ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

LUẬN VĂN THẠC SĨ

ĐỀ TÀI

Mô hình đồ thị và ứng dụng đối với bài toán cộng đồng trên mạng xã hội

Giáo viên hướng dẫn : **TS. Vũ Vinh Quang**
Học viên : Hoàng Văn Dũng
Lớp : Cao học K17

Thái Nguyên, tháng 9 năm 2020

LỜI CẢM ƠN

Đầu tiên, em xin gửi lời cảm ơn chân thành và sâu sắc nhất tới thầy Vũ Vinh Quang, người đã trực tiếp hướng dẫn tận tình và đóng góp những ý kiến quý báu trong suốt quá trình em làm luận văn tốt nghiệp này.

Tiếp theo em xin gửi lời cảm ơn đến đến các thầy cô giáo trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên, đã tận tâm truyền đạt những kiến thức quý báu làm nền tảng để em hoàn thành luận văn này.

Học Viên

Hoàng Văn Dũng

LỜI CAM ĐOAN

Tôi xin cam đoan mô hình đồ thị và ứng dụng đối với bài toán cộng đồng trên mạng xã hội được trình bày trong luận văn là do tôi thực hiện dưới sự hướng dẫn của thầy Vũ Vinh Quang

Tất cả những tham khảo từ các nghiên cứu liên quan đều được nêu nguồn gốc một cách rõ ràng từ danh mục tài liệu tham khảo trong luận văn. Trong luận văn không có việc sao chép tài liệu, công trình nghiên cứu của người khác mà không chỉ rõ về tài liệu tham khảo.

Thái Nguyên, ngày tháng năm 2020

Học viên

Hoàng Văn Dũng

MỤC LỤC

DANH SÁCH CÁC HÌNH VẼ

MỞ ĐẦU

Mô hình đồ thị là một mô hình cấu trúc dữ liệu kinh điển trong tin học, đối với mô hình này, chúng ta thường quan tâm đến việc mô tả cấu trúc trong máy tính điện tử và các thuật toán tìm kiếm, tối ưu trên mô hình này. Có thể nói việc nghiên cứu ứng dụng của mô hình đồ thị trong thực tế là một lĩnh vực vô cùng rộng lớn và có ý nghĩa thực tế rất cao. Đa số các nguồn dữ liệu hiện nay đều có thể biểu diễn được dưới dạng cấu trúc dữ liệu đồ thị, thí dụ như dữ liệu mạng internet, mạng xã hội, cấu trúc protein, các hợp chất hóa học, quy trình của một chương trình…Cấu trúc đồ thị được đánh giá là cấu trúc có tính bao quát mô tả các đối tượng so với các cấu trúc tuần tự, cây, mạng ngữ nghĩa…và dựa trên cấu trúc này có thể thiết kế được nhiều thuật toán giải quyết được nhiều bài toán có độ tính toán phức tạp cao. Mỗi dữ liệu lớn trên các hệ thống mạng đều có thể biểu diễn dưới dạng các đồ thị và mối quan hệ của chúng theo các liên kết và kết nối vật lý; kết nối giữa các mạng trong lớp mạng; mối quan hệ trong mạng xã hội; siêu liên kết giữa các trang web và các tương tác phức tạp giữa các thực thể. Các đồ thị chứa đựng những thông tin giá trị cho việc ứng dụng vào hệ thống mạng như những phát hiện từ cộng đồng, những điểm chung; phân lớp; tìm kiếm trên mạng; những hệ thống được đưa ra theo thứ tự nào đó; độ tin cậy và uy tín; tìm kiếm và lấy dữ liệu từ điểm đến điểm. Để đưa các dữ liệu đang xét vào dưới dạng đồ thị, chúng ta cần phải định nghĩa các ma trận mô tả cấu trúc tổng thể của đồ thị; định nghĩa các ma trận mô tả các mẫu đặc trưng của các giao tiếp bên trong đồ thị phải tìm ra các cấu trúc có tính đặc trưng cộng đồng của mạng; hiểu rõ mô hình của việc lấy ra từ các đồ thị; phát triển và ứng dụng những thuật toán hiểu quả nhất để khai thác dữ liệu trong hệ thống mạng. Nhiều dữ liệu có cấu trúc đều có thể được biểu diễn bằng mạng, bằng tập hợp các đỉnh được nối với nhau theo cặp bằng các liên kết, như mạng sinh học, sự hợp tác của các nhà nghiên cứu,… Đặc điểm quan trọng gắn kết các mạng lại với nhau đó là cấu trúc cộng đồng, trong đó có các nhóm có mật độ kết nối mạnh của các đỉnh trong nhóm và các kết nối yếu giữa các nhóm. Do đó, việc xác định cộng đồng có thể được xem như là việc tìm kiếm các cụm đỉnh phân nhóm. Một cộng đồng là một nhóm của các thực thể dùng chia sẻ những tài sản tương tự nhau hoặc kết nối với nhau thông qua mối quan hệ được lựa chọn. Việc xác định các kết nối và định vị các thực thể trong cộng đồng khác nhau được coi là mục tiêu chính của nghiên cứu phát hiện cộng đồng. Bằng trực quan có thể dễ dàng tìm ra những nhóm cộng đồng có

độ tập trung cao, nhưng không phải cộng đồng nào cũng được hình thành bằng các mối liên hệ chặt chẽ và dễ thấy vì một số cộng đồng được hình thành dưới dạng ẩn. Điều quan trọng là phải tìm được sự phân phối của các cạnh giữa các đỉnh, từ đó phát hiện và đưa ra các cộng đồng tồn tại bên trong mạng xã hội.

Như vậy, khai phá dữ liệu đồ thị và phát hiện cấu trúc cộng đồng trên mạng xã hội là một lĩnh vực nghiên cứu khá hấp dẫn và được các nhà khoa học quan tâm nghiên cứu. Xuất phát từ một đồ thị mạng xã hội, người ta tìm ra các cụm, các nhóm cộng đồng có mối liên hệ chặt chẽ với nhau, từ đó phát hiện và tìm ra đặc tính của cấu trúc mạng.

Mục tiêu chính của luận văn đặt ra là nghiên cứu các đặc trưng cơ bản về mô hình đồ thị, một số các thuật toán tìm kiếm tối ưu trên mô hình đồ thị. Khái niệm về bài toán cộng đồng và một số thuật toán xác định cộng đồng trên mạng xã hội. Dự kiến nội dung báo cáo của luận văn gồm: Phần mở đầu, 3 chương chính, phần kết luận, tài liệu tham khảo, phụ lục. Bố cục được trình bày như sau:

Phần mở đầu: Nêu lý do chọn đề tài và hướng nghiên cứu chính của luận văn

Chương 1: Trình bày tổng quan về mô hình đồ thị bao gồm: Các khái niệm cơ bản, các phương pháp mô tả đồ thị, các thuật toán duyệt đồ thị và trọng tâm là các thuật toán xác định đường đi ngắn nhất trên đồ thị [1, 2, 3, 4, 6, 7].

Chương 2. Trình bày các khái niệm cơ bản về mô hình mạng xã hội và bài toán cộng đồng bao gồm: Tổng quan về mạng xã hội, khái niệm về độ đo đỉnh và độ đo cạnh, các thuật toán tương ứng. Khái niệm về cộng đồng và một số thuật toán xác định cộng đồng, Họ thuật toán Girvan_Newman, thuật toán CONGA[5, 8, 9].

Chương 3 Đưa ra một số kết quả cài đặt các thuật toán: thuật toán xác định độ đo đỉnh, thuật toán xác định độ đo cạnh, thuật toán xác định cộng đồng.

Các kết quả cài đặt các thuật toán được thực hiện trên môi trường Matlab version 7.0

Chương 1

MỘT SỐ KIẾN THỨC CƠ BẢN VỀ MÔ HÌNH ĐỒ THỊ

Nội dung chương 1 đưa ra các khái niệm cơ bản về lý thuyết đồ thị, các thuật toán cơ bản trên mô hình đồ thị. Các kiến thức được tham khảo trong các tài liệu [1] [2]

1.1 Một số khái niệm cơ bản [1]

1.1.1 Định nghĩa về đồ thị

Đồ thị là một cấu trúc rời rạc bao gồm các đỉnh và các cạnh nối các đỉnh này, các loại đồ thị khác nhau được phân biệt bởi kiểu và số lượng cạnh nối hai đỉnh nào đó của đồ thị. Giả sử V là tập hữu hạn, không rỗng các phần tử nào đó. Bộ $G = (V,E)$ được gọi là đồ thị hữu hạn. Mỗi phần tử của V gọi là một đỉnh và mỗi phần tử $u = (x,y)$ của E được gọi là một cạnh của đồ thị $G = (V,E)$.

Xét một cạnh u của E khi đó tồn tại hai đỉnh x, y của V sao cho $u = (x,y)$, ta nói rằng x nối với y hoặc x và y phụ thuộc u.

- Nếu cạnh $u = (x,y)$ mà x và y là hai đỉnh phân biệt thì ta nói x, y là hai đỉnh kề nhau.
- Nếu $u = (x,x)$ thì u là cạnh có hai đỉnh trùng nhau ta gọi đó là một *khuyên*.
- Nếu $u = (x,y)$ mà x, y là cặp đỉnh có phân biệt thứ tự hay có hướng từ x đến y thì u là một cung, khi đó x là gốc còn y là ngọn hoặc x là đỉnh vào, y là đỉnh ra.
- Khi giữa cặp đỉnh (x,y) có nhiều hơn một cạnh thì ta nói rằng những cạnh cùng cặp đỉnh là những cạnh song song hay là cạnh bội.

Trong thực tế ta có thể gặp nhiều vấn đề mà có thể dùng mô hình đồ thị để biểu diễn, như sơ đồ mạng máy tính, sơ đồ mạng lưới giao thông, sơ đồ thi công một công trình.

Định nghĩa 1. *Đơn đồ thị vô hướng $G = (V,E)$ bao gồm V là các tập đỉnh và E là các tập các cặp không có thứ tự gồm hai phần tử khác nhau của V gọi là các cạnh.*

Định nghĩa 2. *Đa đồ thị vô hướng $G = (V,E)$ bao gồm V là tập các đỉnh, và E là họ các cặp không có thứ tự gồm hai phần tử khác nhau của V gọi là các cạnh. Hai cạnh e_1 và e_2 được gọi là cạnh lặp nếu chúng cùng tương ứng với một cặp đỉnh.*

Rõ ràng mỗi đơn đồ thị là đa đồ thị, nhưng không phải đa đồ thị nào cũng là đơn đồ thị, vì trong đa đồ thị có thể có hai (hoặc có nhiều hơn) cạnh nối một cặp đỉnh nào đó.

Định nghĩa 3. *Giả đồ thị vô hướng G = (V,E) bao gồm V là các tập đỉnh, và E là họ các cặp không có thứ tự (không nhất thiết phải khác nhau) của V gọi là các cạnh. Cạnh e được gọi là khuyên nếu nó có dạng e = (u,u).*

Định nghĩa 4. *Đơn đồ thị có hướng G = (V,E) bao gồm V là các tập đỉnh và E là các cặp có thứ tự gồm hai phần tử khác nhau của V gọi là các cung.*

Nếu trong mạng có thể có đa kênh thoại một chiều, ta sẽ phải sử dụng đến khái niệm *đa đồ thị có hướng:*

Định nghĩa 5. *Đa đồ thị có hướng G = (V,E) bao gồm V là các tập đỉnh và E là họ các cặp có thứ tự gồm hai phần tử khác nhau của V gọi là các cung. Hai cung e_1, e_2 tương ứng cùng với một cặp đỉnh được gọi là cung lặp.*

Trong các phần tử tiếp theo chủ yếu chúng ta sẽ làm việc với đơn đồ thị vô hướng và đơn đồ thị có hướng. Vì vậy, để ngắn gọn, ta bỏ qua tính từ đơn khi nhắc đến chúng.

1.1.2 Các thuật ngữ cơ bản

Định nghĩa 6. *Hai đỉnh u và v của đồ thị vô hướng G được gọi là kề nhau nếu (u,v) là cạnh của đồ thị G. Nếu e = (u,v) là cạnh của đồ thị thì ta nói cạnh này là liên thuộc với hai đỉnh u và v, hoặc cũng nói là cạnh e là nối đỉnh u và đỉnh v, đồng thời các đỉnh u và v sẽ được gọi là các đỉnh đầu của cạnh (u,v).*

Định nghĩa 7. *Ta gọi bậc của đỉnh v trong đồ thị vô hướng là số cạnh liên thuộc với nó và sẽ ký hiệu là deg(v).*

Định nghĩa 8. *Nếu e = (u,v) là cung của đồ thị có hướng G thì ta nói hai đỉnh u và v là kề nhau, và nói cung (u,v) nối đỉnh u với đỉnh v hoặc cũng nói cung này là đi ra khỏi đỉnh u và đi vào đỉnh v. Đỉnh u(v) sẽ được gọi là đỉnh đầu(cuối) của cung (u,v).*

Định nghĩa 9. *Ta gọi bán bậc ra (bán bậc vào) của đỉnh v trong đồ thị có hướng là số cung của đồ thị đi ra khỏi nó (đi vào nó) và ký hiệu là $deg^+(v)(deg^-(v))$.*

1.1.3 Đường đi, chu trình. Đồ thị liên thông.

Định nghĩa 10. *Đường đi độ dài n từ đỉnh u đến đỉnh v, trong đó n là số nguyên dương, trên đồ thị vô hướng G = (V,E) là dãy $x_0, x_1, \ldots, x_{n-1}, x_n$ trong đó $u = x_0$, $v = x_n$, $(x_i, x_{i+1}) \in E$, $i = 0,1,2,\ldots, n-1$.* Đường đi nói trên còn có thể biểu diễn dưới dạng dãy các cạnh: $(x_0,x_1), (x_1,x_2), \ldots, (x_{n-1},x_n)$.

Đỉnh u gọi là đỉnh đầu, còn đỉnh v gọi là đỉnh cuối của đường đi. Đường đi có đỉnh đầu trùng với đỉnh cuối (tức là u = v) được gọi là chu trình. Đường đi hay chu trình