Journal of the Korean Data & Information Science Society
2015, **26**(5), 1061–1069

http://dx.doi.org/10.7465/jkdi.2015.26.5.1061
한국데이터정보과학회지

LDA 기법을 이용한 버스 승객의 잠재적 이동패턴 분석[†]

조아[1] · 이경희[2] · 조완섭[3]

[12]충북대학교 비즈니스데이터융합학과 · [3]충북대학교 경영정보학과

접수 2015년 8월 20일, 수정 2015년 9월 21일, 게재확정 2015년 9월 23일

요 약

최근 교통 분야에서 발생하는 교통 빅데이터 (교통카드 데이터, ATMS 데이터 등)의 분석결과를 교통 정책에 활용하는 사례가 늘어나고 있는 추세이다. 또한 교통 데이터 분석 기법을 기존의 단순 빈도 분석 기법에서 다양한 데이터 마이닝 기법으로 확장하여 교통 데이터 속에 숨어있는 의미를 파악하려는 연구도 진행되고 있다. 본 연구에서는 교통카드 데이터에 대하여 토픽모델링 기법 중의 하나인 LDA (Latent Dirichlet Allocation) 기법을 적용하여 청주시 버스 승객들의 이동패턴을 분석한다. 이를 위해 교통카드 데이터의 하차 결측치를 추정하고, LDA 기법을 적용하여 이동패턴을 추출하였다. 또한 LDA 분석으로 도출된 값을 측정값으로 하여 다차원적 분석을 함으로써 청주시 버스 승객들의 이동패턴 특징을 파악할 수 있다. 분석 결과, 청주시의 경우 크게 1) 시외지역에서 터미널을 이용해 청주시에서 유입되는 패턴, 2) 주거지역에서 상업지역으로 이동하는 패턴, 3) 청주 인근 학교에서 상업 지역 (청주 중심가)로 이동하는 패턴을 발견할 수 있었다. 이동패턴은 도시 계획, 대중교통 서비스 향상, 버스 노선 신설 등 다양한 교통정책의 수립에 활용될 수 있을 것으로 기대된다.

주요용어: 교통 카드, 다차원 분석, 빅데이터, LDA (Latent Dirichlet Allocation)

1. 머리말

빅데이터를 활용한 다양한 분석사례가 발표되고 있다. Cho (2015)는 관광 분야에서 빅데이터를 적용한 사례이고, Kang (2015)는 스포츠 분야에서 빅데이터를 적용한 사례이다. 최근 들어서는 ICT (Information Communication Technology) 기술이 발전함에 따라 교통 빅데이터 (교통카드 데이터, ATMS 데이터 등)를 의사결정에 활용하는 사례가 늘어나고 있는 추세이다. 특히 교통카드 데이터는 대중교통 승객의 특성을 파악하고 다양한 교통정책을 수립할 수 있어 각광받고 있다. 하지만 수도권 외의 지역은 이동거리와 무관한 동일 요금체계 이기 때문에 하차 할 때 승객이 교통카드를 태그하지 않아 승객의 특성을 파악하는데 어려움이 있다.

또한, 교통 데이터 분석 기법을 기존의 단순 분석 기법에서 텍스트 마이닝 기법으로 확장하여 교통 데이터 속에서 숨어있는 의미를 파악하려는 연구도 진행되고 있다. 최근 프랑스에서는 교통 카드 데이터를 텍스트 마이닝 기법인 유니그램 (unigram) 모델을 적용하여 승객의 패턴을 분석한 사례가 있다.

본 논문에서는 청주시 교통카드 데이터에 누락된 하차정보를 추정하는 기법과, 이에 LDA 기법에 적용하여 청주시 버스승객의 주요 이동패턴을 분석하고자 한다. 특히 다차원 분석을 연계함으로써 승객의 주요 이동패턴의 특성을 찾아낸다.

[†] 본 논문은 2014년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었음.
[1] (362-763) 충북 청주시 서원구 충대로 1, 충북대학교 비즈니스데이터융합학과, 석사과정.
[2] (362-763) 충북 청주시 서원구 충대로 1, 충북대학교 비즈니스데이터융합학과, 연구원.
[3] 교신저자: (362-763) 충북 청주시 서원구 충대로 1, 충북대학교 경영정보학과, 교수.
E-mail: wscho@cbnu.ac.kr

본 논문의 구성은 다음과 같다. 제 2절에서는 교통카드를 활용한 연구와 LDA 기법 관련 연구를 설명한다. 제 3절에서는 연구의 주요 내용 및 방법을 제시한다. 제 4절에서는 분석 결과를 설명한다. 마지막으로 제 5절에서 결론을 맺는다.

2. 관련 연구

본 장에서는 교통카드 데이터를 활용한 연구와 LDA 기법을 적용한 연구들을 소개한다. 이러한 연구는 아직 연구논문으로 발표된 것은 많지 않으며, 일부 기업이나 공공기관의 연구 보고서로 발표되고 있는 실정이다.

Kim (2007)에서는 교통 카드 데이터에서 필연적으로 발생하게 되는 데이터의 오류 및 결측 현황을 분석하고 보정할 수 있는 방안을 제시하였다. 특히 교통 데이터의 신뢰도 향상 및 전수화 관련 연구의 필요성을 강조하였다.

Kim 등 (2013)에서는 서울시 지하철에서 수집된 실 데이터를 이용하여 이동패턴을 추출하고, 지리적으로 유사하면서 동일한 기능을 수행하는 Zone을 발견하고 이들 간의 연관성을 파악하였다. 또한 추출된 이동패턴을 정량적으로 평가하기 위한 지표를 제안하였다.

Chu 등 (2014)에서는 GPS를 통해 수집되는 택시 트랜잭션 데이터를 LDA 기법을 적용하여 택시의 토픽 (이동패턴)을 분석하였고, 추출된 이동 패턴을 시각화하였다. 그 외에도 Mohamed 등 (2014)에서는 프랑스의 르넨 지역의 교통카드 데이터를 유니그램 모델을 적용하여 토픽을 추출하고 추출된 토픽을 지역의 특성과 결합하여 분석하였다. 또한 Park(2014)에서는 각 지역의 연령별, 교육정도별, 주택 유형별, 사업구분별 통계적 수치에 주성분분석 (PCA) 기법을 적용하여 특성을 추출하고, 확률적인 토픽 모델링 기법 (hLDA)을 적용하여 추출한 지역의 의미와 이동패턴을 분석하였다. 본 논문과는 확률적 토픽 모델링 기법을 적용했다는 점에서는 유사하나, 추출된 이동패턴을 분석하는 데 있어서 본 논문은 DW구축을 통해 다차원적으로 분석을 했다는 점에서 차이를 보인다.

3. 연구 내용 및 방법

3.1. 전체 프로세스

본 연구의 전체 프로세스는 Figure 3.1과 같다. 개별 승객 식별이 가능한 암호화된 고유번호가 포함된 한달 간의 교통카드 데이터를 수집하였다. 그 다음 하차 정보를 추정해 승객의 승/하차 DB를 구축하였다. LDA 분석을 위해 문서-단어관계로 데이터구조를 변환하였으며 LDA 분석모듈을 통해 승객들의 주요 이동패턴을 도출하고 해당 승하차 패턴이 해당 이동 토픽에 해당될 확률을 도출하였다. 또한 도출된 결과를 해석하기 위해 DW를 구축하여 분석하였다.

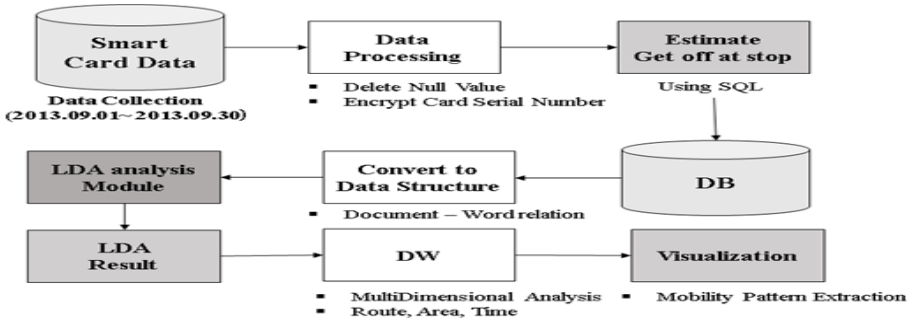

Figure 3.1 Total process of the research

3.2. 하차 추정

승객의 하루 통행은 집에서 출발하여 집으로 돌아온다는 기본적인 개념에 근거하여 두 통행은 서로 보정이 가능하다는 보정방안을 참고하였다 (Kim, 2007). Figure 3.2와 같이, 특정인 X가 A정류장에서 승차하여, B정류장에 하차했을 때, B정류장을 다음 승차정류장 (C)으로 보정하였다. 또한, D정류장 (마지막 하차 정류장)은 첫 번째로 승차한 정류장 (A)로 보정하였다. 단, 하차 정보가 결측된 통행에 이어 다음 통행이 존재하는 경우 이 방식의 적용이 가능하다.

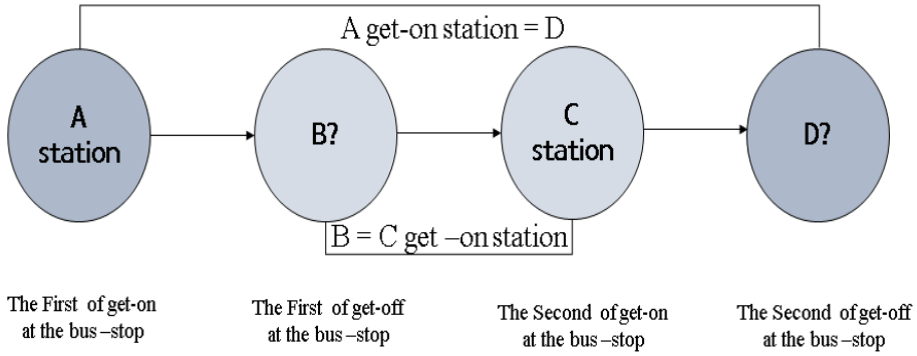

Figure 3.2 The Basic Concept of get-off estimation

3.3. 이동패턴 추출

본 논문은 버스 승객의 이동패턴을 파악하기 위해 LDA 분석 기법을 이용하였다. 여기서 LDA는 대용량의 문서에서 주제를 추출하기 위한 기법으로, 동시 발생하는 확률을 기반으로 유사한 문서들을 클러스터링 하는 기법이다. 기존의 LDA 분석 기법은 대용량의 텍스트 문서에 나타난 토픽 (주제)을 추출하는 기법으로 사용되었으나, 본 논문에서는 승객 1명의 거래데이터를 하나의 문서로 간주하고 승객의 교통카드 거래데이터에 포함된 승하차 패턴을 단어로 인식하여 LDA 분석 기법을 적용하였다. 본 논문에서는 Park (2014) 논문을 참고하여 '토픽(주제)'를 '이동 토픽'으로 명시하였다. '이동 토픽'은 유사한 승하차패턴을 나타내는 승객들을 그룹화한 결과를 의미한다. Table 3.1은 기존의 LDA 기법에 사용된 용어와 본 논문에서 사용한 용어 표기를 나타낸 것이다.

Table 3.1 The Concept of LDA

The notation of LDA Algorithm	The notation of this paper
Document (D)	Transaction data (D)
word (ω)	Get-on/off pattern (ω)
Topic (T)	Mobility topic (T_N) ($0 < N < 11$)
Distribution over words ($p(w\|T)$)	Distribution over get-on/off pattern ($p\,(t_n/T_N)$)

이동토픽 T_N은 여러 개의 이동패턴 t_n으로 구성되며, 본 논문에서는 1개월간의 교통카드데이터로부터 10개의 이동토픽을 찾아내기 위해 N을 10으로 제한하였다.

Algorithm 1 Mobility topic

```
conn = connectdb(host, user, passwd, database)
cur = conn.cursor()
pasDoc = []
Doc=[]
cur.execute('SELECT passengerID, ODpattern FROM table')
for row in cur:
    Doc.append(row['passengerID'])
    pasDoc.append(row['ODpattern'])
for pasdoc in pasDoc :
    pasdoc2 = pasdoc[:-1]
    pasword = pasdoc2.lower().split(",")
    word_list.append(pasword)
dictionary = corpora.Dictionary(word_list)
dictionary.save('pas_dic_pattern.txt')
doc1 = ""
for stringtoken in padDoc:
    doc1 += stringtoken
new_vec = dictionary.doc2bow(doc1.lower().split(","))
corpus= [dictionary.doc2bow(text) for text in word_list]
tfidf = models.TfidfModel(corpus)
corpus_tfidf = tfidf[corpus]
ida = gensim.model.LdaModel(num_topic =10)
topic_word = []
for i in lda.show_topics(num_words = len(dictionary), num_topics=10):
      topic_word.append(i)
      i = 0, j=0
w = []
topic_id = 0
topic_id_list = []
for topic in topic_word :
      topic_id = topic_id + 1
      topic1 = topic.replace("+",",")
      topic2 = topic1.replace(" ", "")
      topic3 = topic2.split(",")
      for word in topic3 :
        w_array = word.split('*')
        topic_id_list.append(topic_id)
        w.append(w_array)
for token_ram in w :
      cur.execute("INSERT INTO result_tab VALUES(%d,%d%s)", topic_id_list[topic_count],
weight, str(pattern)))
conn.commit()
```

3.4. 다차원 분석

LDA 기법은 비지도 학습 (unsupervised learning)이기 때문에, 추출된 토픽이 무엇을 의미하는 지 알 수 없다. 본 장에서는 추출된 이동 토픽들의 특징을 이해하기 위해 다차원 분석을 수행하였다. 교통카드 데이터에서 발생하는 시간, 정류장명, 노선 컬럼과 LDA 분석을 통해 도출된 승하차 패턴별 각 이동 토픽의 확률 (측정값)로 구성된다.

측정값은 승하차 패턴 (승차정류장-하차정류장)과 LDA 분석의 결과값인 승하차 패턴별 각 이동 토픽의 확률값을 승하차 패턴을 기준으로 조인 (join)하여 하나의 레코드 (승하차 패턴)에 토픽의 개수 (N) 만큼 각각의 이동 토픽에 할당된 측정값이 나타난다.

차원은 사실 테이블의 발생시간, 정류장명, 노선 컬럼, 이동패턴 컬럼은 시간 테이블, 노선 테이블, 지역 특성 테이블, 토픽 테이블 3개의 차원과 연계하여 이동 토픽을 다차원적으로 분석하였다.

4. 분석 결과

Figure 4.1부터 Figure 4.5는 주요 이동토픽을 주요 노선과 연계하여 표현한 것이다. 각각 왼쪽의 그림은 이동토픽별 주요 이동패턴을 지도상에 표현한 것이며, 오른쪽 그래프는 버스노선별 다차원 분석한 결과이다.

(1) 이동토픽 1번

이동토픽 1번은 다른 이동토픽과 비교하여 502번 노선의 성격이 가장 강하게 나타났다. 502번 노선은 조치원역 (외곽지역)에서 동부종점 (청주 시내)로 이동하는 노선이므로, 추출된 패턴이 조치원역에서 시외버스터미널 패턴으로 보아, 시외지역에서 터미널/역/을 이용해 청주시로 유입하는 패턴임을 알 수 있다. 또한, 502번 노선이 경유하는 현대 3차 아파트 앞, 수곡우체국에서 육거리 또는 도청으로 이동하는 패턴이 나타났다. 즉, 이동토픽 1번은 502번 버스에 탑승하여 외곽지역에서 청주시내로 이동하는 패턴과 이들이 청주시내에서 이동하는 패턴이 나타났음을 알 수 있다.

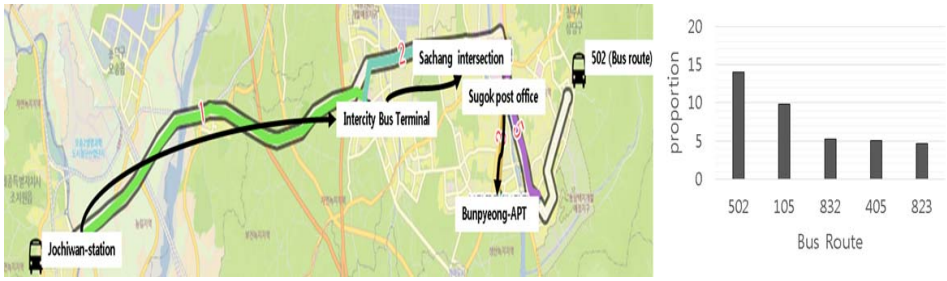

Figure 4.1 The result of mobility topic 1 (Cho Ah, 2015)

(2) 이동토픽 2번

이동토픽 2번은 한국교원대학교에서 시외버스터미널로 이동하는 패턴의 비율이 가장 높게 나타났다. 이를 세부적으로 분석하기 위해 노선별, 지역별, 용도별로 다차원적으로 분석해본 결과, 다른 이동 토픽과 비교하여 이동토픽 2번은 513번 노선의 성격이 가장 강하게 나타났고, 이는 513번이 한국교원대학교에서 청주시 중심가로 유입되는 주요 노선이기 때문인 것으로 보인다. 또한, 이동 토픽 2번의 상위 승하차 패턴에 속하는 청주대학교, 분평 주공 2단지에서 도청으로 이동하는 패턴이 나타나는 것으로 보아, 513번 버스를 타고 청주 중심가로 이동한 후, 다른 지역으로 이동할 때 105번 노선 또는 713번 노선을 타고 이동하는 것으로 보인다. 한국교원대 (외곽)에서 버스를 타고 청주 시내로 유입한 후, 청주대학교, 분평동으로 이동하는 패턴임을 알 수 있다.

Figure 4.2 The result of mobility topic 2 (Cho Ah, 2015)

(3) 이동토픽 3번

이동토픽 3번은 이동토픽들 중 가장 높은 비율을 나타낸 이동토픽이다. 시외버스터미널에서 지하상가, 지하상가에서 사창사거리로 이동하는 패턴이 가장 높게 나타났다. 이를 세부적으로 분석하기 위해 노선별, 지역별, 용도별로 다차원적으로 분석해본 결과, 다른 이동토픽과 비교하여 이동토픽 3번은 105번 노선과 824번 노선의 비율이 가장 높게 나타났다. 이는 시외버스터미널에서 청주 중심가로 이동하는 패턴임을 알 수 있다. 즉, 시외버스터미널에서 청주중심가를 이동할 때 청주시 버스 승객들은 105번 노선과 824번 노선을 타고 이동하는 것을 알 수 있다. 즉, 동일한 출발지 (터미널)와 목적지 (청주중심가)를 버스 승객들이 어떠한 방법으로 이동하는지를 알 수 있다.

Figure 4.3 The result of mobility topic 3 (Cho Ah, 2015)

(4) 이동토픽 4번

이동토픽 4번은 증평우체국에서 한국교통대로의 이동패턴 비율이 가장 높게 나타났다. 이를 세부적으로 분석하기 위해 노선별, 지역별, 용도별로 다차원적으로 분석해본 결과, 다른 이동토픽과 비교하여 학교와 관련된, 충청 대학교, 한국 교통대, 청주교육대, 충북대학교 학교에서 청주의 중심가로 이동하는 패턴과 111번 노선의 성격이 강하게 나타났다. 111번 노선은 청주 중심가에서 한국교통대로 이동하는 주요 노선에 해당하므로 이동토픽 4번은 청주 중심가에서 한국교통대로 이동하는 패턴임을 알 수 있다. 각각 출발지는 시외버스터미널 (105번 노선 경유), 청주교육대학교 방향 (111번 노선)으로 다르지만, 청주 중심가를 지나 증평우체국까지 가는 방향은 동일한 것을 볼 수 있다. 즉, 한국교통대로 이동하는 버스 승객들은 111번 노선과 105번 노선을 이용하고, 105번 노선을 승차한 승객들은 증평우체국에서 하차해서 111번 노선으로 환승해 한국교통대로 이동하는 패턴을 확인 할 수 있다.

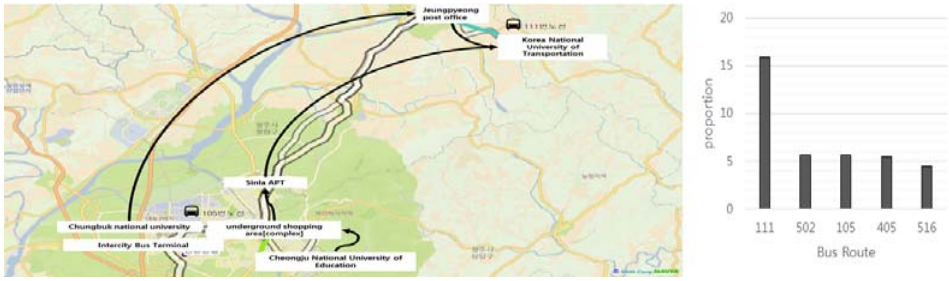

Figure 4.4 The result of mobility topic 4 (Cho Ah, 2015)

(5) 이동토픽 5번

이동토픽 5번은 시외버스터미널에서 신한은행, 육거리에서 한솔초등학교로의 이동패턴 비율이 가장 높게 나타났다. 이를 세부적으로 분석하기 위해 노선별, 지역별, 용도별로 다차원적으로 분석해본 결과,

다른 이동토픽과 비교하여 30-1번 노선 성격이 가장 강하게 나타났다. 이는 중심가로 이동할 때, 수곡동 또는 산남동에서 청주 중심가인 육거리로 이동하는 패턴임을 알 수 있다. 즉, 산남동에서 육거리로 이동하는 버스 승객들이 30-1번 노선을 타고 이동하는 토픽임을 알 수 있다.

Figure 4.5 The result of mobility topic 5 (Cho Ah, 2015)

5. 결론

청주 광역권 대중교통체계 개편전략에 의하면, 청주시 시내버스 노선체계는 전체 노선의 대부분이 사직로와 상당로에 집중되어 있어 과도하게 도심 집중 현상이 나타나며, 이는 노선체계가 수익금 위주로 편성되어 있음을 반영하는 결과라고 한다.

본 논문에서 LDA를 통해 추출된 주요 패턴 또한 사직로와 상당로에 위치한 도청, 지하상가 정류장으로 이동하는 패턴으로 나타났다. 차이점은 각 이동토픽별로 도청, 지하상가 정류장과 같은 청주 중심가로 이동하는 패턴은 동일하나, 중심가로 이동하는 출발지가 다르거나 버스 승객들이 이용하는 노선이 다르게 나타났다. 이동토픽 1번과 2번은 조치원역 또는 한국교원대가 위치한 청주의 외곽지역에서 중심부로 이동하는 패턴, 이동토픽 3번은 시외버스터미널에서 청주 중심부로 이동하는 패턴, 이동토픽 4번은 한국 교통대 (증평)에서 청주 중심가로 이동하는 패턴이 나타났다. 이동토픽 5번은 청주 남부 생활권인 수곡동과 산남동에서 이동하는 패턴이 발견되었다. 즉, LDA 분석으로 도출된 이동토픽들을 다차원적으로 분석함으로써 각각의 이동토픽별로 어떠한 특징을 가지고 있는지 알 수 있었다. 또한, 이동 토픽 4번이 한국교통대로 이동하는 패턴임을 알 수 있듯이, 유사한 패턴을 보이는 승객을 그룹화하여 해당 이동토픽의 이용자의 특성도 유추할 수 있었다.

서울과 수도권을 제외한 지방자치단체는 버스 승객의 하차 정보가 없어 버스 승객의 이동패턴을 찾고자 하는 노력을 많이 기울이고 있다. 이동패턴을 찾게 되면 대중교통 서비스 개선을 통해 시민의 만족도를 향상시키고 이것은 시민의 삶의 질을 높이는 데 활용될 수 있기 때문이다. 본 논문의 연구결과는 실제 교통카드 데이터를 이용하여 승객들이 많이 이용하는 이동 패턴을 분석했다는 점에서 의미가 있다. 이는 도시 계획, 대중교통 서비스 향상, 버스 노선 신설 등에 활용될 수 있을 것으로 기대된다.

References

Cho, A. (2015). *Mobility pattern analysis of bus passengers with LDA*, Master Thesis, Chungbuk National University, CheongJu.

Cho, W. S., Cho, A., Kwon, K. and Yoo, K. H. (2015). Implementation of smart chungbuk tourism based on SNS data analysis. *Journal of the Korean Data & Information Science Society*, **26**, 409-418.

Chu, D., Sheets, D., Zhao, Y., Wu, Y., Yang, J., Zheng, M. and Chen, G. (2014). Visualizing hidden themes of taxi movement with semantic transformation. *Pacific Visualization Symposium (PacificVis)*, 137-144, IEEE.

Kang, B., Huh, M. and Choi, S. (2015). Performance analysis of volleyball games using the social network and text mining techniques. *Journal of the Korean Data & Information Science Society*, **26**, 619-630.

Kim, K. H., Oh, K. H. and Lee, Y. K. (2013). Discovery of travel pattern in seoul metropolitan subway using big data of smart card transaction systems. *Society for e-Business Studies*, **18**, 211-222.

Kim, S. K. (2007). *The estimation and application of origin-destination tables by using smart card data*, Research Report, Seoul Development Institute, Korea.

Mohamed, K., Côme, E., Baro, J. and Oukhellou, L. (2014). *Understanding passenger patterns in public transit through smart card and socioeconomic data*, UrbanComp.

Park, H. S. (2014). *The pattern extraction and mobility analysis of transportaion mobility using Topic modeling*, `http://www.si.re.kr/node/50607`, Seoul 2014 research paper contest using public data.

Journal of the Korean Data & Information Science Society
2015, **26**(5), 1061–1069

http://dx.doi.org/10.7465/jkdi.2015.26.5.1061
한국데이터정보과학회지

Latent mobility pattern analysis of bus passengers with LDA[†]

Cho Ah[1] · Lee Kyung Hee[2] · Cho Wan Sup[3]

[12]Department of Business Data Convergence, Chungbuk National University
[3]Department of Management Information System, Chungbuk National University

Received 20 August 2015, revised 21 September 2015, accepted 23 September 2015

Abstract

Recently, transportation big data generated in the transportation sector has been widely used in the transportation policies making and efficient system management. Bus passengers' mobility patterns are useful insight for transportation policy maker to optimize bus lines and time intervals in a city. We propose a new methodology to discover mobility patterns by using transportation card data. We first estimate the bus stations where the passengers get-off because the transportation card data don't have the get-off information in most cities. We then applies LDA (Latent Dirichlet Allocation), the most representative topic modeling technique, to discover mobility patterns of bus passengers in Cheong-Ju city. To understand discovered patterns, we construct a data warehouse and perform multi-dimensional analysis by bus-route, region, time-period, and the mobility patterns (get-on/get-off station). In the case of Cheong Ju, we discovered mobility pattern 1 from suburban area to Cheong-Ju terminal, mobility pattern 2 from residential area to commercial area, mobility pattern 3 from school areas to commercial area.

Keywords: Big data, LDA (Latent Dirichlet Allocation), multidimensional-analysis, transportation card.

[†] This work was supported by the research grant of the Chungbuk National University in 2014.
[1] Master student, Department of Business Data Convergence, Chungbuk National University, Chungbuk 362-763, Korea.
[2] Researcher, Department of Business Data Convergence, Chungbuk National University, Chungbuk 362-763, Korea.
[3] Corresponding author: Professor, Department of Management Information System, Chungbuk National University, Chungbuk 362-763, Korea. E-mail: wscho@cbnu.ac.kr