ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

TẠ DUY KHÁNH

PHƯƠNG PHÁP LAN TRUYỀN ĐỘ TƯƠNG TỰ TRONG PHÂN CỤM DỮ LIỆU VÀ ỨNG DỤNG

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2015

ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

TẠ DUY KHÁNH

PHƯƠNG PHÁP LAN TRUYỀN ĐỘ TƯƠNG TỰ TRONG PHÂN CỤM 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 NGUYỄN ĐÌNH HÓA

Thái Nguyên - 2015

LỜI CÁM ƠN

Đầu tiên, tôi xin gửi lời cảm ơn sâu sắc đến thầy giáo, cán bộ hướng dẫn khoa học PGS.TS Nguyễn Đình Hóa đã tận tình hướng dẫn tôi từ những buổi đầu tiên khi tiếp cận với đề tài khoa học. Trong quá trình làm luận văn, tôi cũng nhận được sự giúp đỡ rất nhiệt tình từ nhóm nghiên cứu của TS. Lê Hoàng Sơn tại Trung tâm tính toán hiệu năng cao, trường ĐH KHTN và đề tài NCKH cấp ĐHQG, mã số GG.14.60.

Tôi xin bày tỏ lòng biết ơn đến các thầy cô giáo ở trường Đại học Công nghệ thông tin và Truyền thông – Đại học Thái Nguyên, các cán bộ Trung tâm Đông Đô - Hà Nội, đã tận tình giảng dạy và tạo mọi điều kiện cho tôi học tập, nghiên cứu và hoàn thành luận văn này.

Tôi xin chân thành cảm ơn các bạn học viên lớp CK12H, CK13H – Khoa học máy tính đã giúp đỡ, tạo điều kiện cho tôi trong suốt quá trình học tập và thực hiện luận văn.

Cuối cùng, tôi xin gửi lời cảm ơn sâu sắc nhất đến gia đình, đồng nghiệp và bạn bè tôi, những người đã động viên, tạo mọi điều kiện cho tôi lao động và học tập trong suốt thời gian qua.

Tôi xin cam đoan luận văn là công trình nghiên cứu của riêng cá nhân tôi, không sao chép của ai. Luận văn là do tôi tự nghiên cứu, đọc, dịch tài liệu, tổng hợp và thực hiện. Nội dung lý thuyết trong luận văn có sử dụng một số tài liệu tham khảo như đã trình bày trong phần tài liệu tham khảo. Chương trình phần mềm và những kết quả trong luận văn là trung thực và chưa được công bố trong bất kỳ một hệ thống nào khác.

Một lần nữa, xin chân thành cảm ơn!

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

STT	Từ viết tắt	Từ tiếng Anh	Ý nghĩa
1	GIS	Geographical Information System	Hệ thống thông tin địa lý
2	AP	affinity propagation	Thuật toán lan truyền độ tương tự
3	CSDL	Database	Cơ sở dữ liệu
4	SIL	Silhouette	Công thức Silhouette
5	DEM	Digital Elevation Model	Mô hình kỹ thuật số độ cao
6	DTM	Digital Terrain Model	Mô hình kỹ thuật số các địa hình
7	TIN	Triangulated Irregular Network	Lưới tam giác không đều
8	SQL	Structured Query Language	Ngôn ngữ truy vấn có cấu trúc

DANH MỤC CÁC HÌNH VẼ

MỞ ĐẦU

1. Đặt vấn đề

Nguồn dữ liệu dồi dào cung cấp nhiều thông tin, từ đó nhân loại đúc rút thành tri thức trong quá trình phát triển xã hội loài người. Với sự phát triển của công nghệ điện toán và hệ thống lưu trữ dữ liệu thì khối lượng tài nguyên số ngày càng trở nên phong phú và đồ sộ. Trong xã hội hiện đại, thông tin đóng một vai trò then chốt. Nhu cầu xử lý dữ liệu, trích rút thông tin, kịp thời khai thác chúng để mang lại những hiệu quả thiết thực cho công tác quản lý, hoạt động sản xuất kinh doanh,… ngày càng trở nên cấp thiết.

Khai phá dữ liệu nói chung để trích rút thông tin và phân cụm dữ liệu nói riêng là một trong những trọng tâm nghiên cứu của khoa học máy tính. Phân cụm dữ liệu là một trong những biện pháp để tìm kiếm tri thức, khi ta chưa biết nhiều thông tin về miền ứng dụng. Phân cụm được coi như một công cụ độc lập để xem xét phân bố dữ liệu, là bước tiền xử lý cho các bước sau. Phân cụm dữ liệu hiện có nhiều ứng dụng trong hầu hết các lĩnh vực hoạt động kinh tế, xã hội.

Có nhiều phương pháp và thuật toán phân cụm dữ liệu khác nhau, tùy theo cách tiếp cận bài toán dưới góc độ nào. Một phương pháp mới được đề xuất tương đối gần đây là *Phương pháp lan truyền độ tương tự*.

Thuật toán lan truyền độ tương tự (*Affinity Propagation* - AP) là thuật toán phân cụm dữ liệu được đưa ra bởi Frey & Dueck vào năm 2007 dựa trên ý tưởng thuật toán lan truyền độ tin cậy trong suy diễn trên mạng xác xuất Bayes, dựa trên cơ sở toán học của lý thuyết xác suất. Thuật toán lan truyền làm việc dựa trên sự tương đồng (*affinity* nghĩa là sự giống nhau, sự tương thích, sự hấp dẫn) giữa các cặp điểm dữ liệu và đồng thời xem xét tất cả các điểm dữ liệu như các tâm cụm tiềm năng, theo thuật ngữ ở đây là tất cả các điểm dữ liệu đều là hình mẫu (exemplar) tiềm năng, và trao đổi các thông điệp giá trị thực cho đến khi có được tập hình mẫu tốt (phân cụm tương ứng).

Thuật toán phân cụm AP có một số ưu điểm: cho kết quả phân cụm tốt, đặc biệt là trong trường hợp có số lượng lớn các cụm, phát hiện cụm có hình dáng bất kỳ, không yêu cầu phải xác định trước số cụm. Nó cũng cho phép dễ dàng thực hiện phân cụm thỏa mãn một số điều kiện xác định trước nào đó, tức là phân cụm bán giám sát. Đặc tính này thích hợp cho phân cụm dữ liệu trong GIS vì những ràng buộc điều kiện địa hình tự nhiên hoặc quản lý hành chính cần tính đến trong các bài toán thực tế.

Luận văn chọn đề tài "***Phương pháp lan truyền độ tương tự trong phân cụm dữ liệu và ứng dụng***" là hướng nghiên cứu chính, với mục tiêu khám phá những điểm mạnh, điểm yếu của phương pháp này, hiểu biết sâu thêm về một cách tiếp cận phân cụm, đồng thời nâng cao kỹ năng thực hành triển khai ứng dụng.

2. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu là các phương pháp phân cụm dữ liệu, tập trung vào thuật toán lan truyền độ tương tự.

Phạm vi nghiên cứu là các điểm mạnh, điểm yếu và tiềm năng ứng dụng phương pháp lan truyền độ tương tự trong phân cụm dữ liệu địa lý.

3. Hướng nghiên cứu của đề tài

Luận văn dự kiến hướng nghiên cứu là:

Nghiên cứu lý thuyết: tìm hiểu sâu hơn về thuật toán lan truyền độ tương tự, trên cơ sở nắm vững bản chất của phương pháp lan truyền độ tin cậy trong suy diễn trên mạng xác xuất Bayes. dựa trên cơ sở toán học của lý thuyết xác suất.

Nghiên cứu ứng dụng: Cài đặt thử nghiệm thuật toán với dữ liệu mô phỏng để đánh giá, phân tích đánh giá kết quả; thử với dữ liệu thực tế.

4. Những nội dung nghiên cứu chính

Nội dung nghiên cứu của luận văn bao gồm:

Tìm hiểu tổng quan về phân cụm dữ liệu; các điểm đặc thù của bài toán phân cụm dữ liệu địa lý; Một số đặc điểm của thuật toán lan truyền độ tương tự, trên cơ sở lý thuyết toán học hoặc phân tích thực nghiệm.

Về thực hành: Cài đặt thử nghiệm thuật toán với dữ liệu mô phỏng để đánh giá, phân tích đánh giá kết quả; thử với dữ liệu thực tế.

Làm quen với hệ thống thông tin địa lý nguồn mở; cơ sở dữ liệu địa lý; cách viết plugin tích hợp phép phân tích dữ liệu địa lý.

5. Phương pháp nghiên cứu

Phương pháp nghiên cứu lý thuyết: tổng quan, phân tích các kết quả nghiên cứu đã có, nhận biết các ưu nhược điểm, lựa chọn cách tiếp cận phù hợp nhất để giải quyết bài toán ứng dụng.

Nghiên cứu thực nghiệm qua phân tích kết quả thử nghiệm với dữ liệu mô phỏng; dữ liệu thực tế; so sánh đánh giá và kết luận.

6. Ý nghĩa khoa học của đề tài

Đề tài nghiên cứu có ý nghĩa khoa học, góp phần làm hiểu biết sâu sắc hơn phương pháp lan truyền độ tương tự trong phân cụm dữ liệu.

Ứng dụng thực tế: phân cụm dữ liệu môi trường, không khí, thời tiết… nhận được từ các trạm quan trắc khí tượng để xác định những tiểu vùng môi trường khí tượng trong một địa phương, khu vực.

7. Bố cục của luận văn

Luận văn bao gồm 3 chương cùng với phần Mở đầu, phần Kết luận, phần Mục lục, phần Tài liệu tham khảo.

Chương 1: Tổng quan về phân cụm dữ liệu GIS và phân cụm dữ liệu địa lý: Trình bày các khái niệm cơ bản, các cách tiếp cận, phương pháp, thuật toán;