BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THÀNH PHỐ HỒ CHÍ MINH

NGUYỄN THỦY ĐOAN TRANG

PHÂN CỤM DỮ LIỆU KHÔNG GIAN ĐỊA LÝ TRONG KHÔNG GIAN MẠNG

TÓM TẮT LUẬN ÁN TIẾN SĨ

THÀNH PHỐ HỒ CHÍ MINH, 2025

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THÀNH PHỐ HỒ CHÍ MINH

NGUYỄN THỦY ĐOAN TRANG

PHÂN CỤM DỮ LIỆU KHÔNG GIAN ĐỊA LÝ TRONG KHÔNG GIAN MẠNG

TÓM TẮT LUẬN ÁN TIẾN SĨ

Chuyên ngành: **Khoa học máy tính**

Mã số: 9480101

Phản biện độc lập 1:

Phản biện độc lập 2:

Phản biện 1:

Phản biện 2:

Phản biện 3:

NGƯỜI HƯỚNG DẪN

1. PGS.TS Nguyễn Thị Thúy Loan
2. TS Lê Nhật Duy

THÀNH PHỐ HỒ CHÍ MINH, 2025

DANH MỤC CÔNG TRÌNH ĐÃ CÔNG BỐ

I. Tạp chí quốc tế

[CT.4] **T. T. D. Nguyen**, L. T. T. Nguyen, Q.-T. Bui, L. N. Duy, W. Pedrycz and B. Vo, "Efficient strategies for spatial data clustering using topological relations," *Applied Intelligence*, vol. 55, no. 2, p. 203, 2025, doi: 10.1007/s10489-024-05927-8 (Q2, IF 3.4).

[CT.3] **T. T. D. Nguyen,** L. T. T. Nguyen, Q.-T. Bui, L. N. Duy, and B. Vo, "NS-IDBSCAN: An Efficient Incremental Clustering Method for Geospatial Data in Network Space," *Information Sciences*, vol. 690, p. 121526, 2025, doi: 10.1016/j.ins.2024.121526, (Q1, IF 8.1).

[CT.2] **T. T. D. Nguyen**, L. T. T. Nguyen, Q.-T. Bui, U. Yun, and B. Vo, "An efficient topological-based clustering method on spatial data in network space," *Expert Systems with Applications*, vol. 215, p. 119395, 2023, doi: 10.1016/j.eswa.2022.119395 (Q1, IF 8.5).

[CT.1] **T. T. D. Nguyen**, L. T. T. Nguyen, A. Nguyen, U. Yun, and B. Vo, "A method for efficient clustering of spatial data in network space," *Journal of Intelligent & Fuzzy Systems*, vol. 40, pp. 11653–11670, 2021, doi: 10.1016/j.eswa.2022.119395 (Q2, IF 1.737).

II. Kỷ yếu hội nghị

T. T. D. Nguyen, L. T. T. Nguyen, L.N. Duy, "A efficiently method determination the reachable set of geospatial data in network space", *International Conference on Computational Intelligence and Innovative Applications (CIIA2022)*, Industrial University of Ho Chi Minh ity, 2022, p.117 – 186.

III. Dự án nghiên cứu khoa học

Võ Đình Bảy (Chủ trì), Bùi Quang Thịnh, Phạm Thế Anh Phú, **Nguyễn Thủy Đoan Trang**, Nguyễn Ánh, "Phân tích dữ liệu tô-pô với bài toán phân cụm dữ liệu", *Quỹ Phát triển Khoa học và Công nghệ Quốc gia Việt Nam (NAFOSTED)*, số tài trợ 102.05-2021.08 (ID: 7977), Nghiên cứu cơ bản trong Khoa học tự nhiên và Kỹ thuật, 2021, Thành viên nghiên cứu chủ chốt.

PHẦN MỞ ĐẦU

1. Tính cấp thiết

Phân cụm (clustering) là một kỹ thuật khai thác dữ liệu cơ bản với nhiều ứng dụng trong thế giới thực, trong đó có phân cụm dữ liệu không gian địa lý. Tuy nhiên các phương pháp phân cụm không gian hiện có chủ yếu được sử dụng trong không gian Euclid, hiếm khi được áp dụng cho các đối tượng có ràng buộc mặc dù có nhiều hiện tượng trong thế giới thực bị hạn chế bởi không gian mạng (network space) như nơi cư trú của học sinh, địa chỉ khách hàng dọc theo đường phố, vị trí thường xảy ra tai nạn trên đường, động đất dọc theo bờ biển, v.v... và kết quả của các phương pháp này thay đổi đáng kể khi thay đổi khoảng cách từ không gian Euclid sang không gian mạng.

Về mặt lý thuyết, nếu lấy khoảng cách đường đi ngắn nhất làm thước đo khoảng cách thì hầu hết các thuật toán phân cụm trong không gian Euclid có thể mở rộng được sang không gian mạng. Tuy nhiên, trên thực tế, sự dịch chuyển không đơn giản như vậy vì độ phức tạp cao của bài toán tìm độ dài đường đi ngắn nhất.

Do vậy, nghiên cứu và đề xuất các phương pháp hiệu quả nhằm giải quyết vấn đề phân cụm dữ liệu không gian địa lý trong không gian mạng là yêu cầu cấp thiết trong phân cụm dữ liệu.

2. Mục tiêu nghiên cứu

Mục tiêu của luận án là khảo sát các phương pháp phân cụm liên quan, từ đó kế thừa, cải tiến, đề xuất phương pháp phân cụm hiệu quả cho dữ liệu không gian địa lý trong không gian mạng, gồm:

1) Đề xuất thuật toán iNS-DBSCAN nhằm cải thiện thời gian thực thi cho thuật toán phân cụm dữ liệu không gian địa lý trong không gian mạng NS-DBSCAN (Wang và cộng sự, 2019).
2) Đề xuất phương pháp phân cụm dữ liệu không gian địa lý trong không gian mạng dựa trên quan hệ tô-pô NS-TBC nhằm giảm số lượng tham số đầu vào.
3) Đề xuất thuật toán phân cụm gia tăng NS-IDBSCAN để phân cụm hiệu quả dữ liệu không gian địa lý trong không gian mạng có phát sinh dữ liệu mới. Phương pháp này nhằm thay cho giải pháp sử dụng các thuật toán có sẵn thì phải thực hiện phân cụm lại từ đầu trên cả tập dữ liệu cũ và mới mỗi khi có dữ liệu mới phát sinh giúp tăng tốc thời gian thực hiện.

3. Đối tượng và phạm vi nghiên cứu

- Đối tượng nghiên cứu: Kỹ thuật phân cụm dữ liệu không gian trên dữ liệu điểm và dữ liệu về đường đi.
- Phạm vi nghiên cứu: Dữ liệu không gian tỉnh thành ở Việt Nam được tải về từ OpenStreetMap (OSM), ESRI Open Data và Inside Airbnb.

4. Đóng góp chính của luận án

Luận án đề xuất các thuật toán giải quyết hiệu quả vấn đề phân cụm cho dữ liệu không gian địa lý trong không gian mạng, góp phần bổ sung vào nền tảng lý thuyết về phân cụm dữ liệu nói riêng và khoa học máy tính nói chung. Luận án có ba đóng góp chính thông qua ba thuật toán được đề xuất: **iNS-DBSCAN, NS-TBC** và **NS-IDBSCAN**.

Thứ nhất, đề xuất thuật toán **iNS-DBSCAN**, đã được công bố ở công trình **[CT.1]**, phân cụm hiệu quả cho dữ liệu không gian địa lý từ việc đưa ra mệnh đề giới hạn không gian xử lý giúp tăng tốc thời gian thực thi cho thuật toán NS-DBSCAN.

Thứ hai, đề xuất phương pháp **NS-TBC,** đã được công bố ở công trình **[CT.2]**, sử dụng quan hệ tô-pô để phân cụm dữ liệu nhằm giảm sự phụ thuộc vào tham số đầu vào, giúp cải thiện chất lượng nhóm, đồng thời giảm thời gian xử lý cho thuật toán phân cụm dữ liệu không gian có ràng buộc mạng.

Thứ ba, luận án đề xuất phương pháp **NS-IDBSCAN,** đã được công bố ở công trình **[CT.3]**, phân cụm gia tăng cho dữ liệu không gian địa lý trong không gian mạng có phát sinh dữ liệu mới. Đóng góp này giúp giải quyết vấn đề phải phân cụm lại từ đầu mỗi khi có dữ liệu mới được thêm, giúp tăng tốc thời gian xử lý.

Ngoài ra, luận án còn đề xuất các chiến lược nâng cao hiệu quả cho phương pháp phân cụm dựa trên tô-pô được công bố mới nhất tại thời điểm nghiên cứu (Alomari và cộng sự, 2023) qua **[CT.4]** (theo hiểu biết của NCS).

Các đóng góp này được trình bày chi tiết ở các Chương 3 ([CT.1]), Chương 4 ([CT.2] và [CT.4]) và Chương 5 ([CT.3]) trong báo cáo toàn văn của luận án.

Như vậy, luận án đã đóng góp ba thuật toán và các chiến lược phân cụm hiệu quả cho hướng nghiên cứu phân cụm dữ liệu không gian địa lý có ràng buộc mạng. Ngoài ra, các phương pháp đề xuất góp phần vào việc hình thành lớp phương pháp phân cụm mới bên cạnh các lớp phương pháp phân cụm đã có, đó là phương pháp phân cụm dựa trên tô-pô (Topological-Based Clustering).

5. Ý nghĩa khoa học và thực tiễn

- **Ý nghĩa khoa học:**

  Nội dung của luận án là nghiên cứu và đề xuất các phương pháp để giải quyết hiệu quả bài toán phân cụm cho dữ liệu không gian địa lý có ràng buộc mạng. Các đóng góp của luận án góp phần bổ sung nền tảng lý thuyết về phân cụm dữ liệu nói riêng và khoa học máy tính nói chung.

- **Ý nghĩa thực tiễn:**

  + Luận án tập trung giải quyết hạn chế trong phân cụm dữ liệu không gian địa lý là chưa quan tâm đến ràng buộc theo mạng dẫn đến khó ứng dụng trong thực tiễn hoặc ứng dụng chưa hiệu quả. Cách tiếp cận trong không gian mạng với độ đo theo khoảng cách đường đi ngắn nhất giúp phân cụm cho các vấn đề trong thế giới thực mà trong đó khoảng cách theo đường đi được ưu tiên hơn khoảng cách Euclid thông thường. Phân cụm gia tăng cho dữ liệu không gian địa lý trong không gian mạng có phát sinh dữ liệu mới giúp tăng tốc thời gian xử lý.
  + Phân cụm gia tăng dữ liệu không gian trong không gian mạng có rất nhiều ứng dụng, đặc biệt trong các tình huống khẩn cấp như hỗ trợ cứu hộ và phục hồi hoạt động trong hoặc sau thảm họa, dịch bệnh, v.v…
  + Việc khai thác "ngách" không gian mạng (network space) cho dữ liệu không gian địa lý giúp có thể giải quyết hiệu quả bài toán phân tích không gian về chọn vị trí "gần" theo khoảng cách không gian thay vì theo địa giới hành chính. Chọn được vị trí thật sự "gần nhất" sẽ góp phần giải quyết nhiều vấn đề về an sinh xã hội như tiết kiệm, giảm tình trạng kẹt xe, giảm ô nhiễm không khí.
  + Ngoài mạng lưới đường phố, phương pháp đề xuất cũng có thể được áp dụng cho các loại mạng lưới khác như đường ống, đường dây điện. Các ứng dụng tiềm năng khác trên phân cụm dữ liệu không gian địa lý trong không gian mạng gồm phát hiện điểm nóng tai nạn giao thông, tội phạm trên đường phố, ứng phó khẩn cấp, khái quát hóa bản đồ, báo động rò rỉ trong đường ống dẫn dầu, xác định các phần mạng có tỷ lệ lỗi cao, và nhiều ứng dụng khác.

6. Cấu trúc của luận án

Luận án bao gồm phần mở đầu, năm chương nội dung chính, chương kết luận và hướng phát triển, tài liệu tham khảo.

- **Mở đầu:** Trình bày tính cấp thiết, mục tiêu nghiên cứu, đối tượng và phạm vi, các đóng góp chính, ý nghĩa khoa học và thực tiễn của luận án.
- **Chương 1 - Tổng quan về lĩnh vực nghiên cứu:** Giới thiệu tổng quan về lĩnh vực nghiên cứu, bao gồm các nghiên cứu liên quan, những vấn đề còn tồn tại và đề xuất định hướng nghiên cứu.
- **Chương 2 - Cơ sở lý thuyết:** Trình bày kiến thức nền tảng liên quan đến các phương pháp đề xuất.
- **Chương 3 - Phân cụm dữ liệu không gian trong không gian mạng:** Đưa ra mệnh đề nhằm giới hạn không gian duyệt cho thuật toán NS-DBSCAN. Từ đó đề xuất phương pháp **iNS-DBSCAN** để phân cụm hiệu quả dữ liệu không gian địa lý trong không gian mạng giúp tăng tốc thời gian xử lý.
- **Chương 4 - Phân cụm dữ liệu không gian địa lý trong không gian mạng dựa trên tô-pô:** Trình bày phương pháp **NS-TBC** sử dụng quan hệ tô-pô để phân cụm dữ liệu không gian địa lý trong không gian mạng giúp giảm sự phụ thuộc vào tham số đầu vào cho thuật toán phân cụm **iNS-DBSCAN** đồng thời tăng tốc thời gian xử lý và cải thiện chất lượng nhóm.
- **Chương 5 - Phân cụm gia tăng dữ liệu không gian địa lý trong không gian mạng:** Đề xuất phương pháp **NS-IDBSCAN** phân cụm gia tăng cho dữ liệu có phát sinh mới. Giải pháp đề xuất sử dụng bảng băm để tổ chức dữ liệu cũ đã được phân cụm giúp truy cập trực tiếp nên loại bỏ được thao tác tìm kiếm nhằm giảm thời gian xử lý và cải thiện chất lượng nhóm.
- **Chương 6 - Kết luận và hướng phát triển:** Tổng kết các kết quả nghiên cứu đạt được và đề xuất các hướng nghiên cứu trong tương lai.

CHƯƠNG 1 TỔNG QUAN VỀ LĨNH VỰC NGHIÊN CỨU

1.1. Phân cụm dữ liệu không gian địa lý

Luật địa lý đầu tiên của Tobler khẳng định rằng các thực thể không gian gần nhau có mối liên hệ mật thiết hơn so với những thực thể cách xa. Do đó, phân cụm dữ liệu không gian địa lý đóng vai trò quan trọng trong việc khám phá các mối quan hệ địa lý và sự phụ thuộc lẫn nhau cũng như các đặc trưng có thể tồn tại ngầm trong cơ sở dữ liệu không gian. Nhiều thuật toán phân cụm không gian hiệu quả đã được phát triển và không ngừng cải tiến. Nhiều tài liệu nghiên cứu đã đề xuất các kỹ thuật phân cụm không gian khác nhau, đặc biệt đa số được phát triển từ kỹ thuật phân cụm dựa trên mật độ (density-based) vì các ưu điểm của nó trong phân cụm dữ liệu không gian và thuật toán đầu tiên của lớp phương pháp này được thiết kế để phân cụm cho dữ liệu không gian "Density Based Spatial Clustering of Applications with Noise".

Kỹ thuật phân cụm dựa trên mật độ nổi tiếng cho dữ liệu không gian đã được sử dụng rộng rãi, đặc biệt là để xử lý các tập dữ liệu có nhiễu và có thể phát hiện các nhóm có hình dạng bất kỳ. DBSCAN (Density Based Spatial Clustering of Applications with Noise) là thuật toán dựa trên mật độ đầu tiên được đề xuất vào năm 1996 bởi Ester và cộng sự với nhiều ứng dụng trong thế giới thực và được thảo luận trong nhiều công trình. Bên cạnh ưu điểm là không cần biết trước số lượng nhóm, DBSCAN có thể xác định các nhóm có hình dạng bất kỳ và tập dữ liệu có nhiễu. DBSCAN là thuật toán nổi tiếng có tốc độ tính toán nhanh đã được chứng minh trên lý thuyết và thực tế. Vào năm 2014, nó đã nhận được giải thưởng vượt thời gian tại hội nghị hàng đầu về khai thác dữ liệu, ACM SIGKDD. Nhờ vào tầm quan trọng và sự mạnh mẽ của thuật toán DBSCAN, nó đã được nghiên cứu kỹ để phát triển các phiên bản tốt hơn.

Trong hai thập niên qua, lấy cảm hứng từ ý tưởng DBSCAN, nhiều nghiên cứu khác nhau đã được thực hiện để làm phong phú thêm lớp phương pháp phân cụm dựa trên mật độ này, như được tóm tắt bởi các đánh giá gần đây. Năm 2021, khảo sát của Bhattacharjee và cộng sự đã hệ thống hóa và phân loại 32 thuật toán phân cụm dựa trên mật độ. Thuật toán DBSCAN được kế thừa và phát triển bởi nhiều thuật toán để phân cụm cho nhiều loại dữ liệu: không gian (spatial), phi không gian (non-Spatial), hình ảnh đa phương tiện hoặc các loại dữ liệu khác. Riêng với dữ liệu không gian, có nhiều thuật toán đã được nghiên cứu và cải tiến.

Kỹ thuật phân cụm dữ liệu không gian được ứng dụng trong các lĩnh vực như Khoa học Trái đất, Thiên văn học, Địa lý, Đa phương tiện và nhiều lĩnh vực nghiên cứu khác. Các ứng dụng có thể được tìm thấy trong việc xác định phân bố mật độ, phát hiện điểm nóng, khám phá xu hướng và hậu cần, v.v… như phân tích tội phạm, dịch tễ học, địa lý kinh tế, phân tích bán lẻ, phân tích sự cố giao thông và nhân khẩu học. Phân cụm dữ liệu không gian địa lý đóng vai trò quan trọng trong việc định lượng các mẫu biến động địa lý như giám sát dịch bệnh, dịch tễ học không gian, phân tích tội phạm, v.v… Hiện nay, các kỹ thuật phân cụm dữ liệu không gian địa lý được quan tâm nhiều trong nghiên cứu y tế công cộng, dịch tễ học như phân cụm dịch bệnh, trong kinh doanh như tổ chức lại công ty nhằm tăng hiệu quả hoạt động và giảm chi phí do hậu quả của khủng hoảng Covid-19 hay trong các bài toán liên quan đến định tuyến vị trí.

1.2. Phân cụm dữ liệu không gian địa lý trong không gian mạng

Trong ứng dụng dữ liệu không gian địa lý cho việc tuyển sinh đầu cấp không theo địa giới hành chính ở Thành phố Hồ Chí Minh để *"học sinh sẽ được học tại trường ở gần nhà, thay vì phân theo địa giới hành chính của phường như trước đây"* [1] nhằm mục đích chọn trường học gần nơi cư trú của học sinh. Tuy nhiên, để có thể chọn được trường học thực sự gần nhà của học sinh thì nên sử dụng khoảng cách đường đi ngắn nhất thay vì khoảng cách Euclid. Điều này đã đặt ra yêu cầu cần giải quyết bài toán trong không gian mạng (network space) cho dữ liệu không gian địa lý.

Cùng vấn đề này được đặt ra cho bài toán phân cụm. Các thuật toán phân cụm không gian hiện có chủ yếu tập trung vào các điểm phân bố trong không gian Euclid với phép đo theo khoảng cách Euclid, còn các thuật toán trong không gian mạng chưa được nghiên cứu nhiều. Điều này đã hạn chế việc áp dụng các thuật toán phân cụm này vào các bài toán trong thế giới thực có ràng buộc mạng như ràng buộc về khoảng cách đường đi. Trong nhiều ứng dụng thực tế, khả năng tiếp cận các đối tượng không gian bị hạn chế bởi các mạng không gian. Trong không gian đô thị, con người không thể di chuyển theo đường chim bay mà bị giới hạn bởi mạng lưới đường giao thông (gọi là không gian mạng), chẳng hạn như mạng lưới đường và lối đi trong thành phố. Vì vậy, khoảng cách thực tế giữa các đối

[1] https://vtv.vn/xa-hoi/quan-dau-tien-cua-tp-ho-chi-minh-tuyen-sinh-dau-cap-khong-theo-dia-gioi-hanh-chinh-20230417195604477.htm#:~:text=VTV.vn%20%2D%20Qu%E1%BA%ADn%208%20l%C3%A0,theo%20%C4%91%E1%BB%8Ba%20gi%E1%BB%9Bi%20h%C3%A0nh%20ch%C3%ADnh

tượng không thể tính theo khoảng cách Euclid mà phải tính theo khoảng cách đường đi ngắn nhất cho dữ liệu có ràng buộc mạng.

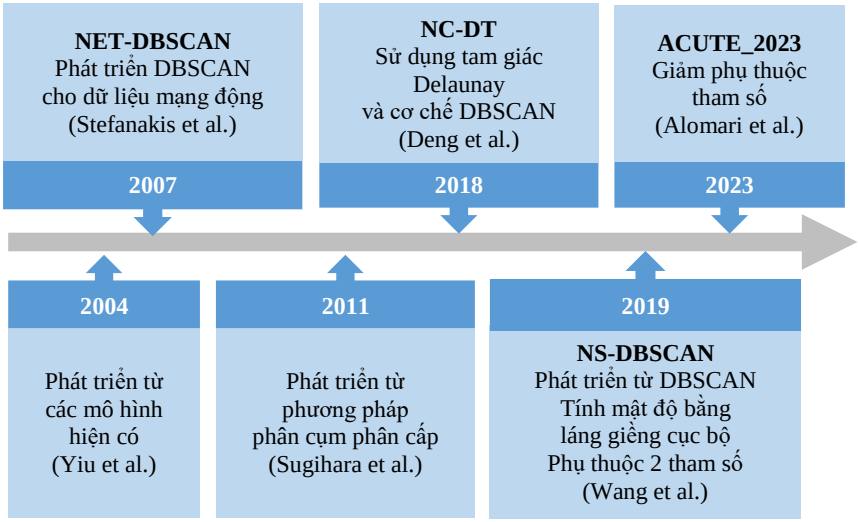
Như vậy nhu cầu đặt ra là cần mở rộng thuật toán phân cụm cho các đối tượng trong không gian mạng, nơi các đối tượng có thể có các yêu cầu ràng buộc về mạng, chứ không chỉ trong không gian Euclid với các phép đo khoảng cách Euclid. Do đó, một số thuật toán phân cụm cho dữ liệu điểm có ràng buộc mạng đã được đề xuất.

Yiu và cộng sự đã đề xuất nghiên cứu đầu tiên về phân cụm với khoảng cách theo mạng vào năm 2004 theo ý tưởng là dựa trên các mô hình phân cụm hiện có và duyệt mạng bằng thuật toán Dijkstra để xác định khoảng cách ngắn nhất giữa các điểm vốn cần nhiều thời gian. Vào năm 2018, một phương pháp phân cụm trong không gian mạng đã được đề xuất bằng cách sử dụng NC_DT (Phương pháp tam giác Delaunay bị ràng buộc bởi mạng) để tính khoảng cách theo mạng giữa các điểm. Phương pháp này cũng mất nhiều thời gian và tài nguyên vì phải xác định tam giác Delaunay. Năm 2019, dựa trên thuật toán DBSCAN, Wang và cộng sự đã phát triển thuật toán NS-DBSCAN để thực hiện phân cụm dữ liệu không gian trong không gian mạng bằng giải pháp tìm láng giềng cục bộ để xác định mật độ, không cần phải tính ma trận khoảng cách, giảm thời gian tính toán. Gần đây nhất, năm 2023, Alomari và cộng sự đã đề cập đến vấn đề phân cụm trong không gian mạng, tuy nhiên cần phải tính ma trận khoảng cách đường đi ngắn nhất giữa các điểm nên mất nhiều thời gian. Một số thuật toán phân cụm dữ liệu không gian trong không gian mạng được trình bày trong Hình 1.1.

Như vậy, các phương pháp phân cụm dữ liệu không gian địa lý trong không gian mạng hiện có gồm hai loại, một loại là phải xác định ma trận khoảng cách giữa các điểm dữ liệu, mà đối với dữ liệu không gian địa lý có ràng buộc mạng là xác định độ dài đường đi ngắn nhất giữa chúng. Thao tác này vốn có độ phức tạp tính toán cao. Loại thứ hai là không tính ma trận khoảng cách này để giảm đáng kể chi phí tính toán.

Câu hỏi đặt ra là: Không tính ma trận khoảng cách, nghĩa là không có khoảng cách giữa các điểm, làm cách nào để thực hiện phân cụm? Câu trả lời là khai thác các đặc điểm ràng buộc mạng của dữ liệu không gian địa lý. Không cần tính khoảng cách giữa tất cả các điểm của toàn bộ tập dữ liệu mà chỉ cần tính khoảng cách giữa các điểm có đường đi (trong không gian mạng). Đối với phương pháp phân cụm dựa trên mật độ có ngưỡng bán kính eps, chỉ cần tính khoảng cách đường đi đến các điểm trong giới hạn bán kính eps, tức là khoảng cách đến các

láng giềng cục bộ. Do đó, không cần phải tính ma trận khoảng cách đường đi ngắn nhất giữa tất cả các điểm trong tập dữ liệu như cách làm thông thường để tránh lãng phí thời gian.



Hình 1.1. Phân cụm dữ liệu không gian trong không gian mạng.

Hướng tiếp cận của luận án là nghiên cứu ưu và nhược điểm của thuật toán thuộc loại thứ hai để phát triển phương pháp hiệu quả cho bài toán phân cụm dữ liệu không gian địa lý trong không gian mạng. Mà thuật toán hiệu quả nhất đến tại thời điểm nghiên cứu (theo hiểu biết của NCS) thuộc loại này là công trình do Wang và cộng sự đề xuất năm 2019 là NS-DBSCAN.

Để cải thiện hiệu suất phân cụm của thuật toán NS-DBSCAN, vào năm 2021, công trình [CT.1] đề xuất phương pháp phân cụm dữ liệu không gian hiệu quả trong không gian mạng nhằm tăng tốc thời gian thực hiện cho NS-DBSCAN (gọi là iNS-DBSCAN). ***Đây là đóng góp đầu tiên của luận án này, được trình bày chi tiết trong Chương 3 của báo cáo toàn văn luận án.***

Thuật toán iNS-DBSCAN giảm đáng kể thời gian xử lý của NS-DBSCAN nhưng kết quả phân cụm vẫn còn phụ thuộc vào hai tham số đầu vào là ngưỡng bán kính (eps) và ngưỡng mật độ ($MinPts$). Vấn đề tìm ngưỡng mật độ phù hợp vốn là một vấn đề khó khăn, và hầu hết các thuật toán phân cụm hiện tại trở nên không hiệu quả khi được cung cấp các tham số không phù hợp. Hầu như tất cả các thuật toán phân cụm nổi tiếng đều yêu cầu các tham số đầu vào khó xác định

nhưng có ảnh hưởng đáng kể đến kết quả phân cụm. Do đó, tìm giải pháp nhằm giảm sự ảnh hưởng của tham số đầu vào đến kết quả phân cụm là mục tiêu của Chương 4.

Năm 2016, dựa trên quan hệ tô-pô, thuật toán phân cụm ACUTE (gọi là ACUTE-2016) được đề xuất ngoài hai ưu điểm là có thể phát hiện các nhóm có hình dạng bất kỳ và xác định nhiễu thì ưu điểm đặc biệt là giảm số lượng tham số đầu vào. Lấy cảm hứng từ ACUTE_2016, công trình [CT.2] đề xuất thuật toán NS-TBC để phân cụm các đối tượng không gian địa lý có ràng buộc mạng dựa trên quan hệ tô-pô nhằm giảm tham số đầu vào cho thuật toán iNS-DBSCAN. ***Đây là đóng góp thứ hai của luận án, được trình bày chi tiết trong Chương 4 của báo cáo toàn văn luận án.*** Ngoài ra, qua công trình [CT.4], chương này còn đề xuất các chiến lược cải tiến thuật toán phân cụm dữ liệu không gian sử dụng quan hệ tô-pô mới nhất (tại thời điểm nghiên cứu, theo hiểu biết của nghiên cứu sinh) ở công trình được đề xuất năm 2023 (ACUTE_2023) (Alomari và cộng sự, 2023, là phiên bản cải tiến của ACUTE_2016).

Phần lớn các kỹ thuật phân cụm thực hiện trên toàn bộ tập dữ liệu hiện có. Khi dữ liệu mới được thêm vào, việc sử dụng các phương pháp hiện có để phân cụm sẽ rất tốn thời gian vì phải xử lý trên toàn bộ tập dữ liệu gồm phần dữ liệu cũ và dữ liệu mới được thêm vào. Mặc dù có nhiều thuật toán phân cụm gia tăng đã được phát triển nhưng cho đến nay vẫn chưa có thuật toán nào được thiết kế cho dữ liệu không gian địa lý trong không gian mạng. ***Vì vậy, đóng góp thứ ba của luận án là đề xuất phương pháp phân cụm gia tăng cho dữ liệu không gian địa lý trong không gian mạng, được trình bày chi tiết ở Chương 5 báo cáo toàn văn luận án.***

1.3. Các vấn đề tồn tại và đề xuất định hướng nghiên cứu

Như vậy, các phương pháp phân cụm không gian hiện có chủ yếu tập trung vào dữ liệu phân bố trong không gian Euclid với phép đo theo khoảng cách Euclid mà trong thực tế các đối tượng có thể bị ràng buộc theo không gian mạng. Hơn nữa, mặc dù nhiều thuật toán phân cụm gia tăng đã được phát triển, nhưng các thuật toán đó không dành cho dữ liệu không gian địa lý có ràng buộc mạng. Và việc tìm giải pháp cải tiến luôn là vấn đề đặt ra cho các thuật toán. Vì vậy, các nội dung chính mà luận án tập trung nghiên cứu gồm:

- Nghiên cứu thuật toán phân cụm dữ liệu trong không gian mạng được công bố và đề xuất giải pháp cải thiện thời gian thực hiện: Chương 3.

- Sử dụng quan hệ tô-pô để tiếp tục cải thiện thời gian thực thi đồng thời giảm sự phụ thuộc của quá trình phân cụm vào tham số đầu vào: Chương 4.
- Đề xuất thuật toán phân cụm gia tăng cho dữ liệu không gian địa lý trong không gian mạng: Chương 5.

CHƯƠNG 2 PHÂN CỤM DỮ LIỆU KHÔNG GIAN TRONG KHÔNG GIAN MẠNG

2.1. Một số khái niệm

Khái niệm 2.1. ***Bài toán phân cụm*** dữ liệu nhằm xác định các nhóm trong dữ liệu, trong đó một nhóm là tập hợp các đối tượng được nhóm lại dựa trên nguyên tắc tối đa hóa sự tương đồng trong nhóm và giảm thiểu sự tương đồng giữa các nhóm.

Khái niệm 2.2. ***Dữ liệu không gian (spatial data)****,* còn được gọi là *dữ liệu không gian địa lý (geospatial data)* là dữ liệu về các đối tượng, sự kiện hoặc hiện tượng *có vị trí* trên bề mặt trái đất[2].

Khái niệm 2.3. ***Dữ liệu không gian địa lý trong không gian mạng*** là dữ liệu không gian địa lý bị ràng buộc theo mạng như các tòa nhà trên mạng đường giao thông hay mạng đường dây điện, đường ống, mạng viễn thông, v.v…

2.2. Một số định nghĩa

2.2.1. *Định nghĩa liên quan đến phân cụm dựa trên mật độ*

Định nghĩa 2.1. ***Láng giềng*** **eps** ***của điểm*** $\boldsymbol{p}$ (**eps** $-$ **neighbors**) là tập hợp các điểm trong khoảng cách $\boldsymbol{eps}$ từ điểm $\boldsymbol{p}$. Kí hiệu: $\boldsymbol{N\_eps(p)}$.

Định nghĩa 2.2. ***(Mật độ của điểm*** **p)** Số đỉnh thuộc tập hợp các điểm láng giềng $\boldsymbol{eps}$ của $\boldsymbol{p}$, $|\boldsymbol{N\_eps}\ (\boldsymbol{p})|$, được định nghĩa là mật độ của điểm $\boldsymbol{p}$. Ký hiệu: $\boldsymbol{Density(p)}$.

Định nghĩa 2.3. ***(Điểm lõi: core points)*** Điểm lõi là điểm có mật độ lớn hơn mật độ tối thiểu $\boldsymbol{MinPts}$. Láng giềng $\boldsymbol{eps}$ của điểm lõi được gọi là điểm biên *(border points).*

Các điểm có mật độ thấp mà không phải là điểm biên được gọi là ***nhiễu – noise*** (hoặc ngoại lệ - outlier).

Định nghĩa 2.4. ***Tiếp cận mật độ trực tiếp (directly density-reachable)*****:** Điểm $\boldsymbol{p}$ tiếp cận mật độ trực tiếp từ điểm $\boldsymbol{q}$ (với ngưỡng bán kính là $\boldsymbol{eps}$ và ngưỡng mật độ là $\boldsymbol{MinPts}$) nếu $\boldsymbol{p}$ là láng giềng $\boldsymbol{eps}$ của $\boldsymbol{q}$ và $\boldsymbol{q}$ có mật độ tối thiểu là $\boldsymbol{MinPts}$:

1) $p \in N\_eps(q)$
2) $|N\_eps(q)| \geq MinPts$

[2] www.sciencedirect.com/topics/computer-science/geospatial-data

Định nghĩa 2.5. ***Tiếp cận mật độ (density-reachable)*** (Hình 2.6): Điểm $\boldsymbol{p_n}$ tiếp cận mật độ từ điểm $\boldsymbol{p_1}$ nếu $\boldsymbol{p_i}$ tiếp cận mật độ trực tiếp từ $\boldsymbol{p_{i-1}}$.

Định nghĩa 2.6. ***Kết nối mật độ (density-connected)***: Nếu hai điểm p và q đều tiếp cận mật độ từ điểm c thì ta gọi p và q là kết nối mật độ với nhau.

Định nghĩa 2.7. ***Cụm (nhóm: Cluster)****:* Một cụm là tập hợp khác rỗng của tất cả các điểm lõi tiếp cận mật độ với nhau và tất cả các điểm là láng giềng của ít nhất một trong các điểm lõi ấy.

2.2.2. Định nghĩa liên quan đến phân cụm gia tăng

Định nghĩa 2.8. ***Điểm lõi mới từ biên,*** ký hiệu $core_{newfb}$, là điểm trước khi thêm điểm x có mật độ bằng $MinPts - 1$ và đã được phân cụm (là biên), khi thêm x vào thì đổi vai trò thành lõi ($|N\_eps(\mathrm{core_{newfb}})| = MinPts$), tức là thỏa 2 điều kiện sau trước khi thêm x:

1) $|N\_eps(\mathrm{core_{newfb}})| = MinPts - 1$
2) Đã được phân cụm (là biên).

Định nghĩa 2.9. ***Điểm lõi mới từ nhiễu***, ký hiệu $core_{newfn}$, là điểm trước khi thêm điểm *x* có mật độ bằng $MinPts - 1$ và chưa được phân cụm (là nhiễu), khi thêm *x* vào thì đổi vai trò thành lõi ($|N\_eps(\mathrm{core_{newfb}})| = MinPts$), tức là thỏa 2 điều kiện sau trước khi thêm x:

1) $|N\_eps(\mathrm{core_{newfn}})| = MinPts - 1$
2) Chưa được phân cụm (là nhiễu).

Những điểm đã là lõi trước khi thêm x gọi là ***lõi cũ***, ký hiệu $core_{old}$.

Định lý 2.1. Hai loại điểm ***lõi mới từ nhiễu*** và ***lõi mới từ biên*** được hình thành khi thêm điểm x đều là láng giềng của x.

Định lý này đóng vai trò quan trọng trong việc ***giới hạn không gian*** của các điểm có khả năng làm thay đổi đáng kể kết quả phân cụm khi có phát sinh dữ liệu mới. Đó là các điểm có sự thay đổi vai trò từ không phải là lõi sang lõi.

2.2.3. Định nghĩa liên quan đến phân cụm dữ liệu không gian địa lý trong không gian mạng (network space)

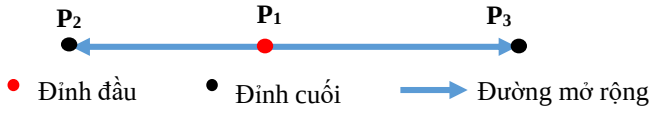
Cho đồ thị có trọng số $G = (V, E, W)$, trong đó V là tập các đỉnh, E là tập các cạnh và W là tập các giá trị trọng số. Cạnh giữa hai đỉnh P và Q được ký hiệu là (P, Q). Trọng số của cạnh (P, Q) được ký hiệu là $W(P, Q)$. Ta có định nghĩa sau:

Định nghĩa 2.10. **Độ dài đường đi** từ điểm $\mathrm{P_0}$ đến điểm $\mathrm{P_n}$ đi qua các điểm $\mathrm{P_1}, \mathrm{P_2}, \ldots, \mathrm{P_{n-1}} \in V$ có giá trị là tổng độ dài tất cả các đoạn của đường đi.

$$\mathrm{d(P_0, P_n)} = \mathrm{W(P_0, P_1)} + \mathrm{W(P_1, P_2)} + \cdots + \mathrm{W(P_{n-1}, P_n)}$$

Định nghĩa 2.11. **Đường đi ngắn nhất** từ điểm P_1 đến P_n là đường đi có tổng trọng số nhỏ nhất trong số tất cả các đường đi có thể có từ P_1 đến P_n trong đồ thị có trọng số $G = (V, E, W)$.

Định nghĩa 2.12. ***Sự mở rộng cơ bản (basic expansion)*** từ đỉnh bắt đầu P_1 đến đỉnh kết thúc P_2 là chuyển động từ P_1 đến P_2 theo cạnh (P_1, P_2). Khi đó, cạnh (P_1, P_2) được gọi là đường mở rộng (expansion path).



Hình 2.1. Mở rộng cơ bản từ đỉnh đầu P_1 đến các đỉnh cuối P_2 và P_3

Định nghĩa 2.13. ***Khoảng cách trong nhóm (Intra-cluster distance)*** là khoảng cách giữa hai đối tượng bất kỳ thuộc cùng một nhóm. Nó là thước đo sự giống nhau giữa các đối tượng trong một nhóm. Nói chung, khoảng cách giữa các nhóm thấp hơn cho thấy các đối tượng trong một nhóm giống nhau hơn.

Định nghĩa 2.14. ***Khoảng cách giữa các nhóm (Inter-cluster distance)*** là khoảng cách giữa hai đối tượng bất kỳ thuộc hai nhóm khác nhau. Nó là thước đo sự khác biệt giữa các nhóm khác nhau. Nói chung, khoảng cách giữa các nhóm cao hơn cho thấy các đối tượng thuộc các nhóm khác nhau có nhiều điểm khác nhau hơn.

Khoảng cách trong nhóm và khoảng cách giữa các nhóm có thể sử dụng để đánh giá chất lượng kết quả phân cụm và so sánh giữa các thuật toán phân cụm.

Hai loại khoảng cách này có thể đo bằng khoảng cách trung bình hoặc khoảng cách tối thiểu giữa các đối tượng. Tiêu chí của bài toán phân cụm là giảm khoảng cách trong nhóm và tăng khoảng cách giữa các nhóm.

Bài toán phân cụm dữ liệu không gian địa lý trong không gian mạng là xác định các nhóm "ẩn" gồm tập hợp các đối tượng dữ liệu "gần nhau" bên trong dữ liệu. Độ "gần nhau" đo theo hàm khoảng cách trong không gian mạng được định nghĩa như sau:

Định nghĩa 2.15. ***Hàm khoảng cách trong không gian mạng***

Cho tập hợp các đoạn đường $R(road) = \{r_i | i = 1..|R|\}$ và tập hợp các điểm $P\ (point) = \{p_i | i = 1..|P|\}$ trong không gian mạng. Khoảng cách giữa hai điểm P_1 và P_n là độ dài đường đi ngắn nhất giữa chúng. Kí hiệu: $dist(P_1, P_n)$

$$dist(P_1, P_n) = W(P_1, P_2) + W(P_2, P_3) + \cdots + W(P_{n-1}, P_n)$$

Trong đó, $W(P_i, P_{i+1})$ là trọng số của cạnh (P_i, P_{i+1}) trong đường đi ngắn nhất từ P_1 đến P_n.

2.2.4. Định nghĩa liên quan đến quan hệ tô-pô

Định nghĩa 2.16. ***Hai đối tượng không gian rời rạc nhau (disjoint)*** nếu các mối quan hệ tô-pô của một đối tượng giao nhau nhiều nhất với mặt ngoài của đối tượng còn lại.

$$\text{A disjoints B} \Leftrightarrow \begin{cases} A^\circ \cap B^\circ = \emptyset \\ A^\circ \cap \partial B = \emptyset \\ \partial A \cap B^\circ = \emptyset \\ \partial A \cap \partial B = \emptyset \end{cases}.$$

Định nghĩa 2.17. ***Hai đối tượng không gian gặp nhau (meet)*** nếu bên trong cả hai (interiors) không giao nhau, nhưng bên trong (interiors) hoặc ranh giới (boundary) của một đối tượng giao với ranh giới (boundary) của đối tượng còn lại.

$$\text{A meets B} \Leftrightarrow \begin{cases} A^\circ \cap B^\circ = \emptyset \\ A^\circ \cap \partial B \neq \emptyset \\ \partial A \cap B^\circ \neq \emptyset \\ \partial A \cap \partial B \neq \emptyset \end{cases}.$$

Định nghĩa 2.17. ***Hai đối tượng không gian chồng nhau (overlap)*** nếu bên trong (interiors) của mỗi đối tượng giao với cả bên trong và bên ngoài của đối tượng còn lại.

$$\text{A overlaps B} \Leftrightarrow \begin{cases} A^\circ \cap B^\circ \neq \emptyset \\ A^\circ \cap B^- \neq \emptyset \\ A^- \cap B^\circ \neq \emptyset \end{cases}.$$

2.3. Phương pháp phân cụm dựa trên mật độ (density-based clustering)

Phương pháp dựa trên mật độ phát hiện các vùng đông đúc bởi các vùng thưa thớt ngăn cách giữa chúng. Ý tưởng cốt lõi của phương pháp phân cụm này là sự kết nối giữa các điểm lõi (core), đó là những điểm có mật độ cao (đạt giá trị ngưỡng nào đó do người dùng chỉ định), các điểm lõi này sẽ kết nối với nhau để hình thành nhóm. Nhờ vào kỹ thuật phân cụm này, DBSCAN có thể phát hiện các nhóm có hình dạng và kích thước tùy ý và lọc nhiễu phù hợp trong dữ liệu không gian mà không cần xác định trước số lượng nhóm.

CHƯƠNG 3 PHÂN CỤM DỮ LIỆU KHÔNG GIAN TRONG KHÔNG GIAN MẠNG

3.1. Giới thiệu

Trong chương này, luận án đề xuất giải pháp cải tiến thuật toán NS-DBSCAN nhằm tăng tốc thời gian xử lý. Thứ nhất, khi tổ chức dữ liệu, luận án đề xuất loại bỏ các cạnh có độ dài vượt quá ngưỡng bán kính esp vì đường đi có chiều dài tối đa là eps thì chắc chắn không chứa các cạnh vượt eps. Điều này giúp thuật toán không phải duyệt qua các cạnh có chiều dài lớn hơn eps. Thứ hai là không đưa vào Bảng thứ tự mật độ các điểm có mật độ thấp khi xây dựng Bảng thứ tự mật độ để giảm thời gian duyệt qua các đỉnh này trong bước tiếp theo là Hình thành nhóm. Thứ ba, luận án cải tiến một số thao tác thuộc về kỹ thuật lập trình nhằm giảm số lượng phép tính trong Thuật toán 3.1, loại bỏ thao tác sắp xếp trong Thuật toán 3.2, bỏ thao tác xác định nhiễu trong Thuật toán 3.3.

3.2. Thuật toán đề xuất iNS-DBSCAN

Mệnh đề 3.1. Một đường đi có độ dài tối đa là eps không thể chứa cạnh có trọng số lớn hơn eps.

Cải tiến đầu tiên áp dụng Định lý 3.1 để loại bỏ các cạnh có chiều dài vượt quá esp vì chắc chắn không mở rộng cơ bản đến các đỉnh kề có chiều dài cạnh lớn hơn eps. Do đó, thuật toán không cần phải kiểm tra các cạnh dài hơn eps để giảm thời gian.

Mệnh đề 3.2. Độ dài cạnh luôn lớn hơn 0.

Áp dụng Mệnh đề 1, khi đường mở rộng đã đạt đến eps thì không thực hiện thao tác kiểm tra để mở rộng cơ bản nữa, điều này giúp giảm thời gian mở rộng cơ bản khi độ dài đường đi đạt đến ngưỡng eps.

Trong Thuật toán 3.2 - Xây dựng bảng thứ tự mật độ có hai đề xuất cải tiến như sau:

Khi chèn một điểm trong hàng đợi Q, thì chèn vào vị trị phù hợp sao cho Q có thứ tự giảm dần mong muốn. Điều này giúp giảm thời gian cần thiết để sắp xếp danh sách Q trong vòng lặp.

Khi xây dựng Bảng thứ tự mật độ, luận án đề xuất không đưa vào các điểm có mật độ thấp để giảm số lượng điểm duyệt trong thao tác Hình thành nhóm ở bước tiếp theo.

Trong Thuật toán 3.3 – Hình thành nhóm: Nhiễu không được xác định trong quá trình hình thành nhóm vì nhiễu là những điểm không thuộc về nhóm nào. Tức là thuật toán không thực hiện các dòng lệnh kiểm tra và xác định nhiễu. Do đó, thời gian thực hiện của thuật toán được giảm xuống.

Ngoài ra, luận án cải tiến một số thao tác thuộc về kỹ thuật lập trình nhằm giảm số lượng phép tính trong Thuật toán 3.1, loại bỏ thao tác sắp xếp trong Thuật toán 3.2 và thao tác xác định nhiễu trong Thuật toán 3.3 của NS-DBSCAN. Sau khi hoàn thành quá trình phân cụm thì các điểm không thuộc nhóm nào được xem là nhiễu.

3.3. Kết quả thực nghiệm

Thuật toán đề xuất iNS-DBSCAN được so sánh với thuật toán gốc NS-DBSCAN sử dụng cùng 8 bộ dữ liệu từ OSM để đánh giá hiệu quả của thuật toán cải tiến. Kết quả thực nghiệm cho thấy rằng iNS-DBSCAN có được kết quả phân cụm không thay đổi và tăng tốc thời gian thực hiện. Thời gian xử lý của thuật toán cải tiến giảm trung bình 16%, có trường hợp giảm được đến 25%, cải tiến được đáng kể thời gian thực thi so với thuật toán ban đầu. Do đó, thuật toán cải tiến đã được chứng minh là một công cụ tiết kiệm thời gian để phân cụm dữ liệu. Thực nghiệm đã chứng minh được rằng thuật toán cải tiến đã giảm được thời gian thực hiện thuật toán đồng thời vẫn bảo đảm được chất lượng nhóm.

3.4. Kết luận

Chương này đề xuất thuật toán iNS-DBSCAN nhằm cải tiến thời gian xử lý của thuật toán phân cụm NS-DBSCAN.

Thuật toán 3.1 (LSPD) xác định láng giềng của điểm trung tâm (cp) trong bán kính eps để có được tập hợp tất cả các điểm trong khoảng cách eps từ cp bằng cách mở rộng cơ bản từ cp đến các điểm lân cận của nó. Việc mở rộng cơ bản đến các điểm lân cận được lặp lại cho đến khi chiều dài đường mở rộng bị chặn, tức là không còn ngắn hơn độ dài hiện tại hoặc vượt quá eps. Như vậy, đường mở rộng chắc chắn sẽ không đến các cạnh có độ dài vượt quá eps. Do đó, khuyến nghị đầu tiên là loại bỏ cạnh có độ dài vượt quá eps khi tổ chức dữ liệu.

Thứ hai, khi tạo Bảng thứ tự mật độ, thuật toán đề xuất không đưa vào các điểm có mật độ nhỏ hơn $ln(n)$, với n là kích thước dữ liệu (Theo heuristic gợi ý $MinPts \approx ln(n)$ của Birant và cộng sự đề xuất). Việc loại bỏ các điểm này giúp giảm số lượng điểm phải duyệt trong Thuật toán 3.3 (Thuật toán ban đầu hoạt động bằng cách tính toán mật độ cho tất cả các điểm sự kiện và đưa tất cả các

điểm vào Bảng thứ tự mật độ, ngay cả những điểm có mật độ bằng 0). Việc loại bỏ này không ảnh hưởng đến tính đúng đắn của kết quả phân cụm vì các điểm có mật độ thấp chắc chắn không thuộc bất kỳ nhóm nào.

Thứ ba, một số cải tiến trong kỹ thuật lập trình được thực hiện để giảm các thao tác tính toán từ đó góp phần giảm thời gian xử lý của thuật toán.

Tuy nhiên, như thuật toán NS-DBSCAN, kết quả phân cụm của thuật toán iNS-DBSCAN cũng phụ thuộc vào hai tham số đầu vào là ngưỡng bán kính và mật độ. Do đó, việc tìm giải pháp giảm số lượng tham số của thuật toán là nội dung sẽ được trình bày trong Chương 4.

CHƯƠNG 4 PHÂN CỤM DỮ LIỆU KHÔNG GIAN ĐỊA LÝ TRONG KHÔNG GIAN MẠNG DỰA TRÊN TÔ-PÔ

4.1. Giới thiệu

Các phương pháp phân cụm hiện tại chủ yếu thực hiện trên không gian Euclid và thường phụ thuộc vào tham số đầu vào. Chẳng hạn như phương pháp dựa trên mật độ phụ thuộc vào hai tham số là ngưỡng bán kính và ngưỡng mật độ. Chương này đề xuất thuật toán phân cụm dựa trên tô-pô trong không gian mạng NS-TBC (Network Space Topological-Based Clustering) bằng cách áp dụng quan hệ tô-pô cho thuật toán phân cụm dựa trên mật độ trong không gian mạng iNS-DBSCAN [CT.1] nhằm giảm sự phụ thuộc vào tham số người dùng trong quá trình phân cụm.

Phương pháp đề xuất khai thác ưu điểm của thuật toán ACUTE-2016, được đề xuất vào năm 2016 trong không gian Euclid, nhưng dành cho các đối tượng bị ràng buộc mạng. Thuật toán NS-TBC được áp dụng cho 12 bộ dữ liệu từ các nguồn OSM[3], ESRI Open Data[4] và Inside Airbnb[5] để chứng minh tính hiệu quả của nó. NS-TBC vượt trội hơn theo chỉ số đánh giá chất lượng nhóm Davies–Bouldin (DB) so với thuật toán ACUTE và iNS-DBSCAN, trong đó iNS-DBSCAN là thuật toán phân cụm hiệu quả dành cho dữ liệu không gian mạng được trình bày ở Chương 3. Thời gian chạy cũng được đo lường cẩn thận cho mục đích đánh giá. Việc đánh giá về thời gian chỉ được thực hiện đối với hai thuật toán NS-TBC và iNSDBSCAN vì ACUTE không được thiết kế để hoạt động trong không gian mạng. Kết quả thực nghiệm cho thấy thuật toán NS-TBC sử dụng ít hơn 50% thời gian tính toán so với thuật toán iNS-DBSCAN.

Thuật toán NS-TBC được đề xuất cung cấp giải pháp giảm số lượng tham số cho thuật toán iNS-DBSCAN đồng thời cải thiện đáng kể thời gian thực hiện và chất lượng nhóm.

4.2. Thuật toán đề xuất NS-TBC

Để phân cụm các đối tượng có ràng buộc mạng, thuật toán đề xuất NS-TBC tiến hành kiểm tra quan hệ tô-pô giữa các cặp điểm, kết hợp với điều kiện kiểm tra độ dài đường đi giữa chúng. Giữa các điểm phải có ít nhất một đường đi có độ dài

[3] http://download.geofabrik.de/asia/vietnam.html
[4] https://hub.arcgis.com/search
[5] http://insideairbnb.com

không vượt quá giá trị $2 * r$. Nếu có nhiều đường đi giữa các điểm thì sẽ sử dụng đường đi ngắn nhất.

Bước 1: Xác định các điểm *gặp nhau* hoặc *chồng nhau*:

Với mỗi điểm p, xác định tập hợp các điểm q *gặp nhau* hoặc *chồng nhau* với điểm p trong bán kính r. Bằng cách xác định tập hợp láng giềng có độ dài đường đi ngắn nhất từ p không vượt ngưỡng $eps = 2r$ (Thuật toán 4.1).

Tập hợp các điểm *gặp nhau* hoặc *chồng nhau* với p thì cùng nhóm với p.

Bước 2: Hình thành nhóm:

Hợp các tập láng giềng có chung ít nhất một điểm (Thuật toán 4.2).

Thuật toán 4.1. Xác định các điểm *meet /overlap* với điểm p theo bán kính r.

Đầu vào: Đồ thị G, điểm p, bán kính r.
Đầu ra: Tập hợp các điểm meet or overlap với điểm p. Ký hiệu: MO(p).

(1) $d(p) = 0, d(other\ vertices) = \infty$, chèn p vào hàng đợi rỗng Q
(2) while Q is not empty do
(3) p is dequeued from Q
(4) if p is an event vertex and not in $MO(p)$ then
(5) add p to $MO(p)$
(6) end if
(7) for each vertex q adjacent to p do
(8) if $d(p) < eps$ then
(9) if $s(p) + w(p, q) < s(q)$ and $s(p) + w(p, q) \leq eps$ then
(10) $s(q) = s(p) + w(p, q)$
(11) *if* q is not in Q then
(12) q is enqueued to Q

Thuật toán 4.2. Hình thành nhóm

Đầu vào: Tập hợp các điểm $P = \{P_i\}$ và $MO(P_i)$
Đầu ra: Các nhóm được hình thành

(1) for each point p in P do
(2) for each point q in P do
(3) if $q != p$ and $MO(p) \cap MO(q) != \emptyset$ then
(4) $MO(q) = MO(q) \cup MO(p)$
(5) $MO(q) = \emptyset$
(6) break

4.3. Kết quả thực nghiệm

Kết quả thực nghiệm của thuật toán đề xuất so với các thuật toán iNS-DBSCAN [CT.1] và ACUTE trên 12 bộ dữ liệu cho thấy thuật toán đề xuất cải thiện hiệu suất phân cụm về thời gian thực hiện và chất lượng nhóm.

Thuật toán đề xuất NS-TBC bảo đảm giữa các điểm trong cùng nhóm đều có đường đi với độ dài không vượt ngưỡng $2 * r$. Trong khi các nhóm được tạo ra từ thuật toán ACUTE-2016 có thể chứa các điểm có đường đi vượt ngưỡng hoặc thậm chí không có đường đi. Kết quả này không thể được sử dụng cho các bài toán trong thế giới thực yêu cầu giữa các điểm trong cùng một nhóm phải có ít nhất một đường đi và cũng có thể yêu cầu giới hạn về độ dài đường đi. Hơn nữa, NS-TBC còn có thể tìm thấy các nhóm trong các khu vực nhỏ nằm ở các khu vực xa xôi mà iNS-DBSCAN không phát hiện được.

Các thực nghiệm đánh giá chất lượng nhóm cho thấy chỉ số DB trung bình của thuật toán đề xuất NS-TBC tốt hơn iNS-DBSCAN và ACUTE trên cả 12 bộ dữ liệu gồm sáu bộ từ nguồn OSM, ba bộ từ ESRI và ba bộ từ Inside Airbnb. Kết quả này chứng minh rằng các nhóm được phát hiện bởi thuật toán đề xuất NS-TBC chặt chẽ hơn và tách biệt tốt hơn. Ngoài ra, các chỉ số DUNN, WB, WSS và BSS cũng chứng tỏ NS-TBC cho kết quả nhóm chất lượng hơn.

Thuật toán được đề xuất NS-TBC không chỉ hiệu quả về chất lượng nhóm được gom mà còn tăng tốc thời gian thực thi. Kết quả thực nghiệm so sánh thời gian thực hiện trên ba nguồn dữ liệu cho thấy NS-TBC cải thiện đáng kể về thời gian thực thi so với thuật toán phân cụm trong không gian mạng iNS-DBSCAN với trung bình thời gian tăng tốc được khoảng 50%, có trường hợp lên đến hơn 57%. Như vậy, thuật toán cải tiến đã giảm đáng kể thời gian xử lý cần thiết cho việc phân cụm dữ liệu. Dùng $Big\ O$ để so sánh thời gian tiêu thụ cũng nhận được kết quả là thuật toán NS-TBC có chi phí thời gian xử lý thấp hơn.

4.4. Kết luận

Chương này đã đề xuất phương pháp NS-TBC sử dụng quan hệ tô-pô để phân cụm dữ liệu trong không gian mạng với bốn đóng góp chính sau đây.

Thứ nhất, NS-TBC giúp giảm số lượng tham số đầu vào cho thuật toán phân cụm không gian có ràng buộc mạng.

Thứ hai, NS-TBC khắc phục hiện tượng các đối tượng có đường đi vượt giới hạn hoặc không có đường đi vẫn được gom vào cùng nhóm của thuật toán ACUTE để phân cụm cho các đối tượng không gian có ràng buộc mạng. Điều

này phù hợp với các vấn đề thế giới thực mà trong đó khoảng cách theo đường đi được ưu tiên hơn khoảng cách Euclid thông thường trong không gian mạng.

Thứ ba, NS-TBC cải thiện đáng kể hiệu suất của thuật toán phân cụm trong không gian mạng. Kết quả thực nghiệm cho thấy ưu điểm của nó về độ chính xác và thời gian thực hiện (nhanh hơn khoảng 50%) so với thuật toán iNS-DBSCAN.

Hơn nữa, NS-TBC có thể phát hiện các nhóm nằm ở các vùng nhỏ xa xôi bị iNS-DBSCAN bỏ qua, giúp khắc phục nhược điểm này được ghi nhận bởi Bhattacharjee và cộng sự.

Ngoài ra, với phương pháp phân cụm dựa trên tô-pô, luận án còn đề xuất năm chiến lược hiệu quả để nâng cao hiệu suất của phương pháp phân cụm dựa trên tô-pô được công bố gần nhất ACUTE_2023 (tính đến thời điểm nghiên cứu), được trình bày chi tiết ở Phần 4.4 trong luận án.

Hầu hết các kỹ thuật phân cụm đều thực hiện trên toàn bộ tập dữ liệu. Mà dữ liệu ngày nay không ngừng tăng trưởng nhanh chóng theo thời gian. Mỗi khi có dữ liệu mới phát sinh thì kết quả phân cụm cũ sẽ không còn chính xác nữa. Việc sử dụng thuật toán sẵn có phải thực hiện phân cụm lại từ đầu sẽ lãng phí thời gian. Vì vậy, giải pháp phân cụm gia tăng cho chỉ phần dữ liệu mới được phát sinh là nội dung sẽ được trình bày trong Chương 5.

CHƯƠNG 5 PHÂN CỤM GIA TĂNG DỮ LIỆU KHÔNG GIAN ĐỊA LÝ TRONG KHÔNG GIAN MẠNG

5.1. Giới thiệu

Chương này đề xuất phương pháp phân cụm gia tăng dựa trên mật độ trong không gian mạng (NS-IDBSCAN) để phân cụm dữ liệu mới phát sinh. Thay vì phải thực hiện phân cụm lại toàn bộ tập dữ liệu với kích thước lớn hơn nhiều so với phần dữ liệu mới được thêm vào, NS-IDBSCAN chỉ thực hiện phân cụm phần dữ liệu mới được thêm. Dựa trên kết quả phân cụm hiện tại của dữ liệu cũ, thuật toán được đề xuất NS-IDBSCAN kiểm tra vai trò của từng điểm mới được thêm và các láng giềng của nó để thực hiện việc phân cụm. Tùy thuộc vào mật độ, điểm được thêm vào có thể là nhiễu (noise), biên (border) hoặc lõi (core). Cách tiếp cận này giúp giảm đáng kể thời gian cần thiết để kịp thời đáp ứng nhu cầu. Hơn nữa, việc sử dụng bảng băm với khóa là chỉ mục của phần tử tương ứng với Id điểm để lưu trữ Id nhóm giúp truy cập trực tiếp nhằm loại bỏ thao tác tìm kiếm, tăng thêm tốc độ xử lý.

Thuật toán NS-IDBSCAN được đề xuất đã cải thiện đáng kể thời gian xử lý và trong một số trường hợp nhất định có thể cho kết quả phân cụm tốt hơn. Đó là những trường hợp các điểm dữ liệu được iNS-DBSCAN gán cho nhiều nhóm [CT.1]. Theo giải pháp lưu trữ dữ liệu phân cụm cũ được đề xuất trong chương này, mỗi điểm chỉ được gán cho một nhóm. Hơn nữa, bằng cách gán các điểm thuộc nhiều nhóm vào nhóm có ít điểm hơn, NS-IDBSCAN đã cải thiện chất lượng nhóm, như đã được chứng minh trong phần thảo luận (Phần 5.4).

Phương pháp đề xuất được so sánh với thuật toán phân cụm dữ liệu không gian dựa trên mật độ trong không gian mạng gốc iNS-DBSCAN [CT.1] và các thuật toán gần đây gồm NS-DBSCAN (2018), NS-TBC (2022, [CT.2]) và ACUTE_2023 (2023).

Kết quả thực nghiệm trên ba nguồn dữ liệu OSM, ESRI Open Data, Inside Airbnb và Chicago cho thấy phương pháp đề xuất NS-IDBSCAN tăng tốc đáng kể thời gian xử lý đồng thời bảo đảm được chất lượng nhóm. Chất lượng của thuật toán đề xuất được đo bằng bảy chỉ số: Silhouette, BSS, WSS, WB, Davis-Bouldin, Dunn và Calinski-Harabasz.

Chương này có ba đóng góp sau đây:

Thứ nhất, luận án đã giới thiệu phương pháp phân cụm gia tăng NS-IDBSCAN để phân cụm hiệu quả cho dữ liệu không gian địa lý mới được phát

sinh trong không gian mạng. Không giống như các thuật toán phân cụm dữ liệu không gian địa lý trong không gian mạng sẵn có, NS-IDBSCAN loại bỏ yêu cầu quét lại toàn bộ cơ sở dữ liệu khi có dữ liệu mới được thêm. Cải thiện hiệu quả này là một lợi thế lớn của thuật toán đề xuất.

Thứ hai, giải pháp sử dụng bảng băm để lưu trữ kết quả phân cụm đã có của dữ liệu cũ với khóa là Id điểm và giá trị là Id nhóm giúp có thể truy cập trực tiếp nhằm loại bỏ thao tác tìm kiếm đã tăng thêm đáng kể tốc độ xử lý.

Thứ ba, NS-IDBSCAN xử lý thành công hiện tượng các điểm biên thuộc về nhiều nhóm bằng việc đề xuất giải pháp gán các điểm đó vào chỉ một nhóm/nhóm có số điểm nhiều hơn đã giải quyết hạn chế của thuật toán iNS-DBSCAN [CT.1] giúp nâng cao chất lượng kết quả phân cụm.

5.2. Thuật toán phân cụm gia tăng trong không gian mạng NS-IDBSCAN

Chương này đề xuất thuật toán phân cụm gia tăng NS-IDBSCAN để phân cụm dữ liệu mới phát sinh dựa trên phương pháp iNS-DBSCAN. Vì NS-IDBSCAN theo phương pháp tiếp cận dựa trên mật độ nên nếu diễn tả theo cách đơn giản nhất thì ý tưởng chính là ***điểm lõi sẽ "hút" láng giềng của nó vào chung nhóm với nó.*** Sau đó, những láng giềng đó nếu là điểm lõi thì sẽ tiếp tục "hút" láng giềng của nó vào cùng một nhóm. Do đó, khi thêm điểm mới x có thể làm thay đổi vai trò của dữ liệu cũ nên có thể xảy ra các trường hợp sau:

- **Trường hợp 1:** Gom x vào nhóm đã có và có thể ghép nhóm: Điểm x được thêm có láng giềng là lõi.

Nếu x không có láng giềng là lõi thì còn lại hai trường hợp sau:

- **Trường hợp 2:** Tạo nhóm mới: Điểm x hoặc láng giềng là nhiễu của x trở thành lõi mới và có $MinPts$ láng giềng chưa được phân cụm.
- **Trường hợp 3:** Nhiễu: Thêm điểm x không làm phát sinh điểm lõi mới nào thì x là nhiễu.

Giải pháp lưu trữ dữ liệu:

Để lưu trữ kết quả của n điểm đã được phân cụm: Sử dụng mảng một chiều có chỉ số mảng là Id điểm, giá trị mảng chứa Id nhóm của điểm. Nếu điểm là nhiễu thì chứa giá trị -1.

Giải pháp lưu trữ này giúp loại bỏ thao tác tìm kiếm vốn tốn nhiều thời gian để kiểm tra kết quả phân cụm của một điểm. Để biết p đã được nhóm vào nhóm nào, chỉ cần truy cập trực tiếp vào phần tử có chỉ số p, giá trị tại phần tử này

chính là chỉ số của nhóm mà điểm p thuộc về. Giá trị -1 có nghĩa là điểm này là nhiễu.

Thuật toán 5.1. Thêm một điểm

Đầu vào:
- Đồ thị phẳng vô hướng, ngưỡng bán kính eps, ngưỡng mật độ $MinPts$.
- Kết quả đã phân cụm cho n điểm.
- Điểm mới thêm x.

Đầu ra: Kết quả được nhóm của $n + 1$ điểm.

(1) calculate N_eps(x) by LSPD algorithm
(2) for each $y \in N\_eps(x)$
(3) if y is core and y is clustered then
(4) $cluster(y) \leftarrow x$
(5) if x is core //merge through core x
(6) $cluster(x) \leftarrow$ noise in $N\_eps(x)$
(7) merge $core_{new} \in N\_eps(x)$ and other points *in the same cluster with* $core_{new}$ *to the same cluster as* x
(8) else if y is new core then // merge through neighbor is the new core
(9) $cluster(y) \leftarrow$ noise in $N\_eps(y)$
(10) merge $core_{newfb} \in N\_eps(y)$ and other points *in the same cluster with* $core_{newfb}$ *to the same cluster as y*
(11) if x *or* x′neighbor *is core has* $MinPts$ *neighbors* are noise then
(12) create new cluster C containing x and those noise points
(13) Otherwise x is noise

Thuật toán 5.2. Thêm m điểm

Đầu vào:
- Đồ thị phẳng vô hướng, eps, $MinPts$.
- n điểm đã được phân cụm.
- m điểm được thêm.

Đầu ra:
Kết quả phân cụm của $n + m$ điểm.

(1) Đọc kết quả phân cụm của n điểm ra mảng.
(2) Thực hiện m lần:
(3) Thêm một điểm (Thuật toán 5. 4).
(4) Cập nhật kết quả phân cụm.

5.3. Kết quả thực nghiệm

Kết quả thực nghiệm trên ba nguồn dữ liệu khác nhau cho thấy thuật toán đề

xuất tăng tốc đáng kể trong khi cải thiện được chất lượng so với các thuật toán phân cụm dữ liệu không gian địa lý gần đây trong không gian mạng gồm NS-DBSCAN (2019, Wang và cộng sự), iNS-DBSCAN (2021, [CT.1]), NS-TBC (2023, [CT.2]) và ACUTE_2023 (2023, Alomari và cộng sự). Mặc dù có nhiều thuật toán phân cụm gia tăng đã được phát triển, nhưng chúng chưa được thiết kế đặc biệt cho dữ liệu không gian có ràng buộc mạng.

NS-IDBSCAN tiết kiệm đáng kể thời gian phân cụm cho dữ liệu gia tăng. Kết quả thực nghiệm cho thấy rằng khi thêm một điểm dữ liệu mới, thuật toán đề xuất NS-IDBSCAN tăng tốc rất nhiều thời gian xử lý, đến 98% với thuật toán gốc. Khi thêm lô gồm nhiều điểm, thời gian thực hiện thuật toán cũng giảm đáng kể, chỉ còn 3%, 4%, 7%, 26% và 42% tương ứng cho các trường hợp thêm 10 điểm, 50 điểm, 100 điểm, 500 điểm và 1000 điểm. Thậm chí khi thêm lô có số điểm vượt kích thước dữ liệu đã được phân cụm 2000 điểm, như thêm 2000, 2500 và 3000 điểm mới thì thời gian xử lý vẫn giảm còn tương ứng là 64%, 69% và 75%.

Kết quả thực nhiệm cũng cho thấy rằng thời gian thêm n điểm ít hơn thời gian thêm một điểm nhân với n. Ví dụ với 1000 điểm thì thời gian 119.82007790 nhỏ hơn rất nhiều so với thời gian thêm một điểm là 3.90629101 nhân với 1000 bằng 3906.29101. Sở dĩ có kết quả này là do khi thêm một điểm thì phải đọc kết quả phân cụm đã có từ tệp. Khi thêm 1000 điểm thuật toán cũng chỉ đọc file 1 lần chứ không phải 1000 lần! Như vậy, để có thể thêm mới hiệu quả, kết quả phân cụm thường không nên cập nhật theo từng điểm mới được thêm mà sau một khoảng thời gian nào đó với một lô gồm nhiều điểm sẽ được cập nhật một lần.

Tóm lại, NS-IDBSCAN tiết kiệm đáng kể thời gian phân cụm cho dữ liệu mới phát sinh và kích thước của dữ liệu cũ được phân nhóm càng lớn thì tỷ lệ thời gian tăng tốc càng nhiều.

5.4. Kết luận

Chương này nhằm hiện thực hóa hướng phát triển của Chương 4 qua việc đề xuất giải pháp hiệu quả để phân nhóm dữ liệu không gian địa lý trong không gian mạng có phát sinh dữ liệu mới.

Khi có dữ liệu mới được thêm vào, thay vì phân cụm trên toàn bộ tập dữ liệu, phương pháp đề xuất thực hiện phân cụm chỉ trên các phần tử dữ liệu mới được thêm dựa trên kết quả phân cụm đã có của dữ liệu cũ. NS-IDBSCAN giải quyết vấn đề lãng phí thời gian thực hiện phân cụm lại từ đầu giúp tăng tốc xử lý để đạt được kết quả nhanh hơn. Đặc biệt, khi kích thước dữ liệu đã được phân cụm

càng lớn thì tỉ lệ thời gian tăng tốc càng nhiều. Hơn nữa, thời gian tỉ lệ thuận với công sức nên thời gian thực hiện giảm thì công sức và chi phí cũng giảm.

Chương 5 đã có những đóng góp chính sau đây:

- Đề xuất thuật toán phân cụm gia tăng dữ liệu không gian địa lý trong không gian mạng NS-IDBSCAN (density-based incremental clustering method in network space) với hiệu quả được cải thiện.
- Giải quyết vấn đề quét lại toàn bộ cơ sở dữ liệu khi một số điểm dữ liệu được thêm vào dữ liệu hiện có.
- Khắc phục những hạn chế của thuật toán iNS-DBSCAN [CT.1] đối với trường hợp các điểm biên thuộc nhiều hơn hai nhóm và đề xuất mệnh đề về việc chọn nhóm đối với điểm biên có khả năng thuộc nhiều nhóm.

CHƯƠNG 6 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết luận

Luận án nhằm nghiên cứu, đề xuất các phương pháp hiệu quả để giải quyết những yêu cầu đặt ra trên bài toán phân cụm dữ liệu không gian địa lý có ràng buộc mạng. Kết quả nghiên cứu được thể hiện dưới dạng ba thuật toán gồm iNS-DBSCAN, NS-TBC và NS-IDBSCAN để giải quyết ba mục tiêu nghiên cứu đã đặt ra ở mục 2 trong phần Mở đầu.

Đóng góp đầu tiên của luận án là phát triển thuật toán iNS-DBSCAN giúp tăng tốc đáng kể thời gian thực thi cho thuật toán NS-DBSCAN. Tuy nhiên, như thuật toán NS-DBSCAN, kết quả phân cụm của thuật toán iNS-DBSCAN cũng còn phụ thuộc vào 2 tham số đầu vào là ngưỡng bán kính và ngưỡng mật độ, vì vậy kết quả đầu ra nhạy cảm với các tham số này. Do đó, tìm giải pháp giảm sự phụ thuộc tham số đầu vào là đóng góp thứ hai của luận án. Luận án lấy ý tưởng từ thuật toán ACUTE_2016, đề xuất phương pháp NS-TBC sử dụng quan hệ tô_pô để phân cụm dữ liệu không gian địa lý trong không gian mạng. NS-TBC cải tiến cho cả hai thuật toán iNS-DBSCAN và ACUTE. Đóng góp chính thứ ba của luận án là đề xuất phương pháp phân cụm cho dữ liệu có phát sinh dữ liệu mới NS-IDBSCAN. Phương pháp NS-IDBSCAN sử dụng kết quả phân cụm cũ đã có và chỉ thực hiện phân cụm trên phần dữ liệu mới được thêm vào, thay vì sử dụng thuật toán sẵn có phải tiến hành phân cụm lại trên cả tập dữ liệu cũ và mới mỗi khi có dữ liệu mới phát sinh, giúp giảm đáng kể thời gian xử lý.
Các kết quả nghiên cứu đã được công bố trên các tạp chí chuyên ngành uy tín.

Hướng phát triển

Trong tương lai, luận án sẽ tập trung nghiên cứu và ứng dụng các kỹ thuật song song hóa để đáp ứng các yêu cầu thời gian thực cho bài toán phân cụm dữ liệu không gian địa lý trong không gian mạng.

Ngoài ra, luận án sẽ mở rộng nghiên cứu về các phương pháp phát hiện dị thường cho dữ liệu không gian địa lý có ràng buộc mạng.