

택시 데이터를 이용한 구간 내 교통량과 차량 속도 간의 상관관계 분석

김호용^{1,2} · 정한민^{1,2*}

¹과학기술연합대학원대학교 · ²한국과학기술정보연구원

Correlation Analysis between Traffic and Speed on the road using Taxi Data

Hoyong Kim^{1,2} · Hanmin Jung^{1,2*}

¹University of Science and Technology · ²Korea Institute of Science and Technology Information

E-mail : khyong0101@ust.ac.kr / jhm@kisti.re.kr

요 약

도로 교통 시스템에 빅데이터 기술이 융합되면서 운전자와 보행자에게 편리하고 도움이 되는 새로운 서비스들이 생겨나고 있다. 다양한 교통 데이터를 수집하고 분석함으로써 교통 시스템에서 흔히 발생할 수 있는 정체 현상 또는 교통량 등을 예측하는 모델이 개발되고 있다.

본 논문에서는 대구시에서 수집한 택시 경로 데이터를 이용하여 대구시의 지도 위에 해당 데이터를 시각화한다. 그 다음으로 대구 시 내의 도로 구간을 설정하고, 택시 데이터의 위치와 속도 데이터, 그리고 도로 구간 정보를 이용하여 해당 구간의 교통량과 구간에서의 차량 평균 속도를 구한다. 이 결과를 이용하여 특정 도로 구간에 대한 교통 시스템 문제 파악에 도움을 주고자 한다.

ABSTRACT

As the convergence of traffic system and Big-data technology, new convenient services which is helpful for drivers and pedestrian are appeared. Recently, the various researches about the traffic system, such as prediction of traffic jam and finding the shortest path, are studied.

In this paper, we collect the data of taxi trips in Daegu City, and visualize them on the map of Daegu City. And then, we select specific sections of roads in the city, and by using the data of location and speed about taxis and the information of the road sections, calculate the traffic of that section and the average speed of cars on that section. As a result of this, we give help solving the problem of the specific road sections.

키워드

taxi, visualization, traffic, speed

1. 서 론

최근 빅데이터의 출현으로 데이터의 중요성이 대두되고 있다. 이러한 추세에 따라 세계 곳곳에서는 여러 공공데이터를 오픈하여 데이터를 활용하

공공서비스 개발에 대한 지원을 하고 있다. 이러한 공공데이터 중 도로 교통 시스템 관련 데이터들이 빅데이터 기술과 융합되면서 도로 교통 시스템의 여러 문제점을 해결하여 운전자와 보행자에게, 그리고 도로 상황에 유익한 결과를 주고 있다.

[1]에서는 미국 캘리포니아주의 고속도로 데이터 와 해당 구간에서 발생한 사고 데이터에 인공 신

* corresponding author

경망 모델(Artificial Neural Network Model)을 사용하여 분석한 다양한 사고율 예측 결과를 사용자에게 제공하여 안전 운전에도움이 될 수 있게 하였다. [2]에서는 대중교통 이용객의 이동 데이터를 기반으로 지역별 대중교통 편의성 지표를 제안하고 이를 실제 데이터에 적용하여 대중교통 편의성을 분석하였다.

빅데이터뿐만 아니라 빅데이터 관련 기술 또한 도로 교통 시스템의 문제를 해결하는 데 사용되고 있다. [3]에서는 지자체 A시에서 발생하는 교통 빅데이터를 활용하여 다차원 분석 및 시각화하여 버스의 호차별 효율성을 분석하였다. [4]에서는 뉴욕 시에서 2010-2013년 사이에 발생한 70억 개의 택시 경로 데이터를 통계 기법으로 분석하여 뉴욕 시의 교통 흐름을 파악하였다. 또한 지도 위에 경로마다의 교통량을 표현하여 시각적 효과를 높였다.

본 논문에서는 약 8개월 간 대구시에서 수집된 택시 데이터를 이용하여 대구시의 지도 위에 해당 데이터를 시각화한다. 그 다음으로 대구 시 내의 도로 상에서 일정 간격으로 구간을 설정하고, 택시 데이터의 위치와 속도 데이터, 그리고 도로 구간 정보를 이용하여 해당 구간의 교통량과 구간에서의 차량 평균 속도를 구하여 교통량과 차량 평균 속도 간의 상관관계를 분석한다. 또한 이 분석 결과를 시각화 자료와 비교하여 교통량과 평균 속도로부터 특정 도로 구간에 대한 교통 시스템 문제를 파악하는 데에 도움을 주고자 한다.

II. 제안 방안

교통량과 도로에서의 차량 평균 속도는 교통 상황을 파악함에 있어 중요한 요소이다. 예를 들면, 교통량이 많고 차량 평균 속도가 느리면 해당 구간에서 정체가 발생하였다고 예상할 수 있다. 하지만 교통량과 차량 평균 속도 간의 상관관계를 분석하지 않고 독립적으로 사용한다면 도로 상황을 파악하는 데 있어 오류를 범할 수 있다. 차선이 많은 도로는 많은 수의 차량을 한 번에 수용할 수 있어 교통량이 높게 측정되더라도 정체가 일어나지 않을 수 있다. 반대로, 차선이 적은 도로는 교통량이 적더라도 수용할 수 있는 차량 수가 적기 때문에 교통량이 낮게 측정되더라도 정체가 발생할 수도 있다. 즉, 정체 상황에서 교통량과 차량 평균 속도가 항상 반비례하지는 않는다는 것이다. 본 논문에서는 도로의 구간을 분할하고 각 구간마다의 교통량과 평균 차량 속도를 구한 다음 상관관계를 분석하여 이를 지도 위에 시각화한다. 시각화 결과로부터 교통량과 차량 평균 속도가 반비례하지 않는 구간을 선정하고 이 구간에 대한 로드뷰나 외부 데이터를 이용하여 도로 정체 현상에 교통량이 아닌 외부 요인(차선 수)이 영향을 줄 수 있음을 확인한다.

먼저, 2017년 5월부터 12월까지 약 8개월 간 대구시에서 수집한 8만 5천여 개의 택시 데이터를

이용하여 시각화를 진행한다. 도로 위의 각 구간을 설정하기 위하여 OpenStreetMap API를 이용하였고 그 결과 대구시의 각 도로 위에 100m 간격으로 구간을 나누어주었다. 그림 1은 대구시 지도 위에 각 구간의 지점들을 시각화한 자료이다.



그림 1. 대구 시 도로 위의 구간

구간을 설정한 다음, 각 구간에 속하는 교통량과 평균 차량 속도를 구한다. 특정 구간 S_i 의 양끝 지점 f_i, f_{i+1} 에 대하여 각 지점의 위도, 경도 데이터를 각각 $(x_{f_i}, y_{f_i}), (x_{f_{i+1}}, y_{f_{i+1}})$ 라고 하고 택시의 특정 위치 데이터 P_j 에 대한 위도, 경도 데이터를 (x_{p_j}, y_{p_j}) 라고 하자. 택시 위치 데이터는 센싱 데이터이기 때문에 어느 정도의 오차가 발생할 가능성이 있다. 따라서 각 구간의 범위를 지정하여 해당 문제를 해결하였다. 구간에서 두 지점 사이의 실제 거리는 100m이고 지도 상에서 확인한 도로의 너비는 13m 정도였다. 따라서 특정 구간 S_i 의 범위를 설정하여 해당 범위 내의 택시 데이터들의 정보를 포함할 수 있게 하였다.

본 논문에서 사용한 교통량과 평균 차량 속도 측정을 위한 수식은 다음과 같다.

$$width_{S_i} = 0.13 \times \sqrt{(x_{f_{i+1}} - x_{f_i})^2 + (y_{f_{i+1}} - y_{f_i})^2} \quad (1)$$

$$a = \frac{y_{f_{i+1}} - y_{f_i}}{x_{f_{i+1}} - x_{f_i}} \quad (2)$$

$$distance(S_i, P_j) = \frac{|ax_{p_j} - y_{p_j} - ax_{f_i} + y_{f_i}|}{\sqrt{a^2 + 1}} \quad (3)$$

$$traffic(S_i, P_j) = \begin{cases} 1 & \text{if } distance(S_i, P_j) \leq width_{S_i} \\ 0 & \text{else} \end{cases} \quad (4)$$

$$traffic_{S_i} = \sum_{j=1}^{N-1} traffic(S_i, P_j) \quad (5)$$

$$Speed_{S_i} = \frac{1}{traffic_{S_i}} \sum_{j=1}^N traffic(S_i, P_j) \cdot Speed_{P_j} \quad (6)$$

수식 (1)은 특정 구간 S_i 의 범위를 나타낸다. 수식 (2)는 특정 구간 S_i 의 양끝 지점을 잇는 직선의 기울기로 수식 (3)에서 구간과 택시 위치 데이터 사이의 거리를 측정하는 데 사용된다. 수식 (4)는 구간 S_i 와 택시 위치 데이터 P_j 에 대하여 택시가 구간 S_i 범위 내에 속하는지 구분하는 수식으로 속하면 1, 아니면 0인 값을 가진다. 최종적으로 수식

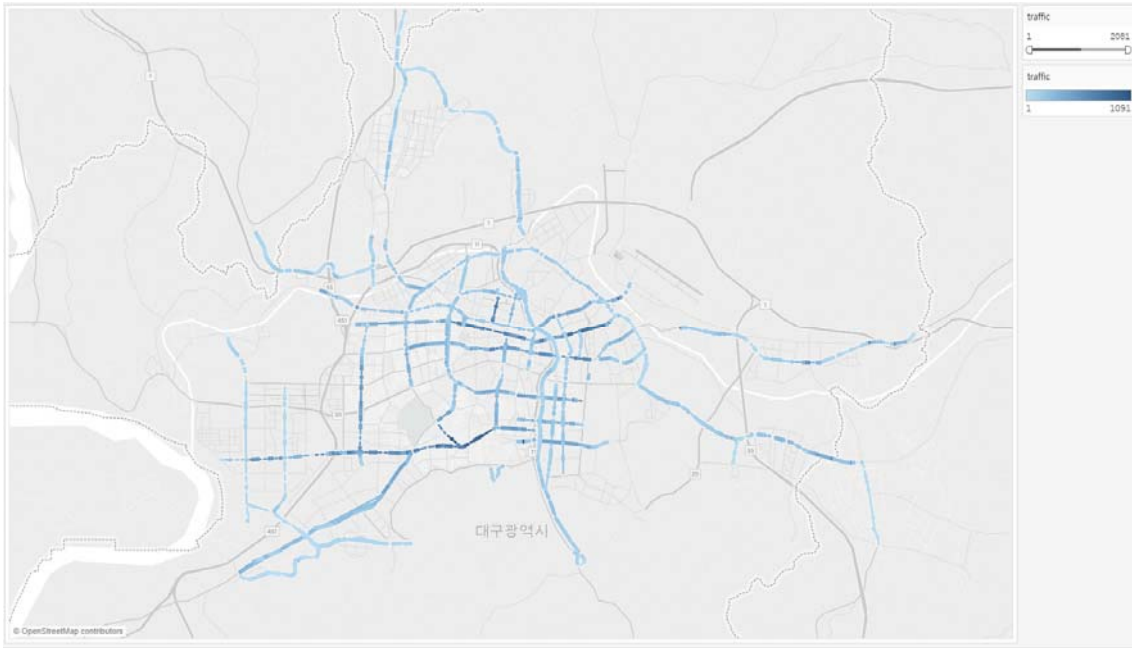


그림 2. 대구시 각 도로 구간의 교통량

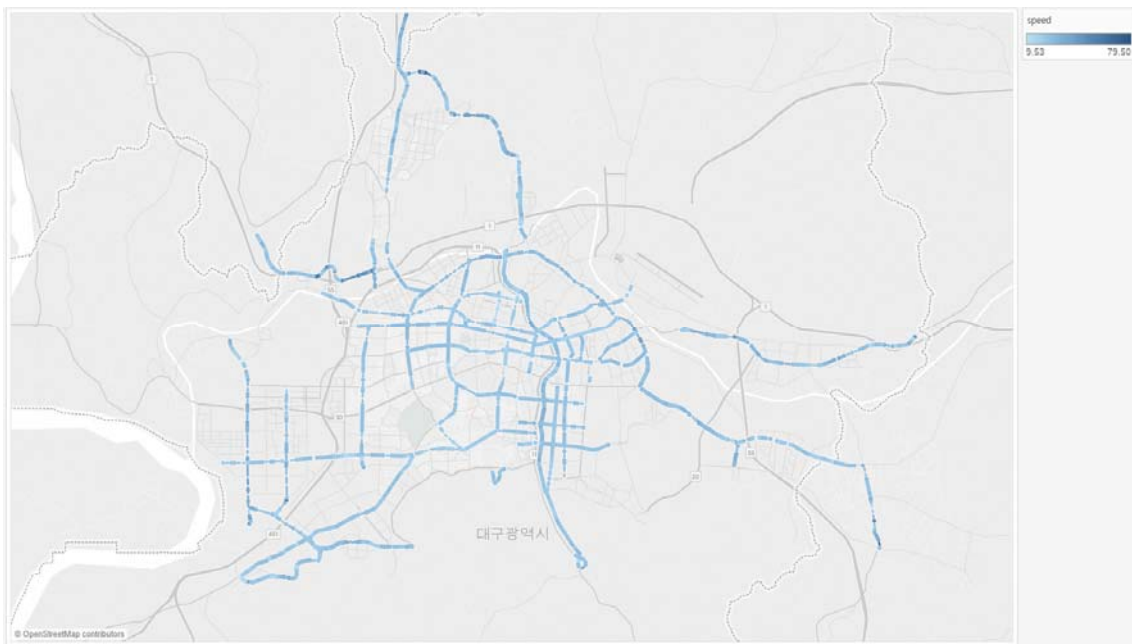


그림 2. 대구시 각 도로 구간의 평균 차량 속도

(5), (6)에서 특정 구간 s_i 에 대한 교통량과 평균 차량 속도를 구한다.

III. 실험 결과

2의 제안 방안에 따라 실험한 결과는 그림 2, 3과 같다. 본 논문에서는 데이터를 시각화하기 위하여 Tableau public 시각화 툴을 사용하였다.

또한 교통량과 평균 차량 속도에 대한 피어슨 상관 계수(Pearson Correlation Coefficient)를 구한 결과, 0.1976이라는 수치를 얻을 수 있었다. 피어슨 상관 계수는 -1 이상 1 이하의 실수 값으로 두 변수 간의 상관관계를 표현한다. -1에 가까울수록 음의 상관관계, 1에 가까울수록 양의 상관관계가 강하다. 따라서 0.1976은 0에 가까운 수치로 두 변수 간의 상관관계가 없다고 볼 수 있다. 또한 시각화한 그림의 결과로부터도 유의미한 결과를 도출

해내기 어렵다.

IV. 결론

본 논문에서는 대구시의 택시 위치 데이터와 속도 데이터를 이용하여 각 도로 구간의 교통량과 평균 차량 속도를 구한 다음 두 변수 간의 상관관계를 분석하여 교통량과 평균 차량 속도 외의 다른 요인이 정체 현상에 어떤 식으로 작용하는지 알아보려고 하였다. 하지만 상관관계 분석 결과 두 변수 간의 상관관계가 없음을 알 수 있었고 시각화를 통한 분석을 통해서도 유의미한 결과를 도출하기 어려웠다. 향후 연구로써 대구시 택시 50대에 대한 약 400만 개의 택시 데이터를 이용하여 분석을 위한 데이터 코퍼스를 확장하고 본 논문에서 제안한 방안을 다시 적용하여 본다. 또한 원본 데이터에 대한 경험적인 피드백을 통하여 혹시 모를 데이터 에러에 대한 예외 처리를 진행한다.

References

- [1] 정덕원, “교통 빅데이터의 실시간 분석 및 예측 서비스 프레임워크”, 학위논문(박사), 건국대학교 대학원, 컴퓨터정보통신공학과 컴퓨터공학, 2014
- [2] 문현구, 김상국, “스마트카드 빅데이터를 이용한 대중교통 편의성 분석”, 학위논문(석사), 경희대학교 일반대학원, 산업공학과, 2015
- [3] 구가영, “효율적 버스운행을 위한 교통데이터 다차원 분석 및 시각화”, 학위논문(석사), 충북대학교 대학원, 비즈니스데이터융합학과, 2015
- [4] Joya A. Deri, Franz Franchetti, and Jose M.F. Moura, “Big Data Computation of Taxi Movement in New York City”, *IEEE International Conference on Big Data (Big Data)*, p2616-2625, 2016