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

PHẠM THỊ LIÊN

BÀI TOÁN ĐỒ THỊ CON ĐẲNG CẤU TRONG KHAI PHÁ DỮ LIỆU ĐỒ THỊ VÀ ỨNG DỤNG PHÁT HIỆN ĐỒ THỊ CON PHỔ BIẾN

LUẬN VĂN THẠC SĨ: KHOA HỌC MÁY TÍNH

Thái Nguyên - 2020

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này do tự bản thân tôi tìm hiểu, nghiên cứu dưới sự hướng dẫn của PGS. TS. Đoàn Văn Ban. Các chương trình thực nghiệm do chính bản thân tôi lập trình, các kết quả là hoàn toàn trung thực. Các tài liệu tham khảo được trích dẫn và chú thích đầy đủ.

TÁC GIẢ LUẬN VĂN

Phạm Thị Liên

LỜI CẢM ƠN

Tôi xin bày tỏ lời cảm ơn chân thành tới tập thể các thầy cô giáo Viện công nghệ thông tin – Viện Hàn lâm Khoa học và Công nghệ Việt Nam, các thầy cô giáo Trường Đại học Công nghệ thông tin và truyền thông - Đại học Thái Nguyên đã dạy dỗ chúng tôi trong suốt quá trình học tập chương trình cao học tại trường.

Đặc biệt tôi xin bày tỏ lòng biết ơn sâu sắc tới thầy giáo PGS.TS. Đoàn Văn Ban đã quan tâm, định hướng và đưa ra những góp ý, chỉnh sửa quý báu cho tôi trong quá trình làm luận văn tốt nghiệp. Cũng như các bạn bè, đồng nghiệp, gia đình và người thân đã quan tâm, giúp đỡ và chia sẻ với tôi trong suốt quá trình làm luận văn tốt nghiệp.

Dù đã có nhiều cố gắng nhưng chắc chắn sẽ không tránh khỏi những thiếu sót vì vậy rất mong nhận được sự đóng góp ý kiến của các thầy, cô và các bạn để luận văn này được hoàn thiện hơn.

Tôi xin chân thành cảm ơn!

Thái Nguyên, tháng 08 năm 2020

Phạm Thị Liên

MỤC LỤC

Trang

THUẬT NGỮ VÀ TỪ VIẾT TẮT

Thuật ngữ	Diễn giải
CAM	Canonical Adjacency Matrix
Community Social Structure	Cấu trúc cộng đồng mạng xã hội
CSDL	Cơ sở dữ liệu
FFSM	FastFrequent Subgraph Mining
GI	Graph isomorphism - Đẳng cấu đồ thị
Graph isomorphism problem	Bài toán đồ thị đẳng cấu
SFI	Viện Santa Fe
SGI	SubGraph Isomorphism – Đẳng cấu đồ thị con

DANH MỤC CÁC HÌNH VẼ

DANH MỤC CÁC BẢNG

MỞ ĐẦU

1. Lý do chọn đề tài

Khai phá dữ liệu là lĩnh vực đang được nhiều người tập trung nghiên cứu và phát triển nhiều ứng dụng phổ biến. Trong đó, bài toán khai phá đồ thị con thường xuyên đã và đang thu hút được nhiều sự quan tâm nghiên cứu bởi phạm vi ứng dụng quan trọng trong nhiều lĩnh vực khác nhau như trong tin-sinh (bioinformatics), tin-hóa (cheminformatics) khai thác đăng nhập trang web, lập chỉ mục video, hay lập chỉ mục cơ sở dữ liệu hiệu quả, ... Vấn đề khai phá mẫu thường xuyên là từ một tập dữ liệu các đối tượng với một ngưỡng độ hỗ trợ tối thiểu minsup. Dữ liệu có thể rất đa dạng từ dữ liệu nhị phân, dữ liệu số nguyên, số thực hoặc các dữ liệu có cấu trúc phức tạp hơn như cấu trúc cây, đồ thị, … Cho đến nay vẫn chưa có lời giải hiệu quả cho bài toán này do độ phức tạp của bài toán là rất lớn khi đồ thị có số đỉnh lớn và mật độ các cạnh dày. Tuy nhiên, sự phức tạp của những vấn đề này sẽ giảm khi cơ sở dữ liệu (CSDL) đồ thị có thêm thông tin về các đỉnh và các cạnh đã được gán nhãn. Có thể sử dụng các nhãn để hạn chế các đỉnh có thể tạo thành các cặp trong quá trình kiểm tra sự đẳng cấu của đồ thị con.

Và đó chính là lý do em lựa chọn đề tài "*Bài toán đồ thị con đẳng cấu trong khai phá dữ liệu đồ thị và ứng dụng phát hiện đồ thị con phổ biến*" để nghiên cứu làm luận văn thạc sĩ của mình.

2. Mục đích nghiên cứu

Nghiên cứu thuật toán đẳng cấu đồ thị và thuật toán phát hiện đồ thị con phổ biến trong CSDL đồ thị.

3. Đối tượng nghiên cứu

- Bài toán đồ thị con đẳng cấu
- Khai phá dữ liệu đồ thị
- Bài toán khai phá đồ thị con phổ biến trong CSDL đồ thị
- Thuật toán FFSM
- Thuật toán SGI Decision Tree.

4. Phạm vi nghiên cứu

+ Lý thuyết:

- Tìm hiểu lý thuyết đồ thị và các phương pháp khai phá dữ liệu đồ thị.
- Nghiên cứu các thuật toán tìm đồ thị con đẳng cấu và các thuật toán khai phá đồ thị con phổ biến.

+ Thực nghiệm:

- Xây dựng chương trình thực nghiệm thuật toán.
- Áp dụng chương trình trên một số bộ cơ sở dữ liệu trong khai phá dữ liệu đồ thị.

5. Ý nghĩa khoa học

Đây là một hướng nghiên cứu lý thuyết, xây dựng các thuật toán tìm kiếm đồ thị con đẳng cấu và đồ thị con phổ biến và kiểm thử trên các bộ dữ liệu về đồ thị.

6. Phương pháp nghiên cứu

- *Phương pháp nghiên cứu lý thuyết*: Sưu tập các tài liệu có liên quan đến đề tài, nghiên cứu để hiểu sâu các nội dung vấn đề cần nghiên cứu.
- *Phương pháp nghiên cứu thực nghiệm*: Cài đặt chương trình thử nghiệm phương pháp tìm kiếm đồ thị con đẳng cấu và thuật toán FFSM để liệt kê các đồ thị con phổ biến trên các bộ dữ liệu và đánh giá hiệu suất của phương pháp.
- *Phương pháp trao đổi khoa học*: Trao đổi nội dung nghiên cứu với giáo viên hướng dẫn, với các đồng nghiệp để đề xuất và giải quyết các nội dung luận văn đề ra.

7. Cấu trúc đề tài

Luận văn được chia thành các phần chính như sau:

Chương 1: Khai phá dữ liệu đồ thị

Chương 2: Đồ thị đẳng cấu và đồ thị con phổ biến

Chương 3: Thử nghiệm thuật toán tìm đồ thị con phổ biến

Kết luận và hướng phát triển.