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

CAO VĂN NGUYÊN

NGHIÊN CỨU MỘT SỐ KỸ THUẬT KHAI PHÁ DỮ LIỆU KHÔNG GIAN SỬ DỤNG CÂY QUYẾT ĐỊNH

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2013

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

CAO VĂN NGUYÊN

NGHIÊN CỨU MỘT SỐ KỸ THUẬT KHAI PHÁ DỮ LIỆU KHÔNG GIAN SỬ DỤNG CÂY QUYẾT ĐỊNH

Chuyên ngành: Khoa học máy tính

Mã số: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. ĐẶNG VĂN ĐỨC

Thái Nguyên – 2013

LỜI CẢM ƠN

*Tôi xin gửi lời cảm ơn chân thành nhất tới **PGS. TS Đặng Văn Đức** - người đã hướng dẫn, chỉ bảo tận tình, cung cấp tài liệu và phương pháp luận nghiên cứu khoa học để tôi hoàn thành bản luận văn này.*

Tôi xin bày tỏ lòng cảm ơn sâu sắc tới thầy cô, bạn bè cùng khóa, cùng lớp đã giúp đỡ tôi trong suốt những năm học qua.

Xin cảm ơn gia đình, bạn bè, những người luôn khuyến khích, động viên và giúp đỡ tôi trong mọi hoàn cảnh khó khăn.

Tôi xin cảm ơn các thầy cô trong trường Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên đã hết sức tạo điều kiện cho tôi trong quá trình học và làm luận văn này.

Luận văn được hoàn thành trong thời gian hạn hẹp nên không thể tránh được những thiếu sót. Tôi xin cảm ơn thầy cô, bạn bè, đồng nghiệp đã có những ý kiến đóng góp chân thành cho nội dung của luận văn, để tôi có thể tiếp tục đi sâu tìm hiểu về lĩnh vực này trong tương lai.

Thái Nguyên, 11/2013

Cao Văn Nguyên

caonguyenvp@gmail.com

LỜI CAM ĐOAN

Tôi xin cam đoan kết quả đạt được trong luận văn là sản phẩm của riêng cá nhân tôi, không sao chép lại của người khác. Trong toàn bộ nội dung luận văn, những điều đã được trình bày hoặc là của riêng cá nhân tôi hoặc là được tổng hợp từ nhiều nguồn tài liệu. Tất cả các nguồn tài liệu tham khảo được dùng đều có xuất xứ rõ ràng, được trích dẫn hợp pháp.

Tôi xin chịu hoàn toàn trách nhiệm và chịu mọi hình thức kỉ luật theo quy định cho lời cam đoan của mình.

Thái Nguyên, 11/2013
Cao Văn Nguyên

MỤC LỤC

TRANG

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

CSDL	Cơ sở dữ liệu
GIS	Geographic information system
Object ID	Identifier of objects
SDT	Spatial Decision Tree
SJI	Spatial Join Index
SJR	Spatial Join Relation
SpatRel	Spatial Relation
SpatMes	Spatial Measure
SQL	Structured Query Language

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

MỞ ĐẦU

1. Đặt vấn đề

Những tiến bộ trong công nghệ CSDL và kỹ thuật thu thập dữ liệu như đọc mã số mã vạch, viễn thám, ghi nhận thông tin từ các vệ tinh,… đã tạo ra một lượng lớn thông tin, dữ liệu. Việc dữ liệu tăng lên nhanh với quy mô lớn đòi hỏi phải được khai phá để trích chọn ra các tri thức hữa ích phục vụ cho công tác chuyên môn. Chính điều này đã dẫn đến sự ra đời của lĩnh vực khai phá dữ liệu hay khai phá tri thức trong các CSDL. Khai phá tri thức trong các CSDL có thể được định nghĩa là khai phá tri thức đáng quan tâm, tiềm ẩn và chưa biết trước trong các CSDL. Khai phá dữ liệu là sự kết hợp của một số lĩnh vực bao gồm học máy, các hệ thống CSDL, thể hiện dữ liệu, thống kê và lý thuyết thông tin.

Đã có nhiều nghiên cứu về khai phá dữ liệu trong các CSDL quan hệ và giao dịch, nhưng đối với các CSDL không gian vấn đề khai phá dữ liệu vẫn còn là những thách thức cần được giải quyết.

Dữ liệu không gian là dữ liệu liên quan đến các đối tượng trong không gian. Một CSDL không gian lưu trữ các đối tượng không gian bao gồm các kiểu dữ liệu không gian và các quan hệ không gian giữa các đối tượng. Dữ liệu không gian mang thông tin hình học và khoảng cách thường được tổ chức theo các cấu trúc chỉ mục không gian và truy cập bằng các phương pháp truy cập không gian. Chính các đặc trưng khác biệt này của các CSDL không gian đã đặt ra nhiều trở ngại nhưng cũng mang đến nhiều cơ hội cho khai phá tri thức từ CSDL không gian. Khai phá dữ liệu không gian hay khai phá tri thức trong CSDL không gian là trích trọn ra các tri thức tiềm ẩn, các quan hệ không gian hay các mẫu chưa rõ lưu trữ trong các CSDL không gian.

Các nghiên cứu trước đây về học máy, các hệ thống CSDL và thống kê đã đặt nền móng cho nghiên cứu khai phá tri thức trong các CSDL. Và những tiến bộ của các CSDL không gian như cấu trúc dữ liệu không gian, lập luận không gian, tính toán hình học,… đã mở đường cho khai phá dữ liệu không gian. Trở ngại lớn nhất trong khai phá dữ liệu không gian là hiệu quả của các thuật toán khai phá dữ liệu không gian do lượng dữ liệu không gian thường có quy mô lớn, các kiểu dữ liệu không gian và các phương pháp truy cập không gian phức tạp.

Các phương pháp khai phá dữ liệu không gian tập trung theo ba hướng chính là khai phá luật kết hợp không gian, phân lớp dữ liệu không gian và phân cụm dữ liệu không gian. Với mong muốn nghiên cứu về phân lớp dữ liệu không gian sử dụng cây quyết định, luận văn đi sâu tìm hiểumột lĩnh vực nhỏ đó là phân lớp dữ liệu không gian sử dụng cây quyết định.