*, kích thước *k* = 1. Do đó độ tương đồng giữa hai tập các hợp ngữ trên là: (2\*1)/(3+3) = 1/3 (~33.33%) nằm trong miền giá trị [0,1].

Áp dụng công thức ước lượng độ tương đồng giữa hai tập các hợp ngữ cho việc ước lượng độ tương đồng giữa hai *bài đăng*.

* **Ước lượng độ tương đồng giữa hai bài đăng**

Mỗi *bài đăng* đặc trưng bởi hai thuộc tính *chủ đề* và *quan điểm* nên ước lượng độ tương đồng giữa các *bài đăng* sẽ dựa trên ước lượng độ tương đồng giữa các thuộc tính *chủ đề* và *quan điểm* của *bài đăng*.

Gọi *ei*, *ej* là hai *bài đăng* có *chủ đề* và *quan điểm* của *bài đăng* lần lượt là *Categoryi*, *Categoryj* và *Sentimenti*, *Sentimentj*. Độ tương đồng giữa hai *bài đăng* *ei*, *ej* được ước lượng như sau:

+ Ước lượng độ tương đồng *chủ đề* của hai *bài đăng* là ước lượng độ tương đồng giữa hai tập các hợp ngữ:

+ Ước lượng độ tương đồng *quan điểm* của hai *bài đăng* theo 3 dạng: tích cực, tiêu cực, trung tính như sau:

+ Ước lượng độ tương đồng *quan điểm* của hai *bài đăng* theo tập các quan điểm như sau:

+ Ước lượng độ tương đồng giữa hai *bài đăng ei*, *ej*:

Với *fe*: [0.1]2->[0,1] là một tương đồng phương thức giữa hai *bài đăng*. Theo điều kiện sau:

+ Trong mô hình của đồ án sẽ sử dụng: .

* **Ước lượng độ tương đồng giữa hai tập các bài đăng**

Xét *E1 = {e11, e12, …, e1m}*, *E2 = {e21, e22, …, e2n}* là hai tập các *bài đăng*. Đồ án sẽ tạo một *tập các* *bài đăng* bao gồm cả hai *tập các* *bài đăng* là *E12* và vector *T* như sau:

*E12 = E1 + E2 = {e1, e2, …, em+n}*, *T = {t1, t2, …, tm+n}*. Trong đó:

Với *Sentry* là độ tương đồng giữa hai *bài đăng* *x* và *y*.

Để ước lượng độ tương đồng giữa hai *tập các* *bài đăng E1* và *E2*, đồ án sử dụng các giả thiết sau:

+ Độ lớn của vector *T* càng lớn thì độ tương đồng giữa *E1* và *E2* càng cao

+ Ước lượng độ tương đồng giữa hai *tập các* *bài đăng E1* và *E2* theo công thức:

Với *fs*: [0,1]k->[0,1] là một tương đồng phương thức giữa hai tập hợp, thoả mãn các điều kiện sau:

Ví dụ: các công thức sau đây là công thức ước lượng độ tương đồng giữa hai *tập các* *bài đăng*:

Trong thử nghiệm đồ án sử dụng công thức .

Ví dụ:

Tiếp theo đồ án sẽ đi vào *ước lượng độ tương đồng giữa các cộng đồng trên mạng xã hội*.

* + 1. **Ước lượng độ tương đồng giữa các cộng đồng**

Một người dùng được gọi là thành viên của *cộng đồng* khi thích trạng hoặc tham gia hội nhóm. Một *cộng đồng* cũng được đặc trưng bởi các thuộc tính của nó. Trong mô hình của đồ án sẽ xem xét các thuộc tính của *cộng đồng* gồm có:

+ *Tên của cộng đồng (name)*: Tên có thể là một câu tóm tắt ngắn. Sau khi loại bỏ các từ tối nghĩa trong tiêu đề, thuộc tính này sẽ trở thành một tập các từ sử dụng để so sánh với các *cộng đồng* khác. Vì vậy, ước lượng độ tương đồng tên của hai *cộng đồng* chính là ước lượng độ tương đồng giữa hai tập các hợp ngữ.

+ *Chủ đề của cộng đồng (category)*: Trên một số *mạng xã hội*, một *cộng đồng* luôn thuộc vào một chủ đề nào đó. Mỗi chủ đề là một từ độc lập (hoặc một tập các từ). Vì vậy, ước lượng độ tương đồng chủ đề của hai *cộng đồng* chính là ước lượng độ tương đồng giữa hai tập các hợp ngữ.

+ *Mô tả của cộng đồng (description)*: Trên nhiều *mạng xã hội*, mỗi *cộng đồng* đều cung cấp một đoạn mô tả ngắn. Mô tả này thường là một văn bản ngắn. Sau khi xoá bỏ các từ tối nghĩa trọng mô tả, thuộc tính này sẽ trở thành một tập các từ được sử dụng để so sánh với các *cộng đồng* khác. Vì vậy, ước lượng độ tương đồng mô tả của hai *cộng đồng* chính là ước lượng độ tương đồng giữa hai tập các hợp ngữ.

* **Độ tương đồng giữa hai cộng đồng**

Độ tương đồng giữa hai *cộng đồng* sẽ được ước lượng bởi tập hợp trung bình các trọng số giống nhau giữa chúng dựa vào các thuộc tính trên.

Gọi *w1*, *w2*, *w3* lần lượt là trọng số của các thuộc tính *tên của cộng đồng (name)*, *mô tả của cộng đồng (description)*, *chủ đề của cộng đồng (category)* và được xác định theo các điều kiện sau:

Xét hai *cộng đồng ci* và *cj* có *tên*, *mô tả*, *chủ đề* lần lượt là *Ni*, *Di*, *Ci* và *Nj*, *Dj*, *Cj*. Độ tương đồng *Scom (ci, cj)* được ước lượng như sau:

Với *Sexp (A, B)* là độ tương đồng giữa hai tập các hợp ngữ *A* và *B*. Các trọng số *w1*, *w2*, *w3* sẽ là khác nhau với mỗi bộ dữ liệu khác nhau. Đồ án sẽ trình bày cách chọn trọng số chi tiết ở chương tiếp theo.

Độ tương đồng *Scom (ci, cj)* nằm trong khoảng [0,1]. Nếu *Scom (ci, cj)* càng gần tới 1 thì độ tương đồng giữa hai *cộng đồng* càng cao và ngược lại, *Scom (ci, cj)* càng gần tới 0 thì độ tương đồng giữa hai *cộng đồng* càng thấp.

* **Độ tương đồng giữa hai tập cộng đồng**

Xét *C1 = {c11, c12, …, c1m}*, *C2 = {c21, c22, …, c2n}* là hai *tập các cộng đồng*. Đồ án sẽ tạo một *tập các* *cộng đồng* bao gồm cả hai *tập các cộng đồng* là *C12* và vector *T* như sau:

*C12 = C1 + C2 = {c1, c2, …, cm+n}*, *T = {t1, t2, …, tm+n}*. Trong đó:

Với *Scom (x, y)* là độ tương đồng giữa hai *cộng đồng* *x* và *y*.

Độ tương đồng giữa hai *tập các* *cộng đồng C1* và *C2* được ước lượng theo công thức sau:

Trong thử nghiệm đồ án sử dụng công thức: .

Tiếp theo đồ án sẽ trình bày ước lượng độ tương đồng từng hành vi của người dùng trên *mạng xã hội* gồm: *đăng bài đăng*, *thích bài đăng*, *nhận xét/thích nhận xét bài đăng*, *tham gia hội nhóm/cộng đồng*.

* + 1. **Ước lượng độ tương đồng hành vi đăng bài đăng**

Gọi *Sip* và *Sjp* lần lượt là *tập các* *bài đăng* đã được đăng bởi người dùng *i* và người dùng *j*. Độ tương đồng dựa trên hành vi đăng bài đăng của người dùng *i* và người dùng *j* là độ tương đồng giữa *tập các* *bài đăng* và được ước lượng theo công thức sau:

Với *Sset (A, B)* là độ tương đồng giữa hai *tập các* *bài đăng A* và *B*. Độ tương đồng dựa trên hành vi *đăng bài đăng Spost (A, B)* sẽ nằm trong khoảng [0,1]. Nếu *Spost (A, B)* càng gần 1 thì độ tương đồng càng cao và ngược lại, *Spost (A, B)* càng gần 0 thì độ tương đồng càng thấp.

* + 1. **Ước lượng độ tương đồng hành vi thích bài đăng**

Gọi *Sil* và *Sjl* lần lượt là *tập các* *bài đăng* đã được thích bởi người dùng *i* và người dùng *j*. Độ tương đồng dựa trên hành vi *thích bài đăng* của người dùng *i* và người dùng *j* là độ tương đồng giữa *tập các* *bài đăng* và được ước lượng theo công thức sau:

Với *Sset (A, B)* là độ tương đồng giữa hai *tập các* *bài đăng A* và *B*. Độ tương đồng dựa trên hành vi *thích bài đăng Slike (A, B)* sẽ nằm trong khoản [0,1]. Nếu *Slike (A, B)* càng gần 1 thì độ tương đồng càng cao và ngược lại, *Slike (A, B)* càng gần 0 thì độ tương đồng càng thấp.

* + 1. **Ước lượng độ tương đồng hành vi nhận xét/thích nhận xét bài đăng**

Đồ án sẽ ước lượng độ tương đồng hành vi *nhận xét/thích nhận xét bài đăng* của người dùng dựa trên các nguyên tắc sau:

+ Mỗi nhận xét được xác định theo một trong các quan điểm: tích cực, tiêu cực, hoặc trung tính.

+ Với mỗi nhận xét được tính phải là nhận xét có quan điểm tích cực hoặc tiêu cực. Nhận xét có quan điểm trung tính sẽ bị loại bỏ.

+ Với mỗi *bài đăng*, nếu số lượng nhận xét tích cực của người dùng lớn hơn số lượng nhận xét tiêu cực của người dùng đó thì *bài đăng* sẽ được xem là tích cực với người dùng này và ngược lại, nếu số lượng nhận xét tiêu cực của người dùng lớn hơn số lượng nhận xét tích cực của người dùng đó thì *bài đăng* sẽ được tính là tiêu cực với người dùng này.

+ Trong trường hợp số lượng nhận xét tích cực của người dùng bằng với số lượng nhận xét tiêu cực của người dùng đó, đồ án sẽ xem xét tiếp tới hành động thích nhận xét của người dùng như sau:

* Nếu số lượng thích nhận xét tích cực lớn hơn số lượng thích nhận xét tiêu cực thì *bài đăng* sẽ được xem là tích cực với người dùng này
* Nếu số lượng thích nhận xét tích cực nhỏ hơn số lượng thích nhận xét tiêu cực thì *bài đăng* sẽ được xem là tiêu cực với người dùng này.
* Nếu số lượng thích nhận xét tích cực bằng với số lượng thích nhận xét tiêu cực thì *bài đăng* sẽ được xem là trung tính với người dùng này. Trường hợp này *bài đăng* sẽ bị loại bỏ khỏi *tập các bài đăng* của người dùng này

Gọi *Sipositive* và *Sinegative* lần lượt là *tập các bài đăng* tích cực và tiêu cực đối với người dùng *i*. Gọi *Sjpositive* và *Sjnegative* lần lượt là *tập các bài đăng* tích cực và tiêu cực đối với người dùng *j*. Để ước lượng độ tương đồng hành vi *nhận xét/thích nhận xét* của người dùng *i* và người dùng *j*, đồ án sử dụng các giả thiết sau:

+ Nếu hai tập *Sipositive* và *Sjpositive* càng có nhiều *bài đăng* tương đồng thì độ tương đồng hành vi *nhận xét/thích nhận xét* của người dùng *i* và người dùng *j* càng cao

+ Nếu hai tập *Sinegative* và *Sjnegative* càng có nhiều *bài đăng* tương đồng thì độ tương đồng hành vi *nhận xét/thích nhận xét* của người dùng *i* và người dùng *j* càng cao

+ Nếu hai tập *Sipositive* và *Sjnegative* càng có ít *bài đăng* tương đồng thì độ tương đồng hành vi *nhận xét/thích nhận xét* của người dùng *i* và người dùng *j* càng cao

+ Nếu hai tập *Sinegative* và *Sjpositive* càng có ít *bài đăng* tương đồng thì độ tương đồng hành vi *nhận xét/thích nhận xét* của người dùng *i* và người dùng *j* càng cao

Độ tương đồng hành vi *nhận xét/thích nhận xét* của người dùng *i* và người dùng *j* được ước lượng như sau:

Độ tương đồng hành vi *nhận xét/thích nhận xét* của hai người dùng *Scmt (i, j)* sẽ nằm trong khoảng [0,1]. Nếu *Scmt (i, j)* càng gần 1 thì độ tương đồng hành vi *nhận xét/thích nhận xét* sẽ càng cao và ngược lại, *Scmt (i, j)* càng gần 0 thì độ tương đồng hành vi *nhận xét/thích nhận xét* càng thấp.

* + 1. **Ước lượng độ tương đồng hành vi tham gia hội nhóm/cộng đồng**

Gọi *Sicm* và *Sjcm* lần lượt là *tập các hội nhóm/cộng đồng* đã tham gia của người dùng *i* và người dùng *j*. Độ tương đồng hành vi *tham gia hội nhóm/cộng đồng* là độ tương đồng giữa các *cộng đồng* và được ước lượng như sau:

Với *Sset (A, B)* là độ tương đồng giữa hai *tập các cộng đồng A* và *B*. Độ tương đồng hành vi tham *gia hội nhóm/cộng đồng Sjoin (i, j)* sẽ nằm trong khoảng [0,1]. Nếu *Sjoin (i, j)* càng gần 1 thì độ tương đồng hành vi *tham gia hội nhóm/cộng đồng* càng cao và ngược lại, *Sjoin (i, j)* càng gần 0 thì độ tương đồng hành vi *tham gia hội nhóm/cộng đồng* càng thấp.

* + 1. **Ước lượng độ tương đồng của người dùng dựa vào hành vi**

Độ tương đồng của người dùng dựa vào hành vi trên *mạng xã hội* sẽ được ước lượng dựa trên độ tương đồng 4 hành vi *đăng bài đăng*, *thích bài đăng*, *nhận xét/thích nhận xét bài đăng* và *tham gia hội nhóm/cộng đồng* của người dùng.

Gọi *w1*, *w2*, *w3*, *w4* lần lượt là trọng số của độ tương đồng dựa trên hành vi *đăng bài đăng*, *thích bài đăng*, *nhận xét/thích nhận xét bài đăng*, *tham gia hội nhóm/cộng đồng* và được xác định với điều kiện sau:

Độ tương đồng của người dùng dựa vào hành vi giữa người dùng *i* và người dùng *j* được ước lượng như sau:

Với *Spost (i, j)*, *Slike (i, j)*, *Scmt (i, j)*, *Sjoin (i, j)* lần lượt là độ tương đồng hành vi *đăng bài đăng*, độ tương đồng hành vi *thích bài đăng*, độ tương đồng hành vi *nhận xét/thích nhận xét bài đăng*, độ tương đồng hành vi *tham gia hội nhóm/cộng đồng* của hai người dùng *i* và *j*. Các trọng số *w1*, *w2*, *w3*, *w4*sẽ là khác nhau với mỗi bộ dữ liệu khác nhau. Đồ án sẽ trình bày cách chọn chi tiết cho mỗi trọng số ở chương tiếp theo.

Độ tương đồng của người dùng dựa vào hành vi *Suser (i, j)* sẽ nằm trong khoảng [0,1]. Nếu *Suser (i, j)* càng gần 1 thì độ tương đồng của người dùng dựa vào hành vi càng cao và ngược lại, *Suser (i, j)* càng gần 0 thì độ tương đồng của người dùng dựa vào hành vi càng thấp.

Như vậy, trong mô hình của đồ án đã *ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội*.

* 1. **Kết luận**

Trong chương 2 đồ án đã trình bày *thuật toán ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội* gồm có mô hình hoá và cụ thể các công thức ước lượng độ tương đồng.

Tiếp theo là chương 3 đồ án sẽ đánh giá *thuật toán ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội*.

**CHƯƠNG 3: ĐÁNH GIÁ THUẬT TOÁN ƯỚC LƯỢNG ĐỘ TƯƠNG ĐỒNG CỦA NGƯỜI DÙNG DỰA VÀO HÀNH VI TRÊN MẠNG XÃ HỘI**

Chương 3 sẽ trình bày các nội dung sau:

* Đánh giá thuật toán *ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội*
* Ứng dụng thuật toán
  1. **Đánh giá thuật toán**

Để đánh giá thuật toán đồ án sẽ sử dụng dữ liệu thực tế của người dùng lấy về được từ trên trang *mạng xã hội* Twitter. Dữ liệu thực tế lấy được bao gồm: *người dùng* và *bài đăng*, cụ thể như sau:

+ *Người dùng*: là các thành viên tham gia *mạng xã hội* Twitter

+ *Bài đăng*: là các *bài đăng* được tương tác bởi các hành vi *đăng bài đăng*, *thích bài đăng*, *nhận xét bài đăng* của *người dùng*. Trong đó:

* Hành vi *đăng bài đăng (post)*: bao gồm hành vi đăng các tweet – dòng trạng thái cá nhân và retweet – dòng trạng thái của người dùng khác
* Hành vi *thích bài đăng (like)*: hành vi thích bài đăng của người dùng khác
* Hành vi *nhận xét bài đăng (comment)*: hành vi reply – nhận xét bài đăng của người dùng khác

Trong chương 2 đồ án đã trình bày ước lượng độ tương đồng của người dùng dựa vào 4 hành vi phổ biến trên mạng xã hội gồm có: *đăng bài đăng*, *thích bài đăng*, *nhận xét/thích nhận xét bài đăng* và *tham gia hội nhóm/cộng đồng*. Trên thực tế, đồ án lấy dữ liệu từ trang mạng xã hội Twitter không có dữ liệu nào cho hành vi tham gia hội nhóm/cộng đồng. Do vậy, đồ án sẽ sử dụng dữ liệu, xây dựng bộ trọng số, ước lượng độ tương đồng và đánh giá thuật toán theo 3 hành vi: *đăng bài đăng*, *thích bài đăng và* *nhận xét bài đăng*. Sơ đồ quy trình đánh giá như sau:

Hình 3.1: Quy trình đánh giá thuật toán

Thuật toán

w3

w2

w1

Xây dựng bộ trọng số

* w1: trọng số độ tương đồng hành vi đăng bài (post)
* w2: trọng số hành vi thích bài đăng (favorite)
* w2: trọng số hành vi nhận xét bài đăng (comment)

Xây dựng bộ dữ liệu đánh giá

Xây dựng bộ dữ liệu học

Xây dựng bộ dữ liệu thử nghiệm

Kịch bản đánh giá

Kết quả và phân tích

* + 1. **Xây dựng bộ dữ liệu thử nghiệm**

Đồ án sử dụng REST API cung cấp bởi Twitter để lấy dữ liệu thực tế người dùng trên trang *mạng xã hội* Twitter. Dữ liệu lấy được chia làm 2 bộ dữ liệu cụ thể như sau:

+ *Bộ dữ liệu đánh giá*: là bộ dữ liệu bao gồm người dùng, các *bài đăng* được tương tác bởi các hành vi *đăng bài đăng (post)*, *thích bài đăng (like/favorite)*, *nhận xét/thích nhận xét bài đăng (comment/like comment)* của người dùng trên *mạng xã hội* Twitter. *Bộ dữ liệu đánh giá* được sử dụng để đánh giá thuật toán.

+ *Bộ dữ liệu học*: là bộ dữ liệu bao gồm các *bài đăng* được đăng, được chia sẻ bởi người dùng trên trang *mạng xã hội* Twitter. *Bộ dữ liệu học* được sử dụng làm dữ liệu đầu vào cho học máy xác định *chủ đề* và *quan điểm* của *bài đăng* trước khi ước lượng độ tương đồng giữa hai *bài đăng*.

Dữ liệu lấy được đã được xử lý để đạt được các tiêu chuẩn như sau:

+ Thông tin người dùng, nội dung *bài đăng* đều được thể hiện bằng các ngôn ngữ sử dụng chữ cái la tinh. Người dùng, các *bài đăng* sử dụng ngôn ngữ khác sẽ loại bỏ.

+ Người dùng phải có tương tác với ít nhất 5 *bài đăng* thông qua 3 hành vi *đăng bài đăng (post)*, *thích bài đăng (like/favorite)*, *nhận xét/thích nhận xét bài đăng (comment/like comment)*. Người dùng không có tương tác nào hoặc có tương tác với dưới 5 *bài đăng* sẽ loại bỏ.

+ Các *bài đăng* được tương tác bởi các hành vi của người dùng trong khoảng thời gian lấy dữ liệu. Các *bài đăng* được tương tác ngoài khoảng thời gian này sẽ loại bỏ.

Cấu trúc lấy và lưu dữ liệu như sau:

+ Người dùng (User): id (mã người dùng), screenname (tên tài khoản người dùng)

+ Bài đăng (Entry): id (mã bài đăng), content (nội dung bài đăng)

+ Hành vi đăng bài đăng (Post):

+ Hành vi thích bài đăng (Like/Favorite):

+ Hành vi nhận xét bài đăng (Comment):

Chi tiết 2 bộ dữ liệu như sau:

* Để xây dựng *bộ dữ liệu đánh giá*, đồ án đã lấy và xử lý được dữ liệu của XXX người dùng và YYY *bài đăng* được tương tác bởi XXX người dùng đó trong khoảng thời gian từ /2016 đến /2016.

Đồ án sẽ sử dụng *bộ dữ liệu học* để xác định *chủ đề* và *quan điểm* cho từng *bài đăng* của *bộ dữ liệu đánh giá*. Sau đó lưu lại dữ liệu dưới dạng file có cấu trúc 4 phần gồm có: thông tin người dùng (id, screenname), các bài đăng được đăng bởi người dùng (mỗi bài đăng ghi trên một dòng), các bài đăng được thích bởi người dùng (mỗi bài đăng ghi trên một dòng), các bài đăng được nhận xét

* Đối với *bộ dữ liệu học*, đồ án sử dụng phương pháp điều tra bằng phiếu đánh giá, tạo mẫu lấy đánh giá của người dùng bằng Google Form. Các bước xây dựng bộ dữ liệu học như sau:
* **Bước 1:** Chuẩn bị bộ dữ liệu bao gồm XXX *bài đăng* sử dụng làm mẫu đánh giá (form), thoả các điều kiện sau:

**+** Đã xử lý theo các tiêu chuẩn như đã nêu trên

**+** Khác với các *bài đăng* thuộc *bộ dữ liệu đánh giá*.

**+** Được tương tác bởi người dùng bất kỳ trong khoảng thời gian từ 2016/ đến /2016.

* **Bước 2:** Chuẩn bị form đánh giá: YYY form. Một form bao gồm 10 *bài đăng* tương ứng với 10 câu hỏi đánh giá. Mỗi câu hỏi bao gồm 2 nội dung đánh giá *bài đăng* là *category* *(chủ đề bài đăng đề cập tới)* và *sentiment* *(quan điểm/trạng thái cảm xúc của bài đăng)*. Ở mỗi nội dung đánh giá sẽ có thể lựa chọn được nhiều hơn một kết quả. Cụ thể danh sách các lựa chọn được liệt kê ở hai bảng 3.1 và 3.2. Với mỗi tài khoản Google của người dùng được phép đánh giá một lần duy nhất để đảm kết quả đánh giá được khách quan. Một kết quả đánh giá hợp lệ là kết quả bao gồm đánh giá cho cả hai nội dung của cả 10 câu hỏi của một form (ví dụ: hình 3.2). Thời gian thu thập kết quả đánh giá của người dùng cho bộ dữ liệu học là từ /2016 đến /2016.
* **Bước 3:** Lưu kết quả đánh giá. Sau khi có được toàn bộ kết quả đánh giá, đồ án thực hiện lưu dữ liệu vào file dưới dạng như sau: Mỗi file sẽ chứa nội dung của một *bài đăng*, các đánh giá *chủ đề* cho *bài đăng*, các đánh giá *quan điểm* cho *bài đăng*. Nội dung *bài đăng*, mỗi đánh giá *chủ đề*, mỗi đánh giá *quan điểm* sẽ cách nhau một dấu “tab”. Tập các file thu được cần được lọc và phân loại ở bước tiếp theo.
* **Bước 4:** Lọc và phân loại ra 2 *bộ dữ liệu học* tách biệt gồm: *bộ dữ liệu học cho xác định chủ đề của bài đăng* và *bộ dữ liệu học cho xác định quan điểm của bài đăng*. Lọc theo các tiêu chí như sau:

+ *Bộ dữ liệu học* bao gồm các *bài đăng* được đánh giá bởi ít nhất 10 người dùng (10 lượt đánh giá)

+ *Bộ dữ liệu học cho xác định chủ đề của bài đăng* bao gồm các *bài đăng* được đánh giá *chủ đề* không thuộc lựa chọn Other (Khác) và được phân loại tương ứng với từng *chủ đề* riêng biệt đã nêu. Các *bài đăng* được đánh giá chủ đề là Other (Khác) sẽ loại bỏ khỏi bộ dữ liệu này.

+ *Bộ dữ liệu học cho xác định quan điểm của bài đăng* bao gồm các *bài đăng* được đánh giá *quan điểm* không thuộc lựa chọn Other (Khác) và được phân loại tương ứng với từng *quan điểm* riêng biệt đã nêu. Các *bài đăng* được đánh giá quan điểm là Other (Khác) sẽ loại bỏ khỏi bộ dữ liệu này.

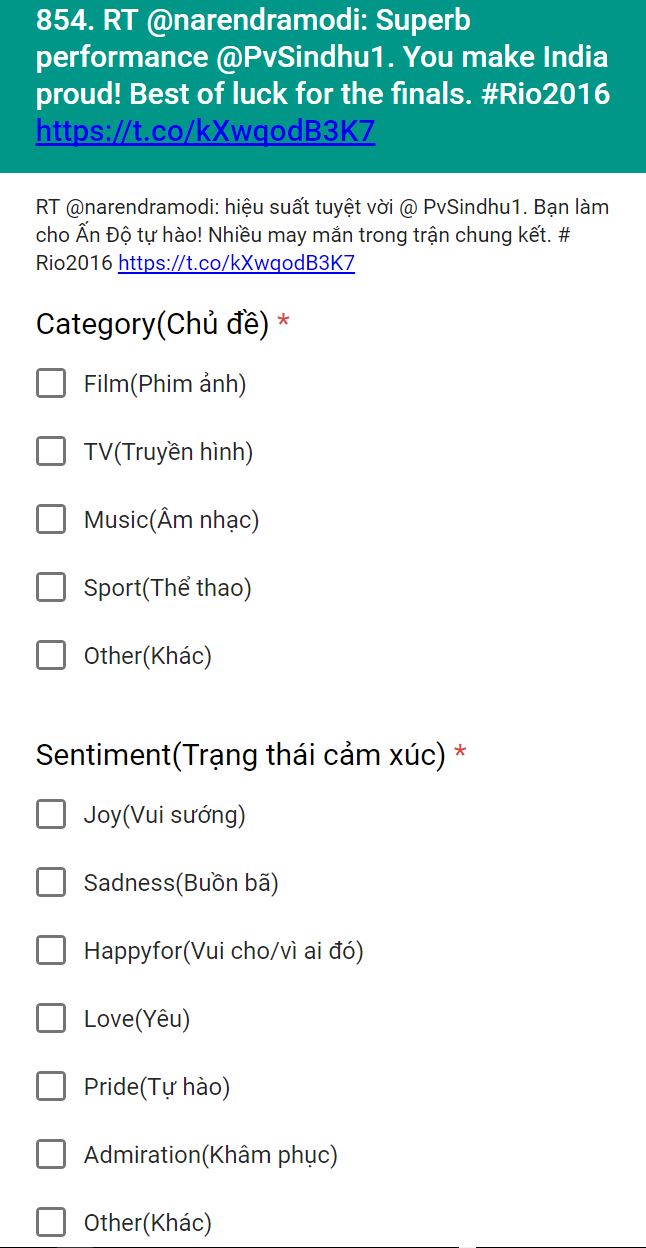
Mỗi *bài đăng* được phân loại tương ứng với từng *chủ đề* sẽ được lưu vào mỗi file chỉ bao gồm nội dung của *bài đăng* và được coi là *một mẫu học* cho *chủ đề* đó của *bộ dữ liệu học cho xác định chủ đề của bài đăng*. Tương tự, mỗi *bài đăng* được phân loại tương ứng với từng *quan điểm* sẽ được lưu vào mỗi file chỉ bao gồm nội dung của *bài đăng* và được coi là *một mẫu học* cho *quan điểm* đó của *bộ dữ liệu học cho xác định quan điểm của bài đăng*.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Category | No | Category | No | Category |
| 1 | Chemistry  (Hoá học) | 9 | Economy  (Kinh tế) | 17 | Music  (Âm nhạc) |
| 2 | Physical  (Vật lý) | 10 | Philosophy  (Triết học) | 18 | Sport  (Thể thao) |
| 3 | Geography  (Địa lý) | 11 | Literary  (Văn học) | 19 | History  (Lịch sử) |
| 4 | Biological  (Sinh vật học) | 12 | Medicine  (Y học) | 20 | Culinary  (Ẩm thực) |
| 5 | Astronomy  (Thiên văn học) | 13 | Religion  (Tôn giáo – Niềm tin) | 21 | Fashion  (Thời trang) |
| 6 | IT (CNTT) | 14 | Architecture  (Kiến trúc) | 22 | Other  (Chủ đề khác) |
| 7 | Politic  (Chính trị) | 15 | Film  (Điện ảnh) |  |  |
| 8 | Education  (Giáo dục) | 16 | TV  (Truyền hình) |  |  |

Bảng 3.1: Tập các chủ đề của bài đăng

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Sentiment | No | Sentiment | No | Sentiment |
| 1 | Joy  (Vui sướng) | 7 | Fear  (Sợ hãi) | 13 | Pride  (tự hào) |
| 2 | Sadness  (Buồn bã) | 8 | Confused  (Bối rối) | 14 | Anger  (Phẫn nộ) |
| 3 | Happyfor  (Vui cho/vì ai đó) | 9 | Love  (Yêu) | 15 | Gratitude  (Biết ơn) |
| 4 | Sorryfor  (Có lỗi với ai đó) | 10 | Disgust  (Thù ghét) | 16 | Admiration  (Khâm phục) |
| 5 | Sorry  (Có lỗi) | 11 | Regret  (Hối tiếc) | 17 | Other  (Quan điểm khác) |
| 6 | Hope  (Hi vọng) | 12 | Disappointed  (Thất vọng) |  |  |

Bảng 3.2: Tập các quan điểm của bài đăng



Hình 3.2: Một câu hỏi trong một form đánh giá

Kết quả cuối cùng đồ án thu được là hai *bộ dữ liệu học* là:

+ *Bộ dữ liệu học cho xác định chủ đề của bài đăng* gồm XXX *mẫu*

+ *Bộ dữ liệu học cho xác định quan điểm của bài đăng* gồm YYY *mẫu*.

* + 1. **Xây dựng bộ trọng số**
    2. **Kịch bản đánh giá**
    3. **Kết quả đánh giá**
  1. **Ứng dụng thuật toán**
     1. **Mô tả ứng dụng**
     2. **Kiến trúc tổng quan ứng dụng**
     3. **Đánh giá kết quả**
  2. **Kết luận**