BỘ CÔNG THƯƠNG

**TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ**

**PHÂN TÍCH MẠNG XÃ HỘI**

**TÌM HIỂU BÀI TOÁN PHÁT HIỆN CỘNG ĐỒNG TRONG MẠNG XÃ HỘI**

**Giảng viên hướng dẫn: T.S Vũ Đức Thịnh**

**Sinh viên thực hiện: 2001210087 – Trần Quốc Kha**

**2044210002 – Bùi Đức Huy**

# PHÂN CÔNG NHIỆM VỤ

|  |  |  |  |
| --- | --- | --- | --- |
| MSSV | Họ và tên | Phân công | Mức độ hoàn thành |
| 2001211738 | Trần Quốc Kha | * Tổng hợp nội dung báo cáo word | 100% |
| 2044210002 | Bùi Đức Huy |  | 100% |

# DANH MỤC CÁC TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **Viết tắt** | **Diễn giải** | **Giải thích** |
| COPRA | Community Overlap Propagation Algorithm | Thuật toán lan truyền nhãn cho phép các đỉnh thuộc về nhiều cộng đồng khác nhau. |
| CONGA | |  | | --- | | CommunityOverla  Newman-Girvan Algorithm | | Thuật toán mở rộng từ Girvan-Newman để xử lý các cộng đồng giao nhau. |
| SC | Spectral Clustering | Thuật toán phân cụm sử dụng ma trận Laplacian và phân tích giá trị riêng để tìm ra cộng đồng. |
| LPA | Label Propagation Algorithm | Thuật toán lan truyền nhãn, hoạt động bằng cách cập nhật nhãn của mỗi đỉnh dựa trên nhãn của láng giềng để tạo thành cộng đồng. |
| GN | Girvan-Newman Algorithm | Thuật toán xác định cộng đồng bằng cách loại bỏ các cạnh có độ trung gian cao nhất trước tiên. |
| KL | Kernighan-Lin Algorithm | Thuật toán tối ưu hóa phân vùng đồ thị bằng cách hoán đổi các đỉnh giữa hai nhóm để giảm số cạnh bị cắt. |

# LỜI MỞ ĐẦU

Trong thời đại công nghệ số hiện nay, mạng xã hội đã trở thành một phần không thể thiếu trong cuộc sống của con người. Các nền tảng như Facebook, Twitter, Instagram hay LinkedIn kết nối hàng triệu người trên toàn thế giới, hình thành nên các cộng đồng với những đặc điểm và mối quan hệ phức tạp.

Việc phát hiện cộng đồng trong mạng xã hội không chỉ giúp hiểu rõ hơn về cấu trúc và hành vi của người dùng mà còn có ý nghĩa quan trọng trong nhiều lĩnh vực như tiếp thị, phân tích dữ liệu, an ninh mạng và nghiên cứu khoa học.

Bài toán phát hiện cộng đồng trong mạng xã hội là một lĩnh vực nghiên cứu quan trọng trong khoa học dữ liệu và trí tuệ nhân tạo. Có nhiều phương pháp và thuật toán đã được đề xuất nhằm xác định các nhóm người có sự tương đồng cao trong một mạng xã hội.

Bài báo cáo này sẽ trình bày tổng quan về bài toán, các phương pháp tiếp cận phổ biến, cũng như công cụ hỗ trợ và kết quả thực nghiệm từ việc áp dụng một số thuật toán vào phân tích cộng đồng.

Thông qua nghiên cứu này, nhóm chúng tôi mong muốn mang lại một cái nhìn tổng quan về vấn đề phát hiện cộng đồng trong mạng xã hội, đồng thời đánh giá hiệu quả của các thuật toán và phương pháp khác nhau. Hy vọng rằng báo cáo này sẽ góp phần giúp ích cho những ai quan tâm đến lĩnh vực phân tích mạng xã hội và ứng dụng của nó trong thực tế.

# MỤC LỤC

[CHƯƠNG 1:GIỚI THIỆU CHUNG 5](#_Toc180228146)

[1.1 Lý do chọn đề tài 5](#_Toc180228157)

[1.2 Mục Tiêu Nghiên Cứu 8](#_Toc180228157)

[CHƯƠNG 2:BÀI TOÁN PHÁT HIỆN CỘNG ĐỒNG TRÊN MẠNG XÃ HỘI 11](#_Toc180228156)

[2.1 Cộng đồng mạng xã hội 11](#_Toc180228157)

[2.2 Phân tích các ràng buộc 12](#_Toc180228157)

[2.3 Cấu trúc thuật toán 13](#_Toc180228157)

[2.4 Các phương pháp tiếp cận 15](#_Toc180228157)

[CHƯƠNG 3:THIẾT KẾ HỆ THỐNG 18](#_Toc180228165)

3.1 [Cấu Trúc Dữ Liệu 18](#_Toc180228157)

[3.2 Sơ Đồ Xử Lý 19](#_Toc180228157)

[3.3 Thuật Toán Giải Quyết Bài Toán 20](#_Toc180228157)

[CHƯƠNG 4:CÀI ĐẶT VÀ THỰC NGHIỆM 20](#_Toc180228165)

[4.1 Môi trường cài đặt 21](#_Toc180228157)

[4.2 Các bước thực hiện 22](#_Toc180228157)

[4.3 Kết quả thực nghiệm 23](#_Toc180228157)

[CHƯƠNG 5: KẾT QUẢ VÀ ĐÁNH GIÁ](#_Toc180228165)

5.[1 Đánh giá kết quả](#_Toc180228157)

[5.2 So sánh với các phương pháp khác](#_Toc180228157)

[CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN](#_Toc180228165)

[6.1 Kết Luận](#_Toc180228157)

[6.2 Hướng Phát Triển Trong Tương Lai](#_Toc180228157)

[TÀI LIỆU THAM KHẢO](#_Toc180228166)

# CHƯƠNG 1: GIỚI THIỆU CHUNG

## 1.1. Lý do chọn đề tài

Mạng xã hội ngày càng phát triển và trở thành một phần quan trọng trong đời sống cá nhân cũng như hoạt động kinh doanh, chính trị, và xã hội. Việc nghiên cứu và phát hiện cộng đồng trong mạng xã hội có nhiều ứng dụng thực tiễn, bao gồm:

* Hỗ trợ các doanh nghiệp xác định nhóm khách hàng tiềm năng, từ đó xây dựng chiến lược tiếp thị phù hợp.
* Phát hiện và ngăn chặn các hành vi tiêu cực như tin giả, thông tin sai lệch, hoặc các hoạt động lừa đảo trên mạng.
* Hỗ trợ nghiên cứu khoa học trong lĩnh vực xã hội học, phân tích hành vi người dùng và dự đoán xu hướng.
* Tăng cường an ninh mạng bằng cách phát hiện các nhóm tội phạm hoặc hành vi nguy hiểm trên không gian mạng.

Ngoài ra, việc phát hiện cộng đồng cũng đóng vai trò quan trọng trong lĩnh vực khoa học dữ liệu và trí tuệ nhân tạo, giúp tối ưu hóa các thuật toán phân tích dữ liệu lớn. Chính vì những lý do trên, nhóm chúng tôi đã chọn đề tài này để nghiên cứu và ứng dụng các phương pháp hiện đại nhằm hiểu rõ hơn về bài toán phát hiện cộng đồng trong mạng xã hội.

## 1.2. Mục tiêu nghiên cứu

### 1.2.1. Phân biệt hai loại cộng đồng trong mạng xã hội

* **Cộng đồng tách rời** là những nhóm đỉnh trong đồ thị mà không có đỉnh nào thuộc về nhiều hơn một cộng đồng. Mỗi đỉnh chỉ có thể nằm trong một tập hợp duy nhất. Điều này thường xảy ra trong các hệ thống mà các thành phần có mối quan hệ độc lập với nhau.
* **Ví dụ** như các phòng ban trong một tổ chức, nơi mỗi nhân viên chỉ thuộc về một phòng duy nhất
* **Cộng đồng giao nhau** là những nhóm đỉnh có thể thuộc về nhiều cộng đồng khác nhau. Điều này phản ánh thực tế rằng một thực thể có thể có nhiều mối quan hệ và tham gia vào nhiều nhóm cùng một lúc.

**Ví dụ** trong mạng xã hội, một người có thể tham gia nhiều nhóm sở thích khác nhau cùng lúc, hoặc một nhà khoa học có thể hoạt động trong nhiều lĩnh vực nghiên cứu khác nhau.

### 1.2.2. Một số thuật toán phổ biến

## Phân cụm phân cấp (Hierarchical Agglomerative Clustering): Thuật toán này xây dựng cây phân cấp bằng cách hợp nhất các đỉnh hoặc chia nhỏ đồ thị theo từng cấp độ. Có hai hướng tiếp cận chính:

## Phân cụm tích hợp (Agglomerative Clustering): Bắt đầu từ từng điểm đơn lẻ và kết hợp chúng dần thành các cụm lớn hơn dựa trên mức độ tương đồng.

## Phân cụm phân tách (Divisive Clustering): Bắt đầu từ toàn bộ đồ thị và dần chia nhỏ thành các cộng đồng con dựa trên các chỉ số như độ trung gian của cạnh.

## Phân cụm theo đồ thị

## Kernighan-Lin Algorithm: Hoán đổi các đỉnh giữa hai nhóm để giảm số cạnh giữa các nhóm, thường được sử dụng trong tối ưu hóa đồ thị hai phần.

## Spectral Bisection: Sử dụng giá trị riêng của ma trận đồ thị để chia đồ thị thành hai phần tối ưu theo cách giảm thiểu số cạnh bị cắt.

## Phân cụm phân hoạch

## K-means: Chia tập hợp đỉnh thành k nhóm dựa trên khoảng cách giữa các điểm dữ liệu, thường được sử dụng với đồ thị có trọng số.

## Fuzzy K-means: Cho phép một điểm có xác suất thuộc về nhiều nhóm khác nhau, phù hợp với mô hình cộng đồng giao nhau.

## Phân cụm theo phổ (Spectral Clustering): Sử dụng ma trận Laplacian của đồ thị để tìm cộng đồng dựa trên phân tích giá trị riêng, giúp xác định các cụm dựa trên cấu trúc tổng thể của đồ thị.

## Các thuật toán phân chia

## Girvan-Newman: Loại bỏ các cạnh có độ trung gian cao để tách cộng đồng, giúp xác định các khu vực tập trung nhiều liên kết.

## CONGA, CONGO: Biến thể của Girvan-Newman dành cho cộng đồng giao nhau, giúp xác định các đỉnh có thể thuộc về nhiều nhóm.

## COPRA: Gán nhãn cho các đỉnh dựa trên quy tắc lan truyền nhãn, cho phép phát hiện cộng đồng mà không cần thông tin trước.

# CHƯƠNG 2: BÀI TOÁN PHÁT HIỆN CỒNG ĐỒNG TRỀN MẠNG XÃ HỘI

## 2.1. Cộng đồng mạng xã hội

* Bài toán phát hiện cộng đồng trong mạng xã hội là bài toán phân tích cấu trúc của một đồ thị để xác định các nhóm đỉnh có mức độ liên kết nội bộ cao hơn so với bên ngoài. Việc phát hiện cộng đồng giúp chúng ta hiểu rõ cách các thực thể trong mạng tương tác với nhau.

Ảnh có chứa hàng, biểu đồ, vòng tròn, thiết kế

Nội dung do AI tạo ra có thể không chính xác.

Hình 2.1: Cộng đồng mạng xã hội đơn giản với 3 cộng đồng

- Trong hình 2.1, tôi đưa ra một ví dụ đơn giản về một cách phân chia mạng xã hội thành 3 cộng đồng con đơn giản.

- Trong xã hội hiện nay xuất hiện nhiều nhóm hoặc tổ chức với kích cỡ khác nhau, ví dụ như gia đình, nhóm các bạn bè hoặc đồng nghiệp, thành phố, quốc gia…

- Sự khuếch tán của Internet ngày nay cũng sinh ra nhiều nhóm ảo trên Web, hay còn được gọi là các cộng đồng trực tuyến. Các cộng đồng xã hội đã được nghiên cứu trong một thời gian rất dài và thường xuyên xuất hiện trong nhiều các hệ thống mạng trong sinh học, khoa học máy tính, công nghệ, chính trị, kinh tế,…

## 2.2. Giới thiệu bài toán phát hiện cộng đồng

## Bài toán: Phát hiện cộng đồng trong mạng xã hội và đưa ra danh sách những nút mạng thuộc từng cộng đồng đó.

**Đầu vào**: Đồ thị mạng xã hội G = (V, E) gồm tập V có đỉnh: v1, v2,…,vn và tập E các liên kết E = {(vi,vj)}.

**Đầu ra**: Tập các cộng đồng Ci và tập hợp các đỉnh thuộc các cộng đồng đó: {C1, C2,...,Cn}

Mục tiêu của bài toán là từ các mạng xã hội cho trước, phát hiện được các cấu trúc cộng đồng nằm trong đó và tìm hiểu về mối liên hệ bên trong các cộng đồng cũng như giữa các cộng đồng với nhau, mối liên hệ đó có ảnh hưởng thế nào đến cấu trúc của toàn mạng xã hội.

Ảnh có chứa biểu đồ, hàng, vòng tròn

Nội dung do AI tạo ra có thể không chính xác.Theo [13], một tập hợp các đỉnh trên đồ thị được coi là một cộng đồng nếu mật độ cạnh bên trong nó cao hơn so với mật độ của các cạnh giữa đỉnh của nó và những cạnh bên ngoài. Dưới đây là một vài ví dụ khá điển hình về mạng xã hội:

Hình 2.2: Mô hình mạng lưới các thành viên Câu lạc bộ Karate

Đây làm một đồ thị nổi tiếng thường xuyên được sử dụng như là một điểm để kiểm tra phát hiện cộng đồng. Nó bao gồm 34 đỉnh tương ứng là các thành viên của một câu lạc bộ Karate tại Hoa Kì. Người ta đã quát sát mạng xã hội này trong khoảng ba năm, các cạnh kết nối cá nhân được quan sát thông qua các tương tác bên ngoài và các hoạt động của Câu lạc bộ. Theo quan sát nhận thấy có sự tập trung mức độ liên kết cao tại một số điểm: Chủ tịch Câu lạc bộ và Huấn luyện viên, điều này đã phân chia mạng ra thành hai nhóm khá rõ ràng.

## 2.3. So sánh

## Nhiều nghiên cứu đã đề xuất các phương pháp phát hiện cộng đồng, trong đó phổ biến nhất là:

## Tổng quan về các phương pháp chính:

## Phương pháp dựa trên đồ thị: Sử dụng thông tin cấu trúc của mạng để xác định cộng đồng.

## Phương pháp dựa trên mô-đun hóa: Tối ưu hóa độ đo modularity để xác định cộng đồng.

## Phương pháp dựa trên phổ: Phân tích giá trị riêng của ma trận Laplacian để phát hiện cộng đồng.

## Phương pháp phân cụm truyền thống: Áp dụng các thuật toán như K-means, DBSCAN.

## Phương pháp dựa trên học máy: Sử dụng deep learning hoặc machine learning để phát hiện cộng đồng.

Ảnh có chứa văn bản, ảnh chụp màn hình, thực đơn, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Hình 2.3: Hình minh họa thuật toán Girvan-Newman

## 2.5. Thuật toán GIRVAN-NEWMAN

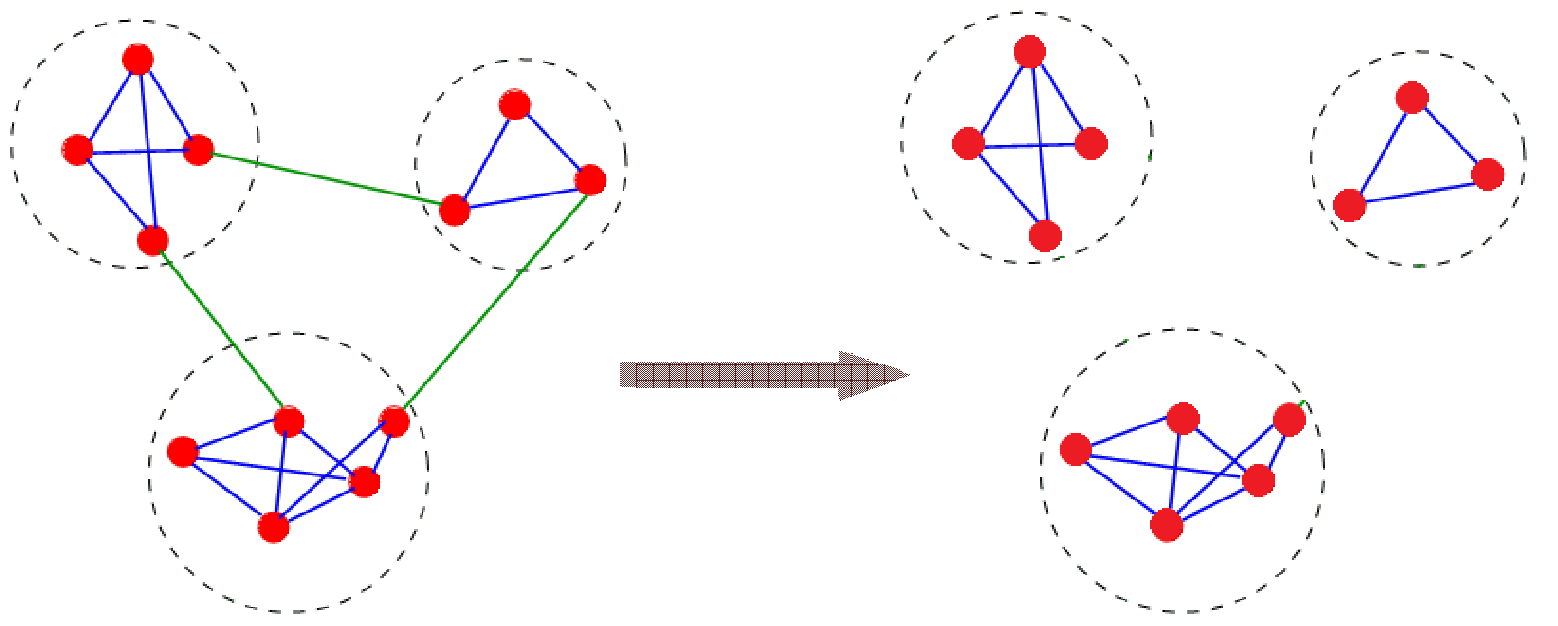
Nếu một mạng lưới bao gồm các cộng đồng hoặc nhóm chúng chỉ được liên kết nối yếu bằng một nhóm cạnh, thì tất cả các đường đi ngắn nhất giữa các cộng đồng khác nhau sẽ phải đi dọc theo một trong số ít các cạnh thuộc nhóm cạnh đó. Vì vậy, các cạnh kết nối các cộng đồng sẽ là cạnh có độ đô trung gian cao. Bằng cách loại bỏ các cạnh, thuật toán Girvan-Newman tách được thành các nhóm riêng biệt. Thuật toán được thực hiện theo các bước sau:

1. Tính độ đo trung gian cho tất cả các cạnh trong mạng.

2. Hủy bỏ các cạnh có độ trung gian cao nhất.

3. Tính lại độ trung gian cho tất cả các cạnh bị ảnh hưởng theo các cạnh đã loại bỏ.

4. Lặp lại từ bước 2 cho đến khi không còn các cạnh trung gian.



Hình 2.4: Ví dụ về phát hiện cộng đồng sử dụng Girven-Newman

* Thuật toán Girvan-Newman đưa lại kết quả tương đối tốt trong nhiều trường hợp, mặc dù vậy nó vẫn gặp phải một số nhược điểm:
* Thuật toán Girvan-Newman sử dụng phương pháp loại trừ đến khi không có cạnh nào vượt qua ngưỡng của độ trung gian cao nhất, vì vậy nên số lượng cộng đồng hoàn toàn không kiểm soát trước được. Bên cạnh đó, thuật toán sử dụng nhiều phép phân vùng, khó có thể xác định được phép phân vùng nào mang lại hiệu quả tốt nhất.
* Do tại mỗi lượt thực hiện, thuật toán tính lại độ trung gian của mỗi cạnh liên quan sau khi xóa đi cạnh có độ trung gian lớn nhất nên độ phức tạp thời gian là khá cao. Giả sử với đồ thị n đỉnh, số cạnh phải xóa đi khỏi đồ thị là m cạnh thì ta cần lượng thời gian tính toán O(mn) cho mỗi lần lặp. Tổng thời gian chạy thuật toán O(m2n). Trong trường hợp xấu nhất, mỗi đỉnh được chia ra thành một cộng đồng riêng rẽ thì độ phức tạp thời gian của thuật toán sẽ lên đến O(n3).
* Trên thực tế, mỗi đơn vị nút mạng có thể thuộc vào rất nhiều cộng đồng khác nhau. Ví dụ với một cá nhân A, đóng góp vai trò là một nút trên mạng xã hội có thể thuộc vào nhiều nhóm: Bạn cùng lớp, đồng nghiệp cùng công ty, anh em họ hàng trong gia đình… Nhưng với cách phân chia của Girvan-Newman không giải quyết được hiện tượng chồng chéo cộng đồng trên.

### 2.6. Thuật toán CONGA

### Trong các thuật toán Girven-Newman, hoạt động cơ bản là loại bỏ từng cạnh. CONGA giới thiệu một phương pháp thứ hai: chia tách một đỉnh. Trong mỗi bước phân chia, một đỉnh v luôn luôn được chia thành hai đỉnh v1 và v2, đồng thời với đó, cạnh nối liền v với một đỉnh khác sẽ được chuyển hướng sang v1 hoặc v2.

### Phương pháp này không bao giờ phân một đỉnh thành hai đỉnh con mà trong đó một trong hai đỉnh con chỉ có duy nhất một đỉnh kề. Điều này là do độ trung gian của cạnh {v1, v2} cũng bằng độ trung gian của {u,v1} như trong hình 2.2. Trong trường hợp này thì việc loại bỏ cạnh {u,v} sẽ thích hợp hơn là việc phân tách đỉnh v. Chính vì vậy, các đỉnh có bậc nhỏ hơn 4 trong đồ thị không bao giờ bị chia nhỏ và tổng quát lại chỉ có 2d(v)-1–d(v)-1 cách phân chia một đỉnh thành hai phần trong một đồ thị.

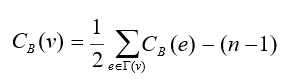
Ảnh có chứa biểu đồ, hàng, Phông chữ, vòng tròn

Nội dung do AI tạo ra có thể không chính xác.

Hình 2.5: Ví dụ trường hợp không phân tách đỉnh v trong đồ thị

Trong phương pháp này, nhóm tác giả định nghĩa của độ trung gian của một đỉnh v trong đồ thị là tổng số đường đi ngắn nhất giữa các cặp đỉnh của đồ thị mà đi qua v.

Ta có thể dễ tính được độ trung gian của đỉnh CB (v) từ các độ đo trung gian của cạnh CB (e) :



**Đánh giá thuật toán:**

**Ưu điểm:**

* Giải quyết được vấn đề chồng chéo cộng đồng bằng cách đặt ra phép phân chia đỉnh, ngoài ra nội dung thuật toán tương đối dễ hiểu và xác định được phép phân chia tối ưu nhất trong các trường hợp, một điều mà thuật toán Girvan-Newman nguyên thủy không làm được.

**Nhược điểm:**

* Thời gian tính toán, với độ phức tạp tính toán O(m3) với m là số cạnh.

### 2.7. Công cụ và ngôn ngữ

### Các công cụ và ngôn ngữ phổ biến để phát hiện cộng đồng trong mạng xã hội:

### Ngôn ngữ lập trình

* + **Python**: Phổ biến nhất với các thư viện hỗ trợ mạnh mẽ.
  + **R**: Thích hợp cho phân tích dữ liệu và trực quan hóa.
  + **C++**: Hiệu suất cao, phù hợp với mạng lớn.

**Thư viện hỗ trợ**

* + **NetworkX**: Hỗ trợ phân tích và trực quan hóa đồ thị trong Python.
  + **igraph**: Xử lý đồ thị hiệu suất cao, có hỗ trợ trong Python và R.
  + **Scikit-learn**: Chứa các thuật toán phân cụm như Spectral Clustering.
  + **Gephi**: Công cụ mạnh để trực quan hóa và phân tích cộng đồng trong mạng lớn

**2.8. CÁC THUẬT TOÁN SỬ DỤNG TRONG PHÁT HIỆN CỘNG ĐỒNG**

**2.8.1. Thuật toán Girvan-Newman**

**Khái niệm**

Thuật toán Girvan-Newman là một phương pháp phân cụm cộng đồng dựa trên việc loại bỏ các cạnh có độ trung gian cao nhất trong đồ thị. Khi các cạnh này bị loại bỏ, đồ thị dần bị chia nhỏ thành các cộng đồng riêng biệt.

**Nguyên lý hoạt động**

1. Tính độ trung gian của tất cả các cạnh trong đồ thị.
2. Loại bỏ cạnh có độ trung gian cao nhất.
3. Cập nhật lại độ trung gian của các cạnh còn lại.
4. Lặp lại quá trình cho đến khi tất cả các cạnh bị loại bỏ hoặc đạt số lượng cộng đồng mong muốn.

**Ưu điểm**

* Phù hợp với mạng nhỏ.
* Cung cấp thông tin chi tiết về cách mạng phân tách.

**Nhược điểm**

* Hiệu suất chậm với mạng lớn.
* Không phù hợp với mạng có sự chồng lấn giữa các cộng đồng.

**Ứng dụng**

* Phân tích mạng xã hội.
* Phát hiện nhóm tội phạm mạng.

**2.8.2. Thuật toán CONGA (Cluster Overlapping Newman-Girvan Algorithm)**

**Khái niệm**

CONGA là một phiên bản mở rộng của Girvan-Newman nhưng cho phép các cộng đồng có sự chồng lấn.

**Nguyên lý hoạt động**

* Thay vì chỉ loại bỏ cạnh có độ trung gian cao, thuật toán còn xem xét các đỉnh có độ trung gian cao.
* Khi loại bỏ cạnh hoặc chia tách đỉnh, thuật toán xác định cách tốt nhất để phân cụm mạng.

**Ưu điểm**

* Xử lý được các cộng đồng chồng lấn.

**Nhược điểm**

* Cần nhiều tài nguyên tính toán hơn so với Girvan-Newman.

**Ứng dụng**

* Mô hình hóa mạng xã hội thực tế với các nhóm có chồng lấn.

**2.8.3. Thuật toán Label Propagation (LPA)**

**Khái niệm**

LPA hoạt động bằng cách gán nhãn cho các đỉnh trong mạng và cho phép nhãn lan truyền theo liên kết.

**Nguyên lý hoạt động**

1. Gán mỗi đỉnh một nhãn duy nhất.
2. Trong mỗi vòng lặp, cập nhật nhãn của mỗi đỉnh bằng nhãn phổ biến nhất từ các láng giềng.
3. Lặp lại cho đến khi nhãn không thay đổi nữa.

**Ưu điểm**

* Thời gian chạy nhanh.
* Phù hợp với mạng lớn.

**Nhược điểm**

* Kết quả có thể không ổn định do phụ thuộc vào thứ tự cập nhật.

**Ứng dụng**

* Phân tích cộng đồng trong mạng xã hội quy mô lớn.

**2.8.4. Thuật toán Spectral Clustering**

**Khái niệm**

Dựa trên phân tích giá trị riêng của ma trận Laplacian để xác định các cộng đồng trong mạng.

**Nguyên lý hoạt động**

1. Xây dựng ma trận kề của đồ thị.
2. Tính toán ma trận Laplacian và tìm các giá trị riêng.
3. Áp dụng thuật toán phân cụm K-means lên các giá trị riêng thu được.

**Ưu điểm**

* Độ chính xác cao.

**Nhược điểm**

* Yêu cầu tính toán lớn.

**Ứng dụng**

* Phân tích mạng sinh học.
* Nhận diện hành vi trong mạng xã hội.

**2.8.5. Thuật toán K-means Clustering**

**Khái niệm**

Phương pháp phân cụm dựa trên khoảng cách giữa các điểm dữ liệu.

**Nguyên lý hoạt động**

1. Chọn K điểm làm tâm cụm ban đầu.
2. Gán mỗi điểm vào cụm gần nhất.
3. Cập nhật lại tâm cụm.
4. Lặp lại cho đến khi không có sự thay đổi đáng kể.

**Ưu điểm**

* Nhanh chóng và dễ triển khai.

**Nhược điểm**

* Không phù hợp với dữ liệu phi Euclid.

**Ứng dụng**

* Phân tích khách hàng trong thương mại điện tử.

**2.8.6. Thuật toán Modularity Optimization**

**Khái niệm**

Dựa trên tối ưu hóa độ đo modularity để xác định các cộng đồng.

**Nguyên lý hoạt động**

* Tính toán modularity của mạng.
* Tối ưu hóa bằng cách gộp hoặc tách các cộng đồng sao cho modularity cao nhất.

**Ưu điểm**

* Kết quả đáng tin cậy.

**Nhược điểm**

* Chạy chậm trên mạng lớn.

**Ứng dụng**

* Phân tích mạng giao thông.

**2.8.7. Thuật toán Louvain Algorithm**

**Khái niệm**

Là một biến thể tối ưu hóa của modularity nhằm phát hiện cộng đồng hiệu quả hơn.

**Nguyên lý hoạt động**

* Sử dụng phương pháp lặp để tối ưu hóa modularity.

**Ưu điểm**

* Tốc độ nhanh hơn so với modularity truyền thống.

**Nhược điểm**

* Không xử lý tốt cộng đồng chồng lấn.

**Ứng dụng**

* Phân tích mạng xã hội và mạng sinh học.

**8. Thuật toán Infomap Algorithm**

**Khái niệm**

Sử dụng lý thuyết thông tin để tối ưu hóa mô tả luồng thông tin trong mạng.

**Nguyên lý hoạt động**

* Xây dựng mô hình đồ thị dựa trên luồng thông tin.
* Áp dụng kỹ thuật nén để xác định cộng đồng.

**Ưu điểm**

* Độ chính xác cao.

**Nhược điểm**

* Tốn tài nguyên tính toán.

**Ứng dụng**

* Phân tích mạng giao thông.

**2.8.9. Thuật toán Deep Learning-Based**

**Khái niệm**

Sử dụng mạng neural để học các đặc trưng cộng đồng trong dữ liệu.

**Nguyên lý hoạt động**

* Xây dựng mô hình học sâu để nhận diện các đặc trưng quan trọng.

**Ưu điểm**

* Xử lý tốt dữ liệu lớn.

**Nhược điểm**

* Yêu cầu nhiều tài nguyên tính toán.

**Ứng dụng**

* Phát hiện cộng đồng trong dữ liệu lớn.

**2.10.10. Thuật toán Hierarchical Agglomerative Clustering**

**Khái niệm**

Dựa trên việc hợp nhất dần các cụm nhỏ thành cụm lớn hơn.

**Nguyên lý hoạt động**

* Gộp các điểm dữ liệu gần nhau lại theo phương pháp bottom-up.

**Ưu điểm**

* Không cần xác định số cụm trước.

**Nhược điểm**

* Không tối ưu với mạng lớn.

**Ứng dụng**

* Nhóm khách hàng trong marketing.

**CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ**

***3.1. Giới thiệu:***

Trong chương này, chúng em sẽ trình bày về quá trình thực nghiệm và đánh giá việc phân tích hành vi người dùng trên mạng xã hội sử dụng công cụ SNAP (Stanford Network Analysis Platform) C++. Mục tiêu chính của chương này là minh họa cách SNAP C++ có thể được áp dụng để hiểu rõ hơn về hành vi, sở thích và các nhóm người dùng trong mạng xã hội.

*Việc phân tích hành vi người dùng trên mạng xã hội mang lại nhiều lợi ích, bao gồm:*

* Hiểu rõ hành vi và phân loại các hành vi của người dùng.
* Phát hiện sự quan tâm của người dùng dựa trên hành vi của họ.
* Xác định các nhóm hoặc cộng đồng người dùng trên mạng xã hội.

***3.2. Mục tiêu thực nghiệm:***

***Thực nghiệm này nhằm đạt được các mục tiêu sau:***

* Xây dựng mô hình biểu diễn người dùng dựa trên hành vi tương tác của họ trên mạng xã hội.
* Sử dụng SNAP C++ để phân tích cấu trúc mạng xã hội và xác định các cộng đồng người dùng.
* Đánh giá hiệu quả của việc sử dụng SNAP C++ trong việc phân tích hành vi người dùng.

***3.4. Phương pháp thực nghiệm:***

***Quá trình thực nghiệm sẽ bao gồm các bước sau:***

1. Biểu diễn dữ liệu mạng xã hội: Dữ liệu mạng xã hội sẽ được biểu diễn dưới dạng đồ thị, trong đó các nút đại diện cho người dùng và các cạnh đại diện cho mối quan hệ hoặc tương tác giữa họ. SNAP C++ hỗ trợ nhiều định dạng đồ thị khác nhau, giúp dễ dàng nhập dữ liệu.
2. Phân tích cấu trúc mạng xã hội: Em sẽ sử dụng các thuật toán có sẵn trong SNAP C++ để phân tích các đặc trưng cấu trúc của mạng, chẳng hạn như:
   * Bậc của nút (Degree): Số lượng kết nối mà một người dùng có.
   * Đường đi ngắn nhất (Shortest Paths): Đo lường mức độ gần gũi giữa các người dùng.
   * Hệ số cụm (Clustering Coefficient): Đo lường xu hướng hình thành các nhóm kín trong mạng.
   * Thuật toán phát hiện cộng đồng: Sử dụng các thuật toán như Louvain Algorithm hoặc Connected Components để xác định các nhóm người dùng có liên kết chặt chẽ với nhau.
3. Mô hình hóa hành vi người dùng: Dựa trên các tương tác (ví dụ: thích, bình luận, chia sẻ), chúng tôi có thể gán trọng số cho các cạnh hoặc tạo các thuộc tính cho nút để biểu diễn hành vi của người dùng.
4. Phân tích hành vi và sở thích: Em có thể phân tích các thuộc tính và cấu trúc cộng đồng để suy ra hành vi và sở thích của người dùng trong từng nhóm.

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phần mềm đa phương tiện

Nội dung do AI tạo ra có thể không chính xác.

* Đoạn code trên tạo một đồ thị vô hướng với 5 nút và 5 cạnh.
* Sau đó, nó in ra tổng số nút và cạnh trong đồ thị.
* Cuối cùng, nó duyệt qua từng nút và in ra bậc (số lượng kết nối) của nút đó.

***3.7. Đánh giá kết quả:***

* **So sánh với dữ liệu thực tế (ground truth):** Nếu có sẵn thông tin về các nhóm người dùng hoặc sở thích đã biết, chúng ta có thể so sánh kết quả phân tích của SNAP C++ với thông tin này để đánh giá độ chính xác.
* **Sử dụng các độ đo hiệu suất:** Đối với các tác vụ cụ thể như phát hiện cộng đồng, có các độ đo như Modularity, Normalized Mutual Information (NMI) có thể được sử dụng để đánh giá chất lượng của các cộng đồng được phát hiện.
* **Phân tích định tính:** Quan sát và phân tích các mẫu hành vi và cộng đồng được phát hiện để hiểu rõ hơn về động lực và cấu trúc của mạng xã hội.

**CHƯƠNG 4: KẾT LUẬN VÀ PHƯƠNG HƯỚNG**

***4.1 Kết luận:***

Chương này đã trình bày về quá trình thực nghiệm và đánh giá việc phân tích hành vi người dùng trên mạng xã hội bằng công cụ SNAP C++. Ví dụ demo code) cho thấy tiềm năng của SNAP C++ trong việc khám phá cấu trúc và hành vi trong mạng xã hội. Việc phân tích này có thể cung cấp những hiểu biết sâu sắc về người dùng, giúp ích cho nhiều ứng dụng khác nhau như marketing, đề xuất nội dung và nghiên cứu xã hội.

***4.2. Phát triển các mô hình hành vi người dùng phức tạp hơn:***

* Kết hợp thông tin ngữ nghĩa: Mô hình hóa hành vi người dùng hiện tại có thể chỉ dựa trên tương tác (like, comment, share). Trong tương lai, có thể tích hợp phân tích nội dung văn bản trong các bài đăng và bình luận để hiểu sâu hơn về sở thích và quan điểm của người dùng. Điều này có thể sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP).
* Phân tích chuỗi hành vi: Nghiên cứu trình tự các hành động của người dùng (ví dụ: xem một chuỗi video, truy cập các trang web cụ thể sau khi thấy một bài đăng) để xây dựng mô hình hành vi chi tiết hơn và dự đoán các hành động tiếp theo.
* Mô hình hóa cảm xúc: Phân tích cảm xúc trong các tương tác (ví dụ: sử dụng sentiment analysis) để hiểu rõ hơn về phản ứng của người dùng đối với các chủ đề hoặc sự kiện khác nhau.

***4.3. Nâng cao hiệu quả của các thuật toán phát hiện cộng đồng:***

* **Phát triển các thuật toán thích ứng:** Nghiên cứu và phát triển các thuật toán có khả năng tự động điều chỉnh các tham số dựa trên đặc điểm của từng mạng xã hội cụ thể, giúp tăng độ chính xác và hiệu quả.
* **Xử lý cộng đồng chồng lấn tốt hơn:** Mặc dù đã có một số thuật toán như CONGA, việc nghiên cứu các phương pháp mới để phát hiện và phân tích các cộng đồng có sự chồng lấn phức tạp vẫn là một hướng đi quan trọng, phản ánh đúng hơn tính chất của mạng xã hội thực tế.
* **Phát hiện cộng đồng động:** Các cộng đồng trong mạng xã hội thường xuyên thay đổi theo thời gian. Phát triển các thuật toán có khả năng theo dõi và phát hiện sự tiến hóa của cộng đồng là một hướng nghiên cứu đầy tiềm năng.

***4.4. Ứng dụng các kỹ thuật học máy và học sâu:***

* **Sử dụng học máy để phân loại hành vi:** Áp dụng các mô hình học máy (ví dụ: cây quyết định, máy vector hỗ trợ) để phân loại người dùng dựa trên các đặc trưng hành vi đã được trích xuất.
* **Ứng dụng học sâu cho biểu diễn đồ thị (Graph Embedding):** Sử dụng các kỹ thuật học sâu như Graph Neural Networks (GNNs) để học các biểu diễn vector cho các nút trong mạng, từ đó có thể sử dụng cho các tác vụ như phát hiện cộng đồng, dự đoán liên kết hoặc phân loại nút.
* **Kết hợp học sâu với SNAP C++:** Tận dụng hiệu suất của SNAP C++ trong việc xử lý đồ thị lớn và kết hợp nó với các thư viện học sâu (ví dụ: TensorFlow, PyTorch) để xây dựng các hệ thống phân tích mạnh mẽ hơn.

***4.5. Mở rộng phạm vi ứng dụng:***

* **Phân tích xu hướng và dự đoán:** Sử dụng các kỹ thuật phân tích hành vi để dự đoán các xu hướng mới nổi trong mạng xã hội hoặc dự đoán hành vi của người dùng trong tương lai (ví dụ: khả năng lan truyền thông tin, ý định mua hàng).
* **Ứng dụng trong các lĩnh vực cụ thể:** Nghiên cứu cách áp dụng các phương pháp phân tích mạng xã hội và hành vi người dùng trong các lĩnh vực cụ thể như marketing (nhắm mục tiêu quảng cáo, tìm kiếm người ảnh hưởng), y tế (theo dõi dịch bệnh), an ninh mạng (phát hiện hoạt động đáng ngờ) và chính trị (phân tích dư luận).
* **Xây dựng hệ thống đề xuất:** Phát triển các hệ thống đề xuất dựa trên hành vi và cộng đồng người dùng để gợi ý nội dung, bạn bè hoặc sản phẩm phù hợp.

***4.6. Nghiên cứu về khía cạnh đạo đức và xã hội:***

* **Bảo vệ quyền riêng tư:** Nghiên cứu các phương pháp phân tích mạng xã hội mà vẫn đảm bảo quyền riêng tư của người dùng.
* **Phát hiện và ngăn chặn thông tin sai lệch:** Áp dụng các kỹ thuật phân tích hành vi để phát hiện và ngăn chặn sự lan truyền của tin giả và thông tin sai lệch trên mạng xã hội.
* **Nghiên cứu về tác động xã hội:** Phân tích tác động của cấu trúc mạng xã hội và hành vi người dùng đối với các hiện tượng xã hội như sự phân cực, hình thành ý kiến đám đông và các phong trào xã hội.

Những phương hướng phát triển này không chỉ mở ra các cơ hội nghiên cứu học thuật mà còn có tiềm năng ứng dụng thực tế to lớn, góp phần hiểu rõ hơn về xã hội và hành vi con người trong kỷ nguyên số.

# TÀI LIỆU THAM KHẢO

1. Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577–8582.
2. Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.
3. Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4), 1118–1123.
4. Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6), 066111.
5. Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512.