

DỰ BÁO TÌNH TRẠNG GIAO THÔNG TRONG ĐÔ THỊ DỰA VÀO DỮ LIỆU TỪ CỘNG ĐỒNG

ESTIMATING TRAFFIC CONDITION BASED ON CROWDSOURCED DATA

¹Trần Minh Quang*, ²Mai Tấn Hà, ³Phạm Nguyễn Hoàng Nam, ⁴Nguyễn Xuân Long
^{1,2,3}Khoa Khoa học và Kỹ thuật máy tính, ⁴Khoa Kỹ thuật xây dựng,
 Trường Đại học Bách Khoa – Đại học Quốc gia Thành phố Hồ Chí Minh
¹*quangtran@hcmut.edu.vn, ²maitanhaksdtvt6@gmail.com,
³phamnguyenhoangnam@gmail.com, ⁴nxlolong@hcmut.edu.vn

Tóm tắt: Ùn tắc giao thông (UTGT) gây ảnh hưởng tiêu cực đến kinh tế, xã hội và là một trong những vấn đề nan giải ở các đô thị lớn. Dự báo tình trạng giao thông (TTGT) một cách chính xác và kịp thời là một trong những giải pháp hữu hiệu góp phần giải quyết bài toán UTGT và hoàn thiện hệ thống giao thông thông minh (ITS) đô thị. Để giải quyết vấn đề này ta cần phải thu thập dữ liệu giao thông liên tục, ở mọi đoạn đường trên hệ thống giao thông của thành phố. Đây là một yêu cầu khó đáp ứng được trong thực tế do các hệ thống cảm biến cố định (ví dụ các hệ thống camera giao thông) khó có thể được triển khai trên diện rộng vì các ràng buộc về kỹ thuật và chi phí, nhất là trong bối cảnh hạ tầng giao thông đô thị chưa phát triển ở các nước đang phát triển như ở Việt Nam. Nghiên cứu này đề xuất giải pháp giải quyết các vấn đề nêu trên, theo đó một khung thức (framework) thu thập dữ liệu giao thông dựa vào cộng đồng được đề xuất và phát triển. Hơn nữa, các giải pháp phân tích dữ liệu trong quá khứ để dự báo TTGT ở những nơi bị thiếu dữ liệu cũng được đề xuất nhằm nâng cao hiệu quả và độ bao phủ của hệ thống. Các kết quả thử nghiệm thực tế cho thấy tính hiệu quả cũng như tính khả thi của các giải pháp đề xuất.

Từ khóa: Giao thông thông minh, ùn tắc giao thông, cảnh báo tình trạng giao thông, dữ liệu từ cộng đồng, khai phá dữ liệu.

Mã phân loại: 8.2

Abstract: Traffic congestion leads to negative effects on social-economic development and is one of the serious problems in large cities. Estimating traffic conditions accurately and timely is one of the essential solutions to resolve this issue, contributing to the implementation of intelligent transportation systems (ITS). In order to reach this objective we need to collect traffic data continuously at every road segment in the road network. This requirement is almost practically impossible since it is hard to deploy fixed sensor systems (e.g., camera systems) at every location because of the technical and operational costs, especially in developing countries like Vietnam where the traffic infrastructure is not developed well. This paper proposes appropriate approaches where a crowdsourced data traffic estimation framework is devised. In addition, traffic prediction based on historical data to resolve the issues of missing real-time data and increasing the system coverage is proposed. Experimental results reveal the effectiveness and the feasibility of the proposed mechanisms.

Keywords: Intelligent transportation systems (ITS), traffic congestion, traffic condition estimation; crowdsourced data, data mining.

Classification code: 8.2

1. Giới thiệu

UTGT luôn là vấn đề thách thức ở các thành phố lớn, nó không chỉ gây thiệt hại về kinh tế mà còn gây ô nhiễm môi trường và các vấn đề xã hội khác [1]. Xe cấp cứu bị kẹt trong vô vọng, những người căng thẳng di chuyển trên những con đường hỗn loạn,... là những ví dụ cụ thể về môi trường giao thông khó khăn, làm suy giảm chất lượng cuộc sống. Vấn đề này càng nghiêm trọng hơn ở

các nước đang phát triển do cơ sở hạ tầng giao thông chưa được phát triển tốt để đáp ứng nhu cầu ngày càng tăng nhanh chóng [2]. Do đó, dự báo và giám sát TTGT nhằm góp phần giảm thiểu UTGT trở thành vấn đề được ưu tiên giải quyết, đặc biệt là để hiện thực hóa hệ sinh thái đô thị thông minh (smart city) mà nhiều thành phố lớn ở Việt Nam như Thành phố Hồ Chí Minh (TP.HCM) và Hà Nội đang quan tâm. Các cơ quan ban ngành đang chi một khoản đầu tư

lớn cho chính sách giao thông vận tải, cơ sở hạ tầng đường bộ cũng như hệ thống máy tính để giải quyết vấn đề này. Trên thực tế, phát triển hệ thống giao thông thông minh (ITS) được chọn làm chủ đề ưu tiên cao nhất trong quá trình hiện thực hóa các ứng dụng đô thị thông minh tại TP.HCM [3, 4].

Việc dự báo nhằm giảm thiểu UTGT đòi hỏi thông tin cập nhật (kịp thời) và chính xác về TTGT xung quanh thành phố để hệ thống phân tích và phổ biến thông tin giao thông hữu ích cho người tham gia giao thông từ đó giúp họ tránh định tuyến đến các vị trí tắc nghẽn. Vì vậy, dữ liệu đầy đủ, ví dụ: thời gian, vị trí (tức là ở đoạn đường nào), mức độ tắc nghẽn, lý do gây ra tắc nghẽn, dự đoán khi nào tắc nghẽn sẽ được giải tỏa, hình ảnh về đoạn tắc nghẽn,... cần được thu thập và phân tích thích hợp. Thông thường, dữ liệu này được kết hợp với các hệ thống đài quan sát cố định như hệ thống cảm biến cố định cài đặt trên các con đường chẳng hạn như hệ thống máy dò vòng lặp (loop detectors), RFID, hệ thống camera,... [5]. Tuy nhiên, các hệ thống này tiêu tốn nhiều chi phí triển khai và bảo trì trong khi phạm vi phủ sóng bị hạn chế vì không thể triển khai các hệ thống như vậy ở khắp mọi nơi trong thành phố. Ở các nước đang phát triển, các hệ thống dựa trên IoT như vậy chưa được triển khai đồng bộ. Do vậy ta cần có cách cách tiếp cận thích hợp, tận dụng nhiều nguồn dữ liệu khác nhau từ cộng đồng bao gồm dữ liệu từ các hệ thống IoT cố định (fixed-IoT), các hệ thống cảm biến gắn kèm với con người để thu thập dữ liệu về giao thông một cách kịp thời, tại chỗ và đầy đủ.

Bài báo này đề xuất một khung thức mới nhằm ước tính TTGT dựa trên nguồn dữ liệu cung cấp từ cộng đồng, ví dụ dữ liệu được chia sẻ bởi người dùng di động và dữ liệu từ các hệ thống cảm biến cố định có sẵn. Để đạt được điều đó, hai thách thức cơ bản phải được giải quyết như sau:

- Đơn giản hóa việc tích hợp nhiều nguồn dữ liệu từ cộng đồng. Do đó, ta cần đề xuất một khung thức (framework) có thể kết hợp hiệu quả và tự động các dữ liệu cần thiết để ước tính TTGT;

- Tổng hợp hiệu quả và phân tích lượng lớn dữ liệu để đưa ra các cảnh báo về TTGT một cách chính xác và kịp thời.

Từ những vấn đề trên, nghiên cứu đề xuất hướng giải quyết như sau:

- Nhóm nghiên cứu đã phân tích kỹ lưỡng và thiết kế một khung thức cho hệ thống dự báo TTGT dựa trên nguồn cung cấp dữ liệu từ cộng đồng;

- Đề xuất phương pháp khai phá dữ liệu để dự báo TTGT ngay tại những nơi dữ liệu thời gian thực về giao thông bị thiếu, từ đó góp phần nâng cao tính bao phủ của hệ thống đề xuất;

- Hệ thống nguyên mẫu bao gồm một ứng dụng trên thiết bị di động nhằm thu thập dữ liệu và phổ biến thông tin kết hợp với các chức năng phân tích và ước tính lưu lượng tại trung tâm đã được phát triển.

Cấu trúc của bài báo được tổ chức như sau: Các công trình nghiên cứu liên quan được khảo sát trong phần 2. Kiến trúc tổng thể và các thành phần được trình bày trong phần 3. Phần 4 mô tả cách tiếp cận dự báo TTGT dựa vào dữ liệu lịch sử, phần 5 trình bày kết quả thử nghiệm và đánh giá hệ thống. Phần 6 là các kết luận rút ra từ nghiên cứu này.

2. Công trình nghiên cứu liên quan

Trong nước đã có một số công trình nghiên cứu liên quan nhằm góp phần giảm thiểu UTGT với những hỗ trợ của công nghệ thông tin và truyền thông (ICT) như chương trình thông tin giao thông của Đài Tiếng nói Việt Nam (VOV) [6] và Đài Tiếng nói Nhân dân TP.HCM (VOH) [7]; cổng thông tin giao thông của Sở Giao thông vận tải TP.HCM (Sở GTVT TP.HCM) [8]; các nghiên cứu từ các trường đại học lớn như nghiên cứu tại trường Đại học Bách khoa - Đại học Quốc gia TP.HCM (ĐH BK TP.HCM) về các giải pháp nền tảng cho giao thông thông minh [9]; hệ thống mạng lưới camera giám sát và điều khiển giao thông [10]; hệ thống giám sát và xử lý vi phạm giao thông [4], [11]. Các công trình này đã bước đầu đem lại những kết quả đáng khích lệ và đưa ra các định hướng nghiên cứu nhằm giảm UTGT cho TP.HCM. Tuy nhiên, các chức năng xử lý

trong các hệ thống hiện có vẫn còn khá thô sơ, và đây cũng là những điểm yếu cần được đầu tư nghiên cứu nhằm nâng cao sức mạnh của các hệ thống giám sát và cảnh báo giao thông ở Việt Nam và đặc biệt là ở TP.HCM. Cụ thể, cơ chế phân tích dữ liệu của hệ thống thông tin giao thông của VOV [6] và VOH [7] còn đơn giản, chủ yếu dựa vào con người (quan sát hệ thống camera ở trung tâm, các phân tích của phóng viên thông qua tương tác với cộng tác viên). Các chức năng hỗ trợ như tìm đường thay thế, tìm đường đi ngắn nhất, cập nhật thông tin trên ứng dụng di động, vẫn chưa được phát triển.

Các hệ thống đề xuất bởi ĐHBK TP. HCM [9] tập trung vào phân tích TTGT từ dữ liệu GPS của xe buýt, dữ liệu này vốn chứa nhiều thông tin nhiễu (do sai số trên GPS). Do vậy, cần phải có nhiều giải thuật phân tích tinh vi để phân tích dữ liệu này. Các giải thuật này nhìn chung là khá phức tạp. Tính hiệu quả của hệ thống sẽ tốt hơn nếu có cơ chế thu thập dữ liệu chi tiết và liên quan mật thiết với TTGT hơn như cơ chế thu thập dữ liệu trên ứng dụng di động về thông tin giao thông của Sở GTVT TP. HCM [8]. Tuy nhiên, các chức năng phân tích chuyên sâu dựa vào thống kê, khai phá dữ liệu nhằm đánh giá TTGT trong điều kiện thiếu dữ liệu hiện thời (real-time data) vẫn chưa được đào sâu nghiên cứu.

Trên thế giới, giải pháp giảm UTGT cũng được nghiên cứu nhiều. Những nghiên cứu này tập trung vào giải quyết các vấn đề lớn từ thu thập dữ liệu (data acquisition), truyền thông dữ liệu (data communications) từ các hệ thống thu thập đến trung tâm xử lý số liệu, phân tích dữ liệu (data analysis) tại trung tâm nhằm cung cấp thông tin chi tiết, kịp thời và đa chiều về UTGT giúp nhà quản lý cũng như người dân có thể triển khai các biện pháp ứng phó kịp thời. Các hệ thống này có thể được kể đến như hệ thống VICS (Vehicle Information and Communication System) [13], hay NAVITIME [14] của Nhật Bản, hệ thống ITS của bang Kansas, Mỹ [15]. Phần lớn các nghiên cứu này sử dụng công nghệ mạng cảm biến cố định được triển khai trên các cung đường cần giám sát. Hệ thống các cảm biến này có thể là hệ thống sử

dụng vòng cảm biến từ trường (loop detector) [16 - 17], hệ thống phát hiện phương tiện dựa vào công nghệ định danh sóng radio RFID [18 - 19], hệ thống mạng camera [20],...

Tương tự như các hệ thống chủ yếu dựa vào các thiết bị cảm biến, dữ liệu từ các phương tiện và tín hiệu vô tuyến đã được nghiên cứu. Nghiên cứu [33] đề xuất một hệ thống giám sát tắc nghẽn giao thông bằng cách kết nối dữ liệu của các phương tiện, bao gồm thu thập dữ liệu, thiết lập cấu trúc phân đoạn, mô hình lưu lượng giao thông, dự đoán tắc nghẽn giao thông cục bộ và dịch vụ tắc nghẽn giao thông đi kèm đến cho người lái xe. M. Lewandowski và cộng sự đề xuất một hệ thống giám sát tắc nghẽn giao thông bằng cách phân tích cường độ tín hiệu vô tuyến mà các thiết bị di động nhận được từ các đèn hiệu được đặt ở hai bên đường [34]. Nghiên cứu [35] đề xuất một hệ thống giám sát giao thông đường bộ sử dụng nhiều phương tiện bay không người lái, đây là một giải pháp thú vị nhằm giám sát giao thông đường bộ. Nghiên cứu [36] đề xuất một hệ thống theo dõi và phổ biến các điều kiện giao thông bằng cơ sở hạ tầng phi tập trung (DOCTraMS) với dữ liệu giao thông được áp dụng từ Sao Paulo, Brazil.

Bên cạnh đó, có nhiều nghiên cứu tập trung vào khảo sát tài liệu học thuật về nguồn dữ liệu được chia sẻ từ đám đông [29], E. Estellés-Arolas và cộng sự đã tìm thấy gần bốn mươi cách hiểu khác nhau về nguồn cung dữ liệu từ đám đông. Song song với đó, những nghiên cứu khác như [30 - 31] cho thấy những tiến bộ của công nghệ di động đã cho phép việc tìm nguồn cung dữ liệu từ đám đông thông qua di động trong các lĩnh vực khác nhau bao gồm sản xuất, cung cấp dịch vụ, chia sẻ dữ liệu. Phương pháp tiếp cận nguồn cung ứng đám đông đã được sử dụng rộng rãi trong quản lý thảm họa và chia sẻ thông tin trong các công ty hoặc cộng đồng [32, 33].

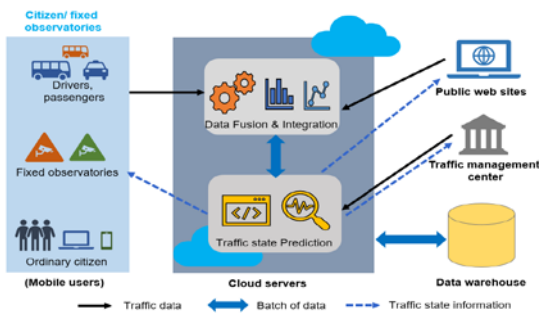
Trong những năm gần đây, những nghiên cứu sử dụng công nghệ mạng di động, thiết bị di động (điện thoại thông minh, máy tính bảng, ...) để đo đạc, thu thập dữ liệu liên quan đến giao thông [21 - 24] đã cho thấy

được sự hiệu quả và tính khả thi khi áp dụng vào thực tế. Do thiết bị di động xuất hiện ở mọi nơi và hầu như đã được phổ cập đến mọi người dân nên các vấn đề về độ bao phủ, chi phí vận hành sẽ được giải quyết nếu có một mô hình thu thập, xử lý và phân tích dữ liệu hiệu quả. Bên cạnh đó, tính hiệu quả của phương pháp sử dụng dữ liệu từ đám đông được thể hiện thông qua các nghiên cứu trong [30– 33].

3. Hệ thống dự báo tình trạng giao thông dựa vào dữ liệu từ cộng đồng

3.1. Tổng quan hệ thống

Hình 1 thể hiện tổng quan về hệ thống dự báo TTGT dựa trên dữ liệu từ cộng đồng được đề xuất trong nghiên cứu này. Kiến trúc tổng thể của hệ thống bao gồm bốn thành phần chính sau: nguồn dữ liệu, tổng hợp và tích hợp dữ liệu, phân tích và dự đoán tình trạng giao thông, và cuối cùng là kho dữ liệu. Các thành phần này được mô tả chi tiết như dưới đây.



Hình 1. Mô hình tổng quan của hệ thống dự báo TTGT dựa vào dữ liệu từ cộng đồng.

3.1.1. Nguồn dữ liệu

Dữ liệu giao thông được thu thập từ nhiều nguồn như sau:

1/ Từ các hệ thống quan sát cố định như hệ thống camera của thành phố được đặt tại các tuyến đường;

2/ Dữ liệu được thu thập từ hệ thống bên thứ 3 như: VoH [7], TomTom [26], và từ các trang web cho phép người dùng truy cập công khai như Google [27];

3/ Dữ liệu được chia sẻ từ cộng đồng thông qua ứng dụng di động mà chúng tôi phát triển để thu thập và cảnh báo TTGT đến người dùng.

3.1.2. Tổng hợp và phân tích dữ liệu

Dữ liệu giao thông được cung cấp bởi cộng đồng sẽ được thu thập, tích hợp và xác định trước khi được phân tích bởi máy tính phân tích dữ liệu ở trung tâm. Các kỹ thuật làm sạch dữ liệu, loại bỏ dữ liệu thừa, bổ sung dữ liệu thiếu, đánh giá tính hợp lệ của dữ liệu sẽ được thực hiện trong bộ phận này.

3.1.3. Phân tích và dự đoán tình trạng giao thông

Sau khi dữ liệu được thu thập và tập hợp về trung tâm, bộ phận phân tích và dự báo lưu lượng giao thông sẽ phân tích các dữ liệu này nhằm đánh giá và dự đoán TTGT một cách chính xác và kịp thời. Thông tin giao thông sẽ được cập nhật trên bản đồ giao thông của thành phố và sẽ được hiển thị trên ứng dụng di động hoặc trên trang web cảnh báo TTGT để phục vụ cho người dân trong thành phố. Khi nắm được tình hình giao thông hiện tại, người dân sẽ có kế hoạch định tuyến phù hợp thông qua ứng dụng di động hoặc ứng dụng web. Thông tin về TTGT này sẽ được lưu trữ trong kho dữ liệu để phục vụ cho việc phân tích dữ liệu, báo cáo trong tương lai.

3.1.4. Kho dữ liệu

Để giảm thời gian phản hồi, chúng tôi chia dữ liệu lưu lượng thành các phiên bản khác nhau: cơ sở dữ liệu thời gian thực và cơ sở dữ liệu đầy đủ. Cơ sở dữ liệu thời gian thực bao gồm dữ liệu TTGT được thu thập trong khung thời gian mới nhất (ví dụ 5 phút), sau đó nó được sử dụng để ước lượng vận tốc của dòng phương tiện trên con đường cụ thể. Cơ sở dữ liệu đầy đủ lưu trữ tất cả dữ liệu được thu thập được từ nhiều nguồn dữ liệu khác nhau và được tích lũy theo thời gian cho đến hiện tại. Cơ sở dữ liệu này được sử dụng để dự đoán vận tốc của các con đường bị thiếu dữ liệu thời gian thực (real-time data) về TTGT tại thời điểm đánh giá. Vì kích thước của cơ sở dữ liệu thời gian thực là khá nhỏ so với cơ sở dữ liệu đầy đủ, thời gian tính toán và hiển thị tình trạng giao thông của một con đường cụ thể sẽ ít hơn rất nhiều so với quá trình tính toán và ước lượng TTGT trên cơ sở dữ liệu đầy đủ.

3.2. Phương pháp phân tích tình trạng giao thông

Nghiên cứu này tập trung vào việc đề xuất khung thức (framework) phân tích dữ liệu cung cấp bởi cộng đồng để đánh giá TTGT. Một yêu cầu thiết yếu là khung thức phải nhẹ nhưng hiệu quả. TTGT liên tục thay đổi theo không gian (các đoạn đường khác nhau) và thời gian. Trong điều kiện lý tưởng nhất, TTGT có thể được cập nhật theo thời gian thực (tức thời) và ở mọi địa điểm. Tuy nhiên, để có được các thông tin chi tiết này, hệ thống sẽ trở nên rất phức tạp trong việc thu thập, tổ chức, quản lý dữ liệu,... dẫn đến chi phí triển khai và vận hành lớn, thậm chí là không khả thi. Để giải quyết những khó khăn trên, ta cần có chiến lược cập nhật TTGT một cách hợp lý, thỏa mãn hai điều kiện sau:

- Hạn chế sự bùng nổ dữ liệu và độ phức tạp tính toán (bằng cách chọn lựa được mức độ chi tiết về không gian và thời gian hợp lý);
- Kết quả trả về của hệ thống mang tính cập nhật, không lỗi thời (tức là đáp ứng gần với thời gian thực - ví dụ sau mỗi 5 phút).

Về không gian: để đảm bảo tính chi tiết của TTGT, ta có thể chia các con đường trên bản đồ giao thông thành các đoạn đường (segment) khác nhau. Các đoạn đường này là đủ ngắn để các thông tin về TTGT trên đoạn đường đó là đồng nhất. Ngược lại, các đoạn đường này cũng đủ dài để hệ thống dữ liệu về các đoạn đường không bị bùng nổ.

Về thời gian: để đảm bảo thông tin về TTGT có ý nghĩa, không bị lỗi thời thì hệ thống cần xử lý dữ liệu liên tục, tức thời (nghĩa là khoảng cách - interval - giữa hai lần tính toán là đủ nhỏ, ví dụ tiến về 0). Tuy nhiên, trong thực tế, TTGT thường không thay đổi quá nhanh và thường tốn một khoảng thời gian đủ lớn (ví dụ 15 phút) để chuyển từ trạng thái này (ví dụ ùn tắc) sang trạng thái khác (hết ùn tắc). Do vậy, trong nghiên cứu này chúng tôi dùng khái niệm chu kỳ tính toán (time_frame) ký hiệu là Δt để chỉ khoảng cách giữa các lần tính toán về TTGT của hệ thống. Chú ý rằng, Δt có thể được tùy biến tùy theo tình trạng thực tế. Ví

dụ thời điểm cao điểm thì Δt nên đủ nhỏ (ví dụ $\Delta t = 5$ phút) trong khi đó ở thời điểm khi TTGT trở nên ổn định, có thể được thiết lập với giá trị lớn hơn (ví dụ $\Delta t = 15$ phút) nhằm tiết kiệm tài nguyên tính toán.

Tất cả các dữ liệu được thu thập trong khoảng thời gian Δt sẽ được tập hợp lại trên hệ thống CSDL của máy chủ để tính toán TTGT và cập nhật lên bản đồ giao thông (trên ứng dụng di động hoặc ứng dụng web). Để đánh giá TTGT, nhóm áp dụng khái niệm "mức độ dịch vụ" (LoS) [25], với điều chỉnh nhỏ phù hợp với các nghiên cứu của nhóm về TTGT đô thị, đặc biệt là ở TP.HCM, một thành phố lớn với khoảng 10 triệu người [12]. Thay vì chia TTGT thành sáu cấp, cụ thể là từ A đến F đại diện cho từ tốt nhất đến kém nhất như trong [25], nhóm kết hợp hai cấp độ tốt nhất là cấp A và B thành một cấp độ (ký hiệu là cấp độ A/B) vì phạm vi vận tốc của chúng gần nhau và phù hợp với cơ sở hạ tầng giao thông tại TP.HCM. Do đó, điều kiện giao thông được phân loại thành một trong năm cấp độ (A/B đến F), mỗi cấp độ được biểu thị bằng một màu khác nhau (từ xanh lục đến đỏ, tương ứng) trên bản đồ trong các ứng dụng di động và trên web của chúng tôi. Các mức LoS được mô tả như sau.

- LoS A/B ($V \geq 30$ km/h) được thể hiện bằng màu xanh lá cây trên bản đồ, tương ứng với mức độ giao thông thông thoáng;
- LoS C (20~30 km/h) được thể hiện bằng màu vàng nhạt trên bản đồ, tương ứng với mức độ giao thông xe đông di chuyển bình thường;
- LoS D (15~20 km/h) được thể hiện bằng màu vàng đậm, tương ứng với mức độ giao thông xe đông di chuyển được;
- LoS E (5~15 km/h) được thể hiện bằng màu cam, tương ứng với mức độ giao thông xe đông di chuyển chậm;
- LoS F ($V < 5$ km/h) được thể hiện bằng màu đỏ, tương ứng với mức độ giao thông ùn tắc.

4. Dự báo tình trạng giao thông dựa vào dữ liệu quá khứ

Như phân tích, hệ thống đề xuất tận dụng nguồn dữ liệu từ cộng đồng nhằm nâng

cao độ bao phủ của hệ thống với nhận định rằng, người dùng di động thường có mặt ở mọi nơi trên thành phố. Tuy nhiên, vấn đề bị thiếu dữ liệu hiện thời (realtime data) vẫn không thể tránh được trong thực tế. Tình huống này xảy ra khi người sử dụng tắt chức năng định vị hoặc khi không có người dùng di động nào cung cấp dữ liệu giao thông tại thời điểm đang xét. Để giải quyết vấn đề về thiếu dữ liệu hiện thời, nghiên cứu đưa ra phương pháp khai phá dữ liệu lịch sử để dự đoán điều kiện giao thông của các đoạn đường thiếu dữ liệu thời gian thực. Một yêu cầu thiết yếu là giải pháp này phải dễ hiện thực nhưng hiệu quả. Cụ thể hơn, mô hình đề xuất có thể sử dụng tốt những thông tin giao thông được trích xuất từ dữ liệu GPS do người dùng di động cung cấp [12]. Mặt khác, trong nhiều kỹ thuật khai phá dữ liệu khác nhau, một kỹ thuật thích hợp như tốc độ dự đoán nhanh và hiệu quả cần được xác định và đưa vào khung đề xuất.

Như mô tả ở mục 3.2, dữ liệu về TTGT dựa vào các nhãn (labels) là LoS của các đoạn đường tương ứng là phù hợp để áp dụng vào các mô hình phân lớp trong kỹ thuật khai phá dữ liệu nhằm dự báo TTGT của các đoạn đường mà dữ liệu thời gian thực bị thiếu. Khung thức đề xuất cho mô hình phân lớp dự báo TTGT được minh họa trong hình 2 và được tóm tắt như sau:

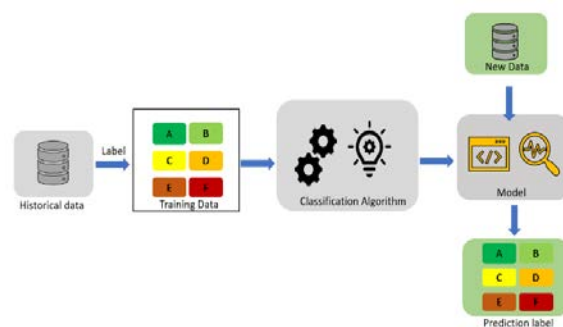
- *Bước 1 - Chuẩn bị dữ liệu huấn luyện:* bước này thực hiện tiền xử lý và gán nhãn cho dữ liệu dựa vào bảng mô tả mức độ phục vụ dịch vụ gồm tập hợp các nhãn sau: A/B, C, D, E, F. Dữ liệu này được sử dụng để huấn luyện mô hình phân lớp.

- *Bước 2 - Huấn luyện mô hình:* mô hình phân lớp dữ liệu sẽ được huấn luyện dựa trên dữ liệu huấn luyện được cung cấp ở bước 1, thông qua các kỹ thuật khác nhau như: Decision Tree, Naïve Bayes, Random Forest, Artificial Neural Network (ANN),...

- *Bước 3 - Áp dụng mô hình vào dữ liệu mới:* mô hình phân lớp xây dựng ở bước 2 sẽ được áp dụng vào phân tích, dự báo dữ liệu mới nhằm xác định nhãn (LoS gồm: A/B, C, D, E, F) của đoạn đường cần được dự báo.

Sau khi có khung thức chung, ta cần xác định các chiều dữ liệu đầu vào cho việc huấn luyện hiệu quả các mô hình khai phá và phân lớp. Một trong những vấn đề ở đây là TTGT khác nhau giữa các đoạn đường ngay cả những đoạn này nằm trên cùng một con đường. Ngoài ra, TTGT cũng thay đổi thường xuyên. Do đó, để cung cấp TTGT phù hợp về mặt không gian và thời gian, chúng ta cần quản lý, tích hợp đầy đủ dữ liệu một cách phù hợp với hai chiều này.

Đối với quản lý chiều không gian, chúng tôi chia mạng lưới đường thành các đoạn đường dựa trên các đề xuất trong Bản đồ số "OpenStreetMap" [28], trong đó mỗi đoạn đường đủ ngắn để xem xét rằng sai số của TTGT tại bất kỳ vị trí nào trên đoạn đường là không đáng kể. Với chiều thời gian, chúng tôi chia thời gian thành các khung/khoảng thời gian theo đó dữ liệu sẽ được thu thập, tích hợp và phân tích. Trong nghiên cứu này, chúng tôi đề xuất thiết lập khung thời gian $\Delta t = 5$ phút, thời gian này đủ dài để tính toán vận tốc trung bình của các đoạn đường một cách hợp lý (có ý nghĩa thống kê) và đủ ngắn để tiết kiệm tài nguyên tính toán trên máy chủ [12], như phân tích ở mục 3.2. Theo phương pháp phân vùng dữ liệu đã đề cập ở trên (không gian và thời gian), định dạng cơ bản của dữ liệu sử dụng cho quá trình huấn luyện bao gồm "segmentID, timeFrame, LoS". Trong đó "segmentID, timeFrame" là đầu vào, và LoS là đầu ra của mô hình dự báo.



Hình 2. Mô hình dự báo TTGT dựa vào dữ liệu lịch sử.

Khung thức khai phá dữ liệu này phù hợp và bổ sung cho mô hình ước tính TTGT dựa trên nguồn dữ liệu thu thập từ cộng đồng được đề xuất trong nghiên cứu trước đây của chúng tôi [12]. Cụ thể hơn, điều kiện giao

thông của các đoạn đường thiếu dữ liệu thời gian thực có thể được ước tính hay dự báo hiệu quả bằng cách sử dụng dữ liệu lịch sử. Chúng tôi đã phát triển một hệ thống nguyên mẫu (prototype) để đánh giá khung thức đề xuất. Chi tiết về hệ thống nguyên mẫu và phân tích đánh giá được trình bày trong phần tiếp theo.

5. Hiện thực và đánh giá

Chúng tôi đã xây dựng hệ thống phần mềm với hai ứng dụng chính: (i) ứng dụng trên di động (mobile app) [12, 38] và (ii) ứng dụng trên web [37] cho phép người dùng thu thập và cung cấp dữ liệu về giao thông. Dựa vào dữ liệu này, hệ thống sẽ tính toán để cung cấp thông tin về TTGT cho người dùng, nhà quản lý giao thông cũng như nhà hoạch định chính sách. Ứng dụng di động được dùng để thu thập dữ liệu thực tế và đánh giá giải pháp đề xuất trên phương diện độ hiệu quả và tính khả thi. Để đánh giá tính khả thi của hệ thống, chúng tôi đề xuất thu thập dữ liệu TTGT tại một số con đường cụ thể như sau: Trường Chinh, Hoàng Văn Thụ, Cộng Hòa (theo cả hai hướng di chuyển). Hình 3 thể hiện bản đồ tuyến đường mà hệ thống thu thập dữ liệu để thử nghiệm. Nhóm đã triển khai thu thập dữ liệu với sự tham gia của 24 tình nguyện viên, thực hiện trong hai tuần vào các ngày thứ 2, thứ 4 và thứ 7 từ khung giờ 7:00 đến 9:00 sáng. Đây là các con đường có các phương tiện tham gia giao thông đông đúc vào buổi sáng, đặc biệt đường Trường Chinh và Cộng Hòa là hai con

đường chính được người dân khu vực các quận Tân Bình, Tân Phú, 12, Học Môn di chuyển vào trung tâm Thành phố làm việc. Đây cũng là các tuyến đường huyết mạch và thường xuyên xảy ra tình trạng UTGT ở phía Tây của Thành phố. Sau quá trình tiền xử lý dữ liệu thu thập được từ ứng dụng di động ta có được bộ dữ liệu với định dạng như bảng 1 và dùng để đưa vào quá trình khai phá dữ liệu bằng mô hình đề xuất. Trong quá trình khai phá, chúng tôi đã sử dụng nhiều thuật toán để tiến hành khai phá như: Random Forest, Decision Tree, Support Vector Machine, ANN, Naïve Bayes, ... Kết quả thực nghiệm cho thấy hai thuật toán đưa ra kết quả tốt nhất là Random Forest và Decision Tree, được biểu diễn ở bảng 2. Kết quả cũng thể hiện F-Sscore của các phương pháp này là tương đối tốt (61.5% và 72.6%) và có thể áp dụng được vào thực tế.



Hình 3. Bản đồ tuyến đường hệ thống tiến hành thu thập dữ liệu.

Bảng 1. Dữ liệu trên ứng dụng di động sau quá trình tiền xử lý dữ liệu.

Đoạn đường segment ID	LOS	Khung giờ (timeframe)	Ngày trong tuần (weekday)	Giờ cao điểm (peak)
24845	D	7:25:00	1	1
24846	C	8:05:00	1	1
...
24847	F	8:30:00	1	1
36296	C	9:50:00	1	0
36298	E	10:00:00	1	0
...
36301	C	10:30:00	1	0
36309	A/B	11:05:00	1	0
36310	C	11:20:00	1	0

Bảng 2. Kết quả sử dụng phân loại dữ liệu (classification) để dự báo TTGT với bộ dữ liệu thu thập từ ứng dụng di động.

Thuật toán	ReCall	F-score	Precision
Random Forest	72.6 %	72.6 %	72.7 %
Decision Tree	61.6 %	61.6 %	61.6 %

6. Kết luận

Nghiên cứu này đã đề xuất một khung thức phân tích, đánh giá tình trạng giao thông dựa vào dữ liệu từ cộng đồng. Theo đó, người sử dụng có thể dễ dàng cung cấp dữ liệu giao thông qua ứng dụng di động và ứng dụng web. Hệ thống phân tích ở trung tâm phân tích các dữ liệu này nhằm đưa ra các đánh giá về TTGT một cách chính xác và kịp thời. Phương pháp dự báo TTGT dựa vào dữ liệu trong quá khứ nhằm giải quyết các vấn đề về thiếu dữ liệu hiện thời cũng như nâng cao độ bao phủ của hệ thống đã được trình bày. Kết quả thực nghiệm cho thấy giải pháp đề xuất có tính hiệu quả và khả thi cao □

Tài liệu tham khảo

- [1] T. Quang, M. Baharudin, and E. Kamioka (2013), *Mc-tes: An efficient mobile phone-based context-aware traffic state estimation framework*, Journal of Information Processing, vol. 21, no. 1, pp. 76-89;
- [2] T. Quang, M. Baharudin, and E. Kamioka (2014), *Synergistic approaches to mobile intelligent transportation systems considering low penetration rate*, Elsevier PMC, vol. 10, no. Part B, pp. 187–202;
- [3] D. of Traffic and Transportation, *Road traffic portal, Ho Chi Minh City*, <http://giaothong.hochiminhcity.gov.vn/>, Truy cập: 01/2021;
- [4] P.T. Vu , *Smart bk traffic*, <http://traffic.hcmut.edu.vn/index.html>, Truy cập: 01/2021;
- [5] M. Ndoye, V. Totten, J. Krogmeier, and D. Bullock (2011), *Sensing and signal processing for vehicle reidentification and travel time estimation*, IEEE Transactions on Intelligent Transportation Systems, vol. 12, no. 1, pp.119–131;
- [6] DH, *Quyết định giao thông thấp, mỗi giờ kẹt xe TP HCM thiệt hại 2,4 tỷ đồng*, <http://ndh.vn/quy-dat-giao-thong-thap-moi-gio-ket-xe-tp-hcm-thiet-hai-2-4-ty-dong-20180112101058307p148c173.news>, Truy cập: 01/2021;
- [7] Sở giao thông vận tải Tp.HCM, <http://sgtvt.hochiminhcity.gov.vn/>, Truy cập: 01/2021;
- [8] Tổng cục thống kê, *Công bố kết quả Tổng điều tra dân số 2019*, <http://tongdieutradaso.vn/cong-bo-ket-qua-tong-dieu-tra-dan-so-2019.html>, Truy cập: 01/2021;
- [9] Giao thông FM91 MHz, *VOV – bản đồ giao thông*, <http://vovgiaothong.vn/>, Truy cập: 01/2021;
- [10] Kênh thông tin giao thông trên Đài Tiếng nói Nhân dân TP. HCM, <http://voh.com.vn/thong-bao/kenh-giao-thong-do-thi-tren-lan-song-fm-95-6-mhz-cua-dai-tnnd-tphcm-111909.html>, Truy cập: 01/2021;
- [11] Công thông tin giao thông TP. HCM của Sở giao thông vận tải TP. HCM, <http://giaothong.hochiminhcity.gov.vn/>, Truy cập: 01/2021;
- [12] Q. T. Minh, H.-N. Pham-Nguyen, H. M. Tan, and N. X. Long, *Traffic congestion estimation based on crowd-sourced data*, in Proceedings of the 2019 International Conference on Advanced Computing and Applications (ACOMP), ACOMP '19, pp. 119–126, 2019;
- [13] P. H. Quang, et al., *Xây dựng mạng camera với hệ thống xử lý hình ảnh thông minh phục vụ điều khiển giao thông và giám sát an ninh*, <http://kc03.vpct.gov.vn/News.aspx?ctl=projectdetail&ID=567>, Truy cập: 01/2021;
- [14] N. T. Thắng, *Nghiên cứu xây dựng Hệ thống giám sát và xử lý vi phạm TTATGT bằng hình ảnh trên hệ thống giao thông đường bộ*, Hội nghị An toàn giao thông, pp. 1-12, 2015;
- [15] T. C. M. Hien (2015), *Xây dựng hệ thống giám sát và xử lý vi phạm TTATGT bằng hình ảnh trên hệ thống giao thông đường bộ*, Hội nghị An toàn giao thông, pp. 1-8;
- [16] Japan Ministry of Land Infrastructure and Transport: *The System Outline of VICS*, <http://www.vics.or.jp>, Ngày truy cập: 12/2020;
- [17] *Route search and location-based consumer services in Japan*, <http://www.navitime.co.jp>,

- Ngày truy cập: 12/2020;
- [18] Kansas ITS,
<http://www.ksdot.org/burtransplan/burovr/inttran.s.asp>,
 Ngày truy cập: 12/2020;
- [19] B. Coifman (2001), *Improved velocity estimation using single loop detectors*, Trans. Res. A, vol. 35, no. 10, pp. 863–880, Dec. 2001;
- [20] S. Tang and F.-Y. Wang (2006), *A PCI-based evaluation method for level of services for traffic operational systems*, IEEE Trans. Intell. Transp. Syst., Vol. 7, No. 4, pp. 494–499;
- [21] J. Wright and J. Dahlgren (2001), *Using vehicles equipped with toll tags as probes for providing travel times*, California PATH Working Paper UCB-ITSPWP-2001-13, Institute of Transportation Studies, University of California, Berkeley, CA;
- [22] X. Ban, Y. Li, A. Skabardonis, and J.D. Margulici (2007), *Performance evaluation of travel time methods for real time traffic applications*, 11th World Conference on Transport Research, Berkeley, CA;
- [23] Y. Cho and J. Rice (2006), *Estimating velocity fields on a freeway from lowresolution videos*, IEEE Trans. Intell. Transp. Syst., Vol. 7, No. 4, pp. 463–469;
- [24] Mobile Millennium Project,
<http://traffic.berkeley.edu/theproject.html>,
 Truy cập: 01/2021;
- [25] Bộ Giao thông vận tải, *Thông tư quy định về tốc độ và khoảng cách an toàn của xe cơ giới, xe máy chuyên dùng tham gia giao thông đường bộ*
<https://luatvietnam.vn/giao-thong/thong-tu-91-2015-tt-bgtvt-bo-giao-thong-van-tai-102206-d1.html>,
 Truy cập: 01/2021;
- [26] I. TomTom, Tomtom, inc,
<https://developer.tomtom.com/products/traffic-api>,
 Truy cập: 01/2021;
- [27] Google map,
<https://www.google.com/maps/@10.7984228,106.656438,15.51z/data=!5m1!1e1,2019>,
 Truy cập: 01/2021;
- [28] OpenStreetMap,
<https://www.openstreetmap.org/#map=3/2.55/66.71>,
 Truy cập: 01/2021;
- [29] E. Estellés-Arolas and F. González-Ladrón-de Guevara (2012), *Towards an integrated crowdsourcing definition*, Journal of Information Science, vol. 38(2), pp. 189–200;
- [30] M. Lease and E. Yilmaz (2012), *Crowdsourcing for information retrieval*, SIGIR Forum, vol. 45, no. 2, pp. 66–75;
- [30] L. See, P. Mooney, G. Foody, L. Bastin, A. Comber, J. Estima, S. Fritz, N. Kerle, B. Jiang, M. Laakso, H.-Y. Liu, G. Mil'cinski, M. Nik'sić, M. Painho, A. P'od'or, A.-M. Olteanu-Raimond, and M. Rutzinger (2016), *Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information*, ISPRS Int. J. Geo-Inf., vol. 5, no. 55;
- [31] B. Deanne, H. Katharine, L. Megan, and O. B. James (2011), *The use of crowd sourcing for gathering information about natural disasters*, Risk Frontier, vol. 11, no. 2, pp. 1–4;
- [32] Ushahidi, *Ushahidi, read the crowd*,
<http://www.ushahidi.com>,
 Truy cập: 01/2021;
- [33] D.-B. Nguyen, C.-R. Dow, and S.-F. Hwang (2018), *An efficient traffic congestion monitoring system on internet of vehicles*, Wireless Communications and Mobile Computing, vol. 2018, pp. 1–17;
- [34] M. Lewandowski, B. P. Iaczek, M. Bernas, and P. Szymala (2018), *Road traffic monitoring system based on mobile devices and bluetooth low energy beacons*, Wireless Communications and Mobile Computing, vol. 2018;
- [35] M. Elloumi, R. Dhaou, B. Escrig, H. Idoudi, and L. A. Saidane (2018), *Monitoring road traffic with a uav-based system*, IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6;
- [36] T.T.deAlmeida, J.A.M.Nacif, F.P.Bhering, and J. G.R.Junior (2019), *Doc-trams: A decentralized and offline community-based traffic monitoring system*, IEEE Transactions on Intelligent Transportation Systems, vol. 20(3), pp.1160–1169;
- [37] Q. T. Minh, et al., *BK Urban Traffic*,
<https://bktraffic.com/utraffic/>,
 Truy cập: 01/2021;
- [38] Q. T. Minh, et al., *Ứng dụng cảnh báo TTGT*,
<https://play.google.com/store/apps/details?id=com.hcmut.admin.bktrafficsystem>,
 Truy cập: 01/2021.

Lời cảm ơn: Nghiên cứu được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (ĐHQG-HCM) trong khuôn khổ Đề tài mã số DS2021-20-01.

Ngày nhận bài: 16/03/2021

Ngày chuyển phản biện: 19/03/2021

Ngày hoàn thành sửa bài: 08/04/2021

Ngày chấp nhận đăng: 15/04/2021

Sản phẩm từ nghiên cứu

UTRAFFIC - HỆ THỐNG ĐÁNH GIÁ TÌNH TRẠNG GIAO THÔNG ĐÔ THỊ DỰA VÀO DỮ LIỆU TỪ CỘNG ĐỒNG (CROWD-SENSED DATA)

Hệ thống Utraffic (Urban Traffic Information System) với nhiều tiềm năng phát triển, kết hợp được sức mạnh của cộng đồng cùng công nghệ thông tin và truyền thông (ICT) vượt trội trong việc giải quyết một bài toán khó ở những đô thị lớn - ùn tắc giao thông (UTGT).

Nền tảng Utraffic bao gồm các ứng dụng nổi bật sau:

- 01** | Ứng dụng trên điện thoại di động cho phép người dùng thông báo TTGT, tìm đường đi phù hợp, nhận được cảnh báo UTGT từ hệ thống phân tích, tích lũy điểm thưởng hoặc sự ghi nhận từ các đóng góp dữ liệu của bản thân;
- 02** | Ứng dụng trên nền tảng web với các chức năng ghi nhận, tra cứu thông tin về TTGT, báo cáo tổng hợp nhằm hỗ trợ nhà quản lý trong việc ra các quyết định phù hợp;
- 03** | Hệ thống API và dữ liệu về TTGT để cộng đồng, các nhà nghiên cứu sử dụng nhằm đưa ra các giải pháp phù hợp, góp phần giải quyết bài toán chống UTGT và xây dựng hệ sinh thái giao thông thông minh (ITS).

Sản phẩm được thực hiện từ nhóm nghiên cứu tại bộ môn Hệ thống thông tin, khoa Khoa học và Kỹ thuật máy tính, Trường Đại học Bách Khoa - ĐHQG TP.HCM.



Acknowledgement: We acknowledge the support of time and facilities from Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for this study.



Truy cập thông tin về Utraffic: <https://bktraffic.com/home/>

Trải nghiệm app: <https://bktraffic.com/home/mobile-app> Hoặc mã QR (trái)

Ngoài hình ảnh, bảng biểu đã chú thích nguồn từ tài liệu tham khảo, những hình ảnh, bảng biểu còn lại đều thuộc bản quyền của tác giả/nhóm tác giả.