


MẪU BÁO CÁO CỦA HỌC VIÊN

| | |
|-----------------------|--|
| Họ và tên (IN HOA) | CHU VŨ THÙY LINH |
| Ảnh |  |
| Số buổi vắng | 3 |
| Bonus | 10 |
| Tên đề tài (VN) | KHAI PHÁ DỮ LIỆU CỦA PHƯƠNG TIỆN GIAO THÔNG CHUYÊN CHỖ VÀ MÔ PHỎNG TÌNH TRẠNG GIAO THÔNG |
| Tên đề tài (EN) | MINING AND DISCOVERING TRUCK DATA AND STIMULATING TRAFFIC STATUS |
| Giới thiệu | <p>Bối cảnh, lí do, bài toán</p> <p>Phân tích dữ liệu giao thông một công việc quan trọng và có nhiều ý nghĩa trong thực tiễn</p> <p>Bài toán này đang thu hút sự quan tâm của các đơn vị quản lý và vận hành hạ tầng giao thông cũng như các nhà khoa học trong lĩnh vực liên quan. Phân tích dữ liệu giao thông giúp ích rất nhiều cho các ngành như ngành vận tải: vận chuyển người và hàng hóa đến đích một cách an toàn, tiết kiệm; ngành giao thông: điều phối lưu lượng giao thông, tránh ùn tắc giao thông; ngành quy hoạch đô thị: đưa ra những giải pháp trong việc quy hoạch các tuyến đường, nhà ga, bến xe [2]</p> |

Trong khoảng thời gian gần đây, các đối tượng kinh doanh vận tải đều bắt buộc gắn thiết bị giám sát hành trình, và cách thức kinh doanh vận tải cũng được hiện đại hóa bằng cách áp dụng công nghệ thông tin, đặc biệt là những thiết bị di động thông minh. Dữ liệu từ những hệ thống giám sát hành trình, hệ thống nghiệp vụ này phần nào cho phép biết được vị trí hiện thời của phương tiện vận tải, biết được những thông tin đi kèm của phương tiện vận tải như vận tốc, người lái, các sai phạm của phương tiện vận tải [1]. Tuy nhiên việc khai thác dữ liệu này còn đang gặp khá nhiều thách thức do lượng dữ liệu lớn, dữ liệu nhiễu nhiễu.

Thông tin

(Cụ thể triển khai input và output từng thuật toán sẽ được nêu ở phần mục tiêu)

1. Thuật toán Traclus

Input: Tập hợp quãng đường $I = \{TR1, \dots, TR_{numtra}\}$

Output: Tập hợp các cụm $O = \{C1, \dots, C_{numclus}\}$ = tập hợp các đoạn đường tiêu biểu

2. Thuật toán Approximate Trajectory Partitioning

Input: $TR_i = p_1 p_2 p_3 \dots p_j \dots p_{len_i}$

Output: Tập hợp các điểm đặc trưng C_{pi}

3. Thuật toán 3: Phân cụm

Input: = Một tập hợp phân đoạn $D = \{L1, \dots, L_{numln}\}$, = Hai tham số ϵ and $MinLns$

Output: Một tập hợp cụm $O = \{C1, \dots, C_{numclus}\}$

Cơ sở dữ liệu

Với cơ sở dữ liệu được cung cấp là nguồn thu thập từ thiết bị giám sát hành trình gắn trên xe taxi và từ ứng dụng gọi xe taxi, ta tiến hành xây dựng hệ thống qua các bước tổng quan như sau:

B1: Chia dữ liệu ra thành các tập bản ghi theo ngày (mỗi ngày là một tập bản ghi), chia phân biệt ngày thường và ngày cuối tuần.

B2: Tiến hành chạy thuật toán phân cụm trên từng tập bản ghi theo ngày ta được các cụm của cung đường di chuyển theo ngày, tiến hành chạy thuật toán phân cụm trên từng khung thời gian ta được các cụm cung đường di chuyển theo khung thời gian.

B3: Chia vùng bản đồ Hà Nội thành các ô (vùng) ta được tọa độ, giới hạn của các ô (vùng).

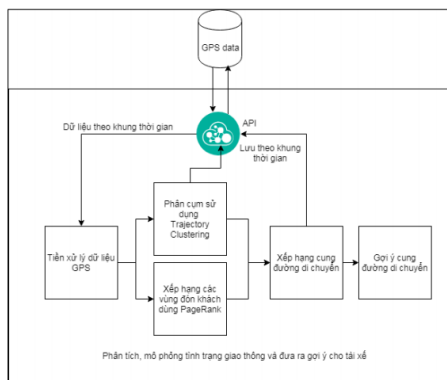
B4: Dựa trên tọa độ của các ô (vùng) và các cụm cung đường di chuyển theo khung thời gian, biểu diễn luồng di chuyển của các phương tiện vận tải theo thời gian.

B5: Dựa vào thuật toán PageRank, với các cách tính điểm ban đầu dựa vào: Số lượng xe; số lượng khách lên xe, xuống xe; vận tốc; ta tính các xếp hạng khác nhau cho các vùng dựa vào PageRank, thu được xếp hạng của các ô (vùng).

B6: Dựa trên vùng và mật độ của vùng hiện tại/ vùng và xếp hạng của vùng hiện tại cùng với mô hình n-MMC [12], chọn các điểm đến tiếp theo là các vùng lân cận, ta xác định vùng đến tiếp theo, được vùng có thể lựa chọn và vùng có xác suất đến nhiều nhất thời điểm tiếp theo.

B7: Dựa trên đưa ra các lựa chọn tốt nhất cho tài xế, dựa trên gợi ý cho tài xế cách di chuyển theo các cung đường khác nhau dựa trên kết nối giữa các vùng

Các thành phần:



- GPS data: Cơ sở dữ liệu của hệ thống, ở hệ thống trong luận văn cơ sở dữ liệu này lưu trữ: Dữ liệu về các bản tin GPS [9, 10] của từng phương tiện (mỗi phương tiện phân biệt bằng id của phương tiện). Dữ liệu về các cung di chuyển đã phân cụm bằng thuật toán TraClus. Dữ liệu về ma trận chuyển dịch qua tập huấn

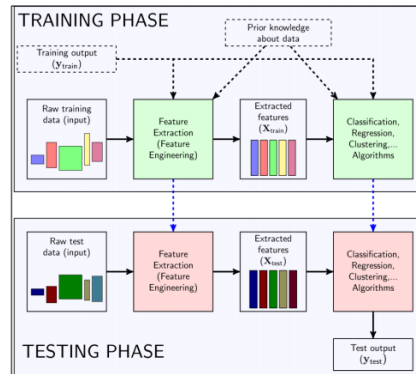
- Tiền xử lý dữ liệu GPS: Module xử lý các dữ liệu nhiễu (kinh độ, vĩ độ, vận tốc không hợp lý)

- Phân cụm sử dụng TrajectoryClustering: Module phân cụm sử dụng thuật toán TrajectoryClustering và lưu trữ dữ liệu đã phân cụm

- Xếp hạng các vùng đón khách bằng PageRank [4]: Module sử dụng thuật toán PageRank để xếp hạng các vùng theo các tiêu chí khác nhau [13,14]

- Xếp hạng và gợi ý cung đường di chuyển: Hai module sử dụng mô hình n-MMC [3] để tập huấn và gợi ý các cung đường di chuyển dựa trên dự đoán về luồng di chuyển, vận tốc

Mô hình chung cho các bài toán dự đoán



Raw input là tất cả các thông tin ta biết về dữ liệu. Với bài toán trong luận văn thì chính là thông tin về dữ liệu GPS của phương tiện vận tải

Trong luận văn phần này chính là ma trận xác suất chuyển dịch dự đoán phương tiện đến tiếp theo với thông tin về vận tốc trung bình trong vùng đó

Prior knowledge about data: Là giả thiết về dữ liệu đang có, ở đây luận văn đưa ra giả thiết vận tốc trung bình của phương tiện ở trong vùng là trung bình cộng vận tốc của các bản ghi

=> Với raw input mới, luận văn sử dụng dữ liệu thu được từ Training phase qua main algorithms để dự đoán output

Mục tiêu

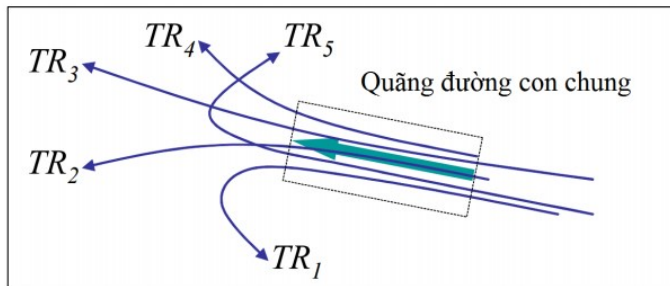
Phương pháp giải quyết của luận văn tập trung vào phân cụm các cung đường di chuyển, xếp hạng các vùng giao thông, dự đoán lưu lượng và điểm đến, trên cơ sở đó gợi ý cung đường di chuyển cho người tham gia giao thông. Dựa trên các nghiên cứu đã có, luận văn đề xuất một số cách áp dụng, kết hợp các nghiên cứu để giải các bài toán thực tiễn. Luận văn đã xây dựng mô hình nhằm giải quyết các bài toán đặt ra và thử nghiệm trên máy tính cá nhân

1. Trình bày quá trình thử nghiệm bao gồm: môi trường thử nghiệm, kết quả thử nghiệm

2. Kết quả thử nghiệm được thực hiện trên hai bộ dữ liệu về taxi từ thiết bị giám sát hành trình và ứng dụng đặt xe taxi

| | |
|--|---|
| | <p>3. Trình bày tổng quan về các kết quả thu được, đưa ra cách đánh giá và đánh giá độ chính xác của mô hình dự báo</p> |
| Nội dung và phương pháp thực hiện | <p>1. Phân vùng và phân cụm các cung đường di chuyển theo thời gian để tìm ra quy luật di chuyển của các phương tiện vận tải: Cụ thể ở đây luận văn tiến hành phân tích dữ liệu của nhiều taxi trong cùng một ngày, trong một khoảng thời gian nhất định để tìm ra các cụm (các cung đường chung), loại bỏ những dữ liệu nhiễu, cụm không đặc trưng, phục vụ cho bài toán mô phỏng luồng di chuyển, tìm ra các đường đi chung, các đường đi tối ưu phục vụ cho bài toán gợi ý di chuyển. Phương pháp phân cụm thường chia thành: không giám sát, giám sát, bán giám sát [7]. Luận văn lựa chọn phương pháp không giám sát, cụ thể là mô hình và thuật toán Trajectory clustering (Jae-Gil Lee và cộng sự) [6] sẽ trình bày bên dưới.</p> <p>2. Mô phỏng luồng di chuyển của các phương tiện vận tải theo vùng: Nhằm đạt mục tiêu khái quát hóa và tăng hiệu năng tính toán luận văn sử dụng tư tưởng chia vùng theo công trình của Naoto [8] và cách chia cung thời gian theo công trình (Xiaomeng Wang và cộng sự) [15] và đề xuất cách biểu diễn mật độ theo vận tốc</p> <p>3. Xếp hạng các khu vực đón, trả khách [5]: Luận văn thực hiện khái quát hóa khu vực đón, trả khách theo tư tưởng chia vùng trong công trình của Naoto và cách chia cung thời gian trong công trình [15]</p> <p>4. Dự đoán luồng giao thông trong các vùng: Luận văn thực hiện dự đoán vùng đến kế tiếp theo công trình (S'ebastien Gambs và cộng sự) [11,12] với cách gán nhãn dựa trên xếp hạng và mật độ, phục vụ cho bài toán gợi ý di chuyển tiếp theo</p> <p>5. Đưa ra gợi ý di chuyển cho tài xế dựa vào mật độ giao thông và kết quả xếp hạng của các vùng: Dựa trên bài toán dự đoán luồng giao thông và xếp hạng đón khách, luận văn thực hiện đưa ra các gợi ý di chuyển cho tài xế, sử dụng các cung đường đã phân cụm để gợi ý cung đường tốt nhất.</p> <p><i>Thuật toán sử dụng:</i></p> <p>1. Thuật toán phân cụm TRACCLUS</p> |

Để làm rõ thuật toán, giả sử có 5 quỹ đạo như trong hình bên dưới, có thể nhìn rõ rằng có một đặc điểm chung, biểu diễn bằng mũi tên trong hình chữ nhật. Tuy vậy, nếu nhóm những quỹ đạo này làm một, chúng ta không thể khám phá đặc điểm chung này khi mà chúng di chuyển đi các hướng khác nhau, vì vậy sẽ bị mất một số thông tin quý giá [6]



Giải pháp ở đây sẽ là phân chia các quỹ đạo thành tập hợp các phân đoạn đường và sau đó nhóm các phân đoạn đường. Công việc này nằm trong khuôn khổ phân vùng và cụm. Mục tiêu chính của việc phân vùng và cụm này là khám phá các quỹ đạo con (phân đoạn đường) chung từ bộ dữ liệu quỹ đạo đầu vào

Phương pháp phân vùng và cụm sẽ gồm 2 giai đoạn:

- Bước phân vùng: Mỗi quỹ đạo được tối ưu phân chia làm các phân đoạn đường. Các phân đoạn đường này sẽ là dữ liệu đầu vào cho bước tiếp theo.
- Bước phân cụm: các phân đoạn đường giống nhau được nhóm vào một cụm. Trong bài báo này, thuật toán phân cụm dựa trên mật độ được sử dụng.

2. Thuật toán Approximate Trajectory Partitioning

Việc phân chia tối ưu cần phải có hai tính chất sau: chính xác và súc tích. Tính chính xác có nghĩa rằng sự khác nhau giữa quỹ đạo và một tập hợp phân đoạn đường càng nhỏ càng tốt. Tính súc tích đồng nghĩa với số lượng phân đoạn càng ít càng tốt

3. Thuật toán Phân Cụm DBSCAN

Trong thuật toán TRACCLUS, thuật toán phân cụm DBSCAN được sử dụng. Đối với thuật toán DBSCAN, chúng ta cần xác định 2 tham số: ϵ (tương ứng với khoảng cách nhỏ nhất giữa 2 điểm để có thể gọi là điểm hàng xóm) và minPts

Kết quả dự

Phần mềm ứng dụng: Node.js, Ngôn ngữ python, Cơ sở dữ liệu Mongo

| | |
|---------------------------|---|
| kiến | <p><i>Thuật toán:</i> Thuật toán phân cụm TRACCLUS, Thuật toán Approximate Trajectory Partitioning, Thuật toán Phân Cụm DBSCAN</p> <p><i>So sánh giữa các phương pháp:</i> Chưa thể khắc phục tình trạng thiếu chính xác do dữ liệu thừa, đặc biệt là dữ liệu từ các ứng dụng di động</p> <p><i>Bộ dữ liệu, etc:</i> Bộ dữ liệu về taxi từ thiết bị giám sát hành trình và ứng dụng đặt xe taxi</p> |
| Tài liệu tham khảo | <p>Tiếng Việt</p> <p>[1]. Nguyễn Văn Tăng (2017) “Phát triển dịch vụ ứng dụng công nghệ GPS trong quản lý, giám sát, điều phối và tối ưu hóa kế hoạch sử dụng phương tiện”, Bộ công thương - Chương trình quốc gia phát triển công nghệ cao đến năm 2020</p> <p>[2]. Viện Khoa học và Công nghệ Giao thông (2016) “Dự thảo về tiêu chuẩn quốc gia cho kiến trúc hệ thống giao thông thông minh ITS”, Bộ Khoa học và Công nghệ</p> <p>Tiếng Anh</p> <p>[3]. A. A. Markov (2006) “Classical Text in Translation An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains”, Science in Context 19(4), pp. 591–600</p> <p>[4]. Bin Jiang (2008) “Ranking Spaces for Predicting Human Movement in an Urban Environment”, Journal International Journal of Geographical Information Science Volume 23 Issue 7, July 2009 pp. 823-837</p> <p>[5]. Daniel Jurafsky & James H. Martin (2006) “Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition”, Chapter 6</p> <p>[6]. Jae-Gil Lee, Jiawei Han, Kyu-Young Whang (2007) “Trajectory clustering: a partition-and-group framework”, Proceedings of the 2007 ACM SIGMOD international conference on Management of data (SIGMOD '07). ACM, New York, NY, USA, pp. 593-604.</p> <p>[7]. Jiang Bian, Dayong Tian, Yuanyan Tang, Dacheng Tao (2018), “A survey on trajectory clustering analysis”</p> <p>[8]. Naoto Mukai (2013) “PageRank-based Traffic Simulation Using Taxi Probe Data”, Procedia Computer Science, 2013. 22: pp. 1156-1163.</p> |

- | | |
|--|--|
| | <p>[9]. Raj Kishen Moloo, Varun Kumar Digumber (2011) “Low-Cost Mobile GPS Tracking Solution”, 2011 International Conference on Business Computing and Global Informatization 5 pp. 10–15</p> <p>[10]. Sameer Darekar, Atul Chikane, Rutujit Diwate, Amol Deshmukh, Prof. Archana Shinde (2012) “Tracking System using GPS and GSM: Practical Approach”, IJSER journal pp. 34–40</p> <p>[11]. S’ebastien Gambs, Marc-Olivier Killijian, Miguel N’uñez del Prado Cortez (2011) “Show Me How You Move and I Will Tell You Who You Are”, transactions on data privacy 4 (2011) pp. 103–126</p> <p>[12]. S’ebastien Gambs, Marc-Olivier Killijian, Miguel N’uñez del Prado Cortez (2012) “Next Place Prediction using Mobility Markov Chains” K.4 COMPUTERS AND SOCIETY MPM '12 Proceedings of the First Workshop on Measurement, Privacy, and Mobility</p> <p>[13]. Sergey Brin, Lawrence Page (1998) “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Computer Networks and ISDN Systems. 30 pp. 107–117</p> <p>[14]. Wenpu Xing, Ali Ghorbani (2004) “Weighted PageRank Algorith Proceedings of the Second Annual Conference on Communication Networks and Services Researchm”</p> <p>[15]. Xiaomeng Wang, Ling Peng, Tianhe Chi, Mengzhu Li, Xiaojing Yao, Jing Shao (2015) “A Hidden Markov Model for Urban-Scale Traffic Estimation Using Floating Car Data”, PLoS ONE 10(12).</p> |
|--|--|