

ĐẠI HỌC THÁI NGUYÊN  
TRƯỜNG ĐẠI HỌC CNTT & TT THÁI NGUYÊN

NGUYỄN THẾ ĐẠT

NGHIÊN CỨU MÔ HÌNH  
PHÂN CỤM CÓ THỨ BẬC CÁC ĐỒ THỊ DỮ LIỆU

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên – 2017

ĐẠI HỌC THÁI NGUYÊN  
TRƯỜNG ĐẠI HỌC CNTT & TT THÁI NGUYÊN

NGUYỄN THẾ ĐẠT

**NGHIÊN CỨU MÔ HÌNH PHÂN CỤM CÓ THỨ BẬC  
CÁC ĐỒ THỊ DỮ LIỆU**

**Chuyên ngành: Khoa học máy tính**

**Mã số: 60 48 0101**

**LUẬN VĂN THẠC SỸ KHOA HỌC MÁY TÍNH**

**NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS ĐOÀN VĂN BAN**

**Thái Nguyên – 2017**

## LỜI CAM ĐOAN

Tên tôi là: Nguyễn Thế Đạt

Sinh ngày: 09/01/1979

Học viên lớp cao học CK14 - Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên.

Hiện đang công tác tại: Trường THCS Hạp Lĩnh – TP Bắc Ninh – Bắc Ninh

Xin cam đoan: Đề tài “**Nghiên cứu mô hình phân cụm có thứ bậc các đồ thị dữ liệu**” do Thầy giáo PGS.TS Đoàn Văn Ban hướng dẫn là công trình nghiên cứu của riêng tôi. Tất cả tài liệu tham khảo đều có nguồn gốc, xuất xứ rõ ràng.

Tác giả xin cam đoan tất cả những nội dung trong luận văn đúng như nội dung trong đề cương và yêu cầu của thầy giáo hướng dẫn. Nếu sai tôi hoàn toàn chịu trách nhiệm trước hội đồng khoa học và trước pháp luật.

*Thái Nguyên, ngày 15 tháng 5 năm 2017*

**Tác giả luận văn**

**Nguyễn Thế Đạt**

## LỜI CẢM ƠN

Sau một thời gian nghiên cứu và làm việc nghiêm túc, được sự động viên, giúp đỡ và hướng dẫn tận tình của Thầy giáo hướng dẫn PGS.TS Đoàn Văn Ban, luận văn với đề tài “**Nghiên cứu mô hình phân cụm có thứ bậc các đồ thị dữ liệu**” đã hoàn thành.

Tôi xin bày tỏ lòng biết ơn sâu sắc đến:

Thầy giáo hướng dẫn **PGS.TS Đoàn Văn Ban** đã tận tình chỉ dẫn, giúp đỡ tôi hoàn thành luận văn này.

Khoa sau Đại học Trường Đại học công nghệ thông tin và truyền thông đã giúp đỡ tôi trong quá trình học tập cũng như thực hiện luận văn.

Tôi xin chân thành cảm ơn bạn bè, đồng nghiệp và gia đình đã động viên, khích lệ, tạo điều kiện giúp đỡ tôi trong suốt quá trình học tập, thực hiện và hoàn thành luận văn này.

*Thái Nguyên, ngày 15 tháng 5 năm 2017*

**Tác giả luận văn**

**Nguyễn Thế Đạt**

## MỤC LỤC

<b>LỜI CAM ĐOAN .....</b>	<b>i</b>
<b>LỜI CẢM ƠN .....</b>	<b>ii</b>
<b>MỤC LỤC .....</b>	<b>iii</b>
<b>DANH MỤC BẢNG .....</b>	<b>v</b>
<b>DANH MỤC CÁC TỪ VIẾT TẮT .....</b>	<b>vi</b>
<b>DANH MỤC CÁC HÌNH VẼ .....</b>	<b>vii</b>
<b>MỞ ĐẦU .....</b>	<b>1</b>
<b>CHƯƠNG 1: PHÂN CỤM DỮ LIỆU VÀ PHÂN CỤM ĐỒ THỊ DỮ LIỆU .....</b>	<b>4</b>
1.1. Phân cụm dữ liệu.....	4
1.1.1. Khái niệm và mục tiêu của phân cụm dữ liệu .....	4
1.1.2. Một số khái niệm cần thiết khi tiếp cận phân cụm dữ liệu.....	7
1.1.3. Một số kỹ thuật trong phân cụm dữ liệu .....	10
1.1.4. Các ứng dụng của phân cụm dữ liệu .....	16
1.2. Phân cụm đồ thị dữ liệu .....	17
1.2.1. Mô hình đồ thị dữ liệu .....	17
1.2.2. Các loại độ đo.....	18
1.2.3. Một số kỹ thuật phân cụm đồ thị dữ liệu.....	23
1.3. Kết luận chương 1 .....	28
<b>CHƯƠNG 2: PHÂN CỤM CÓ THỨ BẬC CÁC ĐỒ THỊ DỮ LIỆU .....</b>	<b>29</b>
2.1. Thuật toán CHAMELEON .....	29
2.2. Thuật toán CURE.....	31
2.3. Thuật toán Girvan-Newman .....	34
2.3.1. Giới thiệu về độ đo modularity .....	34
2.3.2. Độ đo trung gian .....	35
2.3.3. Thuật toán phân cụm Girvan-Newman .....	36
2.4. Thuật toán CNM (Clauset-Newman-Moore).....	39
2.5. Thuật toán Rosvall-Bergstrom.....	42

2.6. Thuật toán INC (Incre-Comm-Extraction) .....	47
2.6.1. Nội dung thuật toán .....	47
2.6.2. Độ phức tạp của thuật toán.....	49
2.6.3. Độ đo chất lượng phân cụm của thuật toán.....	50
2.7. Kết luận chương 2.....	51
<b>CHƯƠNG 3: ỨNG DỤNG THUẬT TOÁN PHÂN CỤM CÓ THỨ BẬC</b>	
<b>TRONG PHÂN CỤM ĐỒ THỊ DỮ LIỆU CÁC MẠNG XÃ HỘI.....</b>	<b>52</b>
3.1. Bài toán phân cụm mạng xã hội.....	52
3.2. Xây dựng chương trình ứng dụng phân cụm đồ thị các mạng xã hội.....	53
3.2.1. Giai đoạn 1: Thu thập dữ liệu.....	53
3.2.2. Giai đoạn 2: Xử lý dữ liệu.....	54
3.2.3. Giai đoạn 3: Xây dựng ứng dụng phân cụm có thứ bậc đồ thị các mạng xã hội .....	55
3.3. Các kết quả thực nghiệm và đánh giá .....	56
3.3.1. Thời gian thực thi thuật toán .....	57
3.3.2. Số cụm được phân chia .....	58
3.3.3. Chất lượng phân cụm .....	58
3.4. Phân cụm đồ thị mạng xã hội dựa trên mối quan tâm của người dùng .....	58
3.4.1. Giới thiệu.....	58
3.4.2. Mô hình hóa dữ liệu .....	60
3.4.3. Xây dựng dữ liệu.....	62
3.4.4. Xây dựng ứng dụng.....	66
3.4.5. Thực nghiệm và đánh giá INC .....	69
3.5. Kết luận chương 3.....	74
<b>KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN .....</b>	<b>75</b>
<b>TÀI LIỆU THAM KHẢO .....</b>	<b>76</b>

**DANH MỤC BẢNG**

Bảng 3.1: Kết quả thực thi các thuật toán.....	57
Bảng 3.2: Kết quả thực thi 2 thuật toán INC và CNM.....	69

## DANH MỤC CÁC TỪ VIẾT TẮT

<b>Từ hoặc cụm từ</b>	<b>Từ tiếng Anh</b>	<b>Từ tiếng Việt</b>
<b>CNM</b>	Clauset-Newman-Moore	Phân cụm có thứ bậc tích tụ
<b>CSDL</b>		Cơ sở dữ liệu
<b>CURE</b>	Clustering Using Representatives	Phân cụm dữ liệu sử dụng điểm đại diện
<b>GN</b>	Girvan-Newman	Phân cụm phân chia
<b>INC</b>	Incre-Comm-Extraction	
<b>MCL</b>	Markov Clustering	Phân cụm theo mô hình Markov
<b>RB</b>	Rosvall-Bergstrom	



## DANH MỤC CÁC HÌNH VẼ

Hình 1.1: Ví dụ về phân cụm dữ liệu .....	4
Hình 1.2: Ví dụ phân cụm các ngôi nhà dựa trên khoảng cách .....	5
Hình 1.3: Ví dụ phân cụm các ngôi nhà dựa trên kích cỡ .....	6
Hình 1.4: Các chiến lược phân cụm có thứ bậc .....	11
Hình 1.5: Ví dụ về phân cụm dựa theo mật độ .....	12
Hình 1.6: Cấu trúc phân cụm dựa trên lưới .....	13
Hình 1.7: Ví dụ về phân cụm dựa trên mô hình .....	14
Hình 1.8: Các cách mà các cụm có thể đưa ra .....	16
Hình 1.9: (a) Tối ưu đường kính cực tiểu hoặc tổng cực tiểu tạo ra cụm B nhưng A lại tốt hơn trên thực tế. (b) Tối ưu K-means tạo ra cụm B nhưng A lại tốt hơn .....	20
Hình 1.10: Minh họa mô hình đồ thị cho bước đi ngẫu nhiên .....	25
Hình 2.1: Phân cụm Chameleon .....	31
Hình 2.2: Sự di chuyển về trung tâm cụm .....	32
Hình 2.3: Sự sáp nhập của các cụm .....	32
Hình 2.4: Cụm dữ liệu khai phá bởi thuật toán CURE .....	33
Hình 2.5: Ví dụ phát hiện cụm sử dụng thuật toán Girvan - Newman .....	38
Hình 2.6: Khung làm việc cơ sở để phân cụm đồ thị như quá trình truyền thông .....	42
Hình 2.7: Ví dụ về mã Huffman .....	43
Hình 2.8: Phân hoạch vào một lượng tối ưu các modul .....	45
Hình 3.1: Các bước thực hiện chương trình .....	53
Hình 3.2: Ví dụ về tập dữ liệu Dolphins.gml .....	54
Hình 3.3: Tập dữ liệu Dolphins.txt .....	54
Hình 3.4: Nạp file dữ liệu đầu vào .....	55
Hình 3.5: Kết quả chạy thuật toán phân cụm CNM cho bộ dữ liệu dolphins.txt.....	56
Hình 3.6: Kết quả chạy thuật toán Girvan-Newman cho bộ dữ liệu dolphins.txt.....	56
Hình 3.7: Biểu đồ so sánh thời gian thực thi thuật toán.....	57
Hình 3.8: Biểu đồ so sánh số lượng cụm .....	58

Hình 3.9: Biểu đồ so sánh chất lượng phân cụm.....	58
Hình 3.10: Đăng tin và bình luận trên Facebook .....	60
Hình 3.11: Một phần danh sách tài khoản Facebook.....	62
Hình 3.12: Giao diện đăng ký một ứng dụng trên Facebook API .....	63
Hình 3.13: Thu thập dữ liệu thủ công với Graph API Explorer.....	63
Hình 3.14: Thu thập dữ liệu tự động với Facebook API.....	64
Hình 3.15: Một phần dữ liệu thu thập được cập nhật trên SQL Server .....	64
Hình 3.16: Một phần dữ liệu về danh sách và số lượng ID người dùng đã bình luận trên các tường Facebook tương ứng.....	65
Hình 3.17: Một phần dữ liệu mạng xã hội dựa trên mối quan tâm của người dùng.....	66
Hình 3.18: Giao diện tự động thu thập bộ dữ liệu .....	67
Hình 3.19: Kết quả chạy chương trình phân cụm với INC và CNM.....	68
Hình 3.20: Một phần biểu đồ dendrogram kết quả phân cụm với INC .....	68
Hình 3.21: Đồ thị so sánh thời gian thực thi INC và CNM .....	69
Hình 3.22: Đồ thị so sánh số lượng cụm theo INC và CNM .....	70
Hình 3.23: Đồ thị tương quan số lượng cụm với giá trị $s$ .....	70
Hình 3.24: Đồ thị so sánh chất lượng phân cụm theo INC và CNM.....	70
Hình 3.25: Đồ thị tương quan chất lượng cụm với giá trị $s$ .....	71
Hình 3.2.6: Kết quả phân chia cụm lớn thành các cụm con (bất động sản, chứng khoán, ô tô, xe máy.....)	72
Hình 3.27: Kết quả phân chia cụm lớn yêu thích đồ nội thất, lưu niệm, thời trang thành các cụm con (giày dép, đồng hồ, hoa tươi, quà lưu niệm, ngân hàng.....)	72
Hình 3.28: Kết quả phân cộng động quan tâm tới Phật giáo .....	73
Hình 3.29: Kết quả phân cộng động quan tâm tới mỹ phẩm, thẩm mỹ, bệnh viện thẩm mỹ đã được phân chia theo INC.....	73