

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

Trang phụ bìa	
LỜI CAM ĐOAN	i
LỜI CẢM ƠN	iii
MỤC LỤC	iv
DANH MỤC CÁC TỪ VIẾT TẮT	vi
DANH MỤC BẢNG	vii
DANH MỤC HÌNH	ixvii
ĐẶT VẤN ĐỀ.....	1
1.1. Sự cần thiết lựa chọn đề tài	1
1.2. Mục tiêu nghiên cứu của đề tài	3
2. Đối tượng và phạm vi nghiên cứu.....	3
2.1. Đối tượng	3
2.2. Phạm vi nghiên cứu.....	3
3. Hướng nghiên cứu của đề tài	3
4. Cấu trúc của luận văn	3
Chương 1. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU ĐỒ THỊ.....	4
1.1. Tổng quan về khai phá dữ liệu đồ thị.....	4
1.1.1. Tại sao cần khai phá dữ liệu:.....	4
1.1.2. Các khái niệm khai phá dữ liệu	4
1.1.3. Các chức năng chính của khai phá dữ liệu.....	5
1.1.4. Các công cụ khai phá dữ liệu.....	6
1.2. Quy trình khai phá dữ liệu đồ thị	7
1.2.1. Hình thành và định nghĩa bài toán	7
1.2.2. Thu thập và tiền xử lý dữ liệu.....	8
1.2.3. Khai phá dữ liệu và rút ra các tri thức	8
1.2.4. Phân tích và kiểm định kết quả	9

1.2.5. Sử dụng các tri thức phát hiện được	9
1.3. Các bài toán trong khai phá dữ liệu đồ thị	9
1.3.1. Khai phá luật kết hợp	9
1.3.2. Phân lớp	9
1.3.3. Phân cụm	10
1.3.4. Dự báo	11
1.3.5. Các mẫu tuần tự	11
1.3.6. Các cây quyết định	12
1.4. Các ứng dụng của khai phá dữ liệu đồ thị	13
1.4.1. Các lĩnh vực liên quan đến phát hiện tri thức và khai phá dữ liệu	13
1.4.2. Ứng dụng của khai phá dữ liệu	13
Chương 2. CÁC PHƯƠNG PHÁP KHAI PHÁ ĐỒ THỊ CON	15
PHỔ BIẾN	15
2.1. Các định nghĩa về đồ thị con phổ biến	15
2.1.1. Giới thiệu về lý thuyết đồ thị	15
2.1.2. Khai phá dữ liệu	19
2.1.3. Một số phương pháp khai phá dữ liệu	21
2.1.4. Khai phá đồ thị con phổ biến	26
2.2. Các phương pháp khai phá đồ thị con phổ biến	27
2.2.1. Thuật toán Apriori để tìm tập con phổ biến	27
2.2.2. Thuật toán FSG (Frequency SubGraph Mining) để phát hiện cộng đồng mạng xã hội	34
2.3. Ứng dụng khai phá đồ thị con phổ biến phát hiện cộng đồng trên mạng xã hội	39
2.3.1. Cộng đồng mạng xã hội	39
2.3.2. Các phương pháp truyền thống	41
2.3.3. Các phương pháp áp dụng thuật toán phân chia:	43

2.3.4. Phát hiện cộng đồng trong mạng xã hội	45
Chương 3. THỬ NGHIỆM, ĐÁNH GIÁ KẾT QUẢ VỚI BÀI TOÁN PHÁT HIỆN CỘNG ĐỒNG MẠNG XÃ HỘI	50
3.1. Mô tả bài toán.....	50
3.2. Mô hình giải quyết bài toán	50
3.2.1. Mô hình đồ thị thông tin.....	50
3.2.2. Hướng của cạnh	50
3.2.3. Trọng số của cạnh	51
3.2.4. Lựa chọn mô hình cho bài toán	51
3.3. Thử nghiệm, đánh giá mô hình (thu thập dữ liệu từ mạng xã hội, biểu diễn dữ liệu, cài đặt, thử nghiệm và đánh giá kết quả)	51
3.3.1. Giới thiệu nhóm Facebook, phân tích nhóm Facebook	51
3.3.2. Phương pháp thu thập dữ liệu từ nhóm Facebook	53
3.3.3. Thử nghiệm bài toán	54
3.3.4. Thuật toán chính	55
3.3.5. Demo bài toán	56
3.3.6. Đánh giá.....	62
KẾT LUẬN VÀ KHUYẾN NGHỊ	64
1. Kết luận	64
2. Khuyến nghị	64
TÀI LIỆU THAM KHẢO	

DANH MỤC CÁC TỪ VIẾT TẮT

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

DANH MỤC BẢNG

Bảng 2.1. Biểu diễn giao dịch.....	30
Bảng 2.2. Biểu diễn giao dịch.....	30
Bảng 2.3. Biểu diễn giao dịch.....	31
Bảng 2.4. Biểu diễn giao dịch.....	31
Bảng 2.5. Biểu diễn giao dịch.....	31
Bảng 2.6. Biểu diễn giao dịch.....	32
Bảng 2.7. Kết quả cuối cùng.....	32
Bảng 3.1. Một dạng format (Ma trận thích (Like)).....	54
Bảng 3.2. Bảng người dùng sau khi đã giải mã.....	57
Bảng 3.3. Mảng chuyển đổi.....	58

DANH MỤC HÌNH

Hình 1.1. Các bước trong khai phá dữ liệu và KDD.....	5
Hình 1.2. Quá trình khai phá dữ liệu (khám phá tri thức).....	7
Hình 1.3. Phân cụm.....	11
Hình 1.4. Cây quyết định	12
Hình 2.1. Mô tả mô hình đồ thị.....	15
Hình 2.2. Các loại đồ thị	16
Hình 2.3. Đơn đồ thị vô hướng	16
Hình 2.4. Đa đồ thị vô hướng.....	17
Hình 2.5. Giá đồ thị vô hướng	18
Hình 2.6. Đơn đồ thị có hướng	18
Hình 2.7. Đa đồ thị có hướng.....	19
Hình 2.8. Minh họa thuật toán FSG.....	35
Hình 2.9. Cộng đồng mạng xã hội đơn giản với 3 cộng đồng.....	40
Hình 2.10. Phương pháp phân vùng đồ thị	41
Hình 2.11. Ví dụ về phép phân chia một đỉnh trong đồ thị	44
Hình 2.12. Một số ví dụ về cộng đồng trên mạng xã hội.....	45
Hình 2.13. Mô hình mạng lưới cộng tác của các nhà khoa học	46
Hình 3.1. Liên kết giữa hai đỉnh (người) trên mạng xã hội	50
Hình 3.2. Quan hệ giữa hai người trên mạng xã hội với trọng số	51
Hình 3.3. Hình ảnh một nhóm Facebook.....	52
Hình 3.5. Ví dụ 3 định dạng worksheet: bạn bè, thích, bình luận	59
Hình 3.6. Đồ thị sau khi xử lý.....	59
Hình 3.7. Bộ dữ liệu sau khi xử lý	60
Hình 3.8. So sánh thuật toán Light-FSG với thuật toán khác	60
Hình 3.9. Giao diện chương trình	61
Hình 3.10. Biểu diễn Mạng đồ thị 2D.....	61
Hình 3.11. Biểu diễn Mạng đồ thị 3D.....	62

ĐẶT VẤN ĐỀ

1.1. Sự cần thiết lựa chọn đề tài

Trong những năm gần đây, khai phá 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-supervised learning), 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 dụng 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.