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

HOÀNG VIỆT DŨNG

KHAI PHÁ ĐỒ THỊ CON PHỔ BIẾN VÀ ỨNG DỤNG

Thái Nguyên, 2018

LỜI CAM ĐOAN

Tôi xin cam đoan số liệu và kết quả nghiên cứu trong luận văn này là trung thực và chưa sử dụng để bảo vệ luận văn của một học vị nào.

Tôi xin cam đoan mọi sự giúp đỡ cho việc thực hiện luận văn này đã được cảm ơn và các thông tin trích dẫn trong luận văn đã được chỉ rõ nguồn gốc.

Hà Nội, tháng 05 năm 2018

Tác giả

Hoàng Việt Dũng

LỜI CẢM ƠN

Để hoàn thành luận văn, tôi đã nhận được sự giúp đỡ rất tận tình, sự đóng góp quý báu của nhiều cá nhân và tập thể.

Trước hết, tôi xin trân trọng cảm ơn Thầy giáo PGS.TS. Nguyễn Long Giang người đã nhiệt tình hướng dẫn, giúp đỡ tôi trong việc hoàn thành luận văn này.

Tôi xin trân trọng cảm ơn sự góp ý chân thành của các Thầy, Cô giáo Viện Công nghệ thông tin, Các thầy giáo, cô giáo Trường Đại học Công nghệ thông tin và truyền thông - Đại học Thái Nguyên, đã tạo điều kiện thuận lợi cho tôi thực hiện và hoàn thành đề tài.

Tôi xin cảm ơn đến gia đình, người thân, các đồng nghiệp và bạn bè đã động viên, giúp đỡ, tạo điều kiện thuận lợi cho tôi trong quá trình thực hiện đề tài này.

Một lần nữa tôi xin trân trọng cảm ơn !

Hà Nội, tháng 5 năm 2018

Tác giả

Hoàng Việt Dũng

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Từ viết tắt	**Ý nghĩa**
KDD	Knowleadge Discovery in Database
CSDL	Cơ sở dữ liệu
CNTT	Công nghệ thong tin
OLAP	On Line Analytical Processing
FSG	Frequency SubGraph Mining
CONGA	Cluster Overlap Newman-Girvan Algorithm
FNCA	Fast Network Community Algorithm

DANH MỤC BẢNG

DANH MỤC HÌNH

ĐẶT VẤN ĐỀ

1.1. Sự cần thiết lựa chọn đề tài

Trong những năm gần đây, khai phá dữ liệu dữ liệu đồ thị là chủ đề thu hút sự quan tâm của cộng đồng nghiên cứu về khai phá dữ liệu và học máy và được ứng dụng rộng rãi trong nhiều lĩnh vực như: phân tích dữ liệu hóa học, sinh học, phân tích mạng máy tính, phân tích mạng xã hội..[4, 5, 6]. Theo tìm hiểu của tác giả, các nghiên cứu liên quan đến khai phá dữ liệu đồ thị thường tập trung vào các bài toán như: liệt kê và đếm (Enumeration and Counting), phân lớp đồ thị (graph classification), phân cụm đồ thị (graph clustering), học bán giám sát (semi-supervisedlearning), tóm tắt đồ thị (graph summarization), khai phá đồ thị phổ biến (frequent graph mining)…

Khai phá đồ thị phổ biến là hướng nghiên cứu sôi động trong mấy năm gần đây trong lĩnh vực khai phá dữ liệu đồ thị. Dựa trên nền tảng của bài toán khai phá luật kết hợp, khai phá đồ thị phổ biến nhằm tìm kiếm các đồ thị con phổ biến (tương ứng với tập mục phổ biến). Các đồ thị con phổ biến là nền tảng để giải quyết bài toán dự báo trên không gian dữ liệu đồ thị và có ứng rộng rãi trong nhiều lĩnh vực khác nhau của đời sống. Một số thuật toán điển hình khai phá đồ thị phổ biến là CMTMiner [7], HSIGRAM, VSIGRAM [8],. Thuật toán CMTMiner thực hiện việc duyệt các cạnh phổ biến và xây dựng cây DFS để tìm các đồ thị con phổ biến.Trong khi đó, HSIGRAM, VSIGRAM là hai thuật toán xác định các đồ thị con phổ biến trong một đồ thị lớn.

Như đã trình bày ở trên, lĩnh vực khai phá dữ liệu đồ thị đã thu được các kết quả quan trọng về lý thuyết và đã có ứng dụng hiệu quả trong việc giải quyết một số bài toán trong thực tiễn.Một trong những bài toán ứng dụng của khai phá đồ thị con phổ biến là phát hiện cộng đồng mạng xã hội.