
Global Report

스마트 운송 서비스 지원을 위한 개방형 빅 데이터에 대한 예측 분석

1. 개요
2. 개방형 빅 데이터에 대한 예측 분석
3. 평가
4. 결론

1. 개요

오늘날 기술 기반 환경에서 엄청난 양의 데이터가 빠른 속도로 생성되고 있다. 기상 데이터, 스포츠 결과 및 주가(주식가격)등과 같은 데이터들을 쉽게 이용할 수 있게 되었으며, 기타 데이터들은 웹이나 소셜 네트워크 등을 통해 수집할 수 있다. 이러한 빅 데이터는 데이터 과학(data science[1]) 또는 데이터 마이닝(data mining[2])을 통해 수집할 수 있는 암시적이고 이전에 알려지지 않은 잠재적 가치를 지닌 유용한 정보와 지식이다[3][4][5][6]. 이러한 지식을 통해 얻은 통찰력은 데이터를 분류하고 분할하는 것과 같은 과학적 이유뿐만 아니라 가설을 형성하는 데에도 이점을 제공하고 있다. 또 다른 통찰력은 더 나은 서비스 제공을 통해 상업적으로 사용될 수 있다. 이러한 오픈 데이터[7][8]가 보다 대중화되고 있다. 과학계/정부/비영리단체 등에서 이러한 빅 데이터 수집이 점점 더 증가하고 있으며, 이러한 개방형 데이터는 다음과 같은 이점이 있다.

- (a)정책 활동에 대한 모니터링을 허용하는 투명성 및 (b)혁신과 같은 부가 가치를 통해 개인이 새로운 응용프로그램을 개발할 수 있는 기회를 제공한다.
- 수집한 정보를 저장하였다가 필요할 때 언제든지 액세스 할 수 있는 통찰력을 제공한다.

따라서 많은 국가들이 모든 사람이 정부정책을 보다 쉽게 이용할 수 있도록 함으로써 개방형 정부 정책에 기여하였다. OECD(Organization for Economic Co-operation and Development : 경제협력개발기구)에 따르면 많은 회원국(한국, 프랑스, 아일랜드, 일본, 캐나다)은 공공부문 데이터를 이용할 수 있고 접근가능하게 하고 데이터 재사용을 지원하기 위해 노력해 왔다. 그 중 상당수는 데이터 재사용을 지원하기 위해 노력해 왔으며, 아울러 공공부문 데이터를 이용할 수 있도록 하고 있다.

캐나다에서는 많은 도시와 지방이 이 과정에서 얻은 가치를 인정하는 개방형 데이터를 제공하는 데 앞장서고 있다. COGWWG(Canada Open Government Working Group)에 따르면 2020년 5월 현재 캐나다 전역에는 Vancouver/Winnipeg/Toronto/Ottawa/Montreal을 포함한 61개 개방형 지방자치단체, Nunavut Geoscience와 같은 12개 개방형 이니셔티브, Manitoba와 2개의 개방된 지역으로 구성된 12개 개방형 지역이 있다. 이러한 개방형 데이터 포털에서 사용가능한 데이터집합의 한 가지 유형은 개방형 데이터 전송과 관련이 있다. 많은 도시에서 버스정류장 번호, 노선번호 등과 같은 대중교통 시스템에 대한 정보를 제공하고 있다. 이 연구에서는 대중교통 시스템의 정시정각 운행에 관심을 둔다. 버스의 지연시간과 같은 정보는 몇몇 도시에서 제공하고 있지만, 소도시에서는 먼저 도착하는 버스에 대한 정보만을 제공하고 있다. 버스의 지연시간을 제공하는 도시 중 하나는 예정된 시간에서 버스편차를 공개적으로 제공하는 Winnipeg시(캐나다의 매니토바주 인구의 절반이상이 거주하고 있는 주 최대도시로 서부 캐나다 대초원지역의 끝에 위치하고 있어 교통/경제/공업/농업/교육 등에서 중요한 역할을 담당하고 있다. 서부와 동부를 연결하는 모든 철도와 고속도로는 이 도시근처를 지나가므로 이 도시를 ‘서부로의 관문’으로 불리기도 한다)이다. 이 편차는 예정된 출발시간을 기준보다 먼저 혹은 늦게 오는지를 몇 초 만에 측정할 수 있다.

Winnipeg Transit(Winnipeg시의 대중교통기관)에서 운영하는 모든 버스에는 기내 컴퓨터와 GPS(Global Positioning System)가 장착되어 있어 버스정류장을 떠날 때 버스의 정시정각 운용성능을 기록한다. 예를 들어, 버스가 10:00:00 정각에 버스정류장을 출발해야 하지만, 실제로 10:03:23에 출발한 경우, 다음 버스정류장에 3분23초 늦게 도착하는 것으로 버스의 온보드 컴퓨터에 기록된다. 또 다른 예로, 버스가 10:00:00에 버스정류장을 떠나야 하지만, 실제로 09:58:23에 출발했다면 다음 버스정류장에 1분37초 일찍 도착하는 것으로 버스의 온보드 컴퓨터에 기록된다. 이처럼 버스의 온보드 컴퓨터에 기록된 데이터는 매일 해당버스에서 수집 및 다운로드하여 매월 오픈데이터 포털로 업데이트된다. 일반적으로(예를 들어, COVID-19 전염병으로 인해 평일 서비스가 감소하기 전에) Winnipeg Transit은 약 5,170개의 버스정류장에서 출발하는 93개 버스노선(조기 또는 늦게)을 통해 640대의 버스를 운행하였다. 이 버스들은 장시간 동안(2018년 1.55M 시간 이상)운영하여 그 결과 오픈 빅 데이터는 잘 알려진 다음과 같은

V's(3V's, ..., 7V's, ..., 42V's, 51V's)를 기록한다.

- 대용량의 버스 정시정각 성능 데이터를 기록하고 수집하고 있다.(예를 들어, 2018년 90M 이상 기록 인스턴스)
- 이러한 개방형 빅 데이터에 포함된 유용한 정보와 지식은 높은 가치를 지니고 있으며, 데이터과학 또는 데이터 마이닝과 같은 기술로 수집할 수 있다.
- 이러한 개방형 빅 데이터는 버스가 모든 버스정류장을 떠날 때마다 매우 빠르게 기록되고 수집된다.
- 이러한 개방형 빅 데이터는 데이터전송 문제, 데이터수집 문제, GPS 및 기타 오류 등으로 인해 기록된 위치(위도 및 경도)가 정확하지 않을 수도 있다.

일반적으로 정시정각 성능은 건설/비상사태/특별 이벤트/교통혼잡/날씨 등 여러 요인에 의해 영향을 받을 수 있다. Winnipeg Transit에서는 정시정각 성능을 다음과 같이 정의하고 있다.

- 예정된 출발시간보다 1분 일찍 정류장을 떠난 버스는 다음 정류장에 그만큼 일찍 도착하는 것으로 고려한다.
- 예정된 출발시간 후 1분~3분 이내에 정차하는 버스는 다음 정류장에 정시에 도착하는 것으로 고려한다.
- 예정된 출발 시간보다 3분 이상 늦게 정류장을 떠난 버스는 다음 정류장에 그만큼 늦게 도착하는 것으로 고려한다.

버스의 지연시간을 예측하는 것이 중요한 데는 이유가 있다. 현대사회에서 많은 사람들은 출퇴근/스포츠 관람/친구만나기/등하교 등 매우 다양한 일상생활을 위해 대중교통 시스템에 많이 의존한다. 이처럼 다른 운송수단 이용자보다 대중교통을 가장 많이 의존하는 버스이용자는 물론 이따금씩 버스를 이용하는 사람도 대중교통을 편리하게 이용할 수 있어야 한다. 전통적으로, 교통지연에 대한 분석은 주로 목적지에 늦게 도착하는 부정적인 측면이 매우 높기 때문에 주목받고 있다[9]. 예를 들어, 고용주나 교사 등은 시간지연에 대해 매우 민감하다. 늦게 출근하는 직원은 급여가 줄어들거나 직장을 잃을 수도 있으며, 시험시간에 늦게 도착하는 학생은 성적에 큰 영향을 미칠 수 있고, 마찬가지로, 일반일 들은 스포츠나 콘서트 시작시간을 놓치고 싶어 하지 않기 때문이다.

버스가 예정시간보다 일찍 출발하는 영향에 대한 것도 중요하다. 버스의 예정 출발시간은 다른 운행노선이나 목적지로 이어지기 때문이다. 두 번째 버스가 예정시간보다 일찍 출발하면 이용자는 교통편을 놓치고 예상보다 훨씬 긴 시간동안 다음 버스를 기다려야 한다. 아울러 예정시간보다 일찍 떠났기 때문에 그날의 마지막 버스를 놓칠 수도 있으며, 이를 피하기 위해 다른 경로를 찾아볼 수도 있어 버스의 정시정각 운행에 대한 정확한 예측은 매우 중요하다. 이에 Winnipeg시에서 운용하고 있는 개방형 빅 데이터 기반의 버스 조기 혹은 지연 가용성을 감안하면, 이 시스템에 대한 분석을 시행하는 것은 당연한 일일 것이다. 우리는 이 시스템에서 자주 발생하는 패턴을 검색하기 위해 데이터를 분석하는 것이 가능한지, 기존 데이터를 기반으로 버스 조기 혹은 지연을 예측하는 것이 가능한지도 궁금한 사항이다.

이러한 궁금 사항(질문)에 답하기 위해 버스의 성능에 대한 개방형 빅 데이터(버스의 조기 및 지연에 따른 다음정류장의 버스 도착상황 등)를 파악해본다. 또한 패턴 마이닝을 통해 빈번하게 데이터를 분석하여 의사결정 트리를 기반으로 분류 예측 한다. 아울러 일러스트레이션을 이용하여 Winnipeg시에서 운용하고 있는 개방형 빅 데이터에 대한 예측분석을 수행한다. 이 연구는 스마트운송 서비스를 지원하기 위한 개방형 빅 데이터에 대한 예측분석에 기여할 수 있을 것이다.

2. 개방형 빅 데이터에 대한 예측 분석

2.1. 데이터 전처리

Winnipeg시에서 운영하고 Winnipeg Transit의 개방형 빅 데이터 포털을 기반으로 i) 자세한 정시정각 성능 데이터, ii) 경로 및 일별 정시정각 성능 데이터 집계사항 등 버스의 정시정각 운행실적을 검토한다. 이들의 자세한 정시정각 성능 데이터는 다음과 같이 9개의 특성을 갖는다. 데이터는 두 개의 컬렉션으로 제공되며, 전처리단계는 분석을 위한 모든 데이터 포인트에 대해 다음과 같이 9가지 특성을 가지는 정시 성능데이터가 두 컬렉션 사이에 있다.

1. 데이터집합의 고유 식별자인 행ID
2. 개별 버스정류장에 할당된 고유한 5자리 정지번호
3. 특정 1~3자리 버스노선번호를 캡처 하는 경로번호
4. 경로번호에 해당하는 공통이름을 캡처하는 경로이름
5. 경로의 종단을 나타내고 버스의 방향(북방/남향)을 결정하는 데 도움이 되는 경로목적지
6. 3가지 값 중 하나를 차지하는 직업 유형
7. 출발날짜와 시간을 나타내는 타임스탬프(예정시간)
8. 예정된 출발시간과 정류장에서 실제 출발시간 사이의 편차(초)를 캡처하는 편차
9. 특정 버스정류장의 위치(경도, 위도)를 나타내는 GPS좌표

이 2가지 데이터 컬렉션<Winnipeg Transit의 개방형 빅 데이터포털 기반 버스의 정시정각 운행실적(정시정각 성능 데이터, 경로 및 일별 정시정각 성능 데이터 집계사항)>을 통해 집계된 성능데이터를 자세히 검토한다. 집계된 데이터에서 일부 세분화된 정보가 손실될 수도 있다. 성능데이터 검토를 통해 더 흥미로운 지식을 발견할 수 있다. 자세한 버스의 정시정각 성능데이터의 9가지 특성 중 데이터전처리 단계에서는 다음 기능을 수행한다.

- 분석의 행ID는 데이터집합의 고유 식별자일 뿐이기 때문에 무시한다.
- 정지번호와 경로번호를 캡처한다.
- 경로번호와 1대1로 대응하는 경로이름을 무시한다.
- 경로를 나타내기 위해 경로대상을 정수 값으로 매핑한다.
- 예정된 시간 동안 타임스탬프에서 요일과 시간을 추출한다.
- 예정된 출발시간과 실제 출발시간이 다른 7가지 심각도를 도출한다.

2.2. 빈번한 패턴 마이닝

자세한 정시정각 성능에 대한 수집데이터는 전처리과정을 거쳐 결과데이터를 검사한다. 데이터분석 프로세스를 용이하게 하기 위해 몇 가지 범주형 특성을 숫자로 매핑한다.

- 속성이 요일일 경우 월~일요일까지 각각 1~7로 매핑한다.
- 속성이 기간일 경우 아침 피크시간을 1, 비성수기 시간을 2, 오후 피크시간을 3, 저녁시간을 4, 야간시간을 5로 매핑한다.
- 속성이 심각도 수준일 경우 극단적으로 늦음을 1, 매우 늦음을 2, 약간 늦음을 3, 정시를 4, 약간 일찍을 5~6, 매우 일찍을 7로 매핑한다.

그런 다음 데이터 마이닝, 머신러닝, 지능형 통계 및 학습엔진을 이용하여 분석을 시행한다. 특히 패턴 마이닝이 빈번하게 발생하는 경우가 많은데, 이처럼 자주 발생하는 패턴은 버스의 정시정각 성능을 나타내는 데 도움이 된다. 특히 정교하고 빈번한 패턴은 다음과 같은 새로운 지식을 얻는 데 도움이 된다:

- 어떤 노선에서 자주 일찍, 정시 또는 늦게 정류장을 출발했는가?

- 어떤 노선과 방향에서 자주 일찍, 정시 또는 늦게 정류장을 출발했는가?
- 어느 주에 자주 일찍, 정시 또는 늦게 정류장을 출발했는가?
- 이 때 정시 또는 늦게 정류장을 자주 출발했는가?
- 노선과 주기(정시 또는 늦게)가 일정하지 않은 버스의 심각도 사이의 관계는 무엇인가?
- 주 중에 주기(정시 또는 늦게)가 일정하지 않은 버스의 심각도 사이의 관계는 무엇인가?
- 주기(정시 또는 늦게)가 일정하지 않은 버스의 기간과 심각도수준 사이의 관계는 무엇인가?
- 6가지 속성 중 가장 빈번하게 나타나는 속성은 무엇인가?

3. 평가

Winnipeg Open Data Portal에서 2017년 7월부터 가장 최근까지의 자세한 정시정각 실적 데이터를 다운로드하였다. 이 연구에서는 사용가능한 모든 데이터와 12개월간(2019년 5월~2020년 4월)의 데이터에서 가장 빈번한 패턴 마이닝을 적용하였다. 2017년 7월부터 2019년 6월까지의 데이터를 7월~12월간 학습하여 분류 및 예측을 수행하였다. 평가는 16GB의 RAM을 갖춘 Intel Core i7-4790 processor에서 실행되었다. 평균 런 타임은 평균 5회에서 계산되었다. 패턴도출에는 약 204초가 걸렸으며, 분류 및 예측은 약 228초가 걸렸다. 이러한 개방형 빅 데이터에 대한 예측 분석은 빈번한 패턴, 연결규칙, 연관분류 규칙, 의사결정 트리 및 예측의 형태에 대한 흥미로운 지식으로 이어졌다. 이 지식에 대한 샘플관찰 결과는 다음과 같다.

- 181번 국도를 운행하는 버스의 53%가 늦게 출발했으며, 이로 인해 113,845개의 버스정류장에서 지연이 발생한 것으로 나타났다.
- 642번 국도를 운행하는 버스의 79%가 정시에 운행하여 27,283대의 버스가 정시정각 운행이 가능한 것으로 나타났다.
- 23번 국도를 운행하는 버스의 79%가 11,940개의 정류장에서 조기에 출발한 것으로 나타났다.
- 2019년 10월 11일에 운행된 버스의 77%가 버스정류장에서 늦게 출발했으며, 이로 인해 197,825개의 정류장에서 지연이 발생한 것으로 나타났다. 이는 기상정보 확인결과, 그 날의 악천후로 인한 것으로 밝혀졌다.

예측정확도 측면에서, 처음에 분류자는 43,195,136개의 예측 중 29,337,870건의 인스턴스에서 지연의 심각도 수준(극도/아주 매우/매우/약간 늦은 시간 또는 정시)을 정확하게 예측하여 약 68%의 정밀도를 이끌어내었다. 이로써 데이터의 정확성을 확인하였고, 예정시간보다 매우 일찍(20분 이상) 출발하는 버스는 신뢰성이 많이 저하되어 노이즈일 가능성이 있는 것으로 판단하였다. 노이즈로 분류된 모델을 학습하면 오버 피팅(over fitting)이 발생하여 모델과 예측정확도에 부정적인 영향을 미칠 수 있다. 이는 일부 파라미터를 미세 조정하여 개선된 것을 확인하였다. 정밀도외에도 RMSE(Root-Mean-Square Error) 측면에서 예상심각도 수준과 실제 버스의 정시정각 성능수준 간의 차이점은 (식 1)과 같이 측정할 수 있다.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \dots\dots\dots (식 1)$$

이는 버스의 정시정각 성능(극히/매우/약간지연 또는 정시)에 대한 예상심각도 수준을 나타내는 것으로 정시정각 인스턴스의 총 수(n)에 따라 초기분류 RMSE가 0.880863으로 도출되어 개선된 분류 RMSE는 0.741962로 감소된 것으로 확인하였다.

4. 결론

개방형 데이터의 고가용성으로 데이터 마이닝은 데이터의 가치를 찾을 수 있는 매우 유용한 방법이 되었다. Winnipeg시는 개방형 데이터 포털을 통해 Winnipeg Transit의 정시정각 성과에 대한 개방형 빅 데이터를 제공하였다. Winnipeg Transit 데이터는 버스의 출/도착 지연에만 초점을 맞춘 유사 데이터와는 달리 늦은 출/도착 버스뿐만 아니라 조기 및 정시 출/도착 버스의 경우 더 넓은 커버리지를 제공한다. 이 연구에서는 이러한 개방형 빅 데이터의 가용성을 감안하여 자주 발생하는 패턴을 검색하기 위해 데이터를 분석할 수 있는지? 과거 데이터를 기반으로 향후 버스의 지연을 예측하는 것이 가능한지?에 대해 답변을 제시하였다. 특히 버스의 운행성능에 대한 개방형 빅 데이터를 빈번한 패턴 마이닝으로 검토하고 분석하여 의사결정트리를 기반으로 분류 예측을 하였다. 실제 Winnipeg시에 대한 예측 분석을 수행한 결과, 빈번한 패턴/연결규칙/연관분류 규칙/의사결정 트리/예측 등 매우 다양한 형태로 흥미로운 지식을 발견하였다. 빈번한 패턴은 자주 발생하는 특성 값을 나타냈으며, 연결규칙은 몇 가지 다른 특성 값을 가질 가능성이 높은 것으로 나타났다. 결과적으로 이 연구에서는 스마트 운송 서비스를 지원하기 위한 개방형 빅 데이터에 대한 예측 분석의 이점을 제시하였다.

기하급수적으로 생성되고 있는 방대한 양의 빅 데이터 수집이 빠르게 증가하면서 개방형 빅 데이터를 공동으로 연구하고 분석하는 추세가 확산되고 있다. 이 연구에서는 대중교통수단(버스)을 이용한 대다수 출퇴근 통근자의 시간지연을 최소화할 수 있는 대중교통 활성화를 위한 개방형 빅 데이터 분석정보를 제시하고 있다. 이 연구결과는 국내 도심지역의 대중교통(버스 등) 활성화를 위한 정책에 벤치마킹 자료로 활용할 수 있을 것이다.

[참고문헌]

- [1] Lakshmanan, L.V.S., C.K. Leung, & R.T. Ng (2000) "The segment support map: scalable mining of frequent itemsets," ACM SIGKDD Explorations 2(2), pp. 21-27.
- [2] Leung, C.K. (2009) "Frequent itemset mining with constraints," in Encyclopedia of Database Systems, pp. 1179-1183.
- [3] Sassi, M.S.H., F.G. Jedidi, & L.C. Fourati (2019) "A new architecture for cognitive internet of things and big data," Procedia Computer Science 159. KES 2019, pp. 534-543.
- [4] Snijders, C., U. Matzat, & U. Reips (2012) "'Big data': big gaps of knowledge in the field of internet," Int. J. Internet Science 7, pp. 1-5.
- [5] Camara, R.C., Cuzzocrea, A., Grasso, G.M., Leung, C.K., Powell, S.B., Souza, J., & Tang, B. (2018) "Fuzzy Logic-Based Data Analytics on Predicting the Effect of Hurricanes on the Stock Market," in FUZZ-IEEE 2018, pp. 1-8.
- [6] Braun, P., Cuzzocrea, A., Leung, C.K., Pazdor, A.G.M., Tanbeer, S.K., & Grasso, G.M. (2018) "An Innovative Framework for Supporting Frequent Pattern Mining Problems in IoT Environments," in ICCSA 2018, pp. 642-657
- [7] Ivancevic, V., & I. Lukovic (2018) "National university rankings based on open data: a case study from Serbia," Procedia Computer Science 126. KES 2018, pp. 1516-1525.
- [8] Kassen, M. (2013) "A promising phenomenon of open data: a case study of the Chicago open data project," Government Information Quarterly 30(4), pp. 508-513
- [9] Audu, A.A., A. Cuzzocrea, C.K. Leung, K.A. MacLeod, N.I. Ohin, & N.C. Pulgar-Vidal (2019) "An intelligent predictive analytics system for transportation analytics on open data towards the development of a smart city," in CISIS 2019, pp. 224-236.

<원문제목> Predictive analytics on open big data for supporting smart transportation services
 <원문출처> Procedia Computer Science 176 (2020) 3009-3018