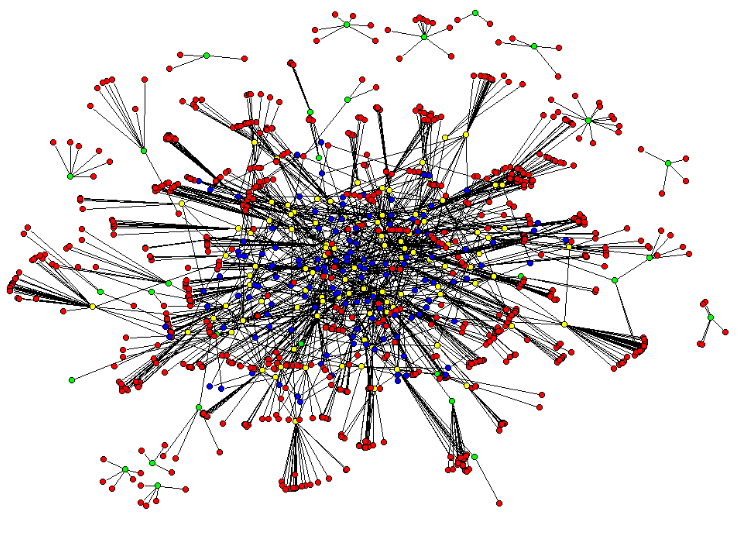
**MỞ ĐẦU**

**Giới thiệu**

Với sự phát triển mạnh mẽ của internet toàn cầu hiện nay, nhu cầu trao đổi, tìm kiếm thông tin online ngày càng bùng nổ. Đặc biệt, sự ra đời của các trang mạng xã hội, các diễn đàn, website cho phép bình luận trực tuyến với số lượng người dùng đông đảo trên toàn thế giới đã trở thành một nơi trao đổi, chia sẻ thông tin không thể thiếu của bất kỳ cá nhân, tổ chức nào. Do đó, các trang mạng xã hội này như một mô hình thu nhỏ của thế giới thực, và là nơi cất giữ lượng thông tin và các mối quan hệ khổng lồ. Một số mạng xã hội phổ biến hiện nay có thể kể đến như: Twitter, Facebook, LinkedIn, Google+... Việc khai thác và phân tích có hiệu quả lượng thông tin khổng lồ này sẽ đem lại rất nhiều lợi ích trong các lĩnh vực như: kinh tế, y tế, chính trị, xã hội, an ninh - quốc phòng… Vì vậy, khai thác thông tin trên mạng xã hội để ứng dụng vào nhu cầu thực tế đang là một hướng đi mới thu hút được nhiều sự chú ý của các nhà khoa học trên toàn thế giới.



Hình 0.1: Mô hình đồ thị mạng xã hội 1

1 http://eclectic.ss.uci.edu/~drwhite/socnetweb/courses/BlockSt.GIF

Mạng xã hội thường được định nghĩa như là một đồ thị được tạo thành bởi các đỉnh (nút) và các cạnh. Các đỉnh của đồ thị là tập các đối tượng và các cạnh thể hiện mối quan hệ hoặc sự tương tác giữa các đối tượng này.

Dự đoán liên kết trong mạng xã hội là một là một bài toán đang rất được quan tâm trong những năm gần đây bởi tính mới mẻ và tính ứng dụng thực tế cao của nó.

Bài toán dự đoán liên kết trong mạng xã hội có thể được chia thành 4 bài toán con như sau:

* + Dự đoán sự tồn tại của liên kết.
  + Dự đoán loại liên kết.
  + Dự đoán trọng số liên kết.
  + Dự đoán số lượng liên kết.

Trong khuôn khổ đề tài *Khóa luận tốt nghiệp,* luận văn chỉ xem xét bài toán *Dự đoán loại liên kết.* Quan điểm đánh giá giữa các người này với người khác được sử dụng phổ biến trong các bài diễn văn, bài luận, cuộc đàm thoại... Các quan điểm này có vai trò quan trọng trong việc xây dựng danh tiếng, các mối quan hệ xã hội và khảo sát lấy ý kiến công chúng. Việc dự đoán một liên kết là "positive" (tích cực) hay "negative" (tiêu cực) có thể sử dụng riêng biệt hai phương pháp là *Phân tích cấu trúc mạng xã hội có dấu* và *Phân tích tâm lý người dùng bằng văn bản text.* Tuy nhiên, các phương pháp tiếp cận này đã làm bỏ qua sự tương tác phong phú giữa ngôn ngữ và bối cảnh xã hội. Phương pháp mới được đề xuất bởi *Robert* có thể nắm bắt được các tương tác như vậy bằng cách tổng hợp các thông tin từ các mạng xã hội có dấu như: *mối quan hệ giữa các người dùng trong mạng xã hội và các văn bản đánh giá liên quan*.

**Phát biểu bài toán**

Mạng xã hội là một mô hình mạng (đồ thị) G được cấu tạo bởi các đỉnh và các cạnh:

* Các đỉnh là tập các đối tượng V.
* Các cạnh là tập các liên kết E.

Luận văn quy định rằng liên kết "positive" sẽ được biểu diễn bằng một cạnh có dấu "+" và liên kết "negative" sẽ được biểu diễn bằng một cạnh có dấu "-" trên đồ thị G.

Bài toán dự đoán loại liên kết trong mạng xã hội có Input và Output như sau:

* Input: Thông tin dấu hiệu cạnh có sẵn của một đồ thị mạng xã hội tại một thời điểm cụ thể.
* Output: Dấu hiệu cạnh của tập chưa được gán nhãn trong đồ thị đó.

**Ứng dụng của bài toán**

Bài toán *Dự đoán loại liên kết trong mạng xã hội* có thể ứng dụng vào trong các lĩnh vực thực tế như:

**Thách thức của bài toán**

Bản thân bài toán *Dự đoán loại liên kết trong mạng xã hội* vốn đã có những thách thức nhất định. Khi xem xét mạng xã hôị trong ngữ cảnh có hai loại liên kết "positive" và "negative", kết hợp với việc *Phân tích văn bản text* sẽ làm tăng độ phức tạp của bài toán so với các phương pháp tiếp cận thông thường chỉ sử dụng một trong hai phương pháp trên. Chẳng hạn như các vấn đề sau:

* *Tính biến động và độ lớn của mạng xã hội*: Mạng xã hội là loại dữ liệu lớn có tính biến động cao và khả năng bùng nổ về kích thước là rất lớn với số lượng người dùng đông đảo. Ngoài ra, các liên kết thêm vào hay bỏ đi cũng thường xuyên có những sự thay đổi. Điều này làm cho việc xem xét các đặc trưng của mạng xã hội phải trả giá cao về độ phức tạp.
* *Tính không đầy đủ của thông tin được chia sẻ:* Giá trị lớn nhất trên các mạng xã hội đó là những thông tin mà người dùng chia sẻ cho nhau, và các thông tin này là yếu tố mang tính quyết định đối với bài toán *Dự đoán loại liên kết*. Tuy nhiên, các thông tin này thường được chia sẻ dưới dạng không đầy đủ, gây khó khăn cho việc phân tích. Vì thế bằng cách nào đó, chúng ta phải suy diễn để làm đầy các thông tin này nhằm đáp ứng nhu cầu bài toán đặt ra, điều này là một trong những yếu tố ảnh hưởng trực tiếp đến độ chính xác và tính hiệu quả của bài toán.
* *Sự khác biệt lớn giữa các mạng xã hội*: Đây cũng là một thách thức rất lớn, bởi có sự khác biệt khá lớn giữa tính chất của những thông tin trong từng mạng xã hội. Các định nghĩa của đỉnh và cạnh trong mỗi mạng cũng sẽ có sự khác nhau. Vì vậy, tìm ra các đặc trưng mang tính tổng quát, không phụ thuộc vào ngữ cảnh riêng của từng mạng xã hội cũng là một vấn đề cần thiết.
* *Các văn bản text thường mang tính chủ quan của người viết*: Đây là một điều dễ hiểu, bởi mỗi người có một phong cách viết khác nhau, các câu đánh giá cũng thường có những sự pha trộn giữa tích cực và tiêu cực hoặc mang nghĩa hàm ý, hay các chữ viết tắt, các ký hiệu hoặc có sự đan xen giữa nhiều ngôn ngữ khác nhau... Do đó, để phân tích cảm xúc của một số lượng lớn người dùng mạng xã hội thông qua văn bản text thu được cần phải tìm ra những đặc trưng chung nhất, đồng thời có các phương pháp làm tăng độ mạnh của quan điểm tạo thuận lợi cho việc đánh giá quan điểm người dùng. Do hạn chế về mặt thời gian, luận văn bỏ qua việc xem xét các từ sai chính tả, các ký hiệu, chữ viết tắt.

Những khó khăn còn tồn tại trên và những điểm khác biệt với cách tiếp cận mới lạ của bài toán dự đoán loại liên kết "positive" và "negative" trong mạng xã hội mà luận văn trình bày so với các hướng tiếp cận truyền thống hứa hẹn sẽ trở thành một bài toán thú vị.

**Mục tiêu của luận văn**

Luận văn hướng đến mục tiêu nghiên cứu đề xuất cải tiến phương pháp dự đoán loại liên kết "positive" và "negative" bằng cách sử dụng đồ thị mạng xã hội kết hợp phân tích văn bản text theo mô hình máy học với mong muốn nâng cao tính hiệu quả theo các tiêu chí sau:

* Nâng cao độ chính xác dự đoán.
* Rút ngắn thời gian rút trích đặc trưng, quá trình huấn luyện và kiểm tra dữ liệu.

Nghiên cứu bộ dữ liệu của mạng xã hội trực tuyến Wikipedia và tiến hành cài đặt thực nghiệm trên mạng xã hội này. Đồng thời, mở rộng cài đặt thực nghiệm với bộ dữ liệu khác là...

**Đối tượng và phạm vi nghiên cứu**

Dấu hiệu của các liên kết trong đồ thị mạng xã hội của mạng trực tuyến Wikipedia.

Dữ liệu của mạng xã hội Wikipedia được công khai và cho phép tải về tại địa chỉ <http://snap.stanford.edu/data/wiki-RfA.html>.

**Đóng góp của luận văn**

**Cấu trúc của luận văn**

Bố cục của luận văn gồm 5 phần với các nội dung chính như sau:

* Chương 1. Tổng quan các Hướng tiếp cận

Chương này trình bày các hướng tiếp cận chính của bài toán dự đoán loại liên kết trong mạng xã hội. Đồng thời, trình bày, phân tích và đánh giá các công trình nghiên cứu tiêu biểu có liên quan đã được công bố. Từ đó, xác định những vấn đề còn tồn tại mà luận văn sẽ tập trung giải quyết.

* Chương 2. Cơ sở lý thuyết

Nội dung của chương này là trình bày các cơ sở lý thuyết cần thiết để xây dựng phương pháp dự đoán loại liên kết "positive" và "negative" kết hợp giữa phân tích cấu trúc đồ thị mạng xã hội và các văn bản đánh giá liên quan. Nội dung trình bày bao gồm:

* + Mô hình tổng quát được dùng cho bài toán dự đoán loại liên kết trong mạng xã hội.
  + Các đặc trưng thường được sử dụng: bao gồm đặc trưng của đồ thị mạng xã hội và đặc trưng văn bản text.
  + Phương pháp máy học (SVM) xây dựng mô hình dự đoán dấu của liên kết.
* Chương 3. Mô hình dự đoán

Nội dung chính của chương 3 là trình bày chi tiết về mô hình dự đoán loại liên kết mà luận văn sử dụng

* Chương 4. Kết quả Thực Nghiệm và Đánh Giá

Trình bày kết quả thực nghiệm mà luận văn đạt được. Từ đó đưa ra các kết luận và định hướng nghiên cứu mở rộng đề tài trong tương lai.

**CHƯƠNG 1. TỔNG QUAN CÁC HƯỚNG TIẾP CẬN**

Mạng xã hội là một mô hình mạng (đồ thị G) được cấu tạo bởi các đỉnh (tập các đối tượng V) và các cạnh (tập các liên kết E). Liên kết "positive" được biểu diễn bằng cạnh có dấu "+". Các nghiên cứu gần đây về bài toán dự đoán loại liên kết "positive" và "negative" trong mạng xã hội có thể chia thành 2 hướng tiếp cận chính:

* Phân tích ma trận
* Máy học

**1.1. Phân tích ma trận**

***1.1.1. Phương pháp lan truyền độ tin cậy***

Đây là phương pháp nghiên cứu đầu tiên được đề xuất bởi Guha và cộng sự, đánh dấu sự phát triển của bài toán dự đoán loại liên kết trong mạng xã hội trong những năm gần đây. Phương pháp này dựa trên ý tưởng của các thuật toán lan truyền trên mạng, từ đó đưa ra mô hình lan truyền độ tin cậy để dự đoán các mức độ tin cậy giữa các cặp đỉnh trong mạng xã hội. ***Bảng 1.1*** thể hiện các ma trận được Guha và cộng sự sử dụng trong mô hình lan truyền độ tin cậy của mình:

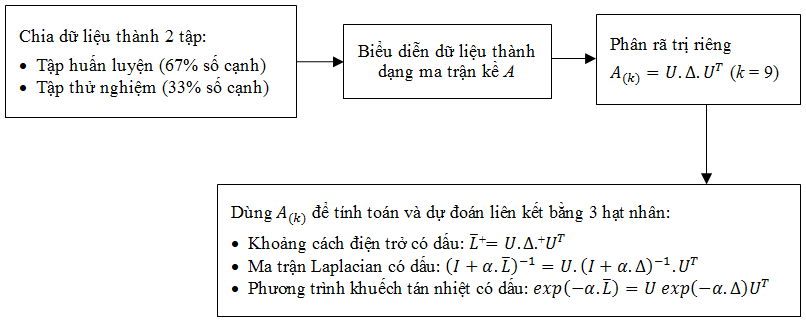
|  |  |  |
| --- | --- | --- |
| Tên | Diễn giải | Giá trị |
| T | Ma trận tin cậy (t\_uv thể hiện mức độ tin tưởng giữa u và v) | [0, 1] |
| D | Ma trận không tin cậy (d\_uv thể hiện mức độ không tin tưởng giữa u và v) | [0, 1] |
| B | Ma trận độ tin cậy kết hợp |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Bảng 1.1.** Danh sách các ma trận sử dụng trong mô hình lan truyền độ tin cậy

Trong mô hình lan truyền độ tin cậy, ma trận C\_(B,α)

***1.1.2. Phương pháp phân tích phổ đồ thị***

Phương pháp tiếp cận phân tích phổ đồ thị được đề xuất bởi tác giả Kenegis và các cộng sự. Trong phương pháp này, Kenegis sử dụng các cá nhân được dẫn xuất từ các biến thể có dấu của ma trận Laplacian (khoảng cách điện trở có dấu, ma trận Laplacian có dấu và phương trình khuếch tán nhiệt có dấu). Quy trình thực hiện được mô hình hóa như **Hình 1.2** dưới đây:



**Hình 1.2.** Mô hình phân tích phổ đồ thị

Tác giả chứng minh sự hiệu quả của phương pháp do mình đề xuất với các phương pháp khác đã được sử dụng là:

* + Rút gọn hạng ma trận
  + Tổng lũy thừa
  + Hàm mũ ma trận

Kết quả thực nghiệm chạy trên mạng xã hội Slashdot cho thấy phương pháp được Kenegis đề xuất cho kết quả tốt hơn các phương pháp trước đó (đạt kết quả khả quan nhất trên ma trận Laplacian có dấu). Tuy nhiên, trong nghiên cứu này, tác giả không đưa ra phương pháp học có giám sát nào để đưa ra hệ số cho mô hình của mình và đồng thời, tính chất về dữ liệu cũng chưa được quan tâm đúng mức.

Đặc điểm chung của các phương pháp tiếp cận theo hướng ma trận là tính toán phức tạp, tốn nhiều thời gian xử lý, đặc biệt khi áp dụng cho các mạng xã hội lớn. Do đó, trong một vài năm trở lại đây, các nhà khoa học bắt đầu quan tâm tới một hướng tiếp cận mới khắc phục được các nhược điểm trên đó là hướng tiếp cận máy học. Tuy vậy, chúng ta không thể phủ nhận vai trò của hướng tiếp cận ma trận đối với sự phát triển chung của bài toán dự đoán loại liên kết trong mạng xã hội.

**1.2. Hướng tiếp cận máy học**

***1.2.1. Phương pháp khai thác đặc trưng dựa vào lí thuyết cân bằng***

Trong các công trình nghiên cứu của Leskovec và cộng sự, các tác giả tập trung nghiên cứu xây dựng các đặc trưng dựa trên lí thuyết cân bằng.

* + *Lí thuyết cân bằng*

Đây là một lí thuyết được ứng dụng rộng rãi trong các lĩnh vực xã hội xuất hiện vào khoảng giữa thế kỉ XX do Heider phát biểu vào năm 1940. Lí thuyết cân bằng sau này được chuyển thành "ngôn ngữ" lí thuyết đồ thị bởi Cartwright và Harray.

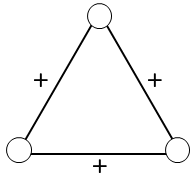
Lí thuyết cân bằng xem xét tính chất cân bằng của mạng xã hội thông qua những cách thức liên kết có thể xảy ra của bộ ba phần tử trong mạng xã hội (Triads). Từ đó, chỉ ra những cách thức liên kết nào là cân bằng hoặc không cân bằng.

Theo lí thuyết này, một bộ ba được gọi là cân bằng khi số liên kết "positive" giữa chúng là số lẻ với các quy luật sau:

* + 1. Bạn của bạn mình là bạn của mình (a là bạn của b, b là bạn của c suy ra a là bạn của c).
    2. Bạn của kẻ thù là kẻ thù của mình (a là bạn của b, b là kẻ thù của c suy ra a là kẻ thù của c)
    3. Kẻ thù của bạn mình là kẻ thù của mình (a là bạn của b, c là kẻ thù của b thì a cũng có kẻ thù là c).
    4. Kẻ thù của kẻ thù của mình là bạn của mình (a là kẻ thù của b, c là kẻ thù của b, do đó a và c là bạn).

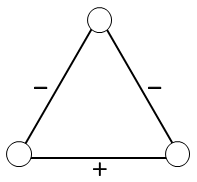
Các quy luật trên được mô hình hóa trên các bộ ba như sau:

**Hình 1.3** mô tả bộ ba cân bằng theo quy luật (1):



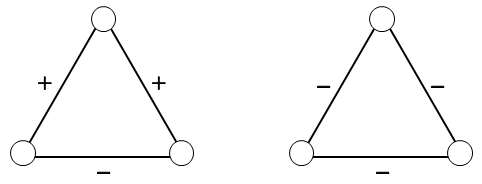
**Hình 1.3**. Bộ ba cân bằng theo quy luật (1)

Các quy luật cân bằng (2), (3), (4) được biểu diễn như **Hình 1.4** dưới đây:



**Hình 1.4.** Bộ ba cân bằng theo quy luật (2), (3), (4)

Các trường hợp còn lại không tuân theo quy luật cân bằng được mô tả trong **Hình 1.5** như sau:



**Hình 1.5.** Các bộ ba không cân bằng

* + *Lí thuyết trạng thái*

Lí thuyết trạng thái được phát triển dựa trên nền tảng là lí thuyết cân bằng. Lí thuyết trạng thái cho rằng một liên kết "positive" hoặc "negative" chứa nhiều ý nghĩa hơn trong lí thuyết cân bằng. Cụ thể:

* + Đối với liên kết "positive" cho cạnh ab:

a và b là bạn (giống với lí thuyết cân bằng).

a có trạng thái thấp hơn b (điểm khác biệt).

* + Tương tự đối với liên kết "negative" cho cạnh ab:

a và b là kẻ thù (giống với lí thuyết cân bằng).

a có trạng thái thấp hơn b (điểm khác biệt).

Bậc của trạng thái có thể được lan truyền nhiều lần thông qua đường đi của mạng xã hội, từ đó, lí thuyết trạng thái tạo ra kết quả dự đoán khác hơn so với lí thuyết cân bằng.