Báo cáo Thực tập tốt nghiệp
Đề tài: Dự đoán mức độ thu hút của địa điểm ăn uống trong tương lai

GVHD: GS.TS Cao Hoàng Trụ

Nguyễn Trọng Nghĩa 51202370
Nguyễn Đăng Trang 51203957

17-5-2016

Danh sách hình vẽ

Mục lục

1 Tóm tắt báo cáo

Bài báo cáo trình bày về chủ đề Phân tích dữ liệu trên mạng xã hội (MXH) và đi sâu vào một bài toán dùng học máy phân tích dữ liệu trên mạng xã hội do nhóm đề xuất, gồm ba phần:

- Phần một nói về MXH và những vấn đề nghiên cứu trên MXH.
- Phần hai trình bày các kiến thức nền tảng liên quan được sử dụng trong phương pháp đề xuất.
- Phần ba nêu lên đề xuất về phương pháp thực hiện bài toán mà nhóm đề xuất: Dự đoán mức độ thu hút của địa điểm ăn uống trong tương lai. Nghiên cứu thu hẹp lại trên các MXH về địa điểm ăn uống, bài báo cáo đi từ các nghiên cứu liên quan đến các MXH về địa điểm ăn uống tới bài toán được đề xuất.

2 Tổng quan về MXH và Phân tích dữ liệu trên MXH

2.1 Tổng quan về MXH

Sự bùng nổ của công nghệ thông tin và sự phổ biến rộng rãi của các thiết bị có khả năng kết nối Internet, đặc biệt là các thiết bị di động như điện thoại thông minh, máy tính bảng đã tạo điều kiện cho các MXH trực tuyến phát triển nhanh chóng

Các MXH trực tuyến như Facebook, Twitter, Flickr, Foody ... dần trở thành một phần quan trọng trong cuộc sống của mỗi cá nhân.

- Nó giả lập những thành phần của thực tế xã hội con người mà ở đó người dùng có thể nói chuyện, đăng bài, kết bạn, đặt relationship ... (Facebook, Twitter)
- Nó cung cấp cho người dùng sự tiện lợi về một nhu cầu nào đó (Flickr – chia sẻ ảnh, Foody – chia sẻ địa điểm ăn uống

Về định nghĩa cho MXH, một MXH có thể được hiểu như là một hệ thống mà chức năng chính là cho phép người dùng thực hiện các tương tác xã hội trên đó (Facebook, Twitter) hoặc là một hệ thống phục vụ chức năng khác (Flickr – chia sẻ ảnh, Foody – chia sẻ thông tin địa điểm ăn uống) nhưng cho phép người dùng thực hiện những tương tác nhất định trên đó.

Nói một cách tổng quát, MXH là một mạng của các tương tác và mối quan hệ, mà trong mạng đó nốt (node) bao gồm các tác nhân và cạnh (edge) bao gồm các tương tác và mối quan hệ giữa các tác nhân. Điều này có nghĩa là khái niệm MXH không chỉ bị giới hạn trong những trang MXH trực tuyến như Facebook, Foody... mà tổng quát hơn, nó là mạng của

bất cứ các tác nhân nào tương tác với nhau, tương tác này có thể được thực hiện qua mạng Internet, qua thư bưu chính hay mặt đối mặt.

2.2 Các vấn đề trong Phân tích dữ liệu trên MXH

Một khía cạnh rất quan trọng khi xét đến các MXH trực tuyến là chúng chứa lượng dữ liệu khổng lồ, cho phép những nghiên cứu trên đó nhằm tìm ra các đặc điểm của MXH, đặc điểm của các tác nhân (actor) trên đó hay đặc điểm các mối liên kết giữa các tác nhân này.

Có hai kiểu dữ liệu chủ yếu thường được phân tích trên MXH trực tuyến:

- Phân tích liên quan đến cấu trúc và liên kết: Xác định các nốt quan trọng, xác định cộng đồng, dự đoán các liên kết trong mạng, dự đoán sự tiến hóa của mạng
- Phân tích liên quan đến nội dung người dùng đưa lên: Lượng dữ liệu khổng lồ mà người dùng trên các trang MXH như Facebook, Youtube, Flickr (hình ảnh) ... là cơ sở cho việc nghiên cứu nó kết hợp với các phân tích liên quan đến cấu trúc và liên kết nhằm tìm ra những đặc điểm của MXH.

Để hiểu hơn về các thành phần của MXH, về dữ liệu trên MXH và những nghiên cứu trong việc phân tích các dữ liệu này, chúng ta sẽ tìm hiểu một số vấn đề nghiên cứu và bài toán quan trọng trong Phân tích dữ liệu MXH.

Cascading Behavior

Cascading Behavior là việc phát hiện và theo dõi một "xu hướng " mới trên MXH. Khi một người nào đó post một bài viết hay, hoặc một sự kiện xảy ra, một sản phẩm được ra mắt, ban đầu, nó không được chú ý nhiều, nhưng dần dần, theo ảnh hưởng của người dùng, mọi người bắt đầu quan tâm, bàn luận đến vấn đề đó. Thì vấn đề đó trở thành một "xu hướng" hay "trào lưu" mới trong MXH. Vấn đề này khác với Phát hiện sự kiện ở chỗ sự kiện diễn ra ngoài đời thực được nhiều người biết đến và đưa nó lên MXH, còn "trào lưu" thì sinh ra và lan truyền trên MXH. Vậy điều quan trọng trong chủ đề này là làm thế nào để xác định được "xu hướng" mới và theo dõi sự lan truyền của nó. "Xu hướng" được lan truyền qua những đối tượng trong MXH, dựa vào sự lan truyền ấy có thể đánh giá đối tượng nào có ảnh hưởng lớn đối với sự lan truyền trong MXH.

Nghiên cứu về Cascading Behavior có thể giúp tìm kiếm, phát triển các ý tưởng mới trên mạng xã hội, giúp cho lĩnh vực truyền thông, phát hiện theo dõi các "xu hướng" mới hoặc ứng dụng để gợi ý, thông báo "trào lưu" mới cho người dùng.

Phát hiện cộng đồng (Communty Discovery)

Cộng đồng (community) trong MXH là tập hợp một nhóm người dùng có chung một đặc điểm nào đó, *ví dụ*: sở thích, hay cùng mối quan tâm, ... Một cộng đồng gồm chủ đề (theme) và tập hợp người dùng trong cộng đồng đó. Chủ đề là điểm chung mà người trong cộng đồng quan tâm.

Xác định cộng đồng trong MXH gồm 2 yêu cầu là xác định chủ đề và tập hợp người dùng trong cộng đồng đó. Một vấn đề nghiên cứu quan trọng trong chủ đề này là nghiên cứu về "sự thay đổi" của cộng đồng theo thời gian: việc thay đổi về kích thước cũng như các giai đoạn của một cộng đồng. *Ví dụ*: khi nào một cộng đồng mới được hình thành hoặc tan rã, một cộng đồng bị tách ra làm 2 cộng đồng nhỏ, 2 cộng đồng hợp lại làm một. Việc nghiên cứu để phát hiện cộng đồng thuộc một lĩnh vực cụ thể có thể cung cấp những thông tin hữu ích, cập nhật đối với những người dùng quan tâm đến lĩnh vực đó. Mặt khác, nó giúp việc nghiên cứu về các vấn đề xã hội trên web liên quan đến cộng đồng và đặc trưng của nó và nó giúp cho việc quảng cáo được hiệu quả hơn dựa vào chủ đề/lĩnh vực liên quan của cộng đồng.

Xác định tác nhân quan trọng trong MXH (Identifying prominent actors in social network)

Những người nổi bật trong MXH là những người có vai trò trung tâm, có liên hệ với nhiều người khác. Trong ngữ cảnh một tổ chức/cộng đồng, những người này được xem là có vai trò quan trọng hơn nhiều so với những người ít giao tiếp quan hệ với người xung quanh.

Khi mô hình hóa MXH bằng đồ thị, ta có thể hình dung đơn giản rằng những người nổi bật sẽ được thể hiện bằng những đỉnh có số lượng đỉnh liền kề lớn. Có nhiều thang đo để đánh giá sự nổi bật của một người trong MXH, trong đó đáng chú ý nhất là Rank Prestige. GIải thuật xếp hạng kết quả tìm kiếm (PageRank) của Google được hiện thực dựa trên Rank Prestige và đã biến Google thành công cụ tìm kiếm đứng đầu thế giới.

Vấn đề cần nghiên cứu chủ yếu trong chủ đề này là đánh giá các actor trong MXH để tìm ra người nổi bật nhất (có thể ứng dụng giải thuật PageRank hoặc các biến thể của nó). Nghiên cứu này mang nhiều ý nghĩa thực tế. *Ví dụ*: Trong lĩnh vực marketing, có thể tìm hiểu những đối tượng nào có khả năng thu hút cộng đồng MXH lớn để truyền thông tin, quảng bá.

Dự đoán liên kết (Link prediction)

Dự đoán liên kết nghiên cứu về mối quan hệ giữa các actor (người dùng) trong MXH. MXH là một cấu trúc xã hội được tạo nên từ tập hợp các chủ thể (cá nhân, tổ chức,...) và mối quan hệ giữa các chủ thể đó. MXH có thể được trực quan hóa thông qua đồ thị, khi đó đồ thị bao gồm: Các đỉnh (vertexs/nodes) đại diện cho các chủ thể và các cạnh, liên kết (edges/links) đại diện cho các mối quan hệ.

Vấn đề khai thác MXH hiện nay đối mặt hai thách thức, đó là: Dữ liệu không đầy đủ và dữ liệu thay đổi liên tục. Thứ nhất, dữ liệu không đầy đủ (Incompletion): Phần lớn dữ liệu về MXH thu thập được không đầy đủ bởi vì chỉ một phần thông tin MXH có thể thu thập được từ các ứng dụng thu thập dữ liệu MXH. Thứ hai, dữ liệu thay đổi liên tục (Dynamic): MXH có đặc trưng là luôn thay đổi liên tục, do đó mỗi đỉnh hay cạnh có thể xuất hiện và biến mất liên tục.

Việc dự đoán các liên kết bị thiếu, không thu thập được trong MXH hiện tại và dự đoán các liên kết sẽ hình thành hay kết thúc trong MXH tương lai rất cần thiết. Vấn đề này gồm các vấn đề nhỏ: dự đoán được những liên kết trong MXH hiện tại mà ta không thể thu thập

được; dự đoán những mối quan hệ sẽ hình thành trong MXH tương lai; dự đoán những mối quan hệ sẽ kết thúc trong MXH tương lai.

Một số ứng dụng của Dự đoán liên kết trên MXH:

- Gợi ý trên MXH: Gợi ý bạn bè, đồng nghiệp, người được theo dõi và những người theo dõi là những điển hình.
- Giải quyết vấn đề hoàn chỉnh mạng: Dữ liệu thu thập không hoàn chỉnh với việc thiếu các đỉnh và các cạnh. Nhiệm vụ có thể suy luận tìm ra những đỉnh và cạnh bị mất đó.
- Dự đoán mối quan hệ xã hội: *Ví dụ* dự đoán mức độ mối quan hệ (mạnh hay yếu) và mô hình mức độ quan hệ đó. Mặc dù nó không liên quan trực tiếp đến dự đoán liên kết, nhưng những kết quả của nó có thể dẫn đến những phương pháp dự đoán liên kết mới.
- Ứng dụng vào các lĩnh vực khác: trong sinh học: dự đoán tương tác protein-protein.

Niềm tin trên MXH (Trust in social network)

Niềm tin trên MXH là lòng tin giữa người dùng trong MXH. Trên MXH, các người dùng có thể không biết nhau ngoài đời thật, nên cần có một công cụ để đánh giá lòng tin giữa các người dùng với nhau, giúp đánh giá về thông tin đăng trên MXH cũng như về bản thân một người dùng (hay tổ chức/ công ty) tham gia vào MXH.

Lòng tin rất khó định nghĩa, ngay cả ở ngoài đời thực. Lòng tin bị ảnh hưởng bởi rất nhiều nhân tố, cả về tâm lí, xã hội,.. nên khó có thể đo được chính xác. Hai điểm lưu ý trong vấn đề lòng tin giữa đời thực và MXH: sự khác nhau giữa văn hóa, thể chế, quan điểm giữa người dùng trong MXH là rất nhiều; danh tính của một người trên MXH có thể không phản ánh đúng như ở ngoài đời thật.

Một số vấn đề nghiên cứu trong chủ đề này: Định nghĩa được lòng tin trong MXH; tìm được một thước đo hợp lí về lòng tin: các nhân tố ảnh hưởng; mô hình hóa được một hệ thống biểu diễn lòng tin giữa các người dùng trong MXH; đánh giá được mức độ tin tưởng giữa 2 người dùng chưa hề có mối quan hệ nào dựa vào mô hình lòng tin đã biết trước.

Niềm tin trên MXH có nhiều ứng dụng quan trọng, bao gồm: giúp đánh giá mức độ tin cậy của một thông tin hay người dùng trên MXH; giúp các công ty/ tổ chức nâng cao được lòng tin với khách hàng và giúp đánh giá, gợi ý kết bạn cho người dùng.

Phân loại nốt (Node classification)

Mỗi người dùng trong MXH có rất nhiều thông tin: profile, các hoạt động tương tác, bình luận, status, các group tham gia. Những thông tin này được phân loại và gán nhãn: ví dụ từ thông tin tuổi gán nhãn cho người đó là trẻ, quan hệ tình cảm có thể gán nhãn là độc thân, kết hôn hay li dị,.. Mỗi người sẽ có nhiều nhãn dựa vào thông tin của họ.

Trong MXH, khi chỉ có một vài người được phân loại nhãn, cần phải có một công cụ giúp phân loại nhãn của các người dùng còn lại dựa vào các tương tác giữa họ. Ngoài ra có thể

đánh giá mức độ liên hệ giữa các nhãn cụ thể với một người dùng.

Trong cách phân loại truyền thống, mỗi thực thể được phân loại được xem là độc lập với nhau, chỉ dựa trên các thuộc tính của nó. Ngược lại trong MXH, việc phân loại một thực thể còn bị ảnh hưởng bởi mối quan hệ giữa nó và các thực thể khác.

Một vài ứng dụng quan trọng của phân loại nốt là: xây dựng hệ thống gợi ý đến người dùng (âm nhạc, phim ảnh, bạn bè,...) dựa trên nhãn mà người dùng được gán; hệ thống hỏi đáp mà những câu hỏi sẽ được trực tiếp gửi đến các chuyên gia có thể trả lời hợp lý; hệ thống quảng cáo: dựa trên việc gán nhãn người dùng, người quảng cáo có thể đưa ra quyết định nhân chóng có nên quảng với người đó hay không.

Tìm kiếm chuyên gia trên MXH (Expert finding in Social Network)

Tìm kiếm chuyên gia trên MXH là một trong những đề tài nghiên cứu quan trọng nhất trong MXH. Mục tiêu của đề tài này là nhằm tìm kiếm những người trên MXH có kiến thức hay kinh nghiệm về một đề tài nhất định, chẳng hạn như toán học, vật lý, công nghệ,... Những người này có vai trò quan trọng vì họ có khả năng chia sẻ, truyền đạt những kiến thức, kinh nghiệm của mình cho những người khác quan tâm đến chủ đề đó.

Trong chủ đề này, ta cần nghiên cứu về thông tin cá nhân của mọi người trên MXH và những mối liên hệ giữa người với người. Sử dụng những thông tin cá nhân để đánh giá "điểm chuyên gia" của mọi người và dùng số điểm này để xếp hạng, rồi chọn ra những người có số điểm cao nhất để làm ứng cử viên. Các ứng cử viên này sẽ được dùng để xây dựng đồ thị. Sau đó, sử dụn mối quan hệ của những ứng cử viên với các chuyên gia (cho trước) về đề tài đó để tìm ra những người cũng là chuyên gia về chủ đề.

Ứng dụng của các nghiên cứu trên chủ đề này có thể kể tới là: về một lĩnh vực nhất định, xác định các người dùng có ảnh hưởng để giúp marketing doanh nghiệp; hoặc gợi ý cho người dùng về những chuyên gia trong lĩnh vực họ hứng thú.

Phân tích dữ liệu do người dùng đưa lên trên MXH (Data Mining)

Người dùng có thể đưa lên các trang MXH một lượng dữ liệu khổng lồ. Các nội dung này cho phép người dùng chia sẻ quan điểm cá nhân, ý kiến cá nhân và là phương tiện để tương tác trên MXH.

Có thể thấy dữ liệu trên các trang MXH rất đa dạng về định dạng: hình ảnh, video, văn bản ... *Ví dụ*: Flickr cho phép người dùng chia sẻ hình ảnh chất lượng cao và kèm theo các mô tả, Youtube là trang chia sẻ video kèm theo các mô tả và cho phép người dùng bình luận lên đó. Facebook cho phép người dùng đăng các post chứa không những dữ liệu dạng văn bản mà có thể kèm theo cả hình ảnh, video.

Dữ liệu ở dạng văn bản đóng vai trò quan trọng trong các trang MXH. Đối với những trang MXH như Facebook, Twitter dữ liệu dạng văn bản được thể hiện qua những bài post và những bình luận của người dùng trên những bài post đó. Đối với những trang MXH mà chủ thể là hình ảnh hoặc video như Flickr hay Youtube, thì dữ liệu ở dạng văn bản cũng đóng

vai trò quan trọng. Nó được dùng để mô tả cho chủ thể hoặc là nội dung phần bình luận của người dùng.

Do đó việc khai phá dạng dữ liệu dạng văn bản là rất hữu ích và có nhiều ứng dụng quan trọng. Khai phá dữ liệu dạng văn bản (Text Mining) có thể hiểu là việc khám phá tri thức được ẩn sau các dữ liệu dạng văn bản. Nó liên quan đến nhiều lĩnh vực khác như Học máy, Xử lý ngôn ngữ tự nhiên, Khai phá dữ liệu.

Quá trình Khai phá dữ liệu dạng văn bản có thể được tóm tắt qua 3 giai đoạn:

- Tiền xử lý (Văn bản preprocessing): Làm sạch và tạo ra đầu vào phù hợp cho giai đoạn Biểu diễn .
- Biểu diễn (Văn bản Presentation): Mô hình hóa dữ liệu. Một cách thông dụng để thực hiện mô hình hóa là chuyển đổi dữ liệu đã tiền xử lý thành dạng vector số. Dưới dạng vector số, ta có thể thực hiện các phép toán đại số tuyến tính lên nó (Dạng biểu diễn này được gọi là Bag of Words hay Vector Space Model).
- Khám phá tri thức (Knowledge Discovery): Áp dụng các kỹ thuật trong Học máy, Khai phá dữ liệu và Xử lý ngôn ngữ tự nhiên lên dạng Biểu diễn để tìm ra thông tin cần thiết.

Nhưng dữ liệu dạng văn bản trên MXH có những đặc trưng riêng và chúng là thách thức cho Quá trình Khai phá đã được chỉ ra trên đây. Những đặc trưng đó là: Tính nhạy cảm theo thời gian, Tính ngắn gọn, Tính phi cấu trúc và Tính phong phú về thông tin. Cụ thể:

- Tính nhạy cảm theo thời gian (Time Sensitivity): Dữ liệu trên các trang MXH được cập nhật từng phút. Các người dùng khác nhau trên Twitter hay Facebook đăng các post hoặc bình luận hàng giờ, hàng phút để tương tác với những người dùng khác. Do đó dữ liệu trên các trang MXH được cập nhật nhanh chóng theo thời gian. Việc truy vấn thông tin từ dữ liệu của các MXH ở thời điểm hiện tại đôi khi cho ra những kết quả đã cũ của nhiều phút trước.
- Tính ngắn gọn (Short Length): Các trang MXH thường có giới hạn nhất định về độ dài của các bài post, bình luận hay review. *Ví dụ*: Twitter giới hạn bài post tối đa 140 ký tự, Picasa giới hạn bình luận tối đa 512 ký tự. Sự ngắn gọn này là thách thức đối với các kỹ thuật khai phá truyền thống trên dạng dữ liệu văn bản.
- Tính phi cấu trúc (Unstructured Phrases):
  - Chất lượng của nội dung được biểu đạt qua một đoạn dữ liệu dạng văn bản không giống nhau tùy thuộc vào người viết. *Ví dụ*: Đối với cùng một địa điểm ăn uống trên Foody, một người có kinh nghiệm ăn uống sẽ viết một bài review "hay" so với một người chưa có kinh nghiệm.
  - Do những đặc trưng riêng của MXH, nội dung các bài post, bình luận, review thường chứa những tiếng lóng, từ viết tắt thiếu ngữ nghĩa ... gây khó khăn cho

các kỹ thuật phân tích. Ví dụ về một bình luận trên Foody: "Hqua mình k ăn món cá ở đó".

- Tính phong phú về thông tin (Abundant Information): Dữ liệu dạng văn bản trên MXH có thể tồn tại dưới nhiều hình thức và mang những ý nghĩa khác nhau. *Ví dụ*: Một bài post trên Facebook có nội dung chính là câu "Du lịch miền Tây", nó đi kèm với hashtag #mientay và một bức ảnh, trong đó mỗi phần của bức ảnh tag tên của những người dùng là bạn bè của người đăng. Những hình thức như hashtag hay tag hình ảnh trong Facebook cho phép dữ liệu dạng văn bản mang thông tin liên quan đến liên kết giữa các nốt trong mạng.

Những chủ đề trình bày tiếp theo là các ứng dụng của Khai phá dữ liệu dạng văn bản trên MXH trực tuyến.

Q&A (Questioning & Answer)

Một lượng lớn các câu hỏi và câu trả lời của người dùng xuất hiện hằng ngày trên các trang hỏi đáp (QA) như Yahoo! Answer, Quora hay StackOverFlow... Các trang QA này cũng là các social network, nơi mà người dùng có thể đặt câu hỏi và tìm kiếm câu trả lời cho các câu hỏi mà mình mong muốn, thu thập các ý kiến về một vấn đề nhất định hoặc tương tác với các user khác ở đó.

Theo thời gian, số lượng các câu hỏi và câu trả lời trên các trang QA trở nên rất lớn và tạo nên Database của riêng nó, do đó thay vì tìm kiếm các document từ Web hoặc post một câu hỏi mới trên các forum, user có thể vào các trang QA này và tìm câu trả lời cho các vấn đề quan tâm.

Một số vấn đề được đặt ra trên các trang QA. Thứ nhất, khi người dùng lên các trang QA để thực hiện truy vấn nhằm tìm kiếm một vấn đề quan tâm, làm thế nào để tìm ra các cặp QA đã có trên trang đó phù hợp với truy vấn mà người dùng thực hiện. Thứ hai, tương ứng với mỗi Question trong các cặp QA được tìm ra, tìm ra các Answer tốt nhất.

Social Tagging (dạng văn bản)

Người dùng có thể upload các object (ảnh, audio, văn bản...) lên MXH và đính kèm theo nó các keyword để mô tả, gán nhãn cho nội dung chính. Những keyword này được gọi là Social Tag (dạng văn bản). Các loại social tag này có nhiều ứng dụng trên MXH:

- Tìm kiếm dựa trên tag: Người dùng trên các trang MXH có thể tìm kiếm các thực thể được gắn tag tương tự hoặc liên quan ngữ nghĩa đến tag mà họ đem tìm kiếm. Vấn đề đặt ra là làm sao có thể tìm được các thực thể liên quan ngữ nghĩa này. *Ví dụ*: Khi người dùng tìm kiếm về một sự kiện gần đây, như "Cuộc cách mạng Ai Cập", hệ thống chỉ trả về các kết quả có tag là "Cuộc cách mạng" và "Ai Cập", những thông tin về các tag khác như "Mubarak từ chức" hay "Phản đối" mặc dù có liên quan đến tag "Cách mạng Ai Cập" bị bỏ qua. Hiện tượng này là do sự "thiếu ngữ nghĩa" trong việc xử lí tag, nguyên nhân là tag quá ngắn.
- Tag recommendation: Khi người dùng đưa một thực thể (ảnh, audio, văn bản, ...)

lên MXH trực tuyến, hệ thống sẽ đưa ra gợi ý cho người dùng những tag nào là phù hợp nhất cho object đó. Vấn đề đặt ra là làm thế nào để tạo ra một hệ thống Tag recommendation có hiệu quả.

- Phân loại các thực thể trên mạng: Cùng với sự bùng nổ của mạng xã hội nơi người dùng có thể đưa lên hàng triệu video, hình ảnh... mỗi giờ, sự phát triển nhanh chóng về số lượng và chủng loại của các thực thể này đặt ra yêu cầu cần thiết phải phân loại chúng một cách hiệu quả. Trên mạng xã hội, người dùng được cho phép gắn tag cho các thực thể. Các tag này mang theo nhiều thông tin mô tả cho thực thể mà nó gắn vào và có thể được dùng hỗ trợ cho việc phân loại.

Phân tích ý kiến trên MXH (Sentiment Analysis)

Phân tích ý kiến liên quan đến việc xác định và phân loại các ý kiến, cảm xúc được biểu đạt qua một đoạn văn bản hoặc một thực thể (hình ảnh, ...) nào đó; xác định thái độ của con người tới một chủ đề nhất định. Đối với các MXH trực tuyến, chủ thể là các blogger, user trên các MXH và họ thể hiện các thực thể là các post, comment, hình ảnh được upload ...

Phân tích ý kiến dựa trên dữ liệu dạng văn bản trên MXH là một hướng nghiên cứu quan trọng. Áp dụng các công cụ và giải thuật cho việc khai phá dữ liệu dạng văn bản trên MXH có nhiều ý nghĩa thực tế. *Ví dụ*: Người chủ kinh doanh có thể dùng Phân tích ý kiến để đánh giá thái độ, ý kiến của người dùng đối với một sản phẩm hoặc sự kiện nào đó. Giả sử, dùng Phân tích ý kiến phân tích các bài review và comment của người dùng cho một địa điểm ăn uống trên mạng Foody.vn. Thông qua các phân tích này, người chủ kinh doanh có thể thu thập được sự quan tâm hoặc khen/chê của người dùng đối với sản phẩm/sự kiện của mình và so sánh với các đối thủ cạnh tranh hoặc đưa ra các giải pháp khắc phục khuyết điểm.

Ngôn ngữ của con người vốn phức tạp, đặc biệt là ngôn ngữ được dùng trên MXH trực tuyến (nó có thể bao gồm cả từ lóng, từ viết tắt, từ viết sai lỗi chính tả... như được chỉ ra ở phần trên) do đó gây khó khăn cho việc phân tích.

Phát hiện sự kiện (Event Detection)

Khi có một sự kiện nổi bật nào đó ngoài thực tế, người dùng có xu hướng đăng các thông tin liên quan lên MXH. Từ những dữ liệu đó, chúng ta có thể phát hiện và theo dõi được những sự kiện nào đang xảy ra (Ở đây chủ yếu nói đến các bài đăng ở dạng văn bản).

Vấn đề đặt ra ở đây là làm thế nào để xác định được sự kiện đang diễn ra và mức độ "hot" của nó, theo dõi về diễn biến và sự lan truyền của sự kiện. *Ví dụ*: Phát hiện, giám sát về sự kiện để gợi ý cho người dùng quan tâm đến nó, Giúp cho truyền thông, báo chí tìm kiếm sự kiên và ranking mức độ "hot" của sự kiện.

Giải quyết những vấn đề này gặp phải khó khăn là dữ liệu về một sự kiện có ảnh hưởng lớn theo thời gian, đòi hỏi phải có giải pháp đối phó với việc xử lí real-time trong việc xác định sự kiện.

3 Kiến thức nền tảng

3.1 Tối ưu tập thuộc tính

Việc tối ưu tập thuộc tính là xác định các thuộc tính quan trọng và loại bỏ các thuộc tính dư thừa, nhiễu để làm giảm số chiều của không gian thuộc tính. Qua đó cải thiện độ chính xác và tốc độ tính toán của hệ thống. Các công cụ và giải thuật đề xuất sử dụng là:

Principal Components Analysis (PCA)

Định nghĩa

PCA là một thuật toán thống kê sử dụng phép biến đổi trực giao để biến đổi một tập hợp dữ liệu từ một không gian nhiều chiều sang một không gian mới ít chiều hơn (2 hoặc 3 chiều) nhằm tối ưu hóa việc thể hiện sự biến thiên của dữ liệu.

Ý tưởng

Tìm ra một không gian mới với số chiều nhỏ hơn không gian cũ với tiêu chí cố gắng phản ánh được càng nhiều thông tin gốc càng tốt, và thước đo cho khái niệm "thông tin" ở đây là phương sai. Đồng nghĩa với việc tìm những thuộc tính mới từ tổ hợp tuyến tính các thuộc tính cũ sao cho mỗi thuộc tính mới có phương sai càng lớn càng tốt.

Kiến thức cơ bản

Độ lệch chuẩn (Standard Deviation), Phương sai (Variance), Hiệp phương sai (Covariance), Ma trận hiệp phương sai (The covariance matrix), Eigenvectors, Eigenvalues.

Phương pháp

1. Xử lý dữ liệu:

   Cho ma trận $X = \left\{x_{ij}\right\} \in \mathbb{R}^{n \times p}$, trong đó n là số mẫu, p là số thuộc tính.

   Mỗi thuộc tính có một giá trị trung bình cộng:

   $$\overline{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$$

   Chuẩn hóa dữ liệu ta được ma trận mới $X^{new} = \left\{x_{ij}^{new}\right\}$, trong đó:

   $$x_{ij}^{new} = x_{ij} - \overline{x}_j$$

2. Tính ma trận hiệp phương sai của ma trận đã chuẩn hóa.

3. Tính Eigenvectors, Eigenvalues của ma trận hiệp phương sai.

   Mỗi eigenvector đại diện cho một thuộc tính, eigenvalue tương ứng càng lớn thì thuộc tính đó càng quan trọng.

Ưu điểm

- Phương pháp học không giám sát, do đó PCA sẽ tự động tìm ra được các thuộc tính đặc trưng mà con người không nhận ra được.
- Làm giảm độ nhiễu của tập dữ liệu.

Nhược điểm

- PCA chỉ làm việc với dữ liệu số.
- PCA phù hợp sử dụng đối với tập dữ liệu tuyến tính, do đó khi dữ liệu là phi tuyến thì dùng PCA thôi là chưa đủ.
- Quy tắc phương sai lớn nhất thì thuộc tính quan trọng nhất không luôn luôn đúng.

Univariate Feature Analysis

Định nghĩa

Univariate Feature Analysis là giải thuật xác định thuộc tính quan trọng dựa trên ý tưởng theo dõi ảnh hưởng của mỗi thuộc tính một cách độc lập đến hiệu suất dự đoán của hệ thống rồi sau đó chọn ra n thuộc tính mang lại hiệu suất tốt nhất làm tập thuộc tính cho mô hình học máy.

Phương pháp

1. Lựa chọn một thuộc tính bất kỳ, xây dựng hệ thống dựa chỉ dựa trên thuộc tính đó, đánh giá hệ thống: đối với bài toán hồi quy sử dụng giá trị *f-regression*.
2. Lựa chọn n thuộc tính có giá trị *f-regression* lớn nhất.

Ưu điểm

Giải thuật Univariate Feature Analysis đơn giản và thời gian chạy nhanh.

Nhược điểm

Đôi khi việc kết hợp các thuộc tính không tốt với nhau cũng có thể đưa ra mô hình dự đoán tốt.

Recursive Feature Elimination

Định nghĩa

Recursive Feature Elimination là giải thuật xác định thuộc tính quan trọng dựa trên ý tưởng áp dụng giải thuật trên tất cả các thuộc tính, sau đó loại bớt các thuộc tính sao cho giải thuật đạt kết quả tối ưu nhất.

Phương pháp

1. Train giải thuật trên tập các thuộc tính và sau đó gán trọng số (*Ví dụ*: hệ số của mô hình tuyến tính) cho mỗi thuộc tính.

2. Thuộc tính có trị tuyệt đối của trọng số tương ứng nhỏ nhất sẽ bị loại khỏi tập thuộc tính. Sau đó lặp lại bước 1. Quá trình lặp kết thúc khi giải thuật đạt được hiệu suất mong muốn hoặc đã đạt đến số thuộc tính mong muốn.

Ưu điểm

Kết quả thu được tốt hơn khi phân tích ảnh hưởng của thuộc tính đến kết quả một cách độc lập.

Nhược điểm

Chi phí tính toán cao.

3.2 Giải thuật học máy SVR

Trong lĩnh vực học máy, Support Vector Regression (SVR) là mô hình học có giám sát, được sử dụng vào bài toán phân tích hồi quy. Mô hình SVR là một mở rộng của giải thuật Support Vector Machine (SVM).

Ý tưởng

Giả sử có tập dữ liệu gồm l mẫu $\{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\}$, với $x_i \in \mathbb{R}^n$ và $y_i \in \mathbb{R}$ với mọi i. Mục đích của mô hình SVR là đi tìm một hàm *f(x)* sao cho số lượng các mẫu (x_i, y_i) thỏa mãn điều kiện $\mid y_i - f(x_i) \mid \leq \epsilon$ (ϵ cho trước) là lớn nhất.

Trường hợp hàm f tuyến tính, biểu diện dạng:

$$f(x) = \langle w, x \rangle + b \; vi \; w \in \mathbb{R}^n, b \in \mathbb{R}$$

$\langle ., . \rangle$: Tích có hướng của hai vector

Hàm f(x) càng tối ưu khi chuẩn của vector w càng nhỏ. Khi đó bài toán đặt ra là:

Phương pháp

Trường hợp hàm f tuyến tính, biểu diện dạng:

$$f(x) = \langle w, x \rangle + b \; vi \; w \in \mathbb{R}^n, b \in \mathbb{R}$$

$\langle ., . \rangle$: Tích có hướng của hai vector

Điều chỉnh w và b thỏa mãn hai điều kiện:

- Chuẩn của vector w nhỏ nhất: $min \parallel w \parallel$
- Hàm tổng khoảng cách từ các điểm dữ liệu đến siêu phẳng là nhỏ nhất. Hàm tính khoảng cách là hàm Vapnik's $\epsilon - insensitive$:

$$\mid y_i - f(x_i) \mid_\epsilon = \begin{cases} 0 & if \mid y_i - f(x_i) \mid \leq \epsilon \\ \mid y_i - f(x_i) \mid - \epsilon & otherwise \end{cases}$$

Kĩ thuật Kernel

Trong nhiều trường hợp, ta không thể tìm được siêu phẳng (hàm *f(x)*) thỏa mãn yêu cầu nằm gần với tất cả các điểm dữ liệu. Khi đó, sử dụng kĩ thuật ánh xạ không gian hiện tại sang không gian nhiều chiều hơn, ở đó ta có thể tìm ra siêu phẳng thỏa mãn. Tuy nhiên, ta không thể tự tạo ra các dữ liệu mới, do đó chỉ có thể suy diễn đặc trưng mới từ các không gian có sẵn.

Trong Hình 7, ta không thể tìm ra được đường thẳng hồi quy thỏa mãn trong không gian hai chiều, sau khi ánh xạ sang không gian nhiều chiều hơn bởi hàm $\phi(x)$ thì ta có thể tìm ra siêu phẳng mong muốn.

3.3 Gán nhãn cụm từ trong Xử lý ngôn ngữ tự nhiên

Trong Xử lý ngôn ngữ tự nhiên, để thực hiện được quá trình phân tích cú pháp thì một trong những vấn đề quan trọng là gán nhãn cho các từ/cụm từ ở trong văn bản, tức là "bẻ" các câu trong văn bản đó ra thành những thành phần nhỏ hơn (từ và cụm từ) sau đó xác định từ loại tương ứng với mỗi từ và cụm từ đó. Bảng 1 minh họa quá trình phân tích cú pháp câu "*Món lẩu ở nhà hàng đó rất ngon*".

Bảng 1: Ví dụ xác định từ loại tương ứng mỗi từ và cụm từ

<table>
<tr><td>Món</td><td>lẩu</td><td>ở</td><td>nhà hàng</td><td>đó</td><td>rất</td><td>ngon</td></tr>
<tr><td>N</td><td>N</td><td>P</td><td>N</td><td>ART</td><td>ADV</td><td>ADJ</td></tr>
<tr><td colspan="2">NP</td><td></td><td colspan="2">NP</td><td colspan="2">VP</td></tr>
<tr><td colspan="2"></td><td colspan="3">PP</td><td></td><td></td></tr>
<tr><td colspan="5">NP</td><td></td><td></td></tr>
</table>

Trong đó, *N*: Danh từ, *P*: Giới từ, *ART*: Mạo từ, *ADV*: Trạng từ, *NP*: Cụm danh từ, *PP*: Cụm giới từ.

3.4 Chuỗi thời gian (Time Series)

Chuỗi thời gian là một chuỗi các điểm dữ liệu, được đo theo từng khoảng khắc thời gian liền nhau theo một tần suất thời gian thống nhất. Công thức chuỗi thời gian:

$$y_t = f(y_{t-1}, y_{t-2}, ..., y_{t-n}) + w(t)$$

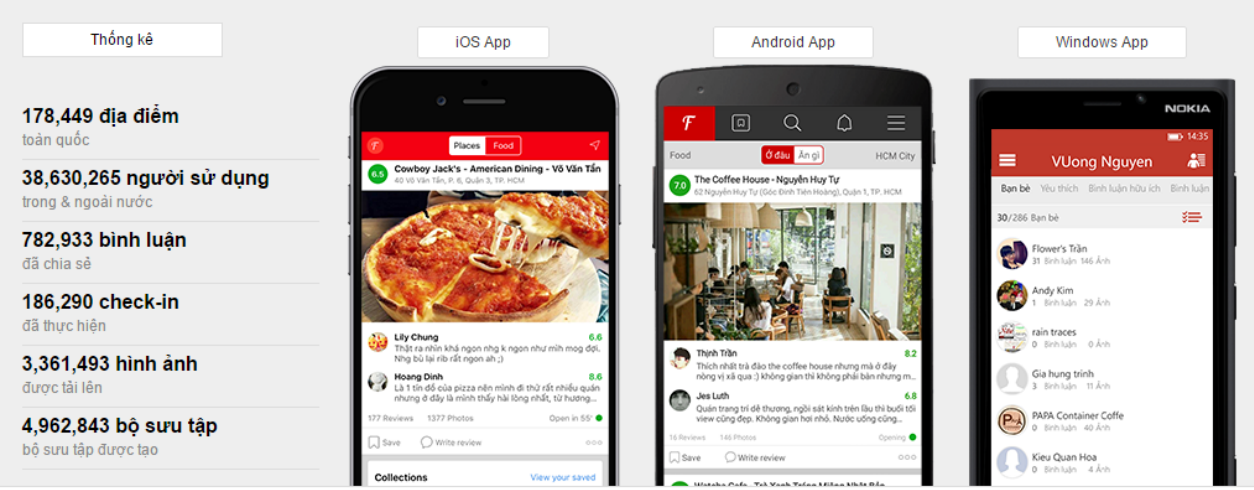
Từ công thức, nhận thấy có thể dự đoán sự kiện thời gian tại t (y_t) dựa vào các sự kiện đã biết trong quá ($y_{t-1}, y_{t-2}, ..., y_{t-n}$); $w(t)$: thông tin chưa biết; thông tin chưa biết này tạo nên độ nhiễu. Vấn đề dự đoán dựa trên một chuỗi giá trị trước có thể xem là giải bài toán hồi quy cơ bản có hướng dẫn (generic supervised learning regression). *Ví dụ*: Bài tóan dự đoán doanh số bán sản phẩm của quí (năm) tới dựa vào doanh số bán của các quí (năm) qua.

4 Tổng quan đề tài

4.1 Giới thiệu đề tài

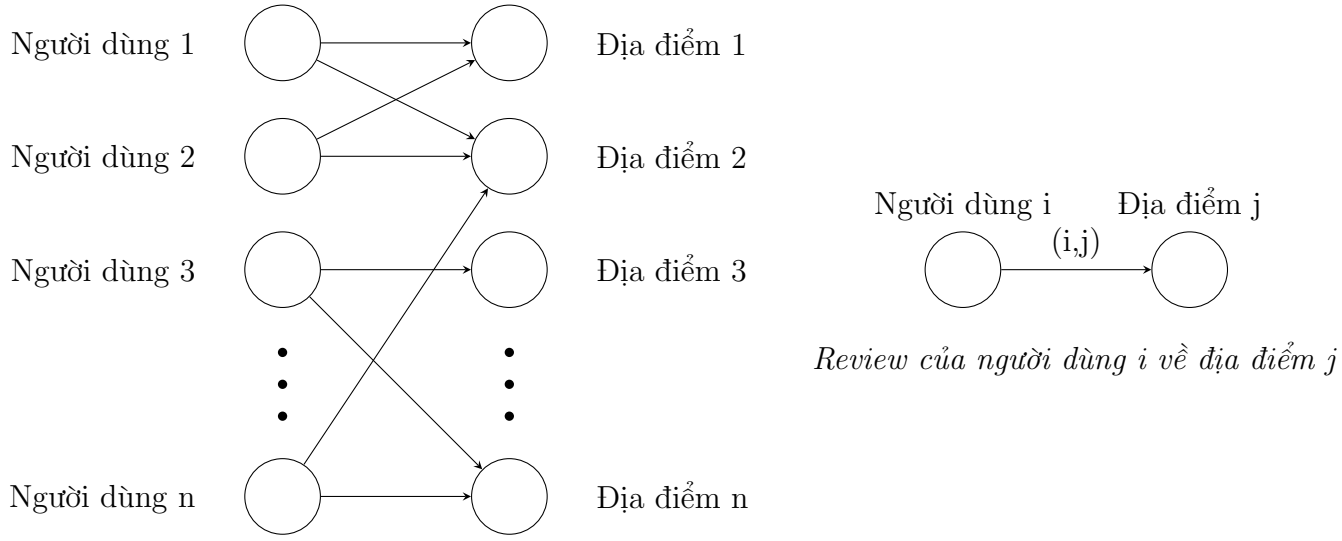
Cũng giống như các MXH trực tuyến nói chung, các trang MXH về ăn uống ngày càng trở nên phổ biến và trở thành một phần quan trọng trong cuộc sống của nhiều người. Ở đó người dùng có thể dễ dàng tìm kiếm ra địa điểm có những món ăn mà họ mong muốn, viết review cho những địa điểm mà họ đã đi hay tìm ra những người bạn có cùng gu ẩm thực. Ở Việt Nam hiện nay một số trang MXH về địa điểm ăn uống đã xuất hiện được một vài năm và phát triển mạnh mẽ với hàng trăm ngàn người dùng. Có thể kể đến đó là các trang MXH Foody, Lozi.

Trang web Foody.vn bắt đầu hoạt động từ năm 2012. Foody chứa thông tin về hàng chục nghìn địa điểm ăn uống và các người dùng trên đó. Mỗi địa điểm có nhiều thông tin liên quan như vị trí địa lý, điểm rating trung bình, số lượng review. Người dùng tới một địa điểm nào đó có thể viết review đánh giá và chấm điểm (rating) cho địa điểm. Đồng thời, người dùng cũng có thể thể hiện việc họ đã tới một địa điểm bằng cách dùng chức năng Checkin hoặc theo dõi những người dùng khác, những người có cùng gu ẩm thực hoặc viết review hay chẳng hạn.



Hình 1: Giao diện và các thống kê về trang foody.vn

Như đã trình bày ở phần 1.1, MXH nói chung có thể biểu diễn bằng đồ thị với nốt là các tác nhân và cạnh đại diện cho mối quan hệ hoặc tương tác giữa các nốt đó. Đối với MXH Foody, mỗi nốt là người dùng hoặc địa điểm, mỗi cạnh giữa một nốt là người dùng và một nốt là địa điểm được thể hiện qua review của người dùng cho địa điểm hoặc comment trên các review. Ngoài ra giữa các nốt là người dùng còn tồn tại các cạnh là quan hệ "theo dõi" giữa họ.



Hình 2: Mô hình hóa mạng xã hội foody.vn

Người chủ địa điểm ăn uống có thể vào trang mạng Foody.vn xem các review và rating của người dùng đối với địa điểm của mình, từ đó đánh giá được độ thu hút của địa điểm ở hiện tại. Sẽ là hữu ích nếu có một hệ thống dự đoán trước được số lượng các review và rating trung bình đối với địa điểm của họ trong tương lai. Những thông tin này sẽ là căn cứ để người chủ địa điểm xem xét mức độ thu hút của địa điểm trong tương lai và có thể đưa ra giải pháp cải thiện. Đề tài được sinh ra là để hướng tới những người chủ địa điểm đó.

Dữ liệu trên trang Foody với các nốt và liên kết đã được giải thích ở trên là cơ sở để thực hiện các phân tích trên nó, nhằm tạo ra một hệ thống học máy giải quyết vấn đề đưa ra. Vấn đề này có thể coi là một dạng của Dự đoán liên kết, ở đó mục tiêu là dự đoán xem trong tương lai sẽ có thêm bao nhiêu review nữa, tức là bao nhiêu cạnh nữa giữa các người dùng và một địa điểm sẽ được hình thành. Đồng thời một mục tiêu khác của bài toán là dự đoán một đặc điểm của nốt trong tương lai, đó là số điểm rating trung bình của một địa điểm. Để đạt được những mục tiêu vừa nêu, cần có thêm những kiến thức về Phân tích dựa trên dữ liệu dạng text và Phân loại nốt. Những sự liên hệ này sẽ được dẫn ra rõ hơn ở những phần tiếp theo.

Hình 3: Mô hình hóa bài toán

4.2 Các nghiên cứu liên quan

Yelp là một trang MXH về địa điểm nổi tiếng ở Mỹ với những chức năng tương tự như trang MXH Foody của Việt Nam. Trang MXH này tổ chức ra một cuộc thi mang tên Yelp Dataset Challenge (bắt đầu từ năm 2014, đến nay đã trải qua 6 vòng thi và đang ở vòng 7), thu hút hàng trăm bài nghiên cứu của các tác giả khắp thế giới. Những bài toán được đưa ra dựa trên tập dữ liệu (*) do cuộc thi này cung cấp bao gồm từ những bài toán Gợi ý các địa điểm phù hợp với sở thích, thói quen của người dùng [10], [16] tới Tóm tắt, trực quan hóa review cho người dùng [14] hay Dự đoán tương lai của địa điểm ăn uống.

Trong đó, nhóm bài toán có nhiều mối liên hệ nhất với đề tài do nhóm đề xuất là những bài toán về Dự đoán tương lai của một địa điểm ăn uống:

- Trong [21], tác giả đã đề xuất ra mô hình học máy nhằm dự đoán xem liệu một địa điểm sẽ tiếp tục hoạt động hay đóng cửa. Đó là một bài toán phân loại nhị phân với đầu vào là các thông tin liên quan tới địa điểm và đầu ra là một trong hai nhãn cho địa điểm: "open" – Địa điểm tiếp tục hoạt động, hoặc "close" – Địa điểm sẽ bị đóng cửa. Tác giả dùng 4 tập thuộc tính khác nhau cho mô hình học máy:
  - Nhóm các thuộc tính liên quan đến đánh giá: Số lượng review tương ứng, điểm rating trung bình, số lượng người dùng checkin.
  - Nhóm các thuộc tính về vị trí địa lý: Kinh độ, vĩ độ, bang.
  - Nhóm các thuộc tính liên quan thời gian: Thời gian (ngày) từ bài review cuối tới hiện tại, thời gian (ngày) từ bài tip cuối cùng tới hiện tại.
  - Nhóm thuộc tính liên quan đến dữ liệu văn bản trong review.

  Kết quả chỉ ra rằng nhóm thuộc tính về vị trí địa lý và nhóm liên quan đến thời gian cho hiệu quả dự đoán cao nhất.
- Trong [8], tác giả đã đề xuất một tập các thuộc tính dựa trên đặc trưng của địa điểm cũng như dữ liệu text trên review đối với địa điểm đó và dùng các mô hình học máy khác nhau (logistic regression, SVM, tree and random forest classifier, SVR) để dự đoán điểm rating cho địa điểm. Bài toán dự đoán điểm rating cho địa điểm được quy về bài toán phân loại cho địa điểm với các nhãn là 0,1,2,3,4,5 chính là điểm rating (được làm tròn) cho địa điểm đó. Kết quả cuối cùng cho thấy các thuộc tính liên quan đến phân tích ý kiến trên review, số lượng review, vị trí của địa điểm và số lượng các địa điểm nằm ở khu vực lân cận là những thuộc tính quan trọng nhất.
- Điểm rating đối với một địa điểm là cơ sở quan trọng để một người dùng đưa ra lựa chọn có tới địa điểm đó hay không. Nhưng đôi khi việc rating của người dùng cũng mang tính chất chủ quan và thiên vị. Ví dụ: Trên Yelp, người A đánh giá nhà hàng D ăn ngon, phục vụ tốt nhưng cho điểm là 4, trong khi người B có những đánh giá tương tự cho D nhưng cho điểm tới 5. Do đó một ý tưởng được đưa ra là dự đoán rating cho một địa điểm chỉ dựa trên dữ liệu văn bản trong review đối với địa điểm. Trong [15],

tác giả tạo một nhóm gồm các từ thông dụng thường xuất hiện nhất trong các bài review và dùng các từ trong nhóm này để tạo tập thuộc tính phục vụ cho hệ thống dự đoán rating. Các mô hình học máy được sử dụng là: Linear Regression, Support Vector Regression, Support Vector Regression với tập thuộc tính được chuẩn hóa và Decision Tree Regression.

- Trong [4], một mô hình học máy dựa trên SVR được đề xuất nhằm dự đoán mức độ thu hút của một địa điểm trong tương lai. Mức độ thu hút này được đánh giá theo thước đo là số lượng review đối với địa điểm trong tương lai (số lượng review mà địa điểm nhận được trong 6 tháng tới). Mô hình này dựa trên hai nhóm thuộc tính chính: Nhóm thuộc tính được trích ra từ tập dữ liệu của Yelp như điểm rating, vị trí của địa điểm, thời gian (ngày) từ bài review gần nhất... và nhóm thuộc tính có được qua quá trình thực hiện các kỹ thuật nhằm Phân tích ý kiến dựa trên review của địa điểm. Các nhóm thuộc tính này là cơ sở quan trọng cho tập thuộc tính trong Phương pháp đề xuất của nhóm.
- Trong [5], tác giả đề xuất ra hai mô hình dự đoán mức độ thu hút của một địa điểm trong tương lai, một mô hình dựa trên mô hình Chuỗi thời gian (Time Series) và một mô hình khác dựa trên Dự đoán liên kết trong đồ thị hai phía (Bipartite Graph). Mức độ thu hút này được thể hiện qua Số lượng review địa điểm nhận được trong tương lai và số điểm rating tại thời điểm đó. Hai mô hình này sử dụng tập thuộc tính tương tự như trong [4] nhưng không có nhóm thuộc tính liên quan tới dữ liệu văn bản trong các review. Một trong số các mô hình được nhóm đề xuất thực hiện cũng được đưa ra dựa trên ý tưởng của mô hình theo Chuỗi thời gian được nói tới trong bài.
- Trong [22], tác giả cũng đưa ra bài toán dự đoán xu hướng của địa điểm dựa vào số lượng review nhận được trong tương lai (Số lượng review nhận được trong tháng tới) sử dụng các phân tích thống kê trên Mô hình chuỗi thời gian.

Ở Việt Nam, từ dữ liệu của trang mạng Foody và các trang mạng xã hội về địa điểm khác, một số đề tài cũng đã được thực hiện:

- Trong [18] và [20], tác giả đã đưa ra bài toán Đề xuất địa điểm dựa trên dữ liệu của mạng Foody và dẫn ra những phương pháp cơ bản để giải quyết bài toán này, đó là phương pháp Lọc – Cộng tác (Collaborative Filtering), lọc dựa trên nội dung (Content-based Filtering) và phương pháp lai (Hybrid) như là sự kết hợp giữa hai phương pháp trên. Dữ liệu dùng cho phân tích được crawl từ trang Foody.vn, giải thuật học máy sẽ học những đặc trưng của địa điểm, hồ sơ của người dùng, và kết hợp với dữ liệu về mặt địa lý để hạn chế các kết quả không hợp lý về mặt khoảng cách.
- Trong [19], tác giả sử dụng các kỹ thuật Khai phá dữ liệu dạng văn bản (Trên tiếng Việt) để phân tích ý kiến dựa vào comment và review của người dùng đối với các địa điểm trên foody.vn và diadiemanuong.com. Tác giả áp dụng các giải thuật xử lý ngôn ngữ tự nhiên chia các câu trong comment và review này thành 5 chủ đề: giá cả, khu vực, dịch vụ, không gian và chất lượng. Phân tích các từ trong mỗi câu để chỉ ra ý kiến của người dùng đối với địa điểm là tích cực hay tiêu cực. Kết quả cuối cùng sẽ có được

bằng cách tính trung bình xem bao nhiêu phần trăm đánh giá là tích cực hay tiêu cực đối với địa điểm, và ứng với từng chủ đề trong số 5 chủ đề trên.

Những thành công nhất định trong các nghiên cứu trên là cơ sở quan trọng để nhóm đưa ra đề xuất cho phương pháp giải quyết thực hiện đề tài.

5 Phương pháp đề xuất

5.1 Tập dữ liệu

Foody.vn là trang web chứa thông tin về các địa điểm ăn uống ở Việt Nam (đặc biệt là TP. Hồ Chí Minh) với lượng thông tin phong phú và đáng tin cậy. Dữ liệu được dùng cho việc phân tích được crawl từ trang web foody.vn dùng thư viện jsoup và crawljax trên ngôn ngữ Java và được lưu trong một cơ sở dữ liệu không quan hệ (MongoDb). Jsoup và crawljax đều là các thư viện mã nguồn mở trên Java hỗ trợ việc trích xuất thông tin từ các trang web với API đơn giản và dễ sử dụng.

5.2 Đề xuất tập thuộc tính

Nhóm đề xuất 4 nhóm thuộc tính: Nhóm thuộc tính của địa điểm ăn uống trong một khoảng thời gian bất kì, Nhóm thuộc tính của địa điểm ăn uống trong một khoảng thời gian từ lúc thành lập đến hiện tại, nhóm thuộc tính trích xuất ra từ các review đối với địa điểm, và nhóm thuộc tính dựa trên phân loại người dùng.

Thuộc tính phụ thuộc khoảng thời gian bất kì

Kí hiệu địa điểm ăn uống là M. Khoảng thời gian (t1,t2) có thời điểm bắt đầu là t1 và thời điểm kết thúc là t2. Khi đó tập thuộc tính đặc trưng cho khoảng thời gian (t1,t2) là:

1. Số lượng review M nhận được trong khoảng thời gian (t1,t2).
2. Số điểm trung bình M nhận được trong khoảng thời gian (t1,t2).
3. Số lượng review M nhận được được yêu thích trong khoảng thời gian (t1,t2).
4. Số điểm cao nhất M nhận được: tính từ khi địa điểm hình thành đến hiện tại.
5. Số điểm thấp nhất M nhận được: tính từ khi địa điểm hình thành đến hiện tại.
6. Số ngày giữa bài review đầu tiên M nhận được đến t2.
7. Số ngày giữa bài review cuối cùng M nhận được đến t2.
8. Số ngày giữa bài review đầu tiên của địa điểm đến bài review cuối cùng địa điểm nhận được tính tới t2.

Thuộc tính phụ thuộc thời gian

Tập hợp thuộc tính đặc trưng cho địa điểm ăn uống M, tính từ khi M thành lập đến hiện tại:

1. Số bài bình luận M nhận được.
2. Số điểm cao nhất M nhận được.
3. Số điểm thấp nhất M nhận được.
4. Số điểm trung bình M nhận được.
5. Số địa điểm ăn uống cùng loại với M.
6. Số địa điểm ăn uống trong bán kính 1km của M.
7. Số bài review về M được yêu thích.
8. Số ngày kể từ bài review cuối cùng về M.
9. Số ngày kể từ bài review đầu tiên về M.
10. Số ngày giữa bài review đầu tiên và cuối cùng về M.
11. Số lượng bài review cao nhất trong 1 ngày M nhận được.
12. Số ngày kể từ bài review cuối cùng về một địa điểm ăn uống cùng loại với M.
13. Số lượng bài review của tất cả nhà hàng cùng loại với M.
14. Số lượng bài review của tất cả nhà hàng trong bán kính 1km của M.
15. Số lượng bài review được yêu thích của tất cả địa điểm ăn uống cùng loại với M.
16. Số lượng bài review được yêu thích của tất cả địa điểm ăn uống trong bán kính 1km của M.
17. Số ngày kể từ bài review cuối cùng về bất kỳ một địa điểm ăn uống trong bán kính 1km.

Thuộc tính trích xuất từ review

Mục tiêu của nhóm là tìm ra K = 100 từ/cụm từ phổ biến nhất (keyword) trong tất cả các bài review trên foody.vn và xem chúng là tập thuộc tính của một bài review. Sau đó với mỗi bài review, ta xem mỗi thuộc tính có bao nhiêu lần được đánh giá tích cực, bao nhiêu lần được đánh giá tiêu cực. Dữ liệu được dùng cho việc trích xuất này là mọi review của mọi địa điểm có trong tập dữ liệu. Thực hiện thứ tự các bước:

1. *Tiền xử lý cho các review*

   Như đã trình bày ở phần 3.3, dữ liệu dạng text trên MXH có đặc trưng là Ngắn gọn và Phi cấu trúc, do đó cần có quá trình tiền xử lý để giải quyết những điều này.

(a) Sửa lỗi chính tả.

(b) Loại bỏ các dấu câu hoặc từ bị thừa, thay thể một số chữ viết tắt, tiếng lóng thông dụng.

2. *Phân tích các câu, từ trong review và gán nhãn*

   Từ các review đã được tiền xử lý, dùng công cụ (đề xuất dùng công cụ VnChunker1.0 – Một công cụ dùng trong Xử lý ngôn ngữ tự nhiên) để phân tích các câu, từ trong review và gán nhãn. Đây chính là giai đoạn 2 và 3 của quá trình phân tích dữ liệu text như đã được đề cập ở trên. Sau khi gán nhãn cho từ và cụm từ, có thể lấy ra được:

   (a) Các cụm danh từ (NP) không chứa tính từ (thường ở dạng N-N).

   (b) Các tính từ.

3. *Xây dựng tập từ khóa*

   Trong các cụm danh từ vừa có được, chọn ra khoảng 100 cụm danh từ có mặt trong hầu hết các review làm tập từ khóa.

4. *Xây dựng tập tính từ và gán nhãn cho chúng*

   Tạo ra một tập các tính từ mang ý nghĩa tích cực và một tập các tính từ mang ý nghĩa tiêu cực của tiếng Việt bằng cách:

   (a) Định nghĩa ra hai danh sách nhỏ, một cho các tính từ tích cực và một cho các tính từ tiêu cực.

   (b) Dựa vào từ điển (?) tìm tất cả các tính từ đồng nghĩa hoặc trái nghĩa với các tính từ có sẵn trong 2 danh sách trên và bổ sung vào một trong hai danh sách trên một cách phù hợp.

5. *Kết hợp các cặp <từ khóa, tính từ>*

   Kết hợp kết quả của bước 2 và bước 3 để thực hiện phép kết hợp các cặp <từ khóa, tính từ>.

   Duyệt qua từng câu trong review và xét thử trong câu đó có những từ khóa nào, kết cặp với tính từ ở vị trí gần nó nhất. Sau khi làm tương tự với mọi câu trong mọi review, đếm số lượng tính từ tương ứng với mỗi từ khóa.

6. *Tính toán giá trị của từng từ khóa*

   Kết hợp với kết quả của bước 4 và 5, ta có thể tạo ra hai danh sách tương ứng với ý nghĩa tích cực và tiêu cực: Có bao nhiêu tính từ tích cực/tiêu cực đi kèm một từ khóa. Xây dựng nên danh sách có dạng:

   (a) Danh sách 1 <từ khóa, số tính từ tích cực đi kèm>

   (b) Danh sách 2 <từ khóa, số tính từ tiêu cực đi kèm>

Thuộc tính phân loại người dùng

Để dự đoán số lượng bình luận và số điểm trung bình về một địa điểm ăn uống, thông tin về người dùng cũng rất quan trọng. Số lượng người dùng rất lớn nên ta chia người dùng thành 6 nhóm cụ thể, sử dụng giải thuật k-means trên các thuộc tính của người dùng là:

1. Số bài bình luận.
2. Số điểm đánh giá trung bình.
3. Số bài bình luận được thích.
4. Thời điểm bài review đầu tiên.
5. Thời điểm bài review cuối cùng.
6. Khoảng cách thời gian giữa bài review đầu tiên và cuối cùng.

Mỗi nhóm người dùng có ảnh hưởng khác nhau đến sự thu hút của địa điểm ăn uống, cho nên được xem là thuộc tính của địa điểm ăn uống. *Ví dụ*: Nhà hàng sau khi được những người review uy tín (số bài review nhiều, số bài review được yêu thích nhiều) thì độ thu hút sẽ tăng lên.

5.3 Mô hình dự đoán

Nhóm đề xuất 3 mô hình dự đoán:

- Mô hình 1: Sử dụng tập thuộc tính 4.5.1
- Mô hình 2: Sử dụng tập thuộc tính 4.5.2
- Mô hình 3: Sử dụng tập thuộc tính 4.5.2 + 4.5.3 + 4.5.4

Với mỗi mô hình, áp dụng giải thuật học máy SVR trên tập thuộc tính:

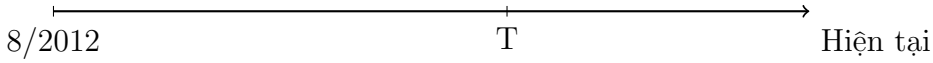
- Thuộc tính chưa giảm số chiều.
- Thuộc tính đã áp dụng giải thuật PCA.
- Thuộc tính đã áp dụng Univariate Feature Analysis.
- Thuộc tính đã áp dụng Recursive Feature Elimination.

Mô hình dự đoán số lượng review và mô hình dự đoán số điểm trung bình có cách thức hoạt động như nhau, nhưng được train và test riêng biệt nhau.

Đề xuất công cụ để áp dụng các giải thuật tối ưu tập thuộc tính PCA, Univariate Feature Analysis, Recursive Feature Elimination: Dùng thư viện scikit-learn (python).

Dữ liệu dùng để học và test

Lựa chọn một thời điểm T và sử dụng tất cả dữ liệu trước thời điểm T để xây dựng hệ thống dự đoán. Những thông tin sau thời điểm T sẽ được dùng để kiểm tra hệ thống, minh họa ở hình 4. Thời điểm T mà nhóm lựa chọn vẫn đang xem xét sao cho dữ liệu xây dựng hệ thống và kiểm tra hệ thống là hợp lý.



Hình 4: Lựa chọn thời điểm T

Mô hình 1: Sử dụng tập thuộc tính 4.5.1

Kí hiệu khoảng thời gian t trong tương lai với thời điểm hiện tại là H. Để dự đoán giá trị trong khoảng thời gian H, sử dụng dữ liệu của n khoảng thời gian trước thời điểm hiện tại với mỗi khoảng thời gian có độ dài bằng độ dài thời gian của H. Theo như trình bày ở phần 4.5.1, mỗi khoảng thời gian có 8 thuộc tính của địa điểm ăn uống tương ứng. Hệ thống dự đoán, một mẫu trong tập học và tập test có dạng:

$$Y_t \quad Y_{t-1}\, Y_{t-2} ... Y_{t-n} \;\; X1_{t-1}\, X1_{t-2} ... X1_{t-n} \; ... \; Xi_{t-1}\, Xi_{t-2} ... Xi_{t-n}$$

Trong đó:

- Y_t : Giá trị mong muốn tại thời điểm t, ở đây có thể là số lượng review hoặc là số điểm trung bình.
- Xi_{t-1} : Thuộc tính i tại thời điểm t-1.
- n : Số lượng thời điểm được sử dụng để đưa ra dự đoán cho thời điểm trong tương lai. Thời điểm được hiểu là thời điểm bắt đầu của một khoảng thời gian.

Giá trị mong muốn tại thời điểm t phụ thuộc vào giá trị mong muốn của n thời điểm trước đó và tất các các thuộc tính của mỗi thời điểm trong n thời điểm đó. Để có một mẫu dùng để học, chọn ngẫu nhiên dữ liệu n+1 tháng liên tiếp trong tập dữ liệu trong tập dữ liệu dùng để học. Nhóm sẽ sử dụng dữ liệu của n tháng trước (thuộc tính) để đưa ra dự đoán cho 1 tháng sau (giá trị cần dự đoán).

Ví dụ: Với bài toán dự đoán giá trị (review hoặc điểm trung bình của địa điểm ăn uống) 1 tháng sau ra sao (H có độ dài 1 tháng). Ta sử dụng dữ liệu của n = 5 tháng trước thời điểm hiện tại. Mỗi tháng sẽ có 8 thuộc tính (tham khảo phần đề xuất thuộc tính A)cộng thêm một giá trị mong muốn của tháng đó. Vậy tổng cộng một mẫu để sử dụng học và test (chưa thu giảm số chiều): có 45 thuộc tính đầu vào và 1 giá trị mong muốn cần dự đoán.

Mô hình 2: Sử dụng tập thuộc tính 4.5.2

Để dự đoán giá trị (số lượng review hoặc điểm số trung bình của địa điểm ăn uống) tại thời điểm *t*, sử dụng toàn bộ thông tin liên quan địa điểm đó từ khi thành lập đến hiện tại. Toàn bộ dữ liệu quá khứ dùng để hình thành tập thuộc tính, trình bày ở phần 4.5.2.

Xây dựng một mẫu dữ liệu để học bằng cách: Lựa chọn thời điểm k ($k < T$). Xem thời điểm k đó chính là giá trị cần dự đoán trong tương lai, và thời điểm k – *(n tháng)* chính là hiện tại. Khi đó: dữ liệu của địa điểm tính tới trước thời điểm k – *(1 tháng)* được sử dụng để tạo nên tập thuộc tính. Dữ liệu tại thời điểm k làm giá trị mong muốn. Vậy mỗi mẫu dữ liệu để học và test (chưa thu giảm số chiều) bao gồm: 17 thuộc tính, 1 giá trị mong muốn cần dự đoán (số lượng review hoặc sô điểm trung bỉnh).

Mô hình 3: Sử dụng tập thuộc tính 4.5.2 + 4.5.3 + 4.5.4

Tương tự ở mô hình 2, chỉ khác về tập thuộc tính. Mỗi mẫu dữ liệu để học và test (chưa thu giảm số chiều) bao gồm: 223 thuộc tính, 1 target đầu ra (số lượng review hoặc sô điểm trung bình).

5.4 Cách đánh giá hệ thống

Như đã trình bày ở mục 4.6, để đánh giá hệ thống, lựa chọn tập dữ liệu có dạng như dữ liệu dùng để học hệ thống, chỉ khác về việc lựa chọn sau thời điểm T.

Giả thuyết:

- y_i : Giá trị thực tế của testcase thứ i. Ở trong bài toán là số lượng review hoặc số điểm trung bình trên thực tế của testcase thứ i.
- $\hat{y}_i$: Giá trị hệ thống dự đoán của testcase thứ i. Ở trong bài toán là số lượng review hoặc số điểm trung bình mà hệ thống dự đoán của testcase thứ i.
- n : Số mẫu. Được hiểu là độ lớn không gian của tập test.

Các chỉ số đánh giá hệ thống

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid \hat{y}_i - y_i \mid$$

Mean Square Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

R-squared (R^2):

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y$$
$$SST = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$SSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$R^2 = 1 - \frac{SSE}{SST}$$

6 Tổng kết

Bài báo cáo đã cho thấy các vấn đề về tổng quan MXH và những chủ đề nghiên cứu lớn trên MXH. Sau đó báo cáo đi vào một bài toán cụ thể do nhóm đề xuất và trình bày về phương pháp thực hiện bài toán.

Tóm lại, trong giai đoạn Thực tập tốt nghiệp, nhóm đã hoàn thành được một số công việc sau:

- Tìm hiểu MXH và những chủ đề nghiên cứu trong Phân tích dữ liệu MXH.
- Tìm hiểu những công trình nghiên cứu liên quan đến MXH về địa điểm ăn uống và đề xuất một bài toán cụ thể.
- Lấy một phần dữ liệu cho bài toán (Từ foody.vn)
- Đề xuất phương pháp thực hiện bài toán.

Trong giai đoạn Luận văn tốt nghiệp, mục tiêu đặt ra là:

- Hoàn thành việc lấy dữ liệu cho bài toán.
- Thực hiện phương pháp đề xuất, cân nhắc thay đổi trong phương pháp để tạo ra hệ thống dự đoán có độ chính xác cao.

7 Tham khảo

[1]. Charu C. Aggawal, "Social Network Data Analytics" – Editor, 2011

[2]. Charu C. Aggawal, "Mining Text Data" – Editor, "Chapter 12 - Text Analytics In Social Media", 2012

[3]. Bing Liu, "Web Data Mining", 2011

[4]. Bryan Wood, Victor Hwang, Jennifer King, "Infering Future Bussiness Attention". Carnegie Mellon University, Yelp Challenge Round One, 2014

[5]. Kapil Jaisinghani, Ravi Todi, Zhengiy Liu, "Modeling Growth and Decline of Businesses in Yelp Network". Stanford University, CS224W Assignment Project, 2013

[6]. Guy Shani, Asela Gunawardana, "Evaluating Recommendation Systems". Microsoft Research, One Microsoft Way, Redmond, WA, 2009

[7]. Alex J. Smola, Bernhard Schölkopf, "A tutorial on Support Vector Regression". Australian National University, Canberra, Australia, 2004

[8]. Kyle Carbon, Kacyn Fujii, Prasanth Veerina, "Applications of Machine Learning to Predict Yelp Ratings". Stanford University, Yelp Dataset Challenge 2014

[9]. Dimitrios Kotzias, Misha Denil, Nando De Freitas, Padhraic Smyth, "From Group to Individual Labels Using Deep Features". Yelp Dataset Challenge Round Five, 2015

[10]. Sumedh Sawant, Gina Pai, "Yelp Food Recommendation System". Stanford University, Yelp Dataset Challenge 2014

[11]. Nabiha Asghar, "Yelp Dataset Challenge: Review Rating Prediction". University of Waterloo, Yelp Dataset Challenge, 2014

[12]. Julian McAuley, Jure Leskovec, "Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text". Stanford University, Yelp Dataset Challenge Round One, ACM RecSys '13 Proceedings, 2014

[13]. James Huang, Stephanie Rogers, Eunkwang Joo, "Improving Restaurants by Extracting Subtopics from Yelp Reviews". University of California, Berkeley, Yelp Dataset Challenge Round One, iConference 2014 Berlin, 2014

[14]. Ji Wang, Jian Zhao, Sheng Guo, Chris North, "Clustered Layout Word Cloud for User Generated Review". Virginia Tech and University of Toronto, Yelp Dataset Challenge Round One, Graphics Interface 2014 Montreal, 2014

[15]. Mingming Fan, Maryam Khademi, " Predicting a Business' Star in Yelp from Its Reviews' Text Alone". University of North Carolina and University of California. Yelp Dataset Challenge Round One, 2014

[16]. Prerna Wivedi, Nikita Chheda, "A Hybrid Restaurant Recommender", International Journal of Computer Applications (0975 – 8887) Volume 55– No.16, October, 2012

[17]. Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[18]. Nguyễn Minh Khôi, Luận văn Thạc sĩ "Hệ Thống Đề Xuất Địa Điểm Dựa Trên Phương Pháp Lai (Hybrid) Trên Dữ Liệu Foody.vn". Đại học Công nghệ TP. Hồ Chí Minh, 2015

[19]. Dang Hung Phan, Tuan Dung Cao, "Applying Skip-gram Word Estimation And SVM-based Classification for Opinion Mining Vietnamese Food Places Text Reviews"

[20]. Lê Quốc Nam, Nguyễn Ngọc Dũng, Luận Văn Tốt Nghiệp "Giới Thiệu Địa Điểm Trên Điện Thoại Thông Minh". Đại học Bách Khoa TP. Hồ Chí Minh, 2014

[21]. Narenkuma Pandian, Vaibhab Aggawal, "Are You Open", Stanford University

[22]. Praveenkumar Kondiloppa, Seung-Jong Park, "Modeling Business Trends Based on Yelp Review", School of Electrical Engineering and Computer Science, Lousiana State University