하둡 맵리듀스와 페이지 랭크를 이용한 서울시 대중 교통 인구 이동 분석

백민석, 오상윤
아주대학교 소프트웨어학과
white0825@ajou.ac.kr, syoh@ajou.ac.kr

Analysis of the population flow of public transportation in Seoul using Hadoop MapReduce and PageRank algorithm

Min-Seok Baek, Sangyoon-Oh
Dept. of Software Engineering, Ajou University

요 약

소셜 네트워크 및 웹 데이터와 같은 대규모 그래프 데이터를 처리하기 위해 병렬 처리 기반의 기법들이 많이 사용되어 왔다. 본 연구에서는 그래프 형식의 대규모 교통 데이터를 하둡 맵리듀스를 이용하여 처리하는 효과적인 기법을 제안한다. 제안하는 방식에서는 도시의 유동 인구 흐름을 가중치로 고려할 수 있도록 Weighted PageRank 알고리즘을 기반으로 하는 병렬 그래프 알고리즘을 사용하며, 해당 알고리즘을 하둡 맵리듀스에 적용하여 주거 및 근무지 등의 지역을 분류하도록 결과를 분석하였다. 제안 기법을 통한 분석 결과를 기반으로 지역 간 유동 인구 그래프 데이터에서 각 도시의 영향력을 측정하는 페이지랭크, 하둡 맵리듀스 기반의 기법을 제시한다.

1. 서론

교통 시스템. 소셜 및 통신 네트워크, 금융 시장 등 여러 분야에서 그래프 형식의 빅 데이터가 사용되고 있다. 순차적인 방법으로 대규모 그래프 데이터를 처리하는 것은 많은 비용과 시간이 소모되기 때문에 이러한 대규모 그래프 데이터의 처리에는, 많은 경우 그래프를 분할하여 병렬로 처리하는 기법을 사용한다. 분할 과정에서 그래프의 연결성이 유지되어야 하기 때문에 기존의 병렬 데이터 처리 방법을 적용했을 때 반드시 효과적이지 않으며, 따라서 병렬 실행 구조에 맞는 기법의 적용이 필요하다. 본 연구에서는 그래프 데이터에 대해 효과적인 분할 및 병렬 처리하는 방법론을 제시하고 이를 구현하여 그 효과성을 증명하였다. 특히 본 연구에서 대상으로 사용한 교통 분야의 데이터에 대한 마이닝 기법은 활발히 연구되고 있는 분야로서, 본 연구의 결과와 같이 인구 이동의 흐름을 분석하는 것을 통해 교통 중심지를 찾는 것에 도움이 될 수 있다.

본 연구에서는 행정동 간 대중교통 유동인구를 노드 간 간선의 가중치로 가지는 그래프 데이터로 모델링하여 어떤 행정동이 인구 이동 관점에서 더 영향력 있는 행정동인지, 그 중심성(Centrality)을 파악하고자 한다. 제시한 방법론의 효과성 확인을 위해 서울시 행정동 단위 대중교통 출발지/도착지 승객 수 정보 데이터[1]를 활용하였다. 데이터는 기준 날짜, 시작 행정동 ID, 종료 행정동 ID, 0~23 시의 각 1 시간 별 승객 수 등의 정보가 각 컬럼으로 정리되어 있었고 2022 년 8 월 12 일 데이터 기준으로 271504 개의 간선을 가진다. 분석과 이를 위한 실험은 데이터를 아침 시간과 저녁 시간으로 나누어 시간에 따른 행정동의 영향력 변화를 파악하고 그것이 지니는 의미를 분석한다.

2. 가중치를 고려한 페이지랭크(PageRank)

페이지랭크는 웹 검색 엔진이 사용자에게 표시되는 웹 페이지의 순위를 정하기 위해 1998 년 래리 페이지 등이 처음 제안한 알고리즘[2]으로 대규모 병렬 데이터 처리에 활용될 수 있는 대표적인 알고리즘이다. 이 페이지랭크 알고리즘은 노드의 영향력을 측정하는 청주시 교통 포인트 분석[3] 등 이전 연구에서 교통 데이터에 활용된 바 있다.

본 연구에서는 실험에 사용된 데이터에서 각 행정동을 노드로 보고 행정동 노드 간의 대중교통 이용 승객 수를 가중치로 한 유향 가중치 그래프(Directed Weighted Graph)로 모델링 할 수 있다. 이때, 가중치가 높은 노드 간의 연결이 더 강한 중심성을 가진다고 가정한다. 여기서 일반적인 Vanila PageRank 가 아닌 Weighted PageRank (WPR) 알고리즘

[4]을 기반으로 하여 중심성을 계산 한다.

$$PR(u) = (1-d) + d(\frac{PR(v_1)}{\sum W^{out}(v_1)} \cdot W(u,v_1) + \cdots + \frac{PR(v_n)}{\sum W^{out}(v_n)} \cdot W(u,v_n))$$

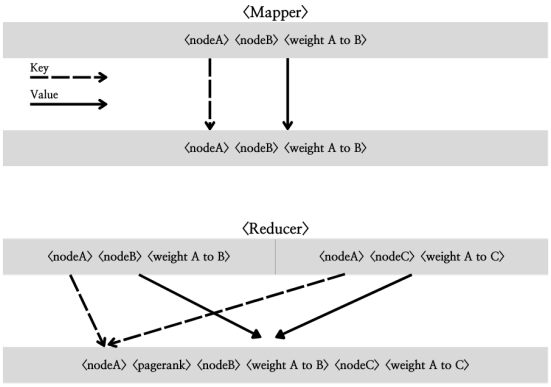
다음은 실험 구현을 위해 설정한 조건 및 정의들이다.

- 노드 u를 가리키는 다른 페이지들이 $v_1, v_2, \cdots, v_n$ 까지 있다고 가정한다.
- 파라미터 d는 damping factor 로 0 에서 1 사이의 값을 갖는다. (여기서는 0.9 로 한다.)
- $\sum W^{out}(v)$는 노드 v에서 밖으로 나가는 링크의 가중치를 모두 합한 값이다. 이때 $W(u,v_1)$ 는 노드 u에서 노드 v로 향하는 간선의 가중치이다.
- 노드 u의 페이지랭크 값을 $PR(u)$로 정의한다.

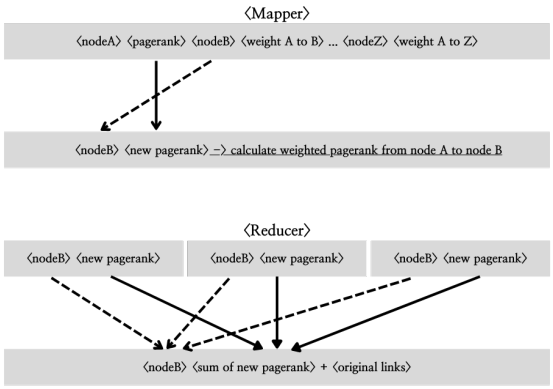
3. 맵리듀스 기반 병렬 그래프 처리

실험에서는 페이지랭크를 병렬 처리하기 위한 프레임워크로 하둡 맵리듀스(MapReduce) [5]를 사용한다. 맵리듀스는 대규모 데이터 셋을 처리하기 위해 사용되는 프로그래밍 모델이며, 하둡은 맵리듀스를 기반으로 작업의 병렬 실행을 가능하게 해주며 동시에 분산 시스템의 자원을 쉽게 사용할 수 있도록 해 주는 프레임워크이다. 사용자는 map 함수를 통해 키/값 쌍 기반의 데이터 셋을 만들고 reduce 함수를 통해서 데이터 셋들을 같은 키 별로 병합한다. 본 연구에서 각 행정동의 페이지랭크를 계산하는 전체 과정은 Job1, Job2, Job3 로 나누어 이루어지며 각 Job 은 다음과 같은 일을 처리한다.

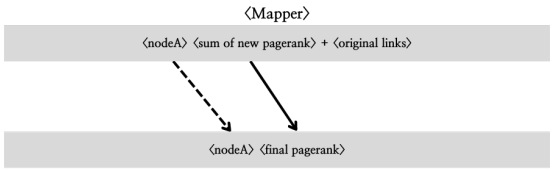
- Job1: 간선의 리스트로 이루어진 데이터를 vertex-based partitioning 할 수 있도록 각 노드에 연결된 모든 노드와 그 가중치의 리스트로 변환한다.
- Job2: 각 노드의 페이지랭크를 여러 번 반복하여 계산한다.
- Job3: 각 노드의 페이지랭크를 정렬하여 출력한다.



(그림 1) Job1: Mapper 와 Reducer 함수



(그림 2) Job2: Mapper 와 Reducer 함수.



(그림 3) Job3: Mapper 와 Reducer 함수

4. 실험 환경

실험은 Hadoop 완전 분산 모드로 진행하였으며, <표 1>과 같이 1 개의 네임 노드와 2 개의 데이터 노드를 할당하였다.

<표 1> 실험 PC 환경

노드	사양
네임 노드	CPU: Inter(R) i7-8700 @ 3.20GZ Memory: 32GB RAM Storage: 1TB HDD
데이터 노드 1	CPU: Inter(R) i7-8700 @ 3.20GZ Memory: 32GB RAM Storage: 1TB HDD
데이터 노드 2	CPU: Inter(R) i7-8700 @ 3.20GZ Memory: 32GB RAM Storage: 1TB HDD

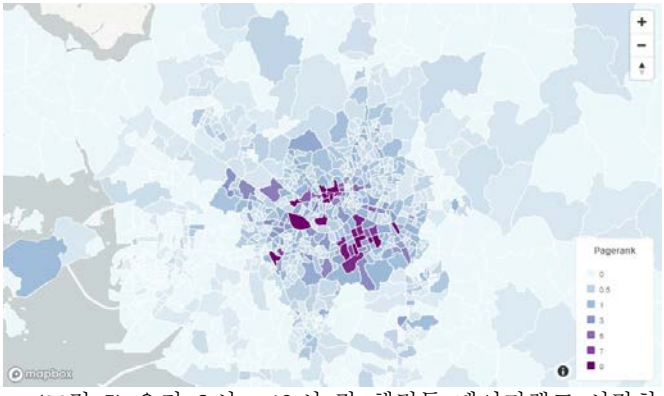
실험 데이터는 2022 년 8 월 12 일 기준으로, 아침 시간대인 오전 6 시부터 10 시 사이와 저녁 시간대인 오후 5 시부터 10 시 사이를 기준으로 두 데이터 셋으로 분할하여 실험을 진행했다. 페이지랭크의 damping factor(d)는 0.9, 반복 횟수는 100 회로 실시하였다. 여기서 d는 대중교통 이용자가 임의로 다른 교통 수단을 이용하지 않을 확률로 가정하였다.

5. 실험 결과

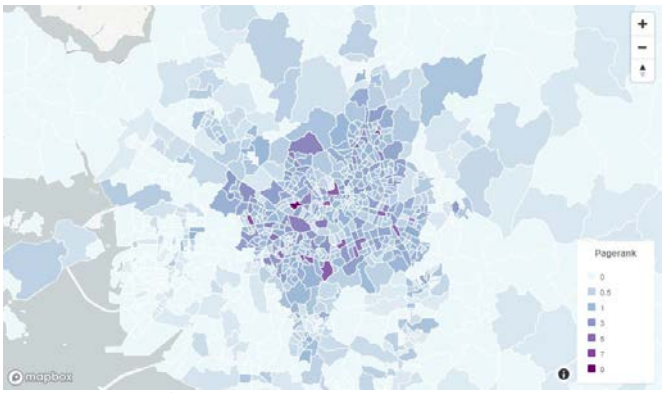
각 행정동의 페이지랭크 값을 효과적으로 확인하기 위해 Mapbox GL JS [6] 를 활용하여 시각화 하였다. Mapbox GL JS 는 Web GL 기술을 이용해, 브라우저에 지도 데이터를 렌더링 해주는 자바스크립트 라이브러리로, 본 연구에서는 이를 Python 에서 사용할 수 있도록 제공하는 API 를 사용하였다. (그림 5)와 (그림 6)은 각각 8 월 12 일 아침과 저녁시간대의 각 행정동 페이지랭크 값을 시각화한 결과이다.

2010 년 수도권교통본부의 조사 [7]에 따르면 서울 시내에서 도착 출근 통행량이 가장 많은 지역은

강남구이며 그 다음은 중구, 서초구, 종로구, 영등포구 순이다. (그림 5)에서 아침 시간대에 페이지랭크가 높은 지역 역시 영등포구 여의동, 종로구 일대, 강남구와 서초구 일대 지역으로 비슷한 양상을 보인다. 이는 아침 시간 출근하는 유동인구가 많은 지역의 페이지랭크가 높다는 것을 의미한다. 반면 저녁 시간대의 (그림 6)에서는 해당 지역들의 페이지랭크 값이 비교적 낮아지고 주변 지역으로 분산됨을 확인할 수 있다. 이는 저녁 시간대 특성상 일터에서 주거 지역으로 돌아가는 인구가 많아 주거 지역 중심으로 페이지랭크 값이 비교적 높게 형성된다고도 유추할 수 있는데, 특히 80 년대 개발 이후부터 전형적인 베드타운으로 발전해 왔던 서울시 동북권 [8]의 경우 아침 시간대보다 저녁 시간대의 페이지랭크가 비교적 크게 높아진 것을 볼 수 있다.



(그림 5) 오전 6 시 ~ 10 시 각 행정동 페이지랭크 시각화



(그림 6) 오후 5 시 ~ 10 시 각 행정동 페이지랭크 시각화

본 실험 결과가 의미하는 바는 연구에서 제시한 페이지랭크 기법이 한 지역의 핵심도시와 그렇지 않은 곳(특히 베드타운)을 구분하는 방법으로서 고려될 수 있다는 것이다. 행정동 간의 대중교통 유동 인구를 반영하여 각 행정동의 중심성을 측정할 수 있었고, 이를 인구의 집결 수준으로 연결시켜 볼 수 있었다.

6. 결론

본 연구에서 교통 데이터를 가중치를 가지는 그래프로 모델링하고, 이를 하둡에서 분산 처리하는 방법론을 제시하고 효과성을 실 데이터를 통해 검증하고자 하였다. 본 분석 및 검증 과정에서는 행정동 간 유동인구를 가중치로 하는 페이지랭크 알고리즘을 맵리듀스를 활용하여 구현 및 실험하였고, 결과로 페이지랭크 값과 각 행정동의 특징에 대한 상관관계를 파악할 수 있었다. 이를 통해 페이지랭크 알고리즘이 유동인구에 따른 지역의 중심성 추론의 도구로서 고려될 수 있다는 점을 알 수 있었으며 이후 다른 영향 요소를 추가 고려한 심층 분석이 가능할 것으로 기대한다.

본 연구에서 제시한 분석 방법론은 도시 계획이나 인구 밀집에 관한 연구, 주거 환경 개선에 기여할 수 있을 것으로 예상된다. 다만 페이지랭크 결과가 파라미터에 의존적일 수 있으며 최적화 문제로부터 성능을 향상 시킬 수 있는 여지가 남아있다. 이에 본 연구팀은 추가 연구를 계획하고 있다

사사문구

"본 연구는 2022 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음" (2022-0-01077)

참고문헌

[1] 서울시 열린 데이터 광장, http://data.seoul.go.kr /dataList/OA-21226/F/1/datasetView.do

[2] Page, Lawrence, et al. "The PageRank citation ranking: Bringing order to the web", Stanford InfoLab, 1999.

[3] Kim, Yong-Yeon, et al. "Analysis on the transportation point in Cheongju City using Pagerank algorithm.", Proceedings of the 2015 International Conference on Big Data Applications and Services, 2015, p. 165-169

[4] Xing, Wenpu, and Ali Ghorbani, "Weighted pagerank algorithm.", Proceedings. Second Annual Conference on Communication Networks and Services Research, Fredericton, 2004, p. 305-314.

[5] Dean, Jeffrey, and Sanjay Ghemawat, "MapReduce: simplified data processing on large clusters." Communications of the ACM, 51.1, p. 107-113, 2008

[6] Kastanakis, Bill, "Mapbox Cookbook", Packt Publishing Ltd, 2016

[7] 국가기록원, "2010 년 (국가교통수요조사 및 DB 구축사업) 전국 여객 기종점통행량조사 2", 건설교통부, 2011

[8] 이영한, "지속가능발전 서울동북권 산업클러스터 구상 연구", 지속가능연구, 제 4, 번호: 2, p. 65-79, 2013.