ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

TRỊNH BÁ QUÝ

PHÂN TÍCH VÀ MÔ PHỎNG TÌNH TRẠNG GIAO THÔNG DỰA VÀO KHAI PHÁ DỮ LIỆU CỦA PHƯƠNG TIỆN VẬN TẢI

Chuyên ngành: Khoa học máy tính

Mã số: 8480103.01

TÓM TẮT LUẬN VĂN THẠC SĨ NGÀNH CÔNG NGHỆ THÔNG TIN

Hà Nội - 2018

Chương 1 Khái quát bài toán khai phá dữ liệu phương tiện vận tải

Ngày nay, với sự phát triển mạnh mẽ và vượt bậc về Công nghệ thông tin, cũng như hạ tầng cơ sở giao thông, việc hiện đại hóa quá trình khai thác, kiểm soát phương tiện vận tải đang được chú trọng triển khai sâu rộng. Điều này thúc đẩy sự gia tăng về dữ liệu của phương tiện vận tải. Các dữ liệu này đến từ các thiết bị giám sát hành trình cũng như các thiết bị đi kèm trong quá trình thực hiện giải quyết các bài toán nghiệp vụ. Vì vậy, nhiều nhà khoa học đã nghiên cứu các công nghệ, thuật toán để giải quyết bài toán về khai phá dữ liệu cách nhanh nhất đáp ứng được những yêu cầu thực tế mà các tổ chức hay doanh nghiệp đưa ra.

1.1 Tổng quan về dữ liệu GPS

GPS - Hệ thống định vị toàn cầu là hệ thống xác định vị trí dựa trên vị trí của các vệ tinh nhân tạo, do Bộ Quốc phòng Hoa Kỳ thiết kế, xây dựng, vận hành và quản lý. Trong cùng một thời điểm, tọa độ của một điểm trên mặt đất sẽ được xác định nếu xác định được khoảng cách từ điểm đó đến ít nhất ba vệ tinh.

GPS sử dụng nguyên tắc hướng thẳng tương đối của hình học và lượng giác học. Mỗi vệ tinh liên tục phát và truyền dữ liệu trong quỹ đạo của nó, do đó, mỗi thiết bị GPS nhận sẽ liên tục truy cập dữ liệu quỹ đạo chính xác từ vị trí của tất cả vệ tinh.

Máy thu tính toán được khoảng cách từ các vệ tinh, giao điểm của các mặt cầu có tâm là các vệ tinh, bán kính là thời gian tín hiệu đi từ vệ tinh đến máy thu nhân vận tốc sóng điện từ là toạ độ điểm cần định vị.

GPS hiện tại gồm 3 phần chính: Phần không gian, phần kiểm soát và phần sử dụng.

1.2 Dữ liệu GPS từ phương tiện vận tải

Dữ liệu định vị của phương tiện vận tải được thiết bị định vị ghi lại và gửi về máy chủ theo một khoảng thời gian cố định. Nếu một phương tiện bật máy (ở trạng thái bật chìa khóa điện), dữ liệu sẽ được gửi lên 15 giây một lần, ngược lại, ở trạng thái tắt máy, dữ liệu sẽ được gửi 30 giây một lần.

1.3 Các ứng dụng của khai phá dữ liệu phương tiện vận tải

Luận văn này tập trung vào mảng ứng dụng "Dịch vụ Giám sát và điều khiển giao thông" – là một nhu cầu bức thiết hiện nay để giải quyết các vấn đề về tắc đường, quy hoạch đô thị với các bài toán cụ thể:

- 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
- Mô phỏng luồng di chuyển của các phương tiện vận tải theo vùng
- Xếp hạng các khu vực đón, trả khách
- Dự đoán luồng giao thông trong các vùng
- Đưa ra gợi ý di chuyển cho tài xế dựa vào mật độ giao thông và kết quả xếp hạng của các vùng

Chương 2 Một số nghiên cứu về phân tích, mô phỏng tình trạng giao thông

Như đã đề cập trong chương 1, luận văn tập trung vào những bài toán cụ thể sau:

- **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[7]: không giám sát, giám sát, bán giám sát. Luận văn lựa chọn phương pháp không giám sát, cụ thể là mô hình và thuật toán Trajectory clustering của Jae-Gil Lee và cộng sự [6] sẽ trình bày bên dưới.
- **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 của Xiaomeng Wang và cộng sự [15] và đề xuất cách biểu diễn mật độ theo vận tốc

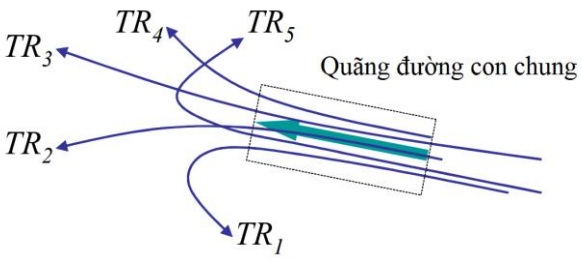
- **Xếp hạng các khu vực đón, trả khách:** Luận văn thực hiện khái quát hóa khu vực đón, trả khách theo tư tưởng chia vùng trong công trình của Naoto[8] và cách chia cung thời gian trong công trình của Xiaomeng Wang và cộng sự [15]
- **Dự đoán luồng giao thông trong các vùng:** Luận văn thực hiện dự đoán vùng đến kế tiếp theo công trình của 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
- **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.

2.1 Thuật toán phân cụm TRACLUS

Phân cụm là là cách nhóm các đối tượng dữ liệu thành các nhóm sao cho các đối tượng trong cùng một nhóm gần nhau hơn và các đối tượng của hai nhóm khác nhau khác nhau rất nhiều. Đối với dự án, phân cụm có thể tích hợp rồi cho phép tìm hiểu các quy luật quãng đường của từng taxi. Các quy luật đường đi của taxi gồm có các đoạn đường được taxi dùng để di chuyển nhiều nhất, các cụm quãng đường sẽ được phân ra dựa trên khoảng cách thực tế.

Để giải quyết hai bài toán trên luận văn sử dụng công trình của Jae-Gil Lee và cộng sự [6], đó là thuật toán TRACLUS.

Để hiểu rõ thuật toán chúng ta giả sử có 5 quãng đường như trong Hình 3.1. Chúng ta có thể nhìn rõ rằng có một đặc điểm chung, biểu diễn bằng mũi tên trong hình chữ nhật. Tuy vậy, nếu chúng ta nhóm những quãng đường 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 chúng ta bị mất một số thông tin quý giá.

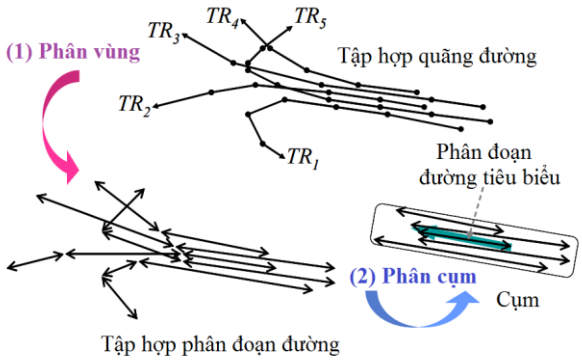


Hình 2.1Mô hình quãng đường con chung

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

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

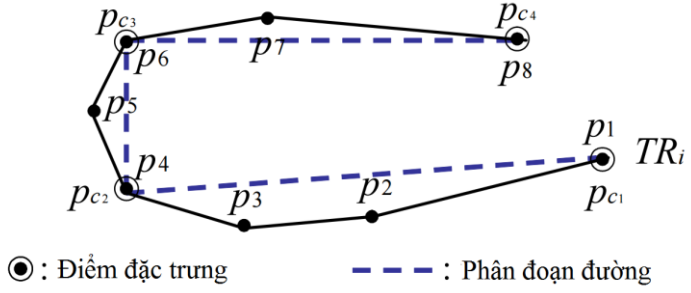
- Bước phân vùng: Mỗi quãng đường được tối ưu phân chia làm các phân đoạn đường. Các phân đoạn đường này sẽ là dữ liệu đầu vào cho bước tiếp theo.
- Bước phân cụm: các phân đoạn đường giống nhau được nhóm vào một cụm. Trong bài báo này, thuật toán phân cụm dựa trên mật độ được sử dụng.



Hình 2.2Ví dụ về phân vùng và cụm quãng đường

2.1.1 *Phân vùng quãng đường*

Chúng ta muốn tìm những điểm mà hành vi của các quãng đường thay đổi nhanh chóng, chúng ta gọi những điểm này là những điểm đặc trưng. Đối với mỗi $TR_i = p_1\ p_2\ p_3 \ldots p_{leni}$, chúng ta xác định một tập hợp các điểm đặc trưng $\{pc_1, pc_2, pc_3, \ldots, pc_{pari}\}$ ($c_1 < c_2 < \ldots < c_{pari}$). Mỗi điểm p_i tương ứng với một tọa độ gồm kinh độ và vĩ độ (X và Y trong tệp dữ liệu đầu vào). Sau đó TR_i được phân vùng tại mỗi điểm đặc trưng, và mỗi vùng được biểu diễn bởi phân đoạn đường. Hình 2.3 miêu tả một ví dụ về quãng đường và cách nó được phân đoạn.



Hình 2.3: Ví dụ về quãng đường và các phân đoạn

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 quãng đường 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

2.1.2 *Phân cụm*

Trong thuật toán TRACLUS, 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 (tương ứng với số lượng điểm hàng xóm).

$N\varepsilon(L)$ được gọi là các hàng xóm của phân đoạn đường $L \in D$ trong khoảng cách bán kính ε: $N\varepsilon(L_i) = \{L_j \in D \mid dist(L_i, L_j) \leq \varepsilon\}$.

Phân đoạn đường $L_i \in D$ được gọi là phân đoạn đường với điều kiệnεvà MinLns thỏa mãn nếu $|N\varepsilon(L_i)| \geq$ MinLns và sẽ gọi là ngoại bên nếu không thõa mãn điều kiện này.

2.2 Mô hình giao thông dựa trên "pagerank"

2.2.1 Xếp hạng bằng duyệt web

Thuật toán PageRank là một trong những thuật toán xếp hạng trang web được sử dụng rộng rãi nhất, dựa trên giả thuyết rằng nếu một trang web có những liên kết quan trọng đến nó, thì liên kết của nó đến các trang khác cũng trở nên quan trọng. Do vậy PageRank tính toán các backlink (liên kết đến trang đó) và chia sẻ xếp hạng thông qua liên kết: Một trang có xếp hạng cao nếu tổng của các trang có liên kết đến nó cao[13].

Thuật toán tính toán dựa trên giả định:

- Mỗi đường link tới trang web sẽ được tính như 1 sự hỗ trợ làm tăng thêm giá trị Pagerank.
- Giá trị Pagerank của trang được định nghĩa đệ quy và phụ thuộc vào số lượng và giá trị của các trang mà có link dẫn đến trang đó (incoming links).
- Một trang web có chứa nhiều link liên kết từ các trang web có giá trị PageRank cao thì giá trị PageRank của trang đó cũng sẽ cao

$$PR(u) = \sum_{v \in B(u)} \frac{PR(v)}{N_v}$$

2.2.2 Damping factor trong PageRank

Có một khái niệm quan trọng trong PageRank gọi là "damping factor" sử dụng trong quá trình chuyển thứ hạng. Khái niệm được sử dụng để tránh vấn đề đường cụt

Khả năng nhảy này trong PageRank đặc trưng bởi hệ số "damping factor" (d). Hệ số này thường được đặt là 0.85. Công thức trở thành:

$$PR(u) = 1 - d + d \sum_{v \in B(u)} \frac{PR(v)}{N_v}$$

2.2.3 PageRank có trọng số

Định nghĩa phía trên của PageRank có một giả định là xếp hạng của một trang được chia đều cho tất cả những trang nó có liên kết. Ví dụ trang A có bốn liên kết in-link đến từ bốn trang B, C, D và E. Theo công thức PageRank [13]

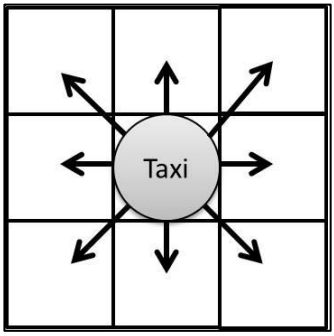
mỗi trang trong bốn trang trên đóng góp cho A xếp hạng như nhau. Tuy nhiên giả định này không đúng trong thực tế. Những trang quan trọng hơn hay phổ biến hơn thường có tỷ lệ chia sẻ xếp hạng cao hơn. Nói cách khác xếp hạng chuyển đến một trang web A từ các trang khác phụ thuộc vào độ phổ biến của các liên kết của nó (in-link và out-link)[14]

PageRank có trọng số được định nghĩa như sau:

$$PR(u) = 1 - d + d \sum_{v \in B(u)} PR(v)\, W^{in}_{(v,u)} W^{out}_{(v,u)}$$

2.2.4 Xếp hạng bằng taxi

Luận văn áp dụng tư tưởng của PageRank có trọng số cho mô hình giao thông bằng cách thay quá trình duyệt web bằng quá trình di chuyển của taxi [8]. Có nghĩa là một chiếc taxi sẽ mang xếp hạng từ một vùng M(i,j) đến một vùng lân cận M(i',j') như Hình 2.11



Hình 2.11 Xếp hạng bởi taxi

2.3 Sử dụng xích Markov trong dự đoán điểm đến tiếp theo

2.3.1 Xích Markov

Xích Markov là một trường hợp đặc biệt của automat hữu hạn có trọng số

Một xích Markov sử dụng một giả định quan trọng trong thứ tự của xích Markov bậc nhất: xác suất của một trạng thái cụ thể chỉ phụ thuộc vào trạng thái trước đó.

Thuộc tính Markov: $P(q_i|q_1 \dots q_{i-1}) = P(q_i|q_{i-1})$

Bởi vì mỗi aij biểu diễn một xác suất $p(q_j|q_i)$, luật xác suất yêu cầu giá trị của tất cả cung đi ra từ một trạng thái phải có tổng là 1:

$$\sum_{j=1}^{n} a_{ij} = 1 \; \forall i$$

2.3.2 Xích Markov di động (Mobility Markov Chain - MMC)

Xích Markov di động (tên tiếng Anh là Mobility Markov Chain, từ bây giờ sẽ ký hiệu là MMC) mô hình hóa hành vi di chuyển của một người như là một quá trình ngẫu nhiên rời rạc. Trong đó xác suất di chuyển đến một trạng thái (Ở đây là một địa điểm) chỉ phụ thuộc vào trạng thái trước đó (địa điểm trước đó) và phân bố xác suất của quá trình chuyển đổi giữa các trạng thái [11,12]. Chính xác hơn một MMC bao gồm:

- Một tập hợp trạng thái $P = \{p_1,\ldots,p_k\}$, ở đây mỗi trạng thái tương ứng với một địa điểm có tần suất cao (Xếp hạng theo thứ tự giảm dần của tầm quan trọng).
- Một tập hợp các chuyển tiếp, như là $t_{i,j}$, đại diện cho việc chuyển từ trạng thái p_i sang trạng thái p_j. Một chuyển đổi từ một trạng thái sang chính nó có thể xảy ra nếu như người đó di chuyển từ một trạng thái sang một địa điểm không thường xuyên rồi quay lại trạng thái đó.

2.3.3 Sử dụng n-MMC để dự đoán điểm đến tiếp theo

Để dự đoán điểm đến tiếp theo dựa trên n vị trí cuối cùng, ta sử dụng ma trận chuyển dịch có thay đổi, mà trong ma trận này hàng đại diện cho n điểm đến cuối cùng – thay đổi so với ma trận chuyển dịch ở nguyên bản là hàng đại diện địa điểm cuối, cột đại diện cho điểm đích. Để minh họa việc dự đoán điểm đến tiếp theo, ở đây sử dụng bảng 1 và hình 2.16 lần lượt cho ma trận chuyển dịch và biểu đồ của 2-MMC. 2-MMC bao gồm 4 trạng thái khác nhau: "Home"(H), "Work"(W) "Leisure"(L) và "Other"(O). Mục tiêu là đoán điểm đến tiếp theo dựa trên 2 điểm phía trước (ở đây n = 2).. Ví dụ, nếu như địa điểm lúc trước là H và địa điểm hiện giờ là W, dự đoán địa điểm tiếp theo sẽ là Home (H) và sự chuyển dịch sẽ chuyển từ trạng thái HW sang WH, bởi vì chúng ta cập nhật vị trí trước đó cho W và vị trí hiện thời cho H.

Source/Dest	H	W	L	O
H W	1,00	0,00	0,00	0,00
H L	1,00	0,00	0,00	0,00
H O	0,64	0,34	0,00	0,00
W H	0,00	0,84	0,08	0,08
L H	0,00	0,50	0,00	0,50
O H	0,00	1,00	0,00	0,00
O W	1,00	0,00	0,00	0,00

Bảng 2.1 Ma trận chuyển dịch

Chương 3 Xây dựng hệ thống phân tích, mô phỏng tình trạng giao thông

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 (1), tiến hành chạy thuật toán phân cụm trên từng khung thời gian ta được các cụm cung đường di chuyển theo khung thời gian(2)
- B3: Chia vùng bản đồ Hà Nội thành các vùng ta được đồ thị của các vùng (3)
- B4: Dựa trên đồ thị của các vùng (3) 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 vùng (4)
- 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 (5)

- B7: Dựa trên (5) đưa ra 3 lựa chọn tốt nhất cho tài xế, dựa trên (1) gợi ý cho tài xế cách di chuyển theo các cung đường khác nhau kết nối giữa các vùng

3.1 Các đề xuất

3.1.1 Đề xuất phân vùng bản đồ Hà Nội

Để khái quát hóa các dữ liệu vận tải trong một khu vực, ta tiến hành chia bản đồ hà nội thành các ô (vùng), số ô này có thể được cài đặt theo các thông số:

- - Kinh độ, vĩ độ của điểm phía trên góc trái (điểm bắt đầu)
- Chiều dài, chiều rộng của mỗi ô
- Số lượng các ô theo chiều ngang
- Số lượng các ô theo chiều dọc

3.1.2 Cách tính xếp hạng cho PageRank có trọng số

Dựa trên kết quả nghiên cứu của Bin Jiang và các cộng sự [4] ta thấy rằng: dữ liệu giao thông và di chuyển phù hợp với mô hình PageRank có trọng số do đặc tính của giao thông là các khu vực gần khu vực phát triển, giao thông thuận lợi có xu hướng phát triển (tương tự với tắc đường) nên ta chọn mô hình PageRank có trọng số để biểu diễn dữ liệu giao thông và tính xếp hạng cho các vùng

Dựa trên mô hình PageRank có trọng số [14] ta thực hiện thuật toán PageRank có trọng số cho các mục đích khác nhau với các in-link, out-link là các luồng di chuyển của taxi:

- Số lượng xe: Ta lấy giá trị khởi tạo là số xe trong mỗi vùng khi bắt đầu chạy thuật toán
- Số lượng khách lên xe, xuống xe: Lấy giá trị khởi tạo là số khách lên xe; xuống xe
- Vận tốc: Lấy giá trị khởi tạo là vận tốc trung bình toàn ngày chia cho vận tốc trung bình của vùng, phần này cần xử lý để tránh các vùng có vận tốc trung bình là 0

3.1.3 Sử dụng mô hình n-MMC với các nhãn về xếp hạng

Dựa trên kết quả nghiên cứu của S´ebastien Gambs và các cộng sự và đặc tính của dữ liệu giao thông, ta nhận thấy:

- Các luồng di chuyển giao thông là có quy luật, dựa vào địa điểm lúc trước của một người (một nhóm người) ta có thể dự đoán được điểm tiếp theo
- Dữ liệu giao thông có tính lan truyền (một vùng tắc đường có thể khiến các vùng tiếp theo của luồng di chuyển bị tắc)

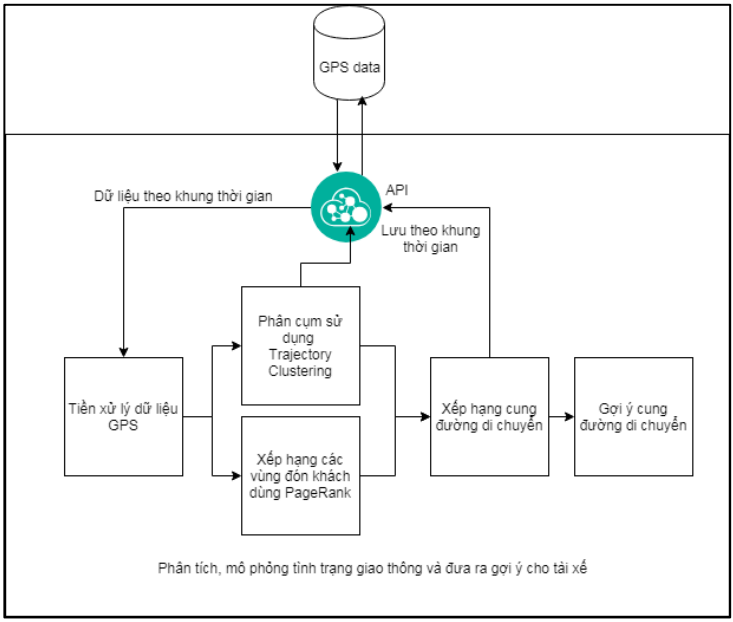
Ta tiến hành gán nhãn các địa điểm của một người (một nhóm người) dựa trên cả vận tốc di chuyển (tắc – thấp – trung bình - cao) hoặc xếp hạng của địa điểm (vùng) đó (thấp – trung bình – cao), cụ thể từ Bảng 2.1 ta tạo thành Bảng chi tiết hơn như sau:

Source/Dest	H thấp	W cao	L thấp	O thấp
H thấp W thấp	1,00	0,00	0,00	0,00
H cao L thấp	1,00	0,00	0,00	0,00
H trung bình O tắc	0,64	0,34	0,00	0,00
W cao H cao	0,00	0,84	0,08	0,08
L trung bình H trung bình	0,00	0,50	0,00	0,50
O cao H thấp	0,00	1,00	0,00	0,00
O thấp W cao	1,00	0,00	0,00	0,00

Bảng 3.1 Bảng ma trận chuyển dịch có thêm nhãn về tốc độ di chuyển

3.2 Tổng quan hệ thống

Hệ thống được thiết kế như sau



Hình 3.1 Hệ thống mô phỏng và đưa ra gợi ý giao thông

Với 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 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:** Module sử dụng thuật toán PageRank để xếp hạng các vùng theo các tiêu chí khác nhau

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

Chương 4 Thử nghiệm và đánh giá

4.1 Tổng quan về dữ liệu sử dụng trong đề tài

4.1.1 Định dạng dữ liệu

Dữ liệu sử dụng trong luận văn là dữ liệu từ các nguồn như sau:

Dữ liệu thiết bị giám sát hành trình của công ty TNHH Phát triển Công nghệ Điện tử Bình Anh với phương tiện là xe taxi (các loại xe khác tương tự ở trạng thái có hàng, không hàng – có khách, không khách) và dữ liệu từ ứng dụng đặt xe, điều phối taxi do chính tác giả luận văn xây dựng

4.1.2 Dữ liệu từ thiết bị GSHT công ty TNHHPTCNĐT Bình Anh

Dữ liệu đầu vào từ thiết bị giám sát hành trình của công ty TNHH Phát triển Công nghệ Điện tử Bình Anh được lưu trong file text, với định dạng như sau:

Đường dẫn đến file text: \<năm: 4 chữ số>\<tháng: 2 chữ số>\<ngày: 2 chữ số>

1 dòng dữ liệu có những thông tin như sau (cách nhau bởi dấu phẩy):

@00:00:17,105.862778,20.992922,0,0,131112,0,0,km(0),vbgt(0),mt()

Trong đó:

@: bắt đầu dòng tin

00:00:17: thời gian trong ngày: giờ: phút: giây

105.862778: Longitude: kinh độ

20.992922: Latitude: vĩ độ

Số 0 ở vị trí thứ 6 là status, sẽ thể hiện trạng thái có khách hay không như sau:

- CÓ KHÁCH = Status & 3 > 0 (phép AND bit)
- KHÔNG KHÁCH = Status & 3 = 0 (phép AND bit)

Dữ liệu tư thiết bị giám sát hành trình của công ty TNHH Phát triển Công nghệ Điện tử Bình Anh gồm 30 ngày, với số xe là 100 xe, tổng dung lượng là 1.33 GB.

4.1.3 Dữ liệu từ ứng dụng đặt taxi, điều phối taxi

Dữ liệu đầu vào từ ứng dụng đặt taxi, điều phối taxi được lưu trong CSDL MongoDB và định dạng như sau:

```
{
    "userPost": "58573bb02714c9029a615c5c",
    "time": 1487138188,
    "lat": 21.0056755,
    "lng": 105.8010069,
    "state": 1,
}
```

Dữ liệu từ ứng dụng đặt xe taxi gồm 23 triệu bản ghi, chiếm dung lượng 3GB, tuy nhiên dữ liệu từ ứng dụng khá rời rạc và nhiều nhiễu

4.1.4 Dữ liệu xử lý trong hệ thống

Sau khi tiền xử lý dữ liệu từ các nguồn dữ liệu, chúng ta thu được dữ liệu đầu vào để chạy thuật toán phân cụm như sau: dữ liệu có 4 cột lần lượt là: vĩ độ (Y), kinh độ (X), ID (ID tương ứng với mỗi taxi), trạng thái khách hàng gồm có 3 trạng thái: 1- không có khách, 2 - trên trường đón khách, và 3- có khách

Vĩ độ (Y)	Kinh độ (X)	ID	Trạng thái khách hàng
21.0300596	105.7889164	0	3
21.0300596	105.7889164	0	3
21.0301935	105.7859652	0	3
21.0301178	105.7896338	0	1
21.0287439	105.7889675	0	1
21.0296401	105.7913306	0	1
21.0671696	105.8348092	0	1
21.0671696	105.8348092	1	1

Bảng 4.1 Dữ liệu đầu vào cho thuật toán phân cụm

Đầu ra sau khi phân cụm trong thuật toán TRACLUS . Dữ liệu sẽ gồm điểm xuất phát (vĩ độ, kinh độ), điểm đích (vĩ độ, kinh độ), ID và ID cụm.

Điểm xuất phát		Điểm đích		ID	ID cụm
Vĩ độ (Y)	Kinh độ (X)	Vĩ độ (Y)	Kinh độ (X)		
21.04617	105.790172	21.046049	105.781655	0	2
21.038296	105.791987	21.032248	105.790092	0	1
21.030695	105.784669	21.030111	105.788194	0	5
21.030111	105.788194	21.03984	105.790379	0	1

Bảng 4.2Ví dụ đầu vào cho bước xử lý kết quả

4.2 Lựa chọn công nghệ

4.2.1 Ngôn ngữ Nodejs

Node.js là một phần mềm mã nguồn mở được viết dựa trên ngôn ngữ JavaScript cho phép lập trình viên có thể xây dựng các ứng dụng chạy trên máy chủ. Ban đầu, Node.js được phát triển bởi Ryan Dahl. Phiên bản đầu tiên của Node.js được cho ra mắt vào năm 2009.

Node.js có thể chạy được trên nhiều nền tảng khác nhau như Windows, Linux hay Mac OS. Node.js được phát triển sử dụng V8 Engine là bộ thư viện JavaScript được Google phát triển để viết trình duyệt web Chrome.

Bản thân Node.js không phải là một ngôn ngữ lập trình mới, thay vào đó Node.js là một nền tảng mã nguồn mở (hay phần mềm mã nguồn mở) được viết dựa trên ngôn ngữ JavaScript.

4.2.2 Ngôn ngữ python

Python là một ngôn ngữ lập trình phổ biến. Được tạo ra bởi Guido van Rossum vào năm 1991.

Ngày nay, Python được sử dụng trong nhiều mục đích, trong luận văn ngôn ngữ python được sử dụng với mục đích phục vụ các tính toán khoa học

Hiện nay, với khả năng xử lý các phép toán phức tạp của mình, Python đang được sử dụng nhiều trong việc phát triển Trí Tuệ Nhân Tạo và các nghiên cứu trong lĩnh vực Machine Learning.

4.2.3 Cơ sở dữ liệu Mongo

MongoDB (bắt nguồn từ "humongous") là một hệ cơ sở dữ liệu NoSQL mã nguồn mở.

Thay cho việc lưu trữ dữ liệu vào các bảng có quan hệ với nhau như truyền thống, MongoDB lưu các dữ liệu cấu trúc dưới dạng giống với JSON(JavaScript Object Notation) và gọi tên là BSON.

Hình 3.2: – So sánh giữa RDBMS và MongoDB

4.3 Kết quả thu được

4.3.1 Môi trường thử nghiệm

Các thuật toán và mô hình hệ thống được xây dựng và thử nghiệm trên các máy tính có cấu hình như sau:

Máy server

- •CPU: Intel(R) Xeon(R) CPU E3-1230 v5 @ 3.40GHz

- RAM: 8 GB
- GPU: Intel HD Graphic
- Hệ điều hành Centos 7

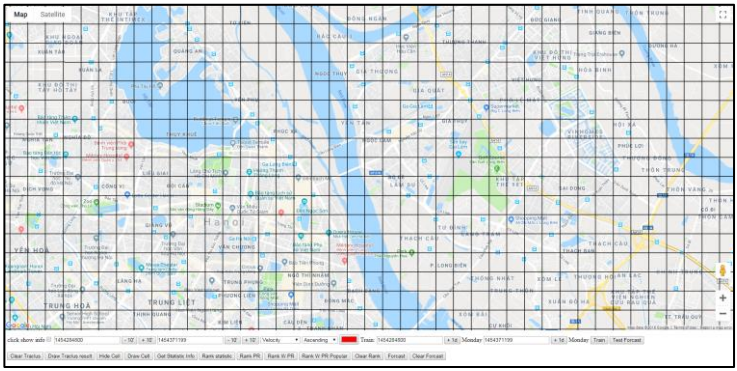
Máy client

- CPU: Intel® Core ™ i5 CPU M520
- RAM: 8 GB
- GPU: ATI mobility Radeaon HD 5730
- Hệ điều hành Win7 Ultimate

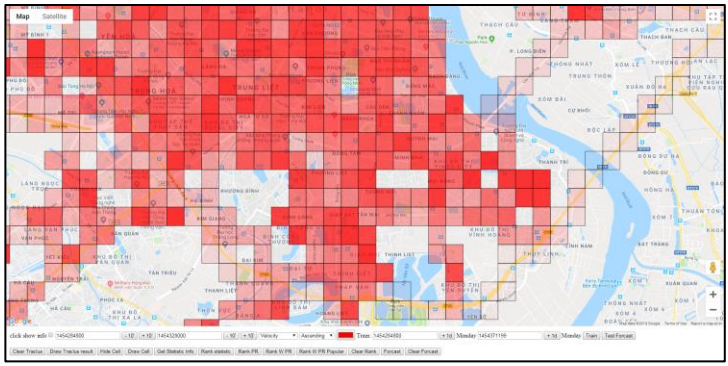
4.3.2 Kết quả thử nghiệm

Các quãng đường mà xe đi qua được phân chia thành các cụm quãng đường nhờ vào thuật toán TRACLUS. Các cụm này sẽ được biểu diễn bởi các màu khác nhau trên Hình 4.3. Nhờ vào view này chúng ta có thể thấy các quãng đường có chung đặc tính (đặc điểm địa lý) sẽ được gom chung vào cùng một cụm. Điều này cho phép phát hiện hành vi cũng hư quy luật di chuyển của taxi.

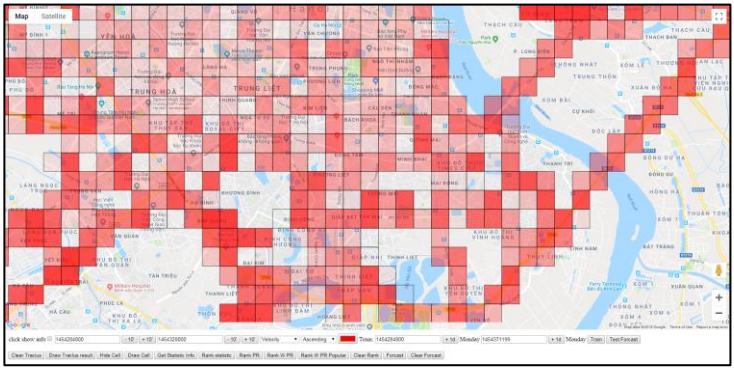
Hình 4.1Kết quả thuật toán TRACLUS trên dữ liệu mẫu



Hình 4.2Hiển thị các tuyến di chuyển và số lượng khách trên mỗi tuyến



Hình 4.8 Xếp hạng các vùng bằng thống kê



Hình 4.3 Xếp hạng các vùng bằng PageRank có trọng số

Hình 4.4 Gợi ý các đoạn đường có thể di chuyển

4.3.3 Tính chính xác của dữ liệu dự đoán

Sử dụng mô hình chung cho các bài toán dự đoán như ở hình 3.2 trên hai nguồn dữ liệu ở mục 4.1 luận văn tiến hành dự đoán điểm đến và mật độ điểm đến tiếp theo dựa trên tập dữ liệu ta thu được kết quả dự đoán chính xác về điểm đến từ 70% - 85% với dữ liệu từ thiết bị giám sát hành trình, và 50 – 73% với dữ liệu từ ứng dụng đặt xe taxi, và chính xác về cả điểm đến và mật độ điểm đến từ 45% - 60% với cả hai bộ dữ liệu.

Với các tham số như trong hình 4.12:

· Miss: Các điểm để dự đoán không nằm trong tập dữ liệu huấn luyện

· Incorrect: Dự đoán sai cả về nhãn và mật độ của điểm đích

· Correct cell: Dự đoán đúng về điểm đến, nhưng sai về mật độ của điểm đích

· Correct: Đúng cả về điểm đến và mật độ

Với cách tính như sau:

· Độ chính xác về cả điểm đến và mật độ = correct/tổng

· Độ chính xác về cả điểm đến = (correct + correct cell)/tổng

KẾT LUẬN

Những vấn đề đã được giải quyết trong luận văn

Luận văn đã tiến hành nghiên cứu giải quyết các bài toán trong Giám sát và điều khiển giao thông. Bài toán này được đánh giá có độ phức tạp cao và có ứng dụng thực tiễn lớn. 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.

Luận văn cũng đã tiến hành xây dựng giao diện trực quan để hiển thị kết quả của các bài toán đặt ra. Luận văn được chạy trên hai bộ dữ liệu thực tế từ hai nguồn dữ liệu khác nhau và đã có một số kết quả nhất định.

Định hướng nghiên cứu trong tương lai

Tiến hành khắc phục tình trạng thiếu chính xác do dữ liệu thưa, đặc biệt là dữ liệu từ các ứng dụng di động. Tiến hành xây dựng hệ thống gợi ý theo hướng tiếp cận học tăng cường.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[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

[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

Tiếng Anh

[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

[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

[5]. Daniel Jurafsky & James H. Martin (2006) "Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition", Chapter 6

[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.

[7]. Jiang Bian, Dayong Tian, Yuanyan Tang, Dacheng Tao (2018), "A survey on trajectory clustering analysis"

[8]. Naoto Mukai (2013) "PageRank-based Traffic Simulation Using Taxi Probe Data", Procedia Computer Science, 2013. 22: pp. 1156-1163.

[9]. Raj Kishen Moloo, Varun Kumar Digumber (2011) "Low-Cost Mobile GPS Tracking Solution", 2011 International Conference on Business Computing

and Global Informatization

[10]. Sameer Darekar, Atul Chikane, Rutujit Diwate, Amol Deshmukh, Prof. Archana Shinde (2012) “Tracking System using GPS and GSM: Practical Approach”, IJSER journal

[11]. S´ebastien Gambs, Marc-Olivier Killijian, Miguel N´u˜nez 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

[12]. S´ebastien Gambs, Marc-Olivier Killijian, Miguel N´u˜nez 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

[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

[14]. Wenpu Xing, Ali Ghorbani (2004) “Weighted PageRank Algorith Proceedings of the Second Annual Conference on Communication Networks and Services Researchm”

[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): e0145348.