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, 12/2016

LỜI CẢM ƠN

Đầu tiên, em xin gửi lời cảm ơn chân thành nhất đến các thầy cô giáo khoa Công nghệ thông tin 1 – Học viện Công nghệ Bưu chính Viễn thông. Các thầy cô đã truyền đạt vào dạy bảo cho em rất nhiều kiến thức hay và bổ ích trong suốt quá trình 4 năm học tập tại Học viện.

Đặc biệt, em xin dành lời cảm ơn sâu sắc nhất tới thầy TS. Nguyễn Mạnh Hùng. Thầy đã tận tình chỉ bảo cho em trong suốt quá trình học tập, dành thời gian xem xét, hướng dẫn, góp ý cho em hoàn thiện đồ án tốt nghiệp được tốt nhất.

Con xin cảm ơn tới gia đình, bố mẹ và anh trai đã luôn ở bên động viên, chỉ bảo trong suốt những năm tháng học tập qua.

Mình xin cảm ơn tất cả những người bạn của mình, những người bạn luôn bên cạnh, giúp đỡ, chia sẻ cùng mình những lúc khó khăn.

Em xin chân thành cảm ơn!

Hà Nội, ngày 10 tháng 12 năm 2016

Sinh viên

Trần Xuân Tiến

MỤC LỤC

MỤC LỤC ..... **ii**
DANH MỤC THUẬT NGỮ ..... **iv**
MỞ ĐẦU **1**
CHƯƠNG 1: BÀI TOÁN ƯỚC LƯỢNG ĐỘ TƯƠNG ĐỒNG CỦA NGƯỜI DÙNG DỰA VÀO HÀNH VI TRÊN MẠNG XÃ HỘI **3**
1.1. Giới thiệu về mạng xã hội **3**
1.2. Giới thiệu bài toán **7**
1.3. Các phương pháp tiếp cận bài toán **10**
1.4. Phương pháp tiếp cận bài toán của đồ án **11**
1.5. Kết luận **11**
CHƯƠNG 2: THUẬT TOÁN ƯỚC LƯỢNG ĐỘ TƯƠNG ĐỒNG CỦA NGƯỜI DÙNG DỰA VÀO HÀNH VI TRÊN MẠNG XÃ HỘI **12**
2.1. Mô hình hoá hành vi của người dùng trên mạng xã hội **12**
2.1.1. Bài đăng 12
2.1.2. Hành vi của người dùng 13
2.2. Ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội **13**
2.2.1. Ước lượng độ tương đồng giữa các bài đăng 13
2.2.2. Ước lượng độ tương đồng hành vi đăng bài đăng 17
2.2.3. Ước lượng độ tương đồng hành vi thích bài đăng 18
2.2.4. Ước lượng độ tương đồng hành vi nhận xét/thích nhận xét bài đăng 18
2.2.5. Ước lượng độ tương đồng của người dùng dựa vào hành vi 20
2.3. Kết luận **20**
CHƯƠNG 3: ĐÁNH GIÁ THUẬT TOÁN **22**
3.1. Xây dựng bộ dữ liệu thử nghiệm **22**
3.1.1. Bộ dữ liệu học cho xác định chủ đề của bài đăng 23

3.1.2. Bộ dữ liệu học cho xác định trạng thái cảm xúc của bài đăng 24
3.1.3. Bộ dữ liệu học cho xác định quan điểm của bài đăng 25
3.1.4. Bộ dữ liệu đánh giá 26
3.1.5. Bộ dữ liệu cho khảo sát trọng số cho ước lượng độ tương đồng giữa các bài đăng 29
3.2. Phương pháp đánh giá 31
3.3. Cài đặt thuật toán 32
3.3.1. Mô hình cài đặt 32
3.3.2. Thư viện hỗ trợ 34
3.4. Khảo sát bộ trọng số 35
3.4.1. Khảo sát bộ trọng số cho ước lượng độ tương đồng giữa các bài đăng 35
3.4.2. Khảo sát bộ trọng số cho ước lượng độ tương đồng của người dùng dựa vào hành vi 37
3.5. Kết quả đánh giá 39
3.6. Kết luận 40
CHƯƠNG 4: ỨNG DỤNG THUẬT TOÁN 41
4.1. Mô tả ứng dụng MyTwitter 41
4.2. Kiến trúc tổng quan ứng dụng 41
4.3. Kịch bản sử dụng của người dùng 42
4.4. Kết quả 44
4.5. Kết luận 47
KẾT LUẬN 48
DANH MỤC THAM KHẢO 50
PHỤ LỤC i

DANH MỤC THUẬT NGỮ

Từ	Từ đầy đủ
Entry	Bài đăng
Community	Hội nhóm/Cộng đồng
WordNet	Cơ sở dữ liệu từ vựng – nhóm các từ thành các tập hợp đồng nghĩa
User	Người dùng
API	Application Programing Interface – Giao diện lập trình ứng dụng
CSDL	Cơ sở dữ liệu

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 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 trong số các dịch vụ trực tuyến và trở thành hoạt động nhận được nhiều sự quan tâm trên Internet hiện nay. Mạng xã hội tạo ra môi trường giao lưu, chia sẻ, kết nối cộng đồng thuận lợi nên thu hút được số lượng 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 lớn đó làm cho mạng xã hội trở thành một kho dữ liệu lớn, cung cấp thông tin 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, đồ án thu thập khối dữ liệu của hai người dùng đó trên mạng xã hội họ sử dụng, sau đó thực hiện phân tích hành vi của họ dựa trên dữ liệu thu thập được và thực hiện so sánh. Việc phân tích tập trung vào các câu hỏi như: dữ liệu đó đề cập tới vấn đề gì? Các vấn đề được đề cập tới giống hay khác nhau? Dữ liệu đó thể hiện quan điểm, cảm xúc như thế nào? Những quan điểm, cảm xúc đó có giống nhau

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
  - 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 hành vi của người dùng trên mạng xã hội
  - Ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội
- **Chương 3: Đánh giá thuật toán**
  - Xây dựng bộ dữ liệu thử nghiệm
  - Cài đặt thuật toán
  - Khảo sát bộ trọng số
  - Phương pháp đánh giá
  - Kết quả đánh giá
- **Chương 4: Ứng dụng thuật toán**
  - Mô tả ứng dụng
  - Kiến trúc tổng quan
  - Kịch bản sử dụng của người dùng
  - 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

1.1. Giới thiệu về mạng xã hội

Hình 1.1: Giao diện mạng xã hội Facebook www.facebook.com

- **Mạng xã hội:**

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 mối quan hệ, mục đích, giao tiếp như: bạn bè, họ hàng, trao đổi tài chính, kiến thức, mua bán…

- **Lịch sử phát triển**

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) cho cá nhân mình cũng như các thành viên khác dùng.

- **Cấu trúc chung của mạng xã hội**

Hình 1.2: Minh hoạ mô hình cấu trúc mạng xã hội (www.genengnews.com)

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, quản lý thông tin cá nhân, thiết lập quyền kiểm soát tài khoản.

+ 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 đăng, thích bài đăng, bình luận bài đăng, tham gia các nhóm yêu thích, gửi email, hội thoại, xem ảnh, xem 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**

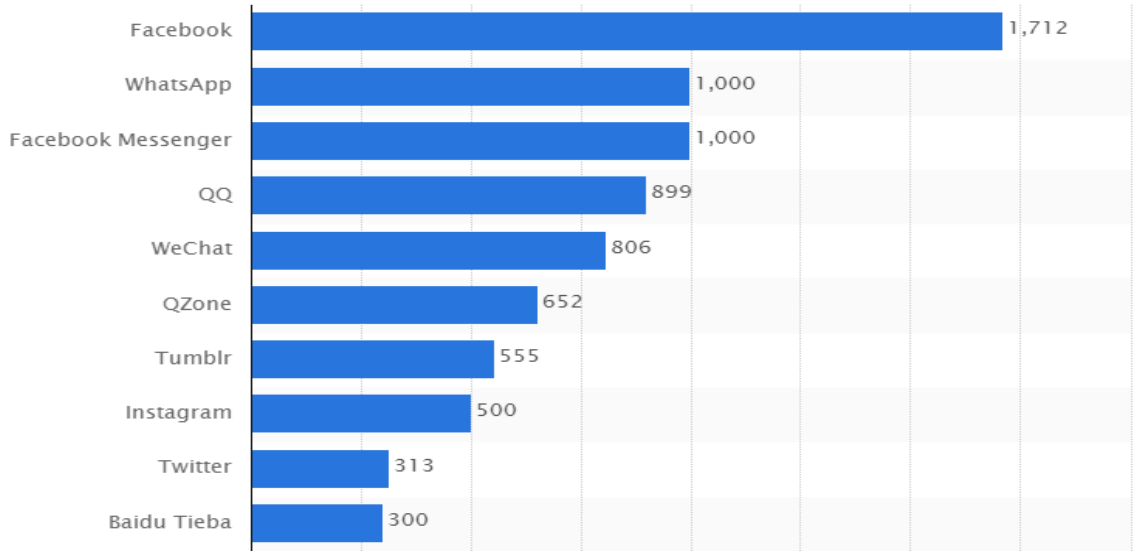
Điểm nổi bật của *mạng xã hội* là tính kết nối và khả năng chia sẻ rất mạnh mẽ qua mạng Internet không phân biệt không gian và thời gian, phá vỡ ngăn cách về địa lý, ngôn ngữ, giới tính lẫn quốc gia.

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 sẽ tiếp tục tăng, *mạng xã hội* ngày càng trở nên thân thiện và phổ cập với mọi người dùng.

Dưới đây là thống kê số lượng người dùng của 10 trang mạng xã hội phổ biến nhất gồm: Facebook, WhatsApp, QQ, Facebook Messenger, Qzone, WeChat, Tumblr, Instagram, Twitter, Baidu Tieba tính đến 9/2016.



Hình 1.3: Biểu đồ thống kê người dùng - Đơn vị: Triệu người dùng (www.statista.com, 2016)

- **Mạng xã hội Twitter:**

Twitter [15] 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 (www.twitter.com)

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

Tính năng của người dùng trên Twitter: để sử dụng được các tính năng Twitter cung cấp, 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:

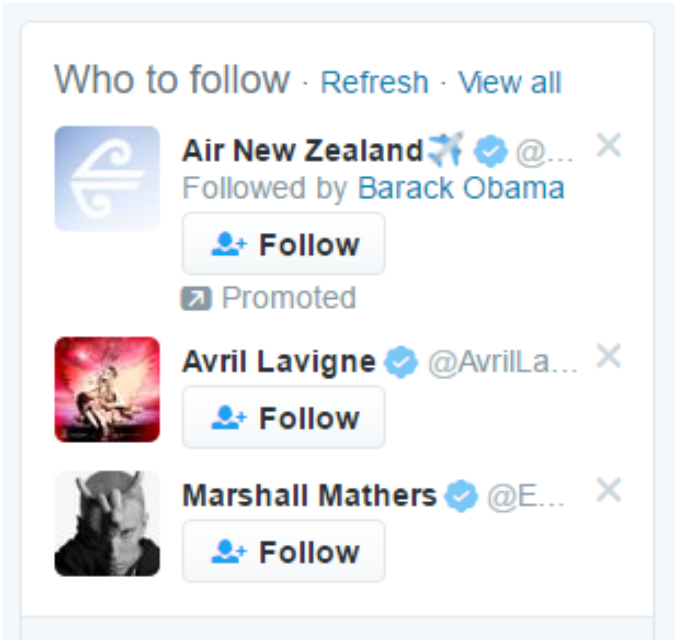
+ Theo dõi theo "*hashtag*" – một dạng như từ khoá. Ví dụ: *#Rio2016*.
+ Theo dõi một Url nhất định.
+ Theo dõi theo các *retweet – bài đăng được chia sẻ*.

Để trích xuất được dữ liệu, Twitter cũng cung cấp REST API (Twitter API) [14] cho tương tác với CSDL thông qua các lời gọi cung cấp sắn, và sử dụng cơ chế xác thực bảo mật OAuth.

1.2. Giới thiệu bài toán

Mạng xã hội đã và đang ngày càng phát triển, nó 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 1.7 tỷ thành viên, theo sau là WhatsApp với 1 tỷ 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 đa dạng 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, gợi ý 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 hội nhóm/cộng đồng (www.facebook.com)



Hình 1.6: Tính năng gợi ý theo dõi (www.twitter.com)

- **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… và sử dụng các tính năng mạng xã hội cung cấp như xem video, xem ảnh, gửi tin nhắn, trò chuyện hội nhóm, chia sẻ file…

Như vậy có thể rút ra được, một người dùng trên mạng xã hội sẽ có hai đặc trưng chính là: các thông tin cá nhân và các hành vi cá nhân. Việc nghiên cứu, đánh giá, nhận xét, so sánh giữa người dùng với nhau sẽ dựa trên các thông tin cá nhân và hành vi cá nhân 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 thông tin 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 – "Post"* bài đăng lên trang cá nhân, lên trang cá nhân của bạn bè, lên các hội nhóm/cộng đồng đã tham gia. Bài đăng có thể là tin tức, hình ảnh, video hay bất kỳ trạng thái, nội dung nhận xét nào của người dùng.

+ *Xem* các bài đăng trên trang cá nhân, trang cá nhân của bạn bè, trên các hội nhóm/cộng đồng đã tham gia.

+ *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 để tạo danh sách bạn bè, tạo vòng kết nối.

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

+ *Tìm kiếm* các nội dung mong muốn trên mạng xã hội.

+ *Sử dụng các ứng dụng mở rộng mà mạng xã hội cung cấp* như: trò chơi, ứng dụng mua sắm online…

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 đăng, các bài đăng đã thích, các bài đăng đã nhận xét/thích nhận xét, các bài đăng đã chia sẻ, các hội nhóm/cộng đồng đã tham gia, các người dùng cùng theo dõi, các sự kiện cùng tham gia.

1.3. Các phương pháp tiếp cận bài toán

Độ tương đồng của hai người dùng dựa vào hành vi sẽ ước lượng theo độ tương đồng giữa các khối dữ liệu tạo ra bởi các hành vi của hai người dùng đó trên mạng xã hội họ 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.

Do người dùng trên mạng xã hội có nhiều hành vi khác nhau nên dữ liệu tạo ra bởi các hành vi này rất đang dạng. Dữ liệu này được phân tích, đánh giá theo các phương pháp như:

+ Phân tích, đánh giá theo ngữ nghĩa: hai khối dữ liệu được so sánh với nhau dựa trên sự tương đồng ngữ nghĩa. Ngữ nghĩa sau đó được so sánh dựa trên WordNet như đã được nghiên cứu bởi Buscaldi et al. [1] hay Lee et al. [7]

+ Phân tích đánh giá theo thông tin của dữ liệu gồm cấu trúc từ vựng của từ, câu và/hoặc các số liệu thống kê của các từ trong văn bản, như đã được nghiên cứu bởi Proisl et al. [10]

Các phương pháp này xem xét tới nội dung văn bản, chưa xem xét tới các thông tin có liên quan như các thẻ, danh mục, tiêu đề, từ khoá.

1.4. Phương pháp tiếp cận bài toán của đồ án

Đồ án sẽ thực hiện ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội bằng cách phân tích, đánh giá độ tương đồng giữa các khối dữ liệu dạng văn bản (text) tạo ra bởi các hành vi của người dùng theo các thông tin mà dữ liệu thể hiện.

Tiêu chí đánh giá độ tương đồng trong mô hình của đồ án sẽ sử dụng 3 hành vi phổ biến trong số các hành vi của người dùng trên mạng xã hội để mô hình hoá tổng quát bài toán bao gồm: *đăng bài đăng – post*, *thích bài đăng – like* và *nhận xét/bình luận bài đăng – comment*. Với mỗi bài đăng tạo ra bởi các hành vi của người dùng như: *đăng bài đăng*, *thích/không thích bài đăng*, *nhận xét/bình luận bài đăng* đều thể hiện ít nhất một vấn đề, quan điểm, trạng thái cảm xúc về vấn đề nào đó. Do đó, bài đăng sẽ được mô hình hoá và ước lượng độ tương đồng theo các yếu tố chủ đề, quan điểm và trạng thái cảm xúc. Mỗi yếu tố của bài đăng sẽ có trọng số khác nhau phụ thuộc dữ liệu, tương tác thực tế của người dùng. Việc ước lượng độ tương đồng của các bài đăng được tạo ra bởi các hành vi của hai người dùng là một trong quá 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 người dùng Y cùng tham gia một mạng xã hội, dựa trên các hành vi của họ lần lượt theo thứ tự: *đăng bài đăng*, *thích/không thích bài đăng*, *nhận xét/bình luận bài đăng* là P_X, L_X, C_X và P_Y, L_Y, C_Y, là quá trình đánh giá độ tương đồng giữa các tập bài đăng tương tác bởi người dùng X và người dùng Y là: P_X và P_Y, L_X và L_Y, C_X và C_Y.

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

1.5. 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. Đồ án xác định phương pháp tiếp cận bài toán là đánh giá độ tương đồng của người dùng dựa trên việc phân tích dữ liệu dạng văn bản được tạo ra bởi 3 hành vi phổ biến trong số các hành vi của người dùng trên mạng xã hội bao gồm: hành vi *đăng bài đăng – post*, hành vi *thích bài đăng – like*, hành vi *nhận xét/bình luận bài đăng – comment*.

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 hành vi của người dùng trên mạng xã hội
- Ước lượng độ tương đồng bài đăng, hành vi, người dùng trên mạng xã hội

2.1. Mô hình hoá 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 sẽ trình bày chi tiết mô hình hoá và ước lượng độ tương đồng dưới đây.

2.1.1. Bài đăng

Một người dùng mạng đã thực hiện đăng ký tài khoản cá nhân trên một trang mạng xã hội là một người dùng (user) của mạng xã hội đó. Một user có thể đăng, bình luận/nhận xét các trạng thái, ảnh, video lên trang cá nhân hoặc đăng trên các hội nhóm/cộng đồng đã tham gia. Trạng thái, ảnh, video hay các bình luận/nhận xét đều là các nội dung được người dùng đăng lên, gọi là *bài đăng (entry)*. Mô hình của đồ án chỉ xem xét tới các *bài đăng* dạng *văn bản (text)* như trạng thái, bình luận/nhận xét của người dùng.

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 chung cuộc của một trận đấu bóng đá… Một *bài đăng* có thể thể hiện sở thích, ý kiến cá nhân nào đó của người dùng hoặc cũng có thể không thể hiện một vấn đề hay ý kiến nào cả.

Như vậy, một *bài đăng* sẽ được mô tả bởi một hay nhiều thuộc tính khác nhau. Mô hình của đồ án sẽ xem xét *bài đăng* dựa trên thuộc tính của nó.

2.1.2. Hành vi của người dùng

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

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 các nhận xét* của một bài đăng.

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

Mô hình của đồ án sẽ xem xét trên 3 hành vi phổ biến là: *đăng bài đăng, thích bài đăng* và *nhận xét/thích nhận xét bài đăng*. Tiếp theo đồ án sẽ trình bày tới ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội.

2.2. Ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội

2.2.1. 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.

Một bài đăng có thể có một số thuộc tính, bao gồm các thuộc tính tường minh như *nội dung (content),* và các thuộc tính không tường minh như *chủ đề (category), quan điểm (sentiment), trạng thái cảm xúc (emotion).* Mô hình của đồ án sẽ xem xét bài đăng có 3 thuộc tính không tường minh là *chủ đề (category), quan điểm (sentiment)* và *trạng thái cảm xúc (emotion),* cụ thể như 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. Việc xác định *chủ đề* của bài đăng có thể được thực hiện bằng cách áp dụng các phương pháp của Joachims [5], Ko và Seo [6], Cong et al. [3].

+ *Quan điểm của một bài đăng (sentiment)*: Qua việc đăng bài, người dùng có thể thể hiện *quan điểm* của họ thuộc 3 dạng: tích cực, tiêu cực, trung tính. Do vậy, ước lượng độ tương đồng về *quan điểm* của bài đăng sẽ dựa vào các giá trị *quan điểm* đó. Việc xác định *quan điểm* của bài đăng có thể được thực hiện bằng cách áp dụng các phương pháp của Ohana and Tierney [9], được cải tiến và phát triển bởi Hung and Lin [4].

+ *Trạng thái cảm xúc (emotion)*: Một bài đăng có thể thể hiện: không, một, hoặc nhiều hơn một *trạng thái cảm xúc* của người dùng. Với mỗi *trạng thái cảm xúc* là một tập các hợp ngữ, như: vui sướng, buồn bã, tự hào… Do vậy, ước lượng độ tương đồng về *trạng thái cảm xúc* của bài đăng chính là ước lượng độ tương đồng dữa hai tập các hợp ngữ.

Mỗi thuộc tính *chủ đề, quan điểm* hay *trạng thái cảm xúc* 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 $A_1 = \{a_1^1, a_1^2, \ldots, a_1^m\}, A_2 = \{a_2^1, a_2^2, \ldots, a_2^n\}$ là hai tập các hợp ngữ. Với *m, n* lần lượt là kích thước của A_1, A_2. Gọi *k* là kích thước của giao hai tập hợp A_1, A_2. Theo phương pháp tính toán độ tương đồng giữa các đối tượng được đề xuất bởi H. M. Nguyen và H. T. Nguyen [8], độ tương đồng của A_1, A_2 được ước lượng theo công thức sau:

$$S_{exp}(A_1, A_2) = \frac{2 * |A_1 \cap A_2|}{|A_1| + |A_2|} = \frac{2 * k}{m + n} \quad (2.1)$$

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) = 0.33, kết quả này 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 trạng thái cảm xúc 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 trạng thái cảm xúc của bài đăng.

Gọi e_i, e_j là hai bài đăng có *chủ đề*, *quan điểm* và *trạng thái cảm xúc* của bài đăng lần lượt là Cat_i, Cat_j, Sen_i, Sen_j và Emo_i, Emo_j. Độ tương đồng giữa hai bài đăng e_i, e_j 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:

$$S_{cat}(e_i, e_j) = S_{exp}(Cat_i, Cat_j) \tag{2.2}$$

+ *Ước lượng độ tương đồng quan điểm* của hai bài đăng:

$$S_{sen}(e_i, e_j) = \begin{cases} 1 \; if \; Sen_i = Sen_j \\ 0.5 \; if \; Sen_i = 0.5 \; or \; Sen_j = 0.5 \\ 0 \; otherwise \end{cases} \tag{2.3}$$

+ *Ước lượng độ tương đồng trạng thái cảm xúc* của hai bài đăng là ước lượng độ tương đồng giữa hai tập các hợp ngữ:

$$S_{emo}(e_i, e_j) = S_{exp}(Emo_i, Emo_j) \tag{2.4}$$

+ Ước lượng độ tương đồng giữa hai bài đăng e_i, e_j:

$$\begin{aligned} S_{entry}(e_i, e_j) = {} & w_{cat} * S_{cat}(e_i, e_j) \\ & + w_{sen} * S_{sen}(e_i, e_j) \\ & + w_{emo} * S_{emo}(e_i, e_j) \end{aligned} \tag{2.5}$$

Với w_{cat}, w_{sen}, w_{emo} lần lượt là trọng số của các thuộc tính *chủ đề của bài đăng, quan điểm của bài đăng, trạng thái cảm xúc của bài đăng*, và được xác định theo điều kiện sau:

$$(i) \quad w_{cat} + w_{sen} + w_{emo} = 1 \tag{2.6}$$

Các trọng số w_{cat}, w_{sen}, w_{emo} 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 $S_{entry}(e_i, e_j)$ nằm trong đoạn [0,1]. Nếu $S_{entry}(e_i, e_j)$ càng gần tới 1 thì độ tương đồng giữa hai bài đăng càng cao và ngược lại, $S_{entry}(e_i, e_j)$ càng gần tới 0 thì độ tương đồng giữa hai bài đăng càng thấp.

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

Xét $E_1 = \{e_1^1, e_1^2, \ldots, e_1^m\}, E_2 = \{e_2^1, e_2^2, \ldots, e_2^n\}$ là hai tập các bài đăng. Đồ án sẽ tạo một *tập các bài đăng* bao gồm cả hai *tập các bài đăng* là E_{12} và vector T như sau:

$E_{12} = E_1 + E_2 = \{e^1, e^2, \ldots, e^{m+n}\}$, $T = \{t^1, t^2, \ldots, t^{m+n}\}$. Trong đó:

$$t^i = \min\left(\max\left(S_{entry}(e^i, e_1^k)\right), \max\left(S_{entry}(e^j, e_2^v)\right)\right), \quad k = 1..m, v = 1..n \tag{2.7}$$

Với S_{entry} là độ tương đồng giữa hai bài đăng x và y.

Để ước lượng độ tương đồng giữa hai *tập các bài đăng* E_1 và E_2, đồ án sử dụng các giả thiết sau:

+ Độ lớn của vector T càng lớn thì độ tương đồng giữa E_1 và E_2 càng cao

+ Ước lượng độ tương đồng giữa hai *tập các bài đăng* E_1 và E_2 theo công thức:

$$S_{set}(E_1, E_2) = f_s(T) = f_s(t^1, t^2, \ldots, t^{m+n}) \tag{2.8}$$

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

$$\begin{aligned} &(i) && f_s(0,0,\ldots,0) = 0 \\ &(ii) && f_s(1,1,\ldots,1) = 1 \\ &(iii) && f_s(X_1) \le f_s(X_2) \text{ nếu } ||X_1|| \le ||X_2|| \end{aligned} \tag{2.9}$$

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

$$(1)\ f(x_1, x_2, \dots, x_n) = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.10}$$

$$(2)\ f(x_1, x_2, \dots, x_n) = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} \tag{2.11}$$

Trong thử nghiệm, đồ án sử dụng công thức $f(x_1, x_2, \dots, x_n) = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$.

Ví dụ: Xét 2 *tập các bài đăng* $E_1 = \{e^1, e^2, e^3\}$, $E_2 = \{e^4, e^5\}$. Độ tương đồng giữa mỗi cặp bài đăng của 2 tập hợp thể hiện dưới bảng sau:

Bảng 2.1: So sánh độ tương đồng giữa 3 bài đăng E_1 và 2 bài đăng E_2

	e^1	**e^2**	**e^3**	**max(e^i, E_1)**
e^4	0.15	0.20	0.30	**0.30**
e^5	0.40	0.15	0.15	**0.40**
max(e^i, E_2)	**0.40**	**0.20**	**0.30**	

Tính toán độ tương đồng giữa E_1 và E_2 như sau:

- $E_{12} = E_1 + E_2 = \{e^1, e^2, e^3, e^4, e^5\}$
- $T = \{0.40, 0.20, 0.30, 0.30, 0.40\}$
- Độ tương đồng là: 0.32

Tiếp theo đồ án sẽ trình bày ước lượng độ tương đồng hành vi đăng bài đăng.

2.2.2. Ước lượng độ tương đồng hành vi đăng bài đăng

Gọi S_i^p và S_j^p 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:

$$S_{pos}(i, j) = S_{set}(S_i^p, S_j^p) \tag{2.16}$$

Với S_{set} *(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* S_{pos} *(A, B)* sẽ nằm trong đoạn [0,1]. Nếu S_{pos} *(A, B)* càng gần 1 thì độ tương đồng càng cao và ngược lại, S_{pos} *(A, B)* càng gần 0 thì độ tương đồng càng thấp.

2.2.3. Ước lượng độ tương đồng hành vi thích bài đăng

Gọi S_i^l và S_j^l 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:

$$S_{lik}(i,j) = S_{set}\left(S_i^l, S_j^l\right) \tag{2.17}$$

Với S_{set} *(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* S_{lik} *(A, B)* sẽ nằm trong đoạn [0,1]. Nếu S_{lik} *(A, B)* càng gần 1 thì độ tương đồng càng cao và ngược lại, S_{lik} *(A, B)* càng gần 0 thì độ tương đồng càng thấp.

2.2.4. 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. Việc xác định này có thể được thực hiện bằng cách áp dụng phương pháp của Cavnar và Trenkle [2].

+ 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 đối 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 đối 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 đối với người dùng này. Trường hợp này bài đăng sẽ bị loại bỏ khỏi tập các bài đăng đã *nhận xét/thích nhận xét* của người dùng này

Như vậy, mô hình của đồ án chỉ xét tới các bài đăng được người dùng nhận xét là tích cực hoặc tiêu cực đối với người dùng đó.

Gọi $S_i^{positive}$ và $S_i^{negative}$ lần lượt là tập các bài đăng tích cực và tiêu cực đối với người dùng i. Gọi $S_j^{positive}$ và $S_j^{negative}$ 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 $S_i^{positive}$ và $S_j^{positive}$ 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 $S_i^{negative}$ và $S_j^{negative}$ 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 $S_i^{positive}$ và $S_j^{negative}$ 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 $S_i^{negative}$ và $S_j^{positive}$ 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:

$$\begin{aligned} S_{cmt}(i,j) = \min(1, \max(0, & S_{set}(S_i^{positive}, S_j^{positive}) \\ & + S_{set}(S_i^{negative}, S_j^{negative}) \\ & - S_{set}(S_i^{positive}, S_j^{negative}) \\ & - S_{set}(S_i^{negative}, S_j^{positive}))) \end{aligned} \quad (2.18)$$

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

2.2.5. 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 3 hành vi *đăng bài đăng*, *thích bài đăng* và *nhận xét/thích nhận xét bài đăng* của người dùng.

Gọi w_{pos}, w_{lik}, w_{cmt} 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* và được xác định với điều kiện sau:

$$(i)\ \ w_{pos} + w_{lik} + w_{cmt} = 1 \qquad (2.20)$$

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

$$S_{user}(i,j) = w_{pos} * S_{pos}(i,j) + w_{lik} * S_{lik}(i,j) + w_{cmt} * S_{cmt}(i,j) \qquad (2.21)$$

Với $S_{pos}(i, j)$, $S_{lik}(i, j)$, $S_{cmt}(i, j)$ 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* của hai người dùng i và người dùng j. Các trọng số w_{pos}, w_{lik}, w_{cmt} 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 $S_{user}(i, j)$ sẽ nằm trong đoạn [0,1]. Nếu $S_{user}(i, j)$ càng gần 1 thì độ tương đồng của người dùng dựa vào hành vi càng cao và ngược lại, $S_{user}(i, j)$ càng gần 0 thì độ tương đồng của người dùng dựa vào hành vi càng thấp.

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

2.3. 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: giữa các bài đăng, giữa các hội nhóm/cộng đồng, hành

vi *đăng bài đăng*, hành vi *thích bài đăng*, hành vi *nhận xét/thích nhận xét bài đăng* và giữa hai người dùng.

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

Chương 3 sẽ trình bày các nội dung sau:

- Xây dựng bộ dữ liệu thử nghiệm
- Phương pháp đánh giá thuật toán
- Cài đặt thuật toán
- Khảo sát bộ trọng số
- Kết quả đánh giá

3.1. Xây dựng bộ dữ liệu thử nghiệm

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 [15]. Dữ liệu thực tế lấy được bao gồm:

+ *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 của người dùng gồm:

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

Đồ án sử dụng REST API [14] và thư viện hỗ trợ Twitter4J [12] cài đặt với ngôn ngữ Java để lấy dữ liệu thực tế người dùng trên trang mạng xã hội Twitter. Dữ liệu lấy được chia làm 5 bộ dữ liệu cụ thể như sau:

Bảng 3.1: Số lượng mẫu thu thập

Bộ dữ liệu thu thập	Số lượng mẫu
Bộ dữ liệu học cho xác định chủ đề của bài đăng	1010
Bộ dữ liệu học cho xác định trạng thái cảm xúc của bài đăng	1010
Bộ dữ liệu học cho xác định quan điểm của bài đăng	500
Bộ dữ đánh giá	500
Bộ dữ liệu cho khảo sát trọng số	500

Đồ án sử dụng các công cụ và thư viện hỗ trợ trong quá trình xây dựng bộ dữ liệu thử nghiệm như:

(*): sử dụng phương pháp điều tra bằng phiếu đánh giá, tạo mẫu lấy đánh giá của người dùng mạng bằng công cụ Google Form [16]. Thời gian lấy đánh giá của người dùng mạng trong khoảng từ 8/10/2016 đến 28/10/2016.

(**): sử dụng thư viện hỗ trợ là LingPipe [11] cài đặt với ngôn ngữ Java kết hợp với từng bộ dữ liệu học để xác định chủ đề, quan điểm, trạng thái cảm xúc cho bài đăng.

3.1.1. Bộ dữ liệu học cho xác định chủ đề của bài đăng

Đối với bộ dữ liệu này, đồ án xây dựng bộ dữ liệu thực tế gồm 1010 bài đăng của người dùng trên mạng xã hội Twitter trong khoảng thời gian từ 8/7/2016 đến 8/10/2016.

+ Lấy đánh giá cho bộ dữ liệu bởi người dùng mạng (*): Các bài đăng sẽ được người dùng mạng đánh giá theo từng chủ đề riêng biệt như thể hiện ở bảng sau:

Bảng 3.2: Tập các chủ đề 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)		

Các bài đăng được đánh giá chủ đề 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*.

+ *Đánh giá bộ dữ liệu (**):* Đồ án thực hiện phân loại và xác định ít nhất một chủ đề phù hợp với mỗi bài đăng trong bộ dữ liệu. Mỗi bài đăng được coi là một *mẫu đánh giá*, nếu có kết quả xác định được tương đương với kết quả đánh giá bởi người dùng mạng thì được coi là một *mẫu đánh giá mẫu học chủ đề đúng*. Kết quả đánh giá được tính toán theo công thức sau:

$$\text{Kết quả đánh giá} = \frac{\sum \text{số mẫu đánh giá đúng}}{\sum \text{số mẫu đánh giá}} \times 100\% \qquad (3.1)$$

Bảng 3.3: Kết quả đánh giá bộ dữ liệu học cho xác định chủ đề của bài đăng

Số mẫu học	**Số mẫu đánh giá**	**Kết quả đánh giá**
320	**250**	66.8%
540		68.4%
760		**72.4%**

➔ Kết quả đồ án chọn ra được bộ dữ liệu học cho xác định chủ đề của bài đăng có kết quả đánh giá cao nhất *72.4%* là bộ dữ liệu gồm 760 mẫu.

3.1.2. Bộ dữ liệu học cho xác định trạng thái cảm xúc của bài đăng

Đối với bộ dữ liệu này, đồ án sử dụng bộ dữ liệu gồm 1010 bài đăng đã xây dựng được và thực hiện các bước tương tự như khi chuẩn bị bộ dữ liệu học cho xác định chủ đề của bài đăng.

+ *Lấy đánh giá cho bộ dữ liệu bởi người dùng mạng (*):* Các bài đăng sẽ được người dùng mạng đánh giá theo từng trạng thái cảm xúc riêng biệt như thể hiện ở bảng 3.4. Các bài đăng được đánh giá trạng thái cảm xúc là Other (Khác) sẽ loại bỏ khỏi bộ dữ liệu này. Mỗi bài đăng được phân loại tương ứng với từng trạng thái cảm xúc sẽ được lưu vào mỗi file chỉ bao gồm nội dung của bài đăng và được coi là *một mẫu học* cho trạng thái cảm xúc đó của *bộ dữ liệu học cho xác định trạng thái cảm xúc của bài đăng*.

+ Đánh giá bộ dữ liệu (**): Đồ án thực hiện phân loại và xác định ít nhất một trạng thái cảm xúc phù hợp với mỗi bài đăng trong bộ dữ liệu. Mỗi bài đăng được coi là một mẫu đánh giá, nếu có kết quả xác định được tương đương với kết quả đánh giá bởi người dùng mạng thì được coi là một *mẫu đánh giá mẫu học trạng thái cảm xúc đúng*. Kết quả đánh giá được tính toán theo công thức (3.1), được thể hiện ở bảng 3.5.

Bảng 3.4: Tập các trạng thái cảm xúc của bài đăng

No	Emotion	No	Emotion	No	Emotion
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 (Cảm xúc khác)
6	Hope (Hi vọng)	12	Disappointed (Thất vọng)		

Bảng 3.5: Kết quả đánh giá bộ dữ liệu học cho xác định trạng thái cảm xúc của bài đăng

Số mẫu học	Số mẫu đánh giá	Kết quả đánh giá
320	**250**	82.0%
540		82.0%
760		**83.6%**

➔ Kết quả đồ án chọn ra được bộ dữ liệu học cho xác định chủ đề của bài đăng có kết quả đánh giá cao nhất *83.6%* là bộ dữ liệu gồm 760 mẫu.

3.1.3. Bộ dữ liệu học cho xác định quan điểm của bài đăng

Đối với bộ dữ liệu này, đồ án chọn ra 500 bài đăng trong tập 1010 bài đăng đã chuẩn bị và thực hiện các bước tương tự như khi chuẩn bị bộ dữ liệu học cho xác định chủ đề của bài đăng.

+ Lấy đánh giá cho bộ dữ liệu bởi người dùng mạng (*): Các bài đăng sẽ được người dùng mạng đánh giá và được phân loại tương ứng theo 3 dạng quan điểm: tích cực, tiêu cực, trung tính. Mỗi bài đăng 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.*

+ Đánh giá bộ dữ liệu (**): Đồ án thực hiện phân loại và xác định một quan điểm duy nhất, phù hợp nhất với mỗi bài đăng trong bộ dữ liệu. Mỗi bài đăng được coi là một mẫu đánh giá, nếu có kết quả xác định được tương đương với kết quả đánh giá bởi người dùng mạng thì được coi là một *mẫu đánh giá mẫu học quan điểm đúng*. Kết quả đánh giá được tính toán theo công thức (3.1).

Bảng 3.6: Kết quả đánh giá bộ dữ liệu học cho xác định quan điểm của bài đăng

Số mẫu học	**Số mẫu đánh giá**	**Kết quả đánh giá**
300	**100**	78.0%
400		**88.0%**

➔ Kết quả sau các bước đồ án chọn ra được bộ dữ liệu học cho xác định chủ đề của bài đăng có kết quả đánh giá cao nhất *88.0%* là bộ dữ liệu gồm 760 mẫu.

3.1.4. Bộ dữ liệu đánh giá

Đối với bộ dữ liệu này, đồ án xây dựng bộ dữ liệu thực tế gồm 500 *mẫu đánh giá* trong khoảng thời gian từ 8/7/2016 đến 8/10/2016.

Mỗi *mẫu đánh giá* là một file dữ liệu bao gồm dữ liệu thực tế của người dùng và bài đăng được tương tác bằng các hành vi *post*, *like*, *comment*, được lưu lại gồm:

+ Mã số mẫu đánh giá (id): là duy nhất với mỗi *mẫu đánh giá*.

+ Giá trị của mẫu đánh giá (value): nhận giá trị là 1 hoặc 2.

+ Dữ liệu của 3 người dùng khác nhau gọi tên: User A, User B và User C.

Trong đó, *giá trị của mẫu đánh giá (value)* được xác định cụ thể như sau:

- Nếu độ tương đồng giữa hai người dùng User A và User B cao hơn độ tương đồng giữa hai người dùng User A và User C, thì *value* nhận giá trị là 1.
- Nếu độ tương đồng giữa hai người dùng User A và User C cao hơn độ tương đồng giữa hai người dùng User A và User B thì *value* nhận giá trị là 2.

Mô tả cấu trúc *mẫu đánh giá* được thể hiện ở bảng 3.7. Một mẫu đánh giá thực tế được thể hiện ở bảng 3.8.

+ Lấy đánh giá cho bộ dữ liệu bởi người dùng mạng (*): Mỗi *mẫu đánh giá* được người dùng mạng đánh giá theo nội dung câu hỏi là: đánh giá độ tương đồng của 3 người dùng gọi tên *A, B, C* dựa trên các bài đăng theo hành vi của 3 người dùng đó xem người dùng *A* tương đồng với người dùng *B* hơn hay người dùng *C* hơn. Sau khi có được toàn bộ kết quả đánh giá, đồ án thực hiện lưu mỗi *kết quả đánh giá* dưới dạng file, tương tự như đã mô tả ở bảng 3.7. *Giá trị của mẫu đánh giá (value)* được lưu lại với giá trị là 1 nếu kết quả đánh giá là *B* tức người dùng *A* tương đồng với người dùng *B* hơn người dùng *C*, được lưu lại với giá trị là 2 nếu kết quả đánh giá là *C* tức người dùng *A* tương đồng với người dùng *C* hơn người dùng *B*.

➔ Kết quả đồ án thu được 500 *mẫu đánh giá* đã được xác định *giá trị (value).*

+ Tiền xử lý để làm đầu vào cho thuật toán (**): Ở mỗi *mẫu đánh giá*, đồ án thực hiện xác định ít nhất một chủ đề, ít nhất một trạng thái cảm xúc phù hợp cho từng bài đăng, một quan điểm duy nhất, phù hợp nhất cho từng bài đăng của *bộ dữ liệu đánh giá*. Nội dung bài đăng, chủ đề, quan điểm, trạng thái cảm xúc cách nhau một dấu "<>", nếu nhiều hơn một chủ đề, trạng thái cảm xúc thì các chủ đề, trạng thái cảm xúc sẽ cách nhau một dấu "#". Mô tả cấu trúc mẫu đánh giá sau khi xử lý như ở bảng 3.9.

Đồ án áp dụng thuật toán đối với 500 *mẫu đánh giá* sau khi xử lý. Thuật toán sẽ xác định *kết quả của mẫu đánh giá (result)* là 1 nếu độ tương đồng giữa hai người dùng User A và User B cao hơn độ tương đồng giữa hai người dùng User A và User C, xác định *kết quả mẫu đánh giá (result)* là 2 nếu độ tương đồng giữa hai người dùng User A và User C cao hơn độ tương đồng giữa hai người dùng User A và User B.

➔ Kết quả đồ án thu được 500 *mẫu đánh giá* đã được xác định *kết quả (result).*

Bảng 3.7: Mô tả cấu trúc mẫu đánh giá

id	
User A	Post entries
	Like entries
	Comment entries
User B	Post entries
	Like entries
	Comment entries
User C	Post entries
	Like entries
	Comment entries
value	

Bảng 3.8: Một mẫu đánh giá thực tế

id	1
User A	I just got a FREE trial bag of Petcurean premium pet food, I think your furry friend would love one too! https://t.co/SUOASa22Nh … Enter for your chance to win McDonald's for a Year! https://t.co/eS1T78OKlL
	NLB 18u wins game 1 in New Orleans 3-0 behind Nicky Agosto's gem. Grant Rowell earned the save. Game 2 underway vs… https://t.co/R7YZGcXgdD … This hurts my heart so terribly. Such a sweet, bright soul who impacted so many, even his opponents. #JoFez16 https://t.co/Z6TXTKMIfA
	NLB 18u wins game 1 in New Orleans 3-0 behind Nicky Agosto's gem. Grant Rowell earned the save. Game 2 underway vs… https://t.co/R7YZGcXgdD
User B	My all access pass for Michael Jackson's Dangerous Tour Rehearsals in 1992. Such a great quote about #life.… https://t.co/FxLN7cVIV1 … Please check out my interview with IHeart Radio this morning. We chatted about my book #MichaelandMe #MichaelJackson https://t.co/R97KvdIoTn
	@ImShanaMangatal it heart warming to know he is being credited. I'm in Vegas attending Michael Jackson One … @ImShanaMangatal the question is do you put this hat on and proceed to dance like Mike? I would. Legends don't die.#MJforever
	Thank u Shana 4 writing about your relationship w/my son. I support you. Now people will know the real man we all ♥ https://t.co/Mt3fvfiqYE … @ImShanaMangatal why Michael Bush don't know nothing about your relationship with @michaeljackson? Why
User C	Goodbye twitter!!! Im shutting down my account until the debates are over. Good night and good luck. #BlackLivesMatter #ImWithHer … RT @ReignOfApril: Look how much the story has changed. The original said serving a warrant & they saw a gun. Now it's conducting surveillan…
	My favorite happy moment is when my son, without asking, runs over to me and gives me a big hug. I wish I could bottle that feeling. :) … I love seeing guys that be extra obsessed with their girlfriends and make them feel like the only girl that matters, as they should.
	Getting used to playing #HardcoreTDM with my American Pretzel @stephie_05 #BlackOps3 #CallOfDuty… https://t.co/ieakgs0Pre
value	?

Bảng 3.9: Mô tả cấu trúc mẫu đánh giá làm đầu vào cho thuật toán

id							
User A	Post entries	<>	Categories	<>	Sentiment	<>	Emotions
	Like entries	<>	Categories	<>	Sentiment	<>	Emotions
	Comment entries	<>	Categories	<>	Sentiment	<>	Emotions
User B	Post entries	<>	Categories	<>	Sentiment	<>	Emotions
	Like entries	<>	Categories	<>	Sentiment	<>	Emotions
	Comment entries	<>	Categories	<>	Sentiment	<>	Emotions
User C	Post entries	<>	Categories	<>	Sentiment	<>	Emotions
	Like entries	<>	Categories	<>	Sentiment	<>	Emotions
	Comment entries	<>	Categories	<>	Sentiment	<>	Emotions
value							

3.1.5. Bộ dữ liệu cho khảo sát trọng số cho ước lượng độ tương đồng giữa các bài đăng

Từ bộ dữ liệu đánh giá đã xây dựng được như trên, đồ án thực hiện chọn ra 500 bài đăng trong tập các bài đăng của 500 *mẫu đánh giá* làm đại diện để chuẩn bị *bộ dữ liệu cho khảo sát trọng số* sử dụng trong công thức ước lượng độ tương đồng giữa các bài đăng. Sau đó thực hiện tạo 500 *mẫu đánh giá* cho *bộ dữ liệu cho khảo sát trọng số* từ 500 bài đăng đã chọn. Mỗi *mẫu đánh giá* được lưu dưới dạng file bao gồm:

+ *Mã số mẫu đánh giá (id)*: là duy nhất với mỗi *mẫu đánh giá*.

+ *Giá trị của mẫu đánh giá (value)*: nhận giá trị là 1 hoặc 2.

+ Gồm nội dung của 3 bài đăng khác nhau gọi tên: Entry X, Entry Y, Entry Z.

Trong đó, *giá trị của mẫu đánh giá (value)* được xác định cụ thể như sau:

- Nếu độ tương đồng giữa hai bài đăng Entry X và Entry Y cao hơn độ tương đồng giữa hai bài đăng Entry X và Entry Z, thì *value* nhận giá trị là 1.
- Nếu độ tương đồng giữa hai bài đăng Entry X và Entry Z cao hơn độ tương đồng giữa hai bài đăng Entry X và Entry Y thì *value* nhận giá trị là 2.

+ *Lấy đánh giá bộ dữ liệu bởi người dùng mạng (*)*: Mỗi *mẫu đánh giá* được người dùng mạng đánh giá theo nội dung câu hỏi là: đánh giá độ tương đồng của 3 bài đăng gọi tên *X, Y, Z* xem bài đăng *X* tương đồng với bài đăng *Y* hơn hay tương đồng với bài đăng *Z* hơn. Sau khi có được toàn bộ kết quả đánh giá, đồ án thực hiện

lưu mỗi *kết quả đánh giá* dưới dạng file, tương tự như đã mô tả ở bảng 3.10. *Giá trị của mẫu đánh giá (value)* được lưu lại với giá trị là 1 nếu kết quả đánh giá là *Y* tức bài đăng *X* tương đồng với bài đăng *Y* hơn bài đăng *Z*, được lưu lại với giá trị là 2 nếu kết quả đánh giá là *Z* tức bài đăng *X* tương đồng với bài đăng *Z* hơn bài đăng *Y*.

➔ Kết quả đồ án thu được 500 *mẫu đánh giá* đã được xác định *giá trị (value)*, được coi là 500 *mẫu đánh giá thực tế* của *bộ dữ liệu cho khảo sát trọng số*.

Bảng 3.10: Một mẫu đánh giá thực tế của bộ dữ liệu cho khảo sát trọng số

id	2
Entry X	Schoola: Save on kids clothes & 40% funds school programs. $20 free if you use my link. Awesome, right?! https://t.co/qMW59V4bwD
Entry Y	Today we work on the underestimated AB WHEEL - the cheapest tool to use for a six pack! Get one for your gym bag! https://t.co/FNWGdZU5L1
Entry Z	Just tracked down the owner who'd left their dog in their car! Please don't leave your dog in the car- especially on a day like today! ?
value	?

+ Tiền xử lý để làm đầu vào cho khảo sát trọng số (**): Ở mỗi *mẫu đánh giá*, đồ án thực hiện xác định ít nhất một chủ đề, ít nhất một trạng thái cảm xúc phù hợp cho từng bài đăng, một quan điểm duy nhất, phù hợp nhất cho từng bài đăng của *bộ dữ liệu cho khảo sát trọng số*. Nội dung bài đăng, chủ đề, quan điểm, trạng thái cảm xúc cách nhau một dấu "<>", nếu nhiều hơn một chủ đề, trạng thái cảm xúc thì các chủ đề, trạng thái cảm xúc sẽ cách nhau một dấu "#".

Bảng 3.11: Một mẫu đánh giá thực tế sau khi xử lý

id	2
Entry X	Schoola: Save on kids clothes & 40% funds school programs. $20 free if you use my link. Awesome, right?! https://t.co/qMW59V4bwD <>IT#Economy#Religion<>Joy#Hope#Happyfor<>POSITIVE
Entry Y	Today we work on the underestimated AB WHEEL - the cheapest tool to use for a six pack! Get one for your gym bag! https://t.co/FNWGdZU5L1 <>IT#Sport#Economy<>Joy#Hope#Happyfor<>NEUTRAL
Entry Z	Just tracked down the owner who'd left their dog in their car! Please don't leave your dog in the car- especially on a day like today!? <>Religion#Politic#Economy<>Confused#Sadness#Hope<>POSITIVE
value	1

Ở *mẫu đánh giá* mô tả ở bảng 3.11, *value* nhận giá trị là 1 do độ tương đồng giữa hai bài đăng Entry X và Entry Y cao hơn độ tương đồng giữa hai bài đăng Entry X và Entry Z.

Tiếp theo đồ án sẽ trình bày phương pháp đánh giá thuật toán.

3.2. Phương pháp đánh giá

Như đã trình bày ở mục trước, đồ án đã chuẩn bị *bộ dữ liệu đánh giá làm dữ liệu để người dùng mạng đánh giá*. Kết quả đồ án có được sau khi lấy kết quả đánh giá của người dùng mạng là *bộ dữ liệu đánh giá* gồm 500 *mẫu đánh giá* đã được xác định *giá trị (value)*. Tập 500 *giá trị (value)* của *bộ dữ liệu đánh giá* được coi là *tập kết quả đánh giá thực tế*.

Cũng với 500 *mẫu đánh giá* của *bộ dữ liệu đánh giá*, đồ án xử để có được 500 *mẫu đánh giá* làm dữ liệu đầu vào cho thuật toán. Kết quả đồ án thu được 500 *mẫu đánh giá* đã được xác định *kết quả (result)*. Tập 500 *kết quả (result)* của *bộ dữ liệu đánh giá* này được coi là *tập kết quả đánh giá bởi thuật toán*.

Như vậy, đồ án thu được 2 *tập kết quả đánh giá* có cùng số lượng 500 *kết quả đánh giá* là: *tập kết quả đánh giá thực tế (bởi người dùng mạng)* và *tập kết quả đánh giá bởi thuật toán*. Dựa vào *value* và *result* của cùng một *mẫu đánh giá* trong 2 tập kết quả này, đồ án xác định *mẫu đánh giá đúng*.

Một *mẫu đánh giá đúng* là *mẫu đánh giá* có *value* và *result* trong *tập kết quả đánh giá thực tế* và *tập kết quả đánh giá bởi thuật toán* là giống hệt nhau. Ví dụ: xét một *mẫu đánh giá X*, có *value* ở *tập kết quả đánh giá thực tế* là 1, có *result* ở *tập kết quả đánh giá bởi thuật toán* cũng là 1, thì *X* được xác định là một *mẫu đánh giá đúng*.

Đồ án thực hiện so sánh lần lượt 2 *tập kết quả đánh giá* và xác định tổng số *mẫu đánh giá đúng*. Dựa trên số lượng *mẫu đánh giá đúng* và tổng số *mẫu đánh giá*, đồ án xác định *độ chính xác* của thuật toán theo công thức (3.1).

Độ chính xác của thuật toán thể hiện tỷ lệ phần trăm của sự giống nhau giữa kết quả đánh giá thực hiện bởi máy tính và kết quả đánh giá thực tế của con người. Do vậy, *độ chính xác* càng cao thể hiện độ tin cậy của mô hình thuật toán càng cao và có thể áp dụng cho ứng dụng thực tế.

Tiếp theo đồ án sẽ trình bày cài đặt thuật toán ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội.

3.3. Cài đặt thuật toán

3.3.1. Mô hình cài đặt

Đồ án cài đặt mô hình thuật toán gồm có:

+ *Đầu vào (input):* là dữ liệu của 3 người dùng gọi tên User A, User B, User C lấy từ *mạng xã hội* gồm các *bài đăng* được tương tác bởi 3 người dùng này.

+ *Đầu ra (output):* là 1 nếu độ tương đồng giữa User A và User B cao hơn độ tương đồng giữa User A và User C, là 2 nếu độ tương đồng giữa User A và User C cao hơn độ tương đồng giữa User A và User B.

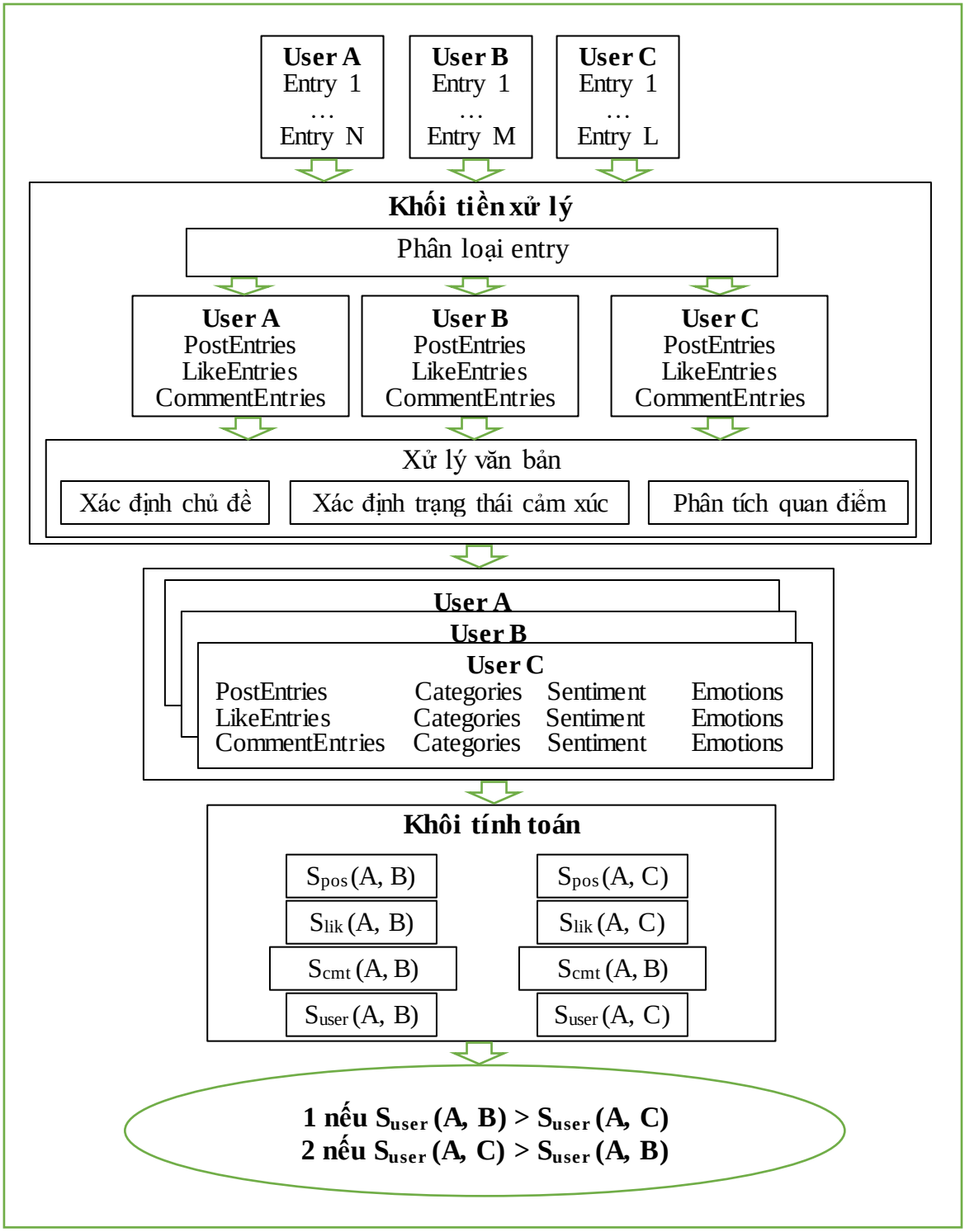
+ *Khối tiền xử lý:* là khối xử lý dữ liệu đầu vào, trong đó:

o Phân loại entry: thực hiện phân loại bài đăng, nhận xét của người dùng theo 3 hành vi *đăng bài đăng*, *thích bài đăng*, *nhận xét bài đăng*. Kết quả thu được 3 khối dữ liệu của 3 người dùng đã được phân loại để tiếp tục xử lý văn bản.

o *Xác định chủ đề:* thực hiện xác định ít nhất một chủ đề phù hợp cho tất cả các bài đăng trong 3 khối dữ liệu của 3 người dùng, sử dụng *bộ dữ liệu học cho xác định chủ đề bài đăng*. Dữ liệu đầu vào là một bài đăng bất kỳ của mỗi người dùng, kết quả đầu ra chuỗi ký tự ghi nội dung bài đăng và ít nhất một chủ đề phù hợp, nếu có nhiều hơn một chủ đề được xác định, các chủ đề cách nhau một dấu "#".

o *Xác định trạng thái cảm xúc:* thực hiện xác định ít nhất một trạng thái cảm xúc phù hợp cho tất cả các bài đăng trong 3 khối dữ liệu của 3 người dùng, sử dụng *bộ dữ liệu học cho xác định trạng thái cảm xúc của bài đăng*. Dữ liệu đầu vào là một bài đăng bất kỳ của mỗi người dùng, kết quả đầu ra chuỗi ký tự ghi nội dung bài đăng và ít nhất một trạng thái cảm xúc phù hợp, nếu có nhiều hơn một trạng thái cảm xúc được xác định, các trạng thái cảm xúc cách nhau một dấu "#".

o *Phân tích quan điểm:* thực hiện xác định quan điểm của các nhận xét của từng bài đăng được người dùng nhận xét để xác định duy nhất một quan điểm phù hợp cho từng bài đăng này. Thực hiện xác định duy nhất một quan điểm phù hợp cho từng bài đăng được người dùng đăng, thích. Sử dụng *bộ dữ liệu học cho xác định quan điểm của bài đăng*. Dữ liệu đầu vào là một bài đăng bất kỳ (và các nhận xét nếu là bài đăng được người dùng nhận xét) của mỗi người dùng, kết quả đầu ra là chuỗi ký tự ghi nội dung bài đăng và quan điểm được xác định.



Hình 3.1: Mô hình cài đặt thuật toán

Khối tiền xử lý cho kết quả là 3 khối dữ liệu của 3 người dùng có các bài đăng đã được xác định chủ đề và trạng thái cảm xúc.

+ *Khối tính toán:* ước lượng độ tương đồng giữa từng cặp người dùng trong 3 người dùng gọi tên User A, User B, User C gồm: ước lượng độ tương đồng hành vi *đăng bài đăng*, ước lượng độ tương đồng hành vi *thích bài đăng*, ước lượng độ tương đồng hành vi *nhận xét bài đăng* và ước lượng độ tương đồng của người dùng dựa vào hành vi. Dữ liệu đầu vào là từng cặp dữ liệu của từng cặp người dùng đã được xử lý, kết quả đầu ra: là 1 nếu độ tương đồng giữa User A và User B cao hơn độ tương đồng giữa User A và User C, là 2 nếu độ tương đồng giữa User A và User C cao hơn độ tương đồng giữa User A và User B.

3.3.2. Thư viện hỗ trợ

- LingPipe [11]: là một thư viện phần mềm sử dụng cho xử lý ngôn ngữ tự nhiên, cung cấp nhiều bộ công cụ hỗ trợ cho việc xây dựng ứng dụng xử lý ngôn ngữ tự nhiên cho nhiều ngôn ngữ, nhiều thể loại, nhiều loại ứng dụng.

Sử dụng LingPipe *phân loại văn bản (phân loại các tài liệu theo chủ đề, quan điểm xác định)*: LingPipe cung cấp một phương tiện dễ dàng cho phân loại, gồm có những ví dụ về phân loại văn bản đã được tạo ra bởi con người, và các phương thức học theo mô hình ngôn ngữ kí tự - một mô hình xây dựng bộ máy phân loại tự động. LingPipe cho kết quả phân loại có thể là một chủ đề, quan điểm phù hợp nhất hoặc nhiều hơn một chủ đề, quan điểm trong tập chủ đề, quan điểm cho trước.

Sử dụng LingPipe *phân tích trạng thái cảm xúc*: LingPipe hỗ trợ gán các nhãn cảm xúc cho một văn bản. Công việc gán nhãn sẽ giống như phân loại văn bản khi coi công việc gán nhãn là công việc phân loại văn bản với tập các cảm xúc cho trước. LingPipe cho kết quả phân tích có thể là một cảm xúc phù hợp nhất hoặc nhiều hơn một cảm xúc trong tập cảm xúc cho trước.

- Twitter4J [12]: là một thư viện Java hỗ trợ lập trình viên tương tác có giới hạn với CSDL của Twitter bao gồm: lấy thông tin người dùng, cập nhật trạng thái… thông qua các lời gọi sẵn có.

Tương ứng với mỗi lời gọi (kèm với tham số) được Twitter cung cấp, Twitter4J cung cấp một phương thức trợ giúp lập trình viên. Một số lời gọi [14] và phương thức tương ứng của Twitter4J [13] mà đồ án đã sử dụng để lấy dữ liệu thực tế của người dùng khi chuẩn bị các bộ dữ liệu thử nghiệm cho thuật toán như:

+ *GET statuses/user_timeline – getUserTimeline (idUser/screenName):* trả về một tập có thể lên tới 3200 *bài đăng* mới nhất được đăng bởi người dùng đang nhắc tới (theo tham số *idUser* hoặc *screenName*).

+ *GET friends/list – getFriendsList (idUser/screenName):* trả về một tập 20 người dùng được theo dõi bởi *người dùng* đang nhắc tới (theo tham số *idUser* hoặc *screenName*) hay được gọi là 20 bạn bè của người dùng đang nhắc tới.

+ *GET favorites/list – getFavorites (idUser/screenName):* trả về một tập 20 bài đăng mới nhất được thích bởi người dùng đang nhắc tới hoặc người dùng quyết định bởi tham số *idUser* hoặc *screenName*.

Tiếp theo đồ án sẽ trình bày cách khảo sát bộ trọng số với bộ dữ liệu đã xây dựng.

3.4. Khảo sát bộ trọng số

Trong chương 2 đồ án đã trình bày công thức (2.5): ước lượng độ tương đồng giữa hai bài đăng, sử dụng bộ trọng số gồm có:

+ w_{cat}: Trọng số thuộc tính *chủ đề của bài đăng*.

+ w_{sen}: Trọng số thuộc tính *quan điểm của bài đăng*.

+ w_{emo}: Trọng số thuộc tính *trạng thái cảm xúc của bài đăng*.

Và công thức (2.21): ước lượng độ tương đồng của người dùng dựa vào hành vi, sử dụng bộ trọng số gồm có:

+ w_{pos}: Trọng số độ tương đồng hành vi *đăng bài đăng (post)*.

+ w_{lik}: Trọng số độ tương đồng hành vi *thích bài đăng (like)*.

+ w_{cmt}: Trọng số độ tương đồng hành vi *nhận xét bài đăng (comment)*.

Các bộ trọng số này là khác nhau đối với các bộ dữ liệu khác nhau. Bộ trọng số tốt nhất là bộ trọng số được áp dụng sẽ cho ra độ chính xác cao nhất khi chạy trên *bộ dữ liệu đánh giá* thực tế. Đồ án sẽ thực hiện khảo sát để xác định bộ trọng số tốt nhất khi chạy trên 500 *mẫu đánh giá* đã thu thập trên mạng xã hội Twitter.

3.4.1. Khảo sát bộ trọng số cho ước lượng độ tương đồng giữa các bài đăng

Đồ án thực hiện xây dựng các bộ 3 giá trị cho w_{cat}, w_{sen}, w_{emo} bằng cách sử dụng 2 vòng lặp lồng nhau với w_{cat}:0.1−>1, w_{sen}:0.1−>1, w_{emo} = 1-w_{cat}-w_{sen}, sau mỗi vòng lặp w_{cat} sẽ tăng lên 0,1, w_{sen} sẽ tăng lên 0.1.

Đối với mỗi bộ 3 giá trị, đồ án áp dụng vào công thức (2.5) để ước lượng độ tương đồng giữa các bài đăng trong 500 *mẫu đánh giá* đã được xử lý, từ đó xác định được từng *kết quả mẫu đánh giá (result)* cho 500 *mẫu đánh giá này*. 500 *mẫu đánh giá* đã được xác định *kết quả mẫu đánh giá (result)* được coi là 500 *mẫu đánh giá bởi thuật toán* của *bộ dữ liệu cho khảo sát trọng số*. Kết hợp với 500 *mẫu đánh giá thực tế* đã trình bày ở mục 3.1.5, đồ án thực hiện tính *độ chính xác* theo công thức (3.1) và thống kê lại kết quả theo từng bộ 3 giá trị ở bảng 3.12, với *mẫu đánh giá đúng* là *mẫu đánh giá* có *value* và *result* trong hai tập 500 *mẫu đánh giá thực tế* và *mẫu đánh giá bởi thuật toán* của *bộ dữ liệu cho khảo sát trọng số* là tương đương nhau.

Bảng 3.12: Bảng thống kê kết quả khảo sát bộ trọng số w_{cat}, w_{sen}, w_{emo}

w_{cat}	w_{sen}	w_{emo}	Độ chính xác (%)
0.1	*0.1*	*0.8*	*80.8*
0.1	*0.2*	*0.7*	*80.6*
0.1	*0.3*	*0.6*	*81.2*
0.1	*0.4*	*0.5*	*81.4*
0.1	*0.5*	*0.4*	*77.6*
0.1	*0.6*	*0.3*	*77.4*
0.1	*0.7*	*0.2*	*77.4*
0.1	*0.8*	*0.1*	*76.4*
0.2	*0.1*	*0.7*	*76.8*
0.2	*0.2*	*0.6*	*81.4*
0.2	*0.3*	*0.5*	*79.6*
0.2	*0.4*	*0.4*	*81.2*
0.2	*0.5*	*0.3*	*78.2*
0.2	*0.6*	*0.2*	*76.6*
0.2	*0.7*	*0.1*	*73.0*
0.3	*0.1*	*0.6*	*77.4*
0.3	*0.2*	*0.5*	*77.8*
0.3	**0.3**	**0.4**	**83.2**
0.3	*0.4*	*0.3*	*80.2*
0.3	*0.5*	*0.2*	*74.2*
0.3	*0.6*	*0.1*	*73.0*
0.4	*0.1*	*0.5*	*76.6*
0.4	*0.2*	*0.4*	*75.2*
0.4	*0.3*	*0.3*	*78.2*

Bảng 3.12: (tiếp)

w_{cat}	w_{sen}	w_{emo}	Độ chính xác (%)
0.4	*0.4*	*0.2*	*77.2*
0.4	*0.5*	*0.1*	*73.6*
0.5	*0.1*	*0.4*	*72.2*
0.5	*0.2*	*0.3*	*72.0*
0.5	*0.3*	*0.2*	*73.2*
0.5	*0.4*	*0.1*	*76.0*
0.6	*0.1*	*0.3*	*69.0*
0.6	*0.2*	*0.2*	*71.4*
0.6	*0.3*	*0.1*	*71.2*
0.7	*0.1*	*0.2*	*68.6*
0.7	*0.2*	*0.1*	*71.0*
0.8	*0.1*	*0.1*	*71.2*

Từ bảng thống kê, đồ án xác định giá trị cho bộ trọng số w_{cat}, w_{sen}, w_{emo} lần lượt là *0.3, 0.3, 0.4* cho *độ chính xác* cao nhất *83.2%* là bộ trọng số tốt nhất. Đồ án tiếp tục sử dụng bộ giá trị của bộ trọng số này cho khảo sát bộ trọng số w_{pos}, w_{lik}, w_{cmt} dưới đây.

3.4.2. Khảo sát bộ trọng số cho ước lượng độ tương đồng của người dùng dựa vào hành vi

Đồ án thực hiện xây dựng các bộ 3 giá trị cho w_{pos}, w_{lik}, w_{cmt} bằng cách sử dụng 2 vòng lặp lồng nhau với w_{pos}:0.1−>1, w_{lik}:0.1−>1, w_{cmt} = 1-w_{pos}-w_{lik}, sau mỗi vòng lặp w_{pos} sẽ tăng lên 0.1, w_{lik} sẽ tăng lên 0.1.

Đối với mỗi bộ 3 giá trị, đồ án áp dụng vào công thức (2.21) để ước lượng độ tương đồng của người dùng trên 500 *mẫu đánh giá* của *bộ dữ liệu đánh giá*, tính *độ chính xác* theo công thức (3.1) và thống kê lại kết quả ở bảng 3.13 như sau:

Bảng 3.13: Mô tả bảng thống kê kết quả độ chính xác theo bộ 3 giá trị trọng số w_1, w_2, w_3

w_{pos}	w_{lik}	w_{cmt}	Độ chính xác (%)
0.1	*0.1*	*0.8*	*86.0*
0.1	*0.2*	*0.7*	*86.2*
0.1	*0.3*	*0.6*	*86.0*
0.1	*0.4*	*0.5*	*85.0*

Bảng 3.13: (tiếp)

w_{pos}	w_{lik}	w_{cmt}	Độ chính xác (%)
0.1	*0.5*	*0.4*	*84.4*
0.1	*0.6*	*0.3*	*85.0*
0.1	*0.7*	*0.2*	*83.8*
0.1	*0.8*	*0.1*	*82.6*
0.2	*0.1*	*0.7*	*84.8*
0.2	*0.2*	*0.6*	*86.6*
0.2	*0.3*	*0.5*	*87.2*
0.2	*0.4*	*0.4*	*87.2*
0.2	*0.5*	*0.3*	*87.6*
0.2	*0.6*	*0.2*	*86.6*
0.2	*0.7*	*0.1*	*85.4*
0.3	*0.1*	*0.6*	*86.2*
0.3	*0.2*	*0.5*	*85.0*
0.3	*0.3*	*0.4*	*87.6*
0.3	**0.4**	**0.3**	**88.4**
0.3	*0.5*	*0.2*	*88.4*
0.3	*0.6*	*0.1*	*86.2*
0.4	*0.1*	*0.5*	*86.2*
0.4	*0.2*	*0.4*	*85.8*
0.4	*0.3*	*0.3*	*87.0*
0.4	*0.4*	*0.2*	*88.2*
0.4	*0.5*	*0.1*	*87.4*
0.5	*0.1*	*0.4*	*86.6*
0.5	*0.2*	*0.3*	*87.6*
0.5	*0.3*	*0.2*	*85.2*
0.5	*0.4*	*0.1*	*87.2*
0.6	*0.1*	*0.3*	*86.8*
0.6	*0.2*	*0.2*	*87.8*
0.6	*0.3*	*0.1*	*86.2*
0.7	*0.1*	*0.2*	*87.6*
0.7	*0.2*	*0.1*	*87.0*
0.8	*0.1*	*0.1*	*86.0*

Từ bảng thống kê, đồ án xác định giá trị cho bộ trọng số w_{pos}, w_{lik}, w_{cmt} lần lượt là *0.3, 0.4, 0.3* cho *độ chính xác* cao nhất *88.4%* làm bộ trọng số tốt nhất. Đây cũng chính là bộ trọng số cuối cùng cho *bộ dữ liệu đánh giá* trong mô hình của đồ án.

Kết hợp 2 bộ trọng số, đồ án có được bộ trọng số cho mô hình của đồ án gồm w_{cat}, w_{sen}, w_{emo}, w_{pos}, w_{lik}, w_{cmt}.

3.5. Kết quả đánh giá

Kết quả đánh giá thuật toán sau khi chạy với 500 *mẫu đánh giá* được trình bảy ở bảng sau:

Bảng 3.14: Kết quả đánh giá thuật toán khi chạy với 500 mẫu đánh giá

Bộ trọng số						**Số mẫu đúng**	**Độ chính xác**
W_{cat}	W_{sen}	W_{emo}	W_{pos}	W_{lik}	W_{cmt}	**442**	**88.4%**
0.3	**0.3**	**0.4**	**0.3**	**0.4**	**0.3**		

Các giá trị trọng số w_c, w_s, w_e, w_1, w_2, w_3 tốt nhất đối với 500 *mẫu đánh giá* đồ án đã thu thập lần lượt là *0.3, 0.3, 0.4, 0.3, 0.4, 0.3*. Đồ án sử dụng bộ trọng số cuối cùng này áp dụng vào các công thức (2.5), (2.21) để ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội khi ứng dụng thuật toán ở chương tiếp theo của đồ án.

Bộ trọng số tốt nhất trong mô hình của đồ án cho *độ chính xác* cao nhất đạt *88.4%* là kết quả cuối cùng khi thực hiện đánh giá thuật toán trên dữ liệu thu thập từ trang mạng xã hội Twitter. Nhìn vào giá trị của bộ trọng số tốt nhất xác định được có thể thấy:

+ Trọng số thuộc tính *trạng thái cảm xúc của bài đăng* là cao nhất, tiếp theo là trọng số thuộc tính *chủ đề của bài đăng* và trọng số thuộc tính *quan điểm của bài đăng*.

+ Trọng số độ tương đồng hành vi *thích bài đăng* là cao nhất, tiếp theo là trọng số độ tương đồng hành vi *đăng bài đăng* và trọng số độ tương đồng hành vi *nhận xét bài đăng*.

Đồ án thu được kết quả đánh giá cho độ chính xác cao nhất đạt *88.4%*, tuy nhiên trong quá trình cài đặt và kiểm thử thì mô hình của đồ án vẫn còn những hạn chế sau:

- Hạn chế về bộ dữ liệu đánh giá và các bộ dữ liệu học cho xác định các thuộc tính của bài đăng: Đồ án xây dựng bộ dữ liệu đánh giá giới hạn gồm các bài đăng trong khoảng thời gian gần nhất của 500 người dùng theo 3 hành vi nhất định, các bộ dữ liệu học cũng giới hạn số lượng <1000 bài đăng. Sự giới hạn này là thưa thớt dữ liệu người dùng, làm giảm tính đa dạng về chủ đề, quan điểm, trạng thái cảm xúc mà

dữ liệu thể hiện, chưa bao hàm hết được hành vi của người dùng từ khi sử dụng mạng xã hội đến nay. Hạn chế này gây ảnh hưởng rõ, thấy rõ ở sự chênh lệch số lượng của bộ dữ liệu học cho xác định từng chủ đề, quan điểm, trạng thái cảm xúc của bài đăng.

- Hạn chế về ngôn ngữ: Thuật toán chưa áp dụng được trên nhiều ngôn ngữ khác nhau, nên chưa thể phát triển và áp dụng vào các ứng dụng trên nhiều ngôn ngữ khác như Tiếng Việt...

3.6. Kết luận

Trong chương 3 đồ án đã trình bày về dựng bộ dữ liệu thử nghiệm cùng với các *kết quả đánh giá* của các bộ dữ liệu. Cụ thể là:

+ *Bộ dữ liệu học cho xác định chủ đề của bài đăng* gồm 760 *mẫu*

+ *Bộ dữ liệu học cho xác định trạng thái cảm xúc của bài đăng* gồm 760 *mẫu*.

+ *Bộ dữ liệu học cho xác định quan điểm của bài đăng* gồm 400 *mẫu*.

+ *Bộ dữ liệu đánh giá* gồm 500 *mẫu*.

+ *Bộ dữ liệu cho khảo sát trọng số cho ước lượng độ tương đồng giữa các bài đăng* gồm 500 *mẫu*.

Đồ án đã trình bày phương pháp đánh giá thuật toán, cách cài đặt thuật toán, và thực hiện khảo sát hai bộ trọng số đối với bộ dữ liệu đã xây dựng gồm:

+ Bộ trọng số w_{cat}, w_{sen}, w_{emo} tương ứng với 3 thuộc tính của bài đăng là *chủ đề của bài đăng*, *quan điểm của bài đăng* và *trạng thái cảm xúc của bài đăng*.

+ Bộ trọng số w_{pos}, w_{lik}, w_{cmt} tương ứng với độ tương đồng hành vi *đăng bài đăng*, hành vi *thích bài đăng*, hành vi *nhận xét bài đăng*.

Đồ án thu được kết quả đánh giá cho độ chính xác cao nhất đạt *88.4%* với bộ giá trị trọng số w_{cat}, w_{sen}, w_{emo}, w_{pos}, w_{lik}, w_{cmt} lần lượt là *0.3, 0.3, 0.4, 0.3, 0.4, 0.3*.

Tiếp theo là chương 4 đồ án sẽ trình bày áp dụng thuật toán vào ứng dụng cụ thể MyTwitter.

CHƯƠNG 4: ỨNG DỤNG THUẬT TOÁN

Chương 4 sẽ trình bày các nội dung sau:

- Mô tả ứng dụng MyTwitter
- Kiến trúc tổng quan ứng dụng
- Kịch bản sử dụng của người dùng
- Kết quả ứng dụng

4.1. Mô tả ứng dụng MyTwitter

MyTwitter là ứng dụng Java Web mô phỏng trang mạng xã hội Twitter, đưa ra gợi ý theo dõi (follow) và gợi ý quảng cáo theo chủ đề, áp dụng trên bộ dữ liệu của 500 người dùng đã xây dựng. Ứng dụng được cài đặt bằng ngôn ngữ Java, HTML, CSS, JavaScript; sử dụng công cụ quản lý Maven, framework Spring Framework, server Apache Tomcat 8.0.27 và công nghệ J2EE.

Ứng dụng gồm có các chức năng chính sau:

- Đăng nhập, Đăng xuất: Người dùng có thể thực hiện đăng nhập vào ứng dụng theo tài khoản gồm tên tài khoản - account và mật khẩu - password. Người dùng có thể đăng xuất sau khi sử dụng ứng dụng.

- Gợi ý theo dõi: là chức năng áp dụng thuật toán ước lượng độ tương đồng của người dùng dựa vào hành vi để gợi ý nhóm người dùng có thể muốn theo dõi. Kết quả gợi ý được hiển thị trên giao diện trang chủ ngay sau khi người dùng đăng nhập và hiển thị trên giao diện trang cá nhân của người dùng.

- Gợi ý quảng cáo theo chủ đề: là chức năng áp dụng thuật toán ước lượng độ tương đồng của người dùng dựa vào hành vi để gợi ý hình ảnh quảng cáo có thể người dùng đang quan tâm, theo chủ đề đang được người dùng và bạn bè quan tâm nhất, hoặc theo chủ đề đang được người dùng và một người bạn cụ thể quan tâm nhất. Kết quả gợi ý được hiển thị trên giao diện trang chủ và giao diện trang cá nhân.

4.2. Kiến trúc tổng quan ứng dụng

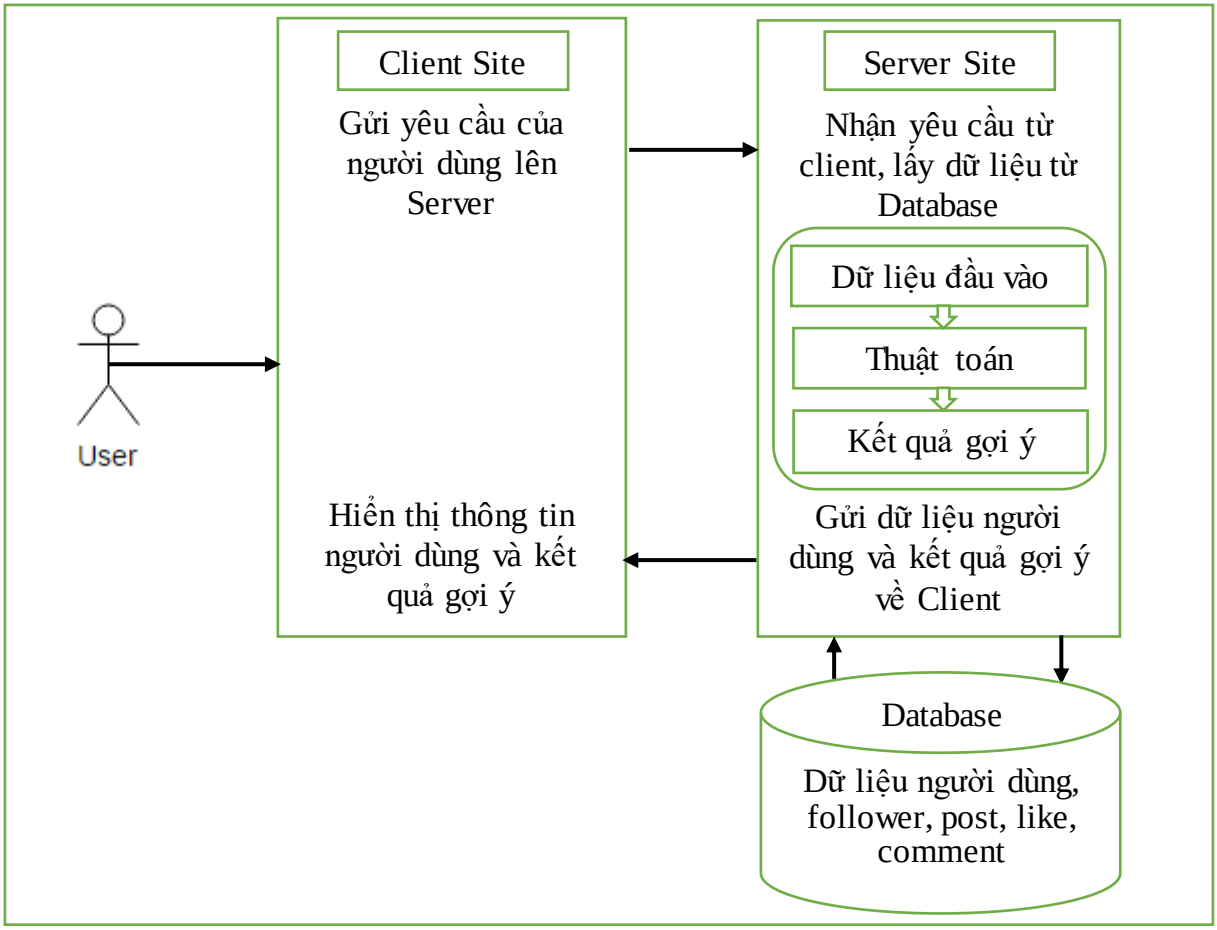
Ứng dụng gồm 3 phần chính:

- Client: gồm 3 trang giao diện của ứng dụng sử dụng giao diện của Twitter: trang đăng nhập, trang chủ và trang cá nhân. Các trang giao diện này sẽ hiển thị dữ liệu người dùng và các kết quả gợi ý.

- Server: nhận các yêu cầu từ client và lấy dữ liệu từ Database, áp dụng thuật toán và trả lại các kết quả gợi ý cho Client.

- Database: lưu dữ liệu người dùng gồm thông tin cá nhân, bạn bè, dữ liệu các bài đăng đã đăng, thích, nhận xét đã được phân tích, xử lý, xác định chủ đề, quan điểm, trạng thái cảm xúc.

Hình 4.1: Mô hình kiến trúc ứng dụng



4.3. Kịch bản sử dụng của người dùng

Kịch bản như sau:

+ Bước 1: Người dùng truy cập trang đăng nhập MyTwitter

+ Bước 2: Trang giao diện đăng nhập hiển thị

+ Bước 3: Người dùng thực hiện nhập tên tài khoản - account và mật khẩu - password tài khoản rồi click chọn Đăng nhập.

+ Bước 4: Hệ thống nhận yêu cầu đăng nhập, cho phép đăng nhập nếu thông tin tài khoản hợp lệ và thực hiện lấy dữ liệu người dùng, tính toán để đưa ra các gợi ý gồm: gợi ý theo dõi, gợi ý quảng cáo theo chủ đề được người dùng và bạn bè quan tâm nhất.

+ Bước 5: Giao diện trang chủ hiện ra với các thông tin gồm có: tài khoản người dùng, danh sách bạn bè (following), danh sách các bài đăng, danh sách gợi ý theo dõi, hình ảnh quảng cáo theo chủ đề.

+ Bước 6: Người dùng click chọn Follow tài khoản userA trong danh sách gợi ý theo dõi.

+ Bước 7: Hệ thống nhận yêu cầu theo dõi, thực hiện cập nhật CSDL, thực hiện xử lý, tính toán để cập nhật lại danh sách gợi ý theo dõi, cập nhật lại danh sách bạn bè, cập nhật lại gợi ý quảng cáo theo chủ đề được người dùng và bạn bè quan tâm nhất.

+ Bước 8: Giao diện trang chủ được làm mới với các thông tin đã được cập nhật.

+ Bước 9: Người dùng click chọn tài khoản userB trong danh sách bạn bè.

+ Bước 10: Hệ thống nhận yêu cầu, thực hiện lấy dữ liệu của tài khoản userB, cập nhật lại dah sách gợi ý theo dõi, tính toán để đưa ra gợi ý quảng cáo theo chủ đề đang được người dùng và userB quan tâm nhất.

+ Bước 11: Giao diện trang cá nhân của tài khoản userB được hiển thị với các thông tin: tài khoản userB, danh sách bạn bè của người dùng và của userB, danh sách các bài đăng của userB, gợi ý theo dõi, gợi ý quảng cáo theo chủ đang được người dùng và userB quan tâm nhất.

+ Bước 12: Người dùng click chọn icon đăng xuất trên giao diện đang hiển thị.

+ Bước 13: Hệ thống thực hiện đăng xuất và hiển thị giao diện trang đăng nhập.

Thuật toán được áp dụng vào mô hình ở các bước 4, 7, 10 như sau:

+ *Gợi ý theo dõi:* Hệ thống lấy dữ liệu của <=100 người dùng ngẫu nhiên trong danh sách người dùng chưa là bạn bè của người dùng đã đăng nhập. Ở mỗi lần tạo danh sách gợi ý theo dõi, hệ thống lấy dữ liệu của <=10 người dùng ngẫu nhiên trong danh sách <=100 người dùng trên để làm đầu vào cho thuật toán. Kết quả ở đầu ra là danh sách gồm <=3 người dùng có độ tương đồng cao nhất với người dùng đã đăng nhập để gợi ý theo dõi.

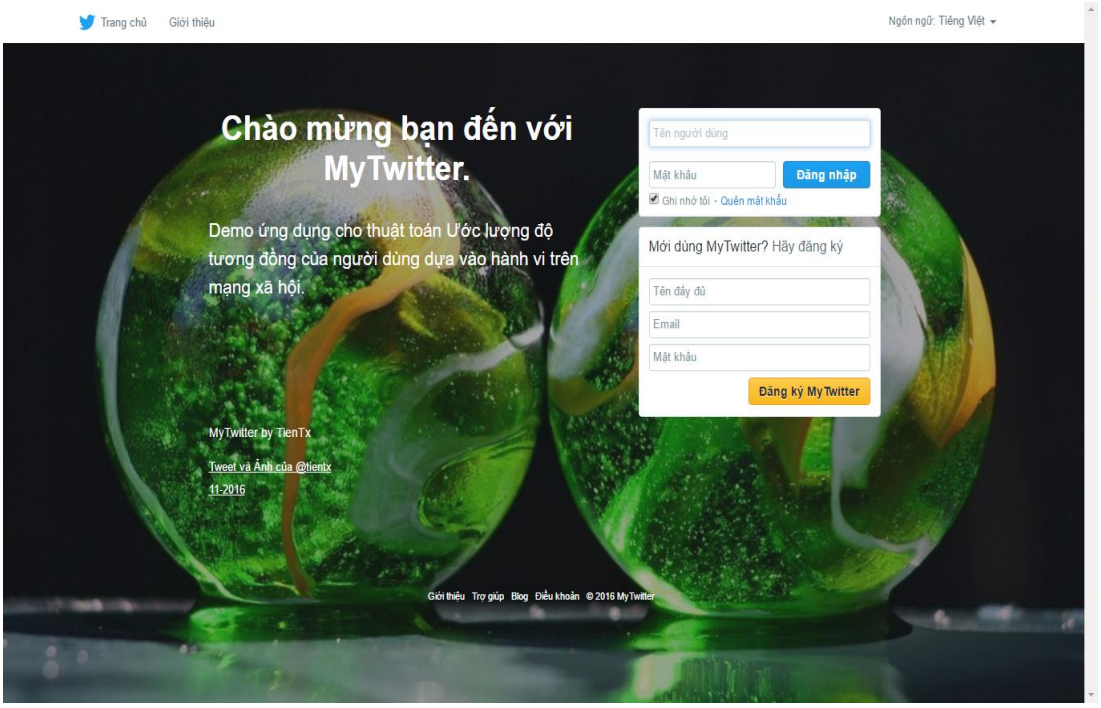
+ Gợi ý quảng cáo theo chủ đề:

- Ở bước 7: Hệ thống gợi ý quảng cáo theo chủ đề đang được người dùng và bạn bè quan tâm nhất: Hệ thống lấy dữ liệu của người dùng và của toàn bộ người dùng trong danh sách bạn bè để lọc ra danh sách gồm toàn bộ các chủ đề đã được xác định của toàn bộ các bài đăng. Từ danh sách chủ đề này, hệ thống tiếp tục lọc ra chủ đề được nhắc tới nhiều nhất, tức đang được quan tâm nhất để đưa ra các hình ảnh quảng cáo theo đó.

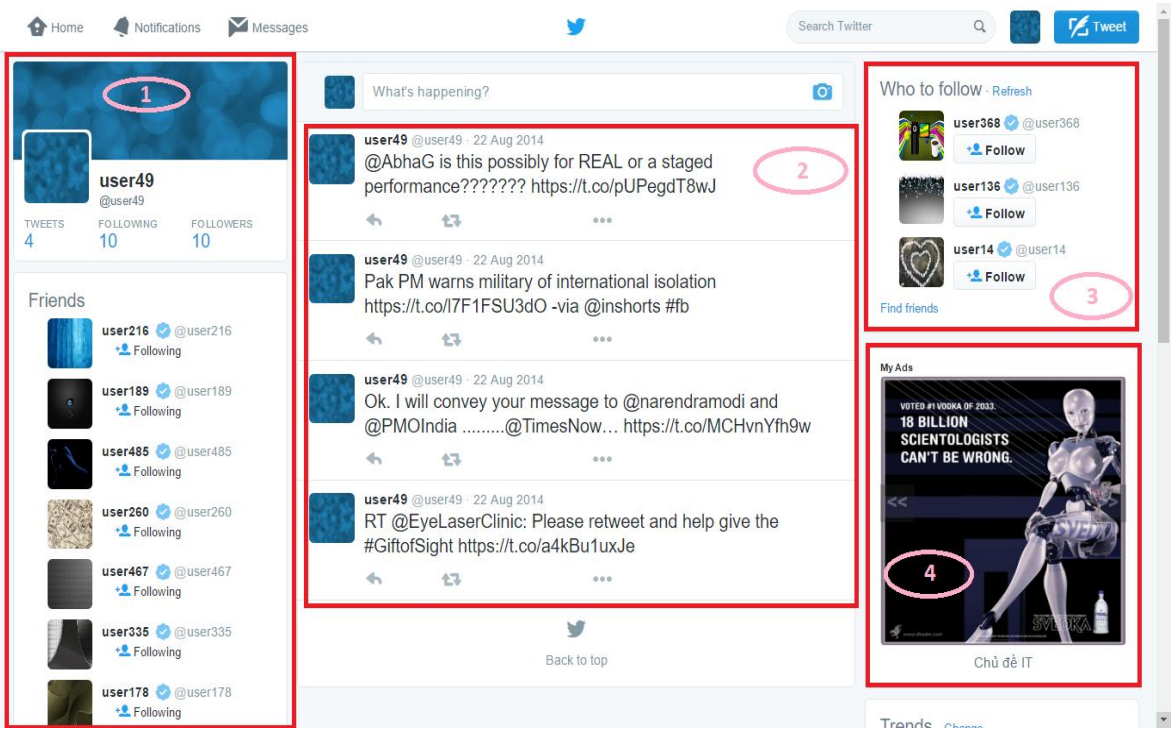
- Ở bước 10: Hệ thống gợi ý quảng cáo theo chủ đề đang được người dùng và userB quan tâm nhất: Hệ thống lấy dữ liệu của người dùng đăng nhập hiện tại và của userB để lọc ra danh sách gồm toàn bộ các chủ đề đã được xác định của toàn bộ các bài đăng. Từ danh sách chủ đề này, hệ thống cũng lọc ra chủ đề đang được quan tâm nhất bởi 2 người dùng này để đưa ra các hình ảnh quảng cáo theo chủ đề đó.

4.4. Kết quả

Một số hình ảnh kết quả của ứng dụng:



Hình 4.2: Giao diện trang đăng nhập hệ thống



Hình 4.3: Giao diện trang chủ của tài khoản user49

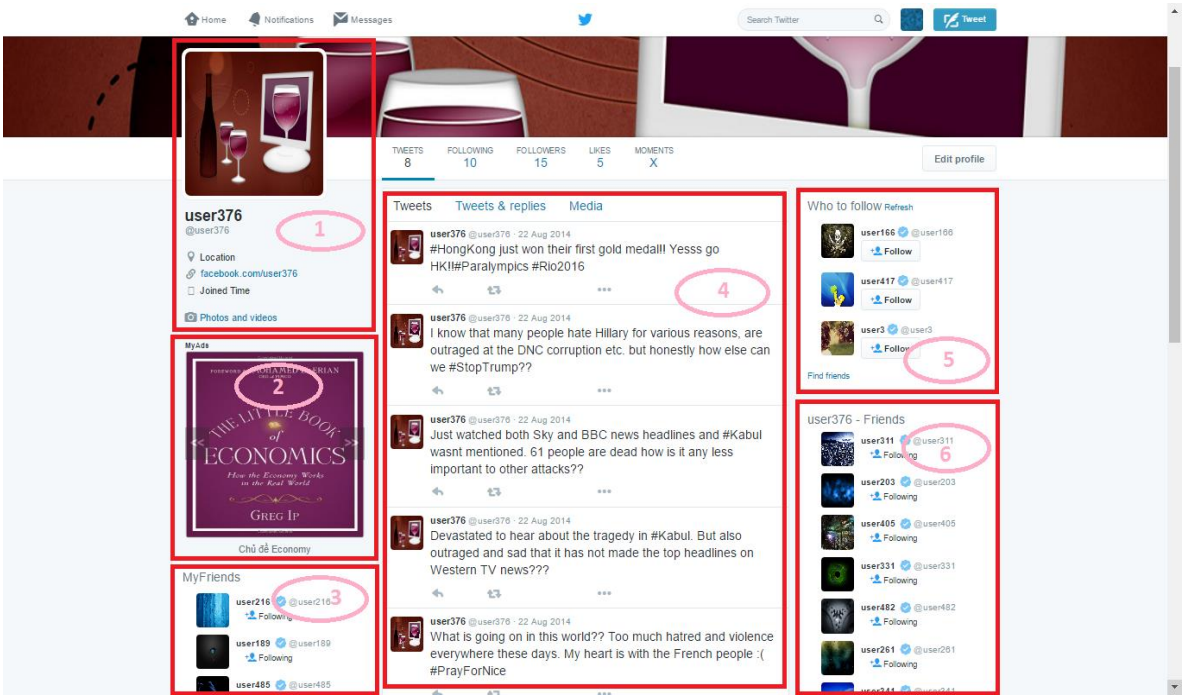
Hình 4.3 mô tả giao diện trang chủ của tài khoản user49, trong đó:

+ Khung 1: Thông tin cá nhân và danh sách bạn bè (follower), danh sách đang theo dõi của user49.

+ Khung 2: Danh sách các bài đăng của user49.

+ Khung 3: Kết quả gợi ý theo dõi cho user49. Danh sách gợi ý gồm 3 tài khoản người dùng khác nhau có độ tương đồng cao nhất được đánh giá bởi thuật toán trong danh sách X người dùng ngẫu nhiên chưa là bạn bè với user49. Kết quả gợi ý là khác nhau tại mỗi lần người dùng cập nhật trang.

+ Khung 4: Kết quả gợi ý quảng cáo theo chủ đề cho user49. Nội dung gợi ý là các ảnh quảng cáo được thể hiện dưới dạng slide. Các ảnh quảng cáo này thuộc một chủ đề nhất định được xác định bởi thuật toán. Chủ đề được gợi ý là chủ đề đang được user49 và bạn bè quan tâm nhất. Kết quả gợi ý theo chủ đề có thể thay đổi khi danh sách bạn bè của người dùng thay đổi. Các ảnh quảng cáo gợi ý như trong hình là nội dung gợi ý quảng cáo cho chủ đề IT, thấy rõ hơn ở hình 4.5.



Hình 4.4: Giao diện trang cá nhân của tài khoản user376 đang được xem bởi user49

Hình 4.4 mô tả giao diện trang cá nhân của tài khoản user376 đang được xem bởi user49, trong đó:

+ Khung 1: Thông tin cá nhân của user376.

+ Khung 2: Kết quả gợi ý quảng cáo theo chủ đề cho user49. Nội dung gợi ý là các ảnh quảng cáo được thể hiện dưới dạng slide. Các ảnh quảng cáo này thuộc một chủ đề nhất định được xác định bởi thuật toán. Chủ đề được gợi ý là chủ đề Economy, chủ đề đang được user49 và user376 quan tâm nhất.

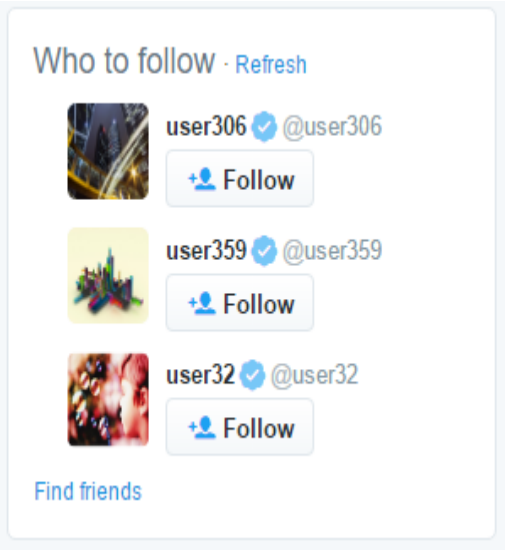
+ Khung 3: Danh sách bạn bè của user49.

+ Khung 4: Danh sách các bài đăng của user376.

+ Khung 5: Kết quả gợi ý theo dõi cho user49. Danh sách gợi ý gồm 3 tài khoản người dùng khác nhau có độ tương đồng cao nhất được đánh giá bởi thuật toán trong danh sách X người dùng ngẫu nhiên chưa là bạn bè với user49. Kết quả gợi ý là khác nhau tại mỗi lần người dùng cập nhật trang.

+ Khung 6: Danh sách bạn bè của user376.

Hình 4.5: Một gợi ý quảng cáo chủ đề IT



Hình 4.6: Gợi ý theo dõi (follow)

- Đánh giá:

+ Kết quả hiển thị các thông tin người dùng đúng như kịch bản.

+ Nội dung gợi ý theo dõi hiển thị đúng như kịch bản và được làm mới ở mỗi lần xem khác nhau.

+ Nội dung gợi ý quảng cáo theo chủ đề được hiển thị đúng, nhưng bị trùng lặp nhiều lần đối với một chủ đề trên nhiều người dùng khác nhau.

4.5. Kết luận

Trong chương 4, đồ án đã áp dụng mô hình thuật toán vào ứng dụng mô phỏng mạng xã hội MyTwitter – ứng dụng gợi ý theo dõi, gợi ý quảng cáo theo chủ đề. Đồ án đã trình bày kiến trúc tổng quan, kết quả ứng dụng và đánh giá theo kịch bản sử dụng của người dùng cùng một số hình ảnh chi tiết của các chức năng trên giao diện ứng dụng.

KẾT LUẬN

Đồ án đã trình bày được về các vấn đề liên quan đến mô hình *ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội*. Cụ thể là:

- Thứ nhất, đồ án đã nghiên cứu và cài đặt mô hình thuật toán ước lượng độ tương đồng của người dùng dựa vào hành vi trên mạng xã hội.

- Thứ hai, đồ án trình bày cách thu thập, xây dựng bộ dữ liệu thử nghiệm, phương pháp đánh giá thuật toán và các kết quả đánh giá bộ dữ liệu, kết quả đánh giá thuật toán. Đồ án đánh giá thuật toán trên bộ dữ liệu được tương tác bởi các hành vi *đăng bài đăng*, *thích bài đăng*, *nhận xét bài đăng* của 500 người dùng thực tế và thu được kết quả cuối cùng: độ chính xác là *88.4%*. Kết quả được xem là đạt với dữ liệu thực tế hiện tại mà đồ án thu thập.

- Thứ ba, đồ án đã áp dụng và cài đặt thành công mô hình thuật toán vào ứng dụng mô phỏng mạng xã hội MyTwitter - ứng dụng gợi ý theo dõi và gợi ý quảng cáo theo chủ đề.

Trong quá trình cài đặt và kiểm thử, mô hình của đồ án vẫn còn những hạn chế như sau:

- Thứ nhất, hạn chế về bộ dữ liệu đánh giá và các bộ dữ liệu học cho xác định các thuộc tính của bài đăng. Do đối với một tài khoản Twitter Developer thông thường bị giới hạn số lần lấy dữ liệu theo thời gian, đồ án xây dựng bộ dữ liệu đánh giá giới hạn gồm các bài đăng trong khoảng thời gian gần nhất của 500 người dùng tương tác theo 3 hành vi nhất định, các bộ dữ liệu học cũng giới hạn số lượng nhỏ hơn 1000 bài đăng. Sự giới hạn này là thưa thớt dữ liệu người dùng, làm giảm tính đa dạng về chủ đề, quan điểm, trạng thái cảm xúc mà dữ liệu thể hiện, chưa bao hàm hết được các hành vi của người dùng từ khi sử dụng mạng xã hội đến nay. Hạn chế này gây ảnh hưởng rõ, thấy rõ ở sự chênh lệch số lượng của bộ dữ liệu học cho xác định từng chủ đề, quan điểm, trạng thái cảm xúc của bài đăng, ở sự trùng lặp chủ đề được gợi ý quảng cáo trong ứng dụng MyTwitter của đồ án.

- Thứ hai, hạn chế về ngôn ngữ: mô hình thuật toán chưa áp dụng được trên nhiều ngôn ngữ khác nhau, nên chưa thể phát triển và áp dụng vào các ứng dụng trên nhiều ngôn ngữ khác như Tiếng Việt...

Một vài hướng phát triển tiếp theo cho mô hình của đồ án trong tương lai như:

- Tiếp tục cải tiến tốc độ tính toán và độ chính xác của thuật toán.

- Phát triển và mở rộng hệ thống với số lượng người dùng lớn hơn, cập nhật các bộ dữ liệu học, bộ dữ liệu đánh giá đa dạng, khách quan hơn.

- Nghiên cứu mở rộng thuật toán trên các ngôn ngữ khác như Tiếng Việt.

- Áp dụng vào nhiều hệ thống hay ứng dụng thực tế có ích cho người dùng về các chức năng như: gợi ý, tư vấn, cảnh báo…

DANH MỤC THAM KHẢO

Tài liệu

1. Davide Buscaldi, Joseph Le Roux, Jorge J. Garca Flores, and Adrian Popescu. Lipn-core: Semantic text similarity using n-grams, wordnet, syntactic analysis, esa and information retrieval based features, 2013.
2. William B. Cavnar and John M. Trenkle. N-gram-based text categorization. Ann Arbor MI, 48113(2):161-175, 1994.
3. Gao Cong, WeeSun Lee, Haoran Wu, and Bing Liu. Semi-supervised text classification using partitioned em. In YoonJoon Lee, Jianzhong Li, Kyu-Young Whang, and Doheon Lee, editors, Database Systems for Advanced Applications, volume 2973 of Lecture Notes in Computer Science, pages 428-493. Springer Berlin Heidelberg, 2004.
4. Chihli Hung and Hao-Kai Lin. Using objective words in sentiwordnet to improve word-of-mouth sentiment classification. IEEE Intelligent Systems, vol.28, no.2:47-54, 2013.
5. Thorsten Joachims. Text categorization with suport vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning, ECML '98, pages 137-142, London, UK, UK, 1998. Spring-Verlag.
6. Youngjoong Ko and Jungyun Seo. Automatic text categorization by unsupervised learning. In Proceedings of the 18th Conference on Computational Linguistics – Volume 1, COLING '00, pages 453-459, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.
7. Ming Che Lee, Jia Wei Chang, and Tung Cheng Hsieh. A grammar-based semantic similarity algorithm for natural language sentences. The Scientific World Journal, 2014:17 pages, 2014.
8. Manh Hung Nguyen and Thi Hoi Nguyen. A general model for similarity measurement between objects. International Journal of Advanced Computer Science and Applications (IJACSA), vol.6, no.2:235-239, 2015.
9. Bruno Ohana and Brendan Tierney. Sentiment classification of reviews using sentiwordnet, 2009.

10. Thomas Proisl, Stefan Evert, Paul Greiner, and Besim Kabashi. Robust semantic similarity at multiple levels using maximum weight matching. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 532–540, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

Website

11. http://alias-i.com/lingpipe/
12. http://twitter4j.org/
13. http://twitter4j.org/javadoc/
14. https://dev.twitter.com/rest/public/
15. http://twitter.com/
16. https://www.google.com/forms/

PHỤ LỤC