Xây dựng hệ thống thời gian thực dự đoán lưu lượng người trong một khu vực

Nguyễn Hoài Phương Uyên, Huỳnh Phan Minh Quang, Lê Thị Hồng Oanh

Trường Đại học Công nghệ Thông tin - Đại học Quốc gia Thành phố Hồ Chí Minh
{18521627,18520140,18521225}@gm.uit.edu.vn

Tóm tắt nội dung Mặc dù thời đại công nghệ số bùng nổ việc tiếp cận khách hàng thông qua internet ngày càng trở nên phổ biến và dễ dàng hơn, nhưng không vì vậy mà quảng cáo ngoài trời dần mất đi thị phần của chính nó. Bởi đối với các hình thức quảng cáo online đòi hỏi doanh nghiệp phải tuân thủ những quy định về thời gian cũng như việc khách hàng có thể bỏ qua quảng cáo. Tuy nhiên đối với quảng cáo ngoài trời dù khách hàng có không thích thì vẫn phải chú ý tới của quảng cáo doanh nghiệp. Nói đến quảng cáo ngoài trời, yếu tố quan trọng nhất để đảm bảo cho sự thành công của chiến dịch chính là vị trí lựa chọn quảng cáo. Nhận thấy vấn đề này, chúng tôi tiến hành làm bài toán về dự đoán số lượng người đi bộ qua một khu vực cụ thể. Trong bài báo này, chúng tôi sẽ mô tả quá trình xây dựng một hệ thống dự đoán số người sẽ đi qua vòng 1 tiếng tại một khu vực. Bài toán này được thực hiện bằng cách ứng dụng công nghệ xử lý dữ liệu bằng cách sử dụng mô hình Linear regression để đào tạo dữ liệu, sau đó tiến hành dự đoán realtime bằng cách ứng dụng framework kafka. Kết quả đào tạo mô hình được đo bởi RMSE đạt 52,48 %. Kết quả dự đoán sẽ được cập nhật lên một trang web mà chúng tôi đã xây dựng.

Keywords: Dự đoán số lượng người · Máy học · Công nghệ dữ liệu lớn.

1 Giới thiệu

Quảng cáo ngoài trời là kênh quảng cáo quốc dân khi có thể tiếp cận với số lượng lớn và đa dạng khách hàng. Người người nhà nhà đều có thể nhìn thấy tiếp nhận những nội dung được truyền tải qua quảng cáo, không phân biệt lứa tuổi, giới tính, sở thích, thu nhập. Những biển quảng cáo tọa lạc tại vị trí đông đúc người qua lại hay những phương tiện giao thông được dán quảng cáo như taxi, xe buýt hằng ngày đi qua biết bao nhiêu cung đường, tiếp cận với rất nhiều người, lên tới hàng trăm hàng nghìn người. Bên canh việc truyền tải hình ảnh, thông điệp một cách sáng tạo, sinh động, bắt mắt, việc lựa chọn vị trí là một trong những yếu tố quyết định thành công của chiến dịch quảng cáo. Việc dự đoán số lượng người đi qua theo khung giờ có thể giúp doanh nghiệp đưa ra phán đoán lựa chọn khung giờ thực hiện chiến dịch quảng cáo.
Do đó trong đề tài này, chúng tôi đã tiến hành xây dựng một hệ thống ứng dụng công nghệ dữ liệu lớn để dự báo lưu lượng người qua từng khung giờ trong tương

lai.
Cụ thể chúng tôi huấn luyện dữ liệu bằng mô hình Linear Regression, sau đó chúng tôi thu thập dữ liệu từ kafka và đưa qua mô hình đã được huấn luyện trước đó để dự báo và cuối cùng cập nhật kết quả dự báo lên trang web mà chúng tôi đã xây dựng.
Mục tiêu đặt ra của bài toán này là xây dựng một hệ thống dự đoán lưu lượng người đi qua khu vực cụ thể từ bộ dữ liệu đã được thu thập sẵn. Từ đó, áp dụng và đánh giá hiệu suất mô hình học máy để tiến hành dự đoán số lượng người.
Nội dung bài báo của chúng tôi gồm 6 phần. Phần 1 giới thiệu đề tài và động lực để nghiên cứu bài toán trên. Phần 2 khảo sát các công trình liên quan về bài toán. Tiếp theo là Phần 3 trình bày chi tiết về bộ dữ liệu. Phần 4 trình bày phương pháp máy học mà chúng tôi sẽ áp dụng trên bộ dữ liệu trên. Phần 5 trình bày kết quả đạt được của chúng tôi khi áp dụng các mô hình máy học trên bộ dữ liệu trên và giới thiệu công nghệ spark và kafka. Cuối cùng là Phần 6 sẽ kết luận lại bài toán và đề xuất ra hướng phát triển tiếp theo.

2 Các công trình nghiên cứu liên quan

Trên thực tế, đề tài dự đoán lưu lượng người qua một khu vực được lấy ý tưởng khi chúng tôi tìm thấy bộ dữ liệu. Sau quá trình tìm kiếm và tham khảo chúng tôi tìm thấy những công trình nghiên cứu dự đoán lưu lượng xe taxi tại một khu vực. Về chủ đề tuy khác nhau nhưng phương thức và quy trình thực hiện lại giống nhau.

2.1 Những nghiên cứu trên thế giới

Vào năm 2017, Jun Xu cùng các cộng sự của mình đã dự đoán lượng taxi trong một khu vực để thực hiện điều phối xe taxi và phân phối một cách hiệu quả theo nhu cầu toàn thành phố. [1]Phân tích dữ liệu ch thấy rằng có sự lặp lại các mẫu trong dữ liệu có thể giúp dự đoán nhu cầu trong một khu vực cụ thể tại một thời điểm cụ thể. Tác giả chia một thành phố lớn vào các khu vực nhỏ hơn và tổng hợp số lượng yêu cầu taxi trong mỗi khu vực trong một khoảng thời gian nhỏ (ví dụ: 20 phút). Trong này theo cách khác, dữ liệu taxi trong quá khứ trở thành một chuỗi dữ liệu về số lượng yêu cầu taxi trong từng khu vực. Sau đó, tác giả đào tạo một thời gian dài ngắn hạn bằng mô hình LSTM) và mô hình RNN. Kết quả dự đoán chính xác lên đến 83%.

3 Bộ dữ liệu

Trong đề tài này chúng tôi sử dụng tập dữ liệu của thành phố Melbourne , Úc. [2] Pedestrian Counting System - Monthly (counts per hour) Tập dữ liệu này chứa số lượng người đi bộ theo giờ. Được thu thập từ năm 2009 từ các thiết bị cảm biến dành cho người đi bộ được đặt trên khắp thành phố. Ở tập dữ liệu này,chúng tôi sử dụng dữ liệu 3 sensor 69,68,67 được thu thập từ năm 2009 đến

ngày 1 tháng 12 năm 2021 để huấn luyện mô hình, trong đó trong đó khu vực sensor69 là khu vực dự đoán và sử dụng dữ liệu từ ngày 1 tháng 12 năm 2021 trở đi làm tập thử nghiệm dự báo theo thời gian thực. Thông tin về 3 sensor:

Sensor$_id$	Tên sensor	Mô tả
67	FLDegS$_T$	Flinders Ln – Degraves St (South)
68	FLDegN$_T$	Flinders Ln – Degraves St (North)
69	FLDegC$_T$	Flinders Ln – Degraves St (Crossing)

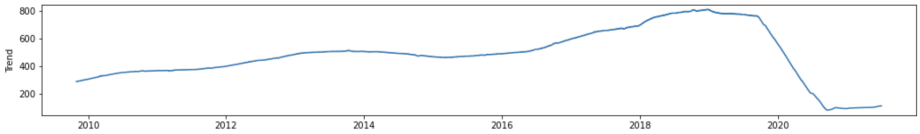
Bảng 1. Thông tin sensor 67, 68, 69

ID	Date_Ti...	Year	Month	Mdate	Day	Time	Sensor_ID	Sensor_...	Hourly_C...
2887628	November 0...	2019	November	1	Friday	17	34	Flinders St-S...	300
2887629	November 0...	2019	November	1	Friday	17	39	Alfred Place	604
2887630	November 0...	2019	November	1	Friday	17	37	Lygon St (East)	216
2887631	November 0...	2019	November	1	Friday	17	40	Lonsdale St-S...	627
2887632	November 0...	2019	November	1	Friday	17	36	Queen St (W...	774
2887633	November 0...	2019	November	1	Friday	17	29	St Kilda Rd-Al...	644
2887634	November 0...	2019	November	1	Friday	17	42	Grattan St-S...	453
2887635	November 0...	2019	November	1	Friday	17	43	Monash Rd-S...	387
2887636	November 0...	2019	November	1	Friday	17	44	Tin Alley-Swa...	27
2887637	November 0...	2019	November	1	Friday	17	35	Southbank	2,691
2887638	November 0...	2019	November	1	Friday	17	45	Little Collins ...	2,173
2887639	November 0...	2019	November	1	Friday	17	46	Pelham St (S)	203
2887640	November 0...	2019	November	1	Friday	17	47	Melbourne C...	2,354

Hình 1. Bộ dữ liệu

Thách thức bộ dữ liệu: Sau khi nghiên cứu bộ dữ liệu, chúng tôi đã trực quan bộ dữ liệu:

Chúng tôi nhận thấy, dữ liệu có xu hướng giảm dần từ năm 2020. Điều này là do tác động của dịch bệnh covid 19. Chính vì vậy việc dự đoán mô hình sẽ bị ảnh hưởng.



Hình 2. Xu hướng dữ liệu

4 Kết luận và hướng phát triển

5 Phương pháp tiếp cận

5.1 Công nghệ dữ liệu lớn

Apache Spark [3] là một framework mã nguồn mở tính toán cụm, được phát triển sơ khởi vào năm 2009 bởi AMPLab. Sau này, Spark đã được trao cho Apache Software Foundation vào năm 2013 và được phát triển cho đến nay. Tốc độ xử lý của Spark có được do việc tính toán được thực hiện cùng lúc trên nhiều máy khác nhau. Đồng thời việc tính toán được thực hiện ở bộ nhớ trong (in-memories) hay thực hiện hoàn toàn trên RAM. Spark cho phép xử lý dữ liệu theo thời gian thực, vừa nhận dữ liệu từ các nguồn khác nhau đồng thời thực hiện ngay việc xử lý trên dữ liệu vừa nhận được (Spark Streaming) Apache Spark có thể chạy nhanh hơn 10 lần so với Haddop ở trên đĩa cứng và 100 lần khi chạy trên bộ nhớ RAM. Apache Spark gồm có 5 thành phần chính : Spark Core, Spark Streaming, Spark SQL, MLlib và GraphX

Apache Kafka Kafka thường được sử dụng trong các kiến trúc dữ liệu phát trực tuyến thời gian thực (real time) để cung cấp các phân tích thời gian thực. Vì Kafka là một hệ thống nhắn tin Pub - Sub nhanh, có thể mở rộng, bền và có khả năng chịu lỗi cao, nên nó được sử dụng trong những trường hợp xử lý khối lượng lớn dữ liệu đến và đáp ứng được khả năng phản hồi ngay lập tức. Kafka có các đặc tính thông lượng, độ tin cậy và sao chép cao hơn, có thể áp dụng cho những thứ như theo dõi các cuộc gọi dịch vụ (theo dõi mọi cuộc gọi). Mô hình cấu trúc Kafka bao gồm các thành phần. sau:Message,Broker,Topic,Partition,Producer,Consumer

5.2 Mô hình

Linear Regression là một phương pháp để dự đoán biến phụ thuộc (Y) dựa trên giá trị của biến độc lập (X). Nó có thể được sử dụng cho các trường hợp chúng ta muốn dự đoán một số lượng liên tục. Ví dụ, dự đoán giao thông ở một cửa hàng bán lẻ, dự đoán thời gian người dùng dừng lại một trang nào đó hoặc số trang đã truy cập vào một website nào đó. Trong bài toán này tôi ứng dụng mô hình Linear Regression của Spark MLlib để dự đoán kết quả dự báo cuối cùng

từ các đặc trưng: số lượng người đi qua khu vực sensor 67,68,69 ở thời điểm hiện tại.

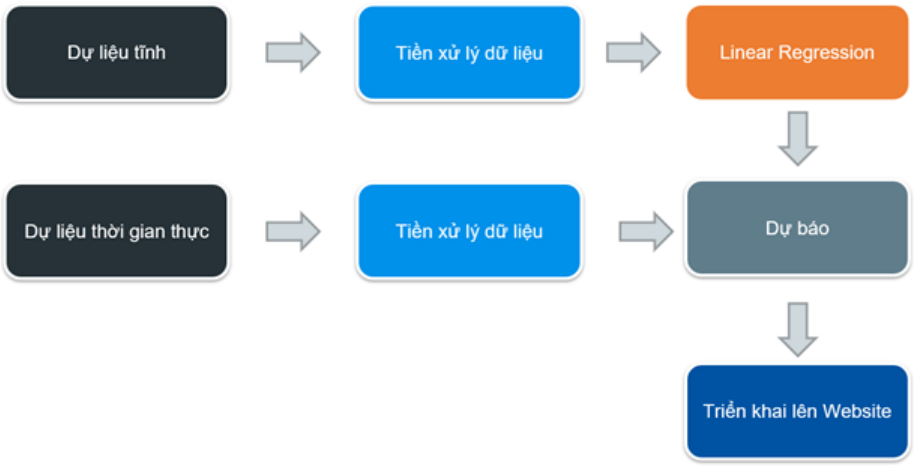
5.3 Tiền xử lý dữ liệu

Dữ liệu khi lấy về chúng tôi sẽ tiến hành xữ lý lọc lấy dữ liệu sensor 67,68,69, những dữ liệu số lượng người không âm.
Dữ liệu thời gian thực chúng tôi xử lý chuyển dữ liệu thời gian sang dạng timestap để làm điều kiện kết hợp các sensor với nhau.

5.4 Lấy thời gian hiện tại

Trong dữ liệu thời gian thực tôi mặc định dữ liệu số lượng người đi qua ba khu vực sensor 67,68,69 là dữ liệu ở thời điểm hiện hiện tại và dùng nó dự báo tương lại. Trong dữ liệu huấn luyện mô hình , để lấy dữ số lượng người đi qua ba khu vực sensor 67,68,69 hiện tại tôi xử lý hạ 1 bậc để lấy dữ liệu ở dòng trước làm dữ liệu hiện tại cho dòng sau.

6 Thiết kế hệ thống



Hình 3. Sơ đồ thiết kế hệ thống

Bước 1: Lấy dữ liệu liệu số lượng người đi qua khu vực sensor 67,68,69
Bước 2: Tiền xử lý dữ liệu
Bước 3: Xây dựng mô hình Linear Regression

- Đầu vào: Số lượng người đi qua khu vực sensor 67,68,69 hiện tại.
- Đầu ra: Số lượng người đi qua khu vực sensor 69

Bước 4: Thu thập dữ liệu số lượng người đi qua khu vực sensor 67,68,69 thời gian thực
Bước 5: Tiền xử lý dữ liệu
Bước 6: Dự báo bằng mô hình Linear Regression đã huấn luyện

- Đầu vào: số lượng người đi qua khu vực sensor 67,68,69 hiện tại.
- Đầu ra: Kết quả dự báo của mô hình Linear Regression

Bước 7: Triển khai lên website.

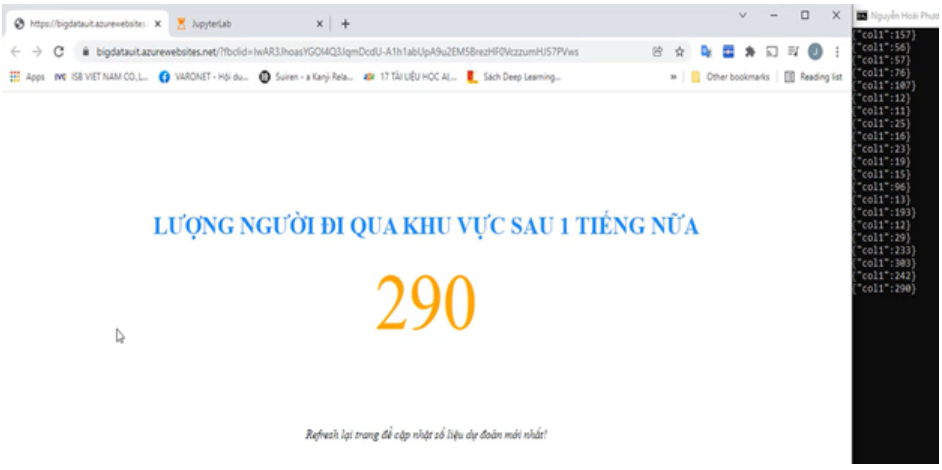
7 Kết quả

Kết quả dự báo trên tập thử nghiệm

Mô hình	RMSE
Linear Regression	52.48

Bảng 2. Kết quả dự báo trên tập thử nghiệm

Kết quả dự báo trên dữ liệu thời gian thực



Hình 4. Kết quả dự đoán số lượng người thời gian thực

8 Kết luận và hướng phát triển

Chúng tôi đã xây dựng thành công hệ thống dự đoán số lượng người đi qua một khu vực có ứng dụng công nghệ dữ liệu lớ.
Trong tương lai chúng tôi sẽ tiếp tục xây dựng hệ thống mở rộng dự báo trên nhiều khu vực và cải thiện độ chính xác của dự báo.

Tài liệu

1. Jun Xu , Rouhollah Rahmatizadeh, Ladislau Bölöni, F.: "Real-Time Prediction of Taxi Demand Using Recurrent Neural Networks" (2007)
2. "Pedestrian Counting System - Monthly (counts per hour)"
3. MLlib: Machine Learning in Apache Spark(2016)