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

TRẦN HÀ PHƯƠNG

PHÂN CỤM ĐỒ THỊ DỮ LIỆU VÀ ỨNG DỤNG

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2016

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

TRẦN HÀ PHƯƠNG

PHÂN CỤM ĐỒ THỊ DỮ LIỆU VÀ ỨNG DỤNG

Chuyên ngành: Khoa học máy tính
Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: PGS. TS. ĐOÀN VĂN BAN

THÁI NGUYÊN - 2016

LỜI CAM ĐOAN

Tên tôi là: Trần Hà Phương

Sinh ngày:

Học viên lớp cao học CHK13 - Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên.

Hiện đang công tác tại:

Xin cam đoan: Đề tài "***Phân cụm đồ thị dữ liệu và ứng dụng***" 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 14 tháng 4 năm 2016

Tác giả luận văn

Trần Hà Phương

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 "***Phân cụm đồ thị dữ liệu và ứng dụng***"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 16 tháng 6 năm 2016

Tác giả luận văn

Trần Hà Phương

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Từ hoặc cụm từ	Từ tiếng Anh	Từ tiếng Việt
BDCM	Binding data Clustering Methods	Phương pháp phân cụm dữ liệu có liên kết
CA	Continuous Attribute	Thuộc tính liên tục
CSDL		Cơ sở dữ liệu
DA	Discrette Attribute	Thuộc tính rời rạc
DBM	Density-Based Methods	Phương pháp dựa trên mật độ
GBM	Grid-Based Methods	Phương pháp dựa trên lưới
HM	Hierarchical Methods	Phương pháp phân cấp
MBCM	Model-Based Clustering Methods	Phương pháp dựa trên mô hình phân cụm
MC	Markov Clustering	Phân cụm theo mô hình Markov
MST	Minimum Spanning Tree	Cây khung nhỏ nhất
PM	Partitioning Methods	Phương pháp phân hoạch
RWA	Random Walk Algorithm	Thuật toán bước đi ngẫu nhiên
SC	Star Clustering	Phân cụm hình sao
SCA	Spectral Clustering Algorithm	Thuật toán phân cụm quang phổ
SOM	Self-Organizing Map	Mạng tự tổ chức

DANH MỤC BẢNG

DANH MỤC CÁC HÌNH ẢNH

MỞ ĐẦU

1. Tính khoa học và cấp thiết của đề tài

Phân cụm là một trong những vấn đề cơ bản phổ biến trong các lĩnh vực nhận dạng mẫu, học máy và khai thác dữ liệu. Hiện tại, trên thực tế có rất nhiều thuật toán phân cụm được công bố. Tuy nhiên, do không tồn tại một thuật toán phân cụm duy nhất cho tất cả các loại bộ dữ liệu, những thuật toán phân cụm mới vẫn liên tục được đề xuất. Kết quả là, người dùng phải chọn thuật toán thích hợp nhất từ nhiều ứng viên để đạt được kết quả chính xác.

Trong thực tế, việc lựa chọn thuật toán phân cụm dữ liệu phù hợp là rất khó khăn do người sử dụng thường không có một kiến thức tiên nghiệm về sự đa dạng và phức tạp của dữ liệu. Để phần nào giảm bớt nhược điểm trên, các thuật toán phân cụm dựa trên đồ thị được đề xuất do ưu điểm ở khả năng xử lý các bộ dữ liệu đa dạng và có cấu trúc. Bản chất của các thuật toán này là biểu diễn dữ liệu dựa trên đồ thị và phân cụm các thành phần theo các thuật toán thiết kế riêng.

Đồ thị là những cấu trúc toán học được sử dụng để đại diện cho mối quan hệ giữa cặp đối tượng từ một tập hợp xác định. Đồ thị chứa đỉnh (đại diện cho các đối tượng) và các cạnh nối các đỉnh (đại diện cho mối quan hệ giữa các đối tượng cặp). Đây là phương pháp cấu trúc dữ liệu quan trọng được sử dụng trong rất nhiều lĩnh vực như khai thác dữ liệu, xử lý ngôn ngữ tự nhiên, tìm kiếm thông tin và khai thác thông tin. Trong phân cụm, sự tương đồng giữa các đối tượng được phân cụm có thể được diễn tả như một đồ thị có trọng số. Trong đó, các đối tượng là các đỉnh và sự tương đồng là trọng số của các cạnh. Bài toán phân cụm sẽ được đơn giản hóa về bài toán phân cụm đồ thị mà nhiệm vụ chính là tách các đồ thị phụ dày đặc và kết nối thưa thớt khỏi nhau dựa trên khái niệm của mật độ nội cụm so với khoảng cách liên cụm.

Với những lý do trên, tác giả đã chọn đề tài ***"Phân cụm đồ thị dữ liệu và ứng dụng"*** làm đề tài nghiên cứu luận văn tốt nghiệp thạc sĩ chuyên ngành Khoa học máy tính.