하둡을 활용한 서울시 공공자전거 '따릉이' 이용 실태 분석 설계

양예슬*, 최재희*, 김윤희**
*숙명여자대학교 컴퓨터과학부
**숙명여자대학교 컴퓨터과학부 교수
e-mail : yulan@sookmyung.ac.kr

A Design of Utilization Analysis of Seoul's Public Bicycle 'Ttareugi' Using Hadoop

Ye-Seul Yang*, Jae-Hee Choi*, Yoonhee Kim**
*Dept of Computer Science, Sookmyung Women's University
**Corresponding author

요 약

서울시에서 2015 년부터 시행한 공공자전거 서비스 '따릉이'에 대한 시민들의 수요가 늘고 있다. 하지만 적합하지 않은 스테이션의 위치 선정으로 인력이 낭비되고 있으며, 시민들의 불만의 소리도 높아지고 있다. 또한, 서울시 측에서 제시한 스테이션의 위치 선정과 설치 규모에 대한 기준이 모호한 상태이다. 본 논문에서는 공공자전거 서비스의 효율적인 운영을 위해, 따릉이의 실제 이용 현황에 대해 분석하고자 하였다. 하둡 분산파일시스템을 활용하여 따릉이 이용률 데이터를 수집하였으며, 수집된 데이터를 바탕으로 서울시 행정구역별로 따릉이 이용 실태를 분석하고 향후 설치되어야 할 스테이션의 규모를 추론하였다. 서울시는 향후 스테이션의 설치 행정구역을 늘려나갈 것이라고 발표한 바 있으므로, 이 예측 결과를 적용한다면 효율적인 스테이션의 설치가 가능할 것이다.

Ⅰ 서론

서울시에서는 교통체증, 대기오염, 고유가 문제를 해결하고 시민들의 삶의 질을 높이기 위한 목적으로, 2015 년 9 월부터 서울 자전거(이하 '따릉이')를 운영하고 있다[2]. 2016 년 8 월, 따릉이는 이용자 수 20 만 1,242 명을 기록하였으며, 서울 시민들의 따릉이에 대한 수요는 매년 증가하고 있다.

기존의 스테이션은 적절하지 못한 위치 선정과 설치 규모로 인하여, 전문인력을 동원하여 따릉이를 빈번하게 재배치해야 한다는 문제가 발생해왔다. 따라서 현재 이용 실태를 조사하고 어떠한 요소들이 이용률에 영향을 미치는가에 관한 분석은 사회비용 차원에서 매우 중요한 문제이다.

앞서 국내 따릉이의 통행 형태나 경제성 등에 관한 연구가 있었다[3.4.5]. 이러한 연구에서는 도시 전체의 관점에서 공공 자전거 규모를 파악하는 것에 대해 주로 이야기하고 있으며, 공공자전거 스테이션의 위치선정과 설치 규모를 산정하는 구체적인 연구는 미미하였다. 또한, 서울시에서는 11 개의 행정구역에 따릉이 스테이션을 설치하였지만, 따릉이의 이용 실태와 추후 설치되어야 할 스테이션의 규모를 제안한 연구활동은 없었다.

따라서, 본 논문에서는 하둡을 활용하여 따릉이 이용률에 영향을 미치는 요소들에 대해 검토하며, 스테이션이 설치 되지 않은 서울시 행정구역에 적절한 스테이션 규모를 제안한다. 이를 위해 먼저 기존의 스테이션 행정구별, 시간대별 이용 현황을 조사한다. 또한 행정구역 내 위치한 통행장소 규모를 각각 측정한다. 이를 위해 본 연구팀에서는 2016 년 12 월 1 일부터 2016 년 12 월 7 일까지 평일 5 일에 걸쳐 실시간 이용률 데이터를 웹 크롤러를 통해 직접 수집하였다. 본 연구팀의 분석은 향후 서울시에서 따릉이 서비스의 활성화를 위한 노력의 초석이 될 것으로 기대된다.

Ⅱ 본론

Ⅱ-1. 측정 방법

본 연구팀은 표 1 과 같이 2016 년 12 월 1 일부터 2016 년 12 월 7 일까지 시간대에 따른 따릉이 이용 현황 데이터를 웹 크롤러를 구축하여 수집하였다. 이 때 수집한 데이터의 양은 9,879KB 이고, 수집에 사용

된 웹 크롤러는 통계 분석 도구 R 을 이용하여 자체 구축하였다. 이 크롤러는 따릉이 웹 페이지에서 제공되는 스테이션 실시간 현황 페이지의 데이터를 15 분마다 수집하며, 데이터는 주기적으로 하둡 분산파일 시스템에 병렬로 저장된다. 데이터의 칼럼으로는 스테이션 이름, 스테이션 규모, 대여 가능 따릉이 수, 스테이션 위치로 구성되어 있다.

표 1. 측정 환경

수집일	2016 년 12 월 1 일 00:00 ~ 2016 년 12 월 7 일 23:00
수집 데이터	스테이션 이름, 스테이션 규모, 대여 가능 따릉이 수, 스테이션 위치
수집 장비	R 로 구현한 자체 웹크롤러
수집 환경	R-3.3.2
수집 대상	서울자전거 따릉이 웹페이지 [2]
수집 데이터 규모	9,879kB

따릉이 이용률에 영향을 미치는 요소들을 조사하기 위해 2016 년 12 월 9 일, 서울열린데이터광장[1]에서 표 2 와 같이 386KB 의 데이터를 수집하였다.

표 2. 수집 데이터

수집일	2016 년 12 월 9 일
수집 데이터	• 서울시 버스 정류소 노선도 다국어 목록정보 [224KB] • 서울메트로 지하철역 주소 및 전화번호 정보 [18KB] • 서울 도시철도공사 지하철 역별 주소 정보 [24KB] • 서울시 교육청 소재구별 학교 현황 [30KB] • 서울시 대학 및 전문대학 DB 정보 [16KB] • 서울시 우체국 정보 [73KB]
수집 대상	서울열린데이터광장
수집 데이터 규모	385KB

기존의 따릉이 스테이션 이용 실태를 수집하고, 이용률에 영향을 미치는 주된 요인을 분석하였다. 더불어, 현재 스테이션이 설치되지 않은 행정구역에 적합한 스테이션 설치 규모를 제안한다.

1) 스테이션 이용률에 따른 적합성을 분석한다. 이용률이 월등하거나 저조한 스테이션의 위치 정보를 행정구역별로 그루핑하여, 웹 인터페이스를 통해 시각화한다.
2) 스테이션 규모에 영향을 미치는 주된 요인을 파악한다. 기존의 스테이션 위치와, 설치된 행정구역 내 위치한 통행장소 규모와의 상관관계를 분석한다.
3) 행정구역별 적합한 스테이션 수를 제안한다. 1)과 2)에서 분석한 내용을 바탕으로, 스테이션이 설치되지 않은 행정구역 내에 설치를 제안한다.

본 논문의 시스템 기능을 구현하고 결과를 도출하기 위해서, 대용량의 데이터 처리에 용이한 하둡 분산처리시스템을 활용하였다. 약 10,000KB 의 따릉이 스테이션 위치 및 실시간 이용 데이터를 바탕으로 표 3 과 같이, 1 대의 Master 서버와 3 대의 Cluster 를 기반으로 분석하였다.

하둡을 활용하여 행정구역별 스테이션 이용률 분석, 각 스테이션 규모의 적합성 분석, 기존 스테이션 설치 규모에 영향을 미치는 요인과 스테이션의 상관관계 분석, 그리고 설치 예정 행정구역에 적합한 스테이션 규모 제안을 하였다.

표 3. 분석 환경

분석 장비	Hadoop-2.7.3
분석 환경	Ubuntu 12.04, 8.00GB RAM 1 대의 Master Server 와 3 대의 Cluster

본 시스템에서는 각 기능에 따라 key 값과 value 값을 적절히 설정하여 Map/Reduce 과정을 통해 얻은 결과를 이용하여 분석을 실시하였다. 또한 스테이션 이용률을 행정구역별로 그루핑하기 위해서 하둡의 정렬을 구현하여 활용하였다. 정렬에는 Composite key, Composite Comparator, Group key Partitioner, Group key Comparator 를 목적에 맞도록 구현하여 결과를 도출하였다.

행정구역별 스테이션 이용률 분석에서는, 앞서 구축한 웹 크롤러를 활용하여 수집한 데이터를 사전데이터로 사용하였다. 행정구역을 key 값으로 하여 스테이션의 고유 이름과 규모를 파악하고, 이용률을 계산하여 이용 현황을 분석하였다. 스테이션의 이용률은 '사용중인 공공자전거 수 / 총 공공자전거 수' 를 백분율로 나타낸 것으로 하였으며, 수식은 그림 1 과 같다.

스테이션 이용률(percentage) = 사용중인 공공자전

거 수 / 총 공공자전거 수 * 100

그림 1. 스테이션 이용률 수식

각 스테이션별 규모 적합성 분석에서는 행정구역과 시간대를 기준으로 스테이션의 하루 평균 이용률을 분석한다. 하루 평균 시간대별 이용률이 80% 이상인 스테이션은 사람들의 이용률이 높은 스테이션으로 분류하고, 20% 이하인 스테이션은 이용률이 낮은 스테이션으로 분류하였다.

기존 스테이션 설치 규모에 영향을 미치는 요인과 스테이션의 상관관계 분석에서는 서울시 버스 정류소 및 주요 지하철역 위치 정보와 서울시 소재구별 학교 현황, 그리고 우체국 등 다양한 관공서 위치와 스테이션 위치와의 상관관계를 분석한다. 행정구역을 key 값으로 하여 각 시설의 규모를 파악하여, 각각의 워드카운트 결괏값이 스테이션 위치와 어떠한 상관관계가 있는지 백분율로 나타내어 계산한다.

스테이션 설치 예정 행정구역에 적합한 스테이션 규모 제안에서는 앞서 계산한 그림 1 의 백분율을 활용하여 스테이션 설치 규모에 가장 많은 영향을 미치는 요인을 분석하고, 이를 통해 서울시의 스테이션을 추후 설치할 예정인 행정구역에 적합한 스테이션 설치 규모를 제안한다.

Ⅱ-2. 측정 결과

그림 2 는 따릉이 평균 이용률을 각 행정구역 별로 분석하여 그래프로 나타낸 것이다. 대여 가능한 따릉이 수는 15 분 단위로 수집하였다. 사전데이터의 '스테이션 위치' 항목에서 행정구역만을 따로 뽑아내도록 전처리 코드를 작성하였다. 이렇게 전처리된 행정구역을 Key 값으로 하여 Map/Reduce 과정을 통해 행정구역별로 스테이션의 평균 이용률을 계산하였다. 측정 결과, 동대문구를 제외한 행정구역은 평균 40%를 넘는 이용률을 보였다. 따라서 동대문구의 낮은 이용률에 대한 요인을 찾고자 하였다. 분석 결과, 이용률이 낮은 이유는 크게 두 가지로 압축되었다. 첫째, 스테이션의 위치 부적합성이다. 많은 사람이 접근하기에는 어려운 위치가 있는 경우이다. 둘째, 위치가 아닌 다른 외부 요인의 영향이다. 인구 규모, 연령대, 주거 형태, 근무 형태, 지리학적으로 자전거를 타기 어려움 등에 의해 이용률이 낮은 것을 알 수 있었다.

또한, 현재 스테이션이 설치된 서울시의 행정구역을 대상으로 스테이션의 평균 이용률을 분석한 결과, 같은 행정구역 내에서 규모의 조정이 필요한 스테이션은 전체 450 개의 스테이션 중 37 개로 나타났다. 그림 2 과 같이, 이 37 개의 스테이션 중에서 서대문구가 차지하는 비율은 18.6%, 광진구 17.2%, 마포구 11.7%, 종로구 10.4%, 동대문구 8.1%, 중구 6.4%, 용산구 4.7%, 성동구 2%, 영등포구 1.4%로 나타났다.

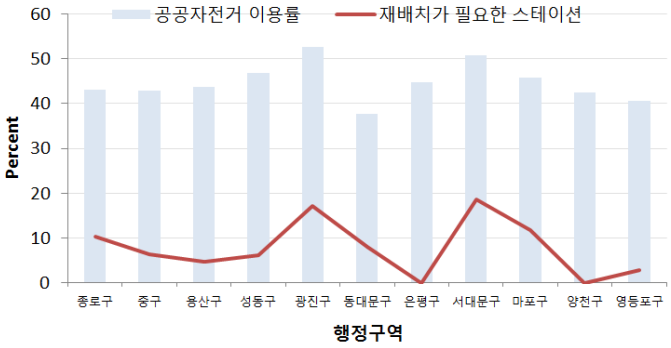

그림 2. 서울시 행정구별 공공자전거 이용률과 재배치가 필요한 스테이션

스테이션 규모와 상관관계를 가지는 요인을 파악하기 위해, 행정구역 내에 있는 통행장소 위치 정보를 바탕으로 각 비율을 계산한다. 이하 '계산비율'이라 칭한다.

각 요인에 대한 계산 비율을 각각 행정구역별 스테이션 비율과 대조해 보면, 행정구역별 스테이션 비율은 상대적으로 버스정류소 비율과 지하철역 비율과 유사한 패턴을 보였다. 따라서 버스 정류소와 지하철역을 한 요소로 묶어 '대중교통 요인'이라는 요소로 구분하고, 행정구역별로 스테이션의 비율을 계산한다. 그 결과 그림 3 과 같이, 과반수의 행정구역에서 대중교통요인 비율과 스테이션 개수의 비율에 관한 꺾은선 그래프가 서로 유사한 모양의 패턴을 보이는 것을 관찰할 수 있었다.

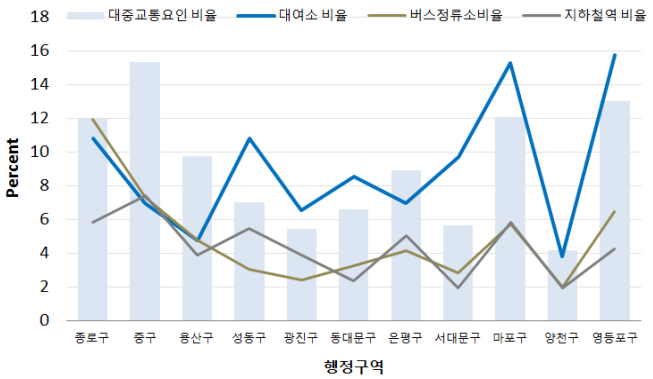

그림 3. 서울시 행정구역별 각 요인과 대조한 스테이션 비율

이를 통해 한 행정구역 내의 스테이션 규모는 대중교통 정류소 규모와 상관관계가 있으며, 스테이션 대비 대중교통 요인은 약 2.24336 의 비율로 존재함을

알 수 있다. 앞서 구한 약 2.24 의 비율을 알 때 대중교통 요인의 비율을 알고 있다면, 현재 따릉이 스테이션이 설치되어 있지 않은 행정구역에 설치되어야 할 따릉이 스테이션의 규모를 추론할 수 있다. 따라서, 다른 요인은 제외하고 대중교통 요인만을 고려하였을 때, 적합한 따릉이 스테이션 설치 규모는 그림 4 와 같이 추론할 수 있다. 강남구는 84 개로 가장 높았으며, 송파구 70 개, 노원구 49 개, 서초구 47 개, 강서구 46 개, 중랑구 40 개, 영등포구 37 개, 성북구 35 개, 강동구 29 개, 구로구 24 개, 도봉구 18 개, 강북구 18 개, 관악구 16 개, 금천구 7 개가 적합하다는 결과를 도출할 수 있었다.

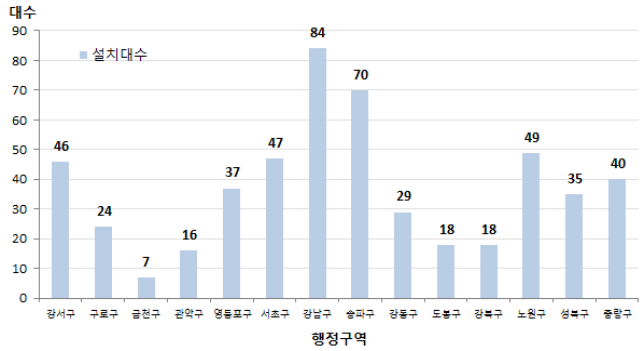

그림 4. 대중교통요인만을 고려한 스테이션 설치 제안

Ⅲ 결론

본 논문에서는 서울시 공공자전거의 이용 실태를 분석하였다. 규모조정이 필요한 스테이션의 수를 행정구역별로 분석하고, 향후 설치되어야 할 스테이션의 개수를 제안하였다. 이를 위해 웹 크롤러를 구축하여 실시간 이용 현황 수집하였으며, 하둡을 사용하여 지하철 출입구, 버스 정류소, 관공서, 학교 등 4 가지 통행장소와 스테이션 규모와의 상관관계를 분석하였다. 실험 결과, 서울시 전체 행정구역 내에서 규모 조정이 필요한 스테이션을 파악할 수 있었다. 같은 행정구역 내에서의 스테이션 규모의 조정은 거리에 따른 비용을 줄일 수 있으므로 유의미하다. 또한, 스테이션의 규모와 상관관계가 있는 요인은 버스 정류소와 지하철역의 개수로 나타났다. 이 분석을 토대로 서울시 공공자전거 스테이션의 향후 설치 규모를 예측하였다. 서울시는 향후 스테이션의 설치 행정구역을 늘려나갈 것이라고 발표한 바 있으므로, 이 예측 결과를 적용한다면 효율적인 스테이션의 설치가 가능할 것이다.

본 연구팀은 하둡을 활용하여 서울시 공공자전거 이용 실태 및 스테이션 규모 분석 시스템의 설계를 진행하였다. 그 결과, 규모의 조정이 필요한 스테이션을 파악하였으며, 추후에 설치되어야 할 스테이션 개수에 대한 추론이 용이하게 되었기 때문에 서울시 공공자전거 운영 정책에 빠르게 적용이 가능할 것으로 보인다.

본 논문의 결과가 서울시 공공자전거의 효율적인 운영 및 관리에 관한 연구활동에 초석이 되기를 기대한다.

참고문헌

[1] 서울열린데이터광장 (http://data.seoul.go.kr/)
[2] 서울자전거 따릉이 (https://www.bikeseoul.com/)
[3] 장재민, 김태형, 이무영. (2016)서울시 공공자전거 이용특성에 관한 연구. 서울도시연구, 17(4), 77-91.
[4] 김상현. (2016). 공공데이터를 이용한 오픈소스 기반 자전거 관리시스템 구현에 관한 연구 : 서울자전거 따릉이 시스템 제안. 숭실대학교 정보과학원
[5] 이문섭. (2017). 서울시 공공자전거 무인대여소의 이용특성 분석 : 동대문구 38 개 대여소의 이용비율 분석을 중심으로. 서울시립대학교