MapReduce 기반 POI를 추출하기 위한 GPS 데이터 분할 방법

오주성*, 전혜지**, 이혜진**, 정민아*, 이성로**
*목포대학교 컴퓨터공학과
**목포대학교 전자공학과
e-mail:ojoos@mokpo.ac.kr

GPS Data Partitioning Method for POI Extraction Based MapReduce

Joo-Seong Oh*, Hye-Ji Jeon**, Hye-Jin Lee**, Min-A Jeong*, Seong-Ro Lee**
*Dept of Computer Engineering, Mokpo University
**Dept of Electronics Engineering, Mokpo National University

요 약

위치 기반 서비스는 여러 분야에서 활용되어지고 있다. 사용자들에게 정확한 정보를 제공하기 위해서는 대량의 위치 데이터를 분석하여 POI를 추출하고 분석해야 된다. 본 논문에서는 POI를 추출하는 방법으로 DBSCAN 클러스터링을 이용하고 이를 MapReduce 환경에서 구현한다. 또한 알고리즘의 수행속도를 향상시키기위해 데이터를 분할하는 방법을 제안한다.

1. 서론

위치 기반 서비스(LBS:Location-Based Service)는 일상생활의 경험을 공유하거나, 스포츠 활동을 분석하고, 멀티미디어 컨텐츠에 위치 정보를 저장하는 등 많은 분야에서 사람들에게 활용되어지고 있다[1].

위치 기반 서비스의 정확도를 높이기 위해서는 많은 양의 위치 데이터가 필요하다. 적게는 수백 MegaByte에서 많게는 수십 GigaByte 크기의 데이터를 분석하는데 단일 프로세서에서 수행하는 것은 한계가 있다. 따라서 본 논문에서는 대용량의 GPS 데이터를 빠르게 분석하기 위해 DBSCAN 클러스터링 방법을 MapReduc 환경에서 구현하였다. 또한 데이터의 분산처리 속도를 향상시키기 위한 데이터 분할 방법을 제안하였다.

2. 관련연구

분산처리환경에서 위치 데이터를 분석하여 POI(Point of Interest)를 추출하는 선행 연구로 다음과 같은 연구들이 있다.

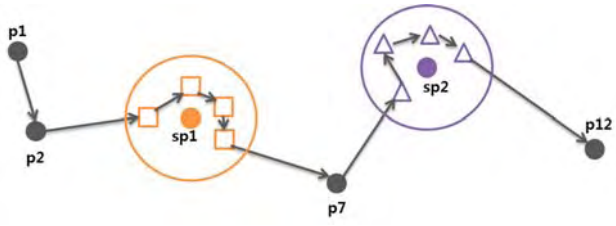
[2]의 연구에서는 Canopy 클러스터링 알고리즘을 이용하여 POI를 추출하였다. GPS 데이터를 입력 받고, 첫 번째 MapReduce 단계에서 Canopy 클러스터링을 수행하여 클러스터의 중심점을 구한다. 두 번째 MapReduce 단계에서는 첫 번째 단계에서 구해진 Canopy를 이용하여 각각의 GPS 포인트들을 클러스터에 할당해 주고 최종적으로 POI를 출력한다.

[3]의 연구는 POI를 추출하고 이동 궤적을 분석하여 다음 방문 장소를 예측해 주는 시스템이다. 이 시스템은 크게 두 개의 단계로 구분된다. 첫 번째 단계는 K-Means Clustering을 이용하여 POI를 추출하고 이동 궤적 데이터를 구축한다. 이동 궤적 데이터는 HMM(Hidden Markov Model)을 학습한다. 두 번째 단계에서는 첫 번째 단계에서 이동 중인 사용자의 이동궤적에 다음 방문 가능한 주요 장소들의 모든 이동 경로를 후보군으로 생성하고 사용자의 학습된 이동 패턴 모델을 적용하여 사용자의 다음 방문 장소를 예측한다.

3. POI 추출을 위한 MapReduce 알고리즘

(그림 1) 시스템 구조

1) SP(Stay Point) 추출 모듈



(그림 2) Stay Point

SP추출 모듈은 각 유저별로 유저가 머문 지점을 추출하는 작업을 수행한다. 이를 위해 거리 임계값과 시간 임계

값을 파라미터로 사용한다.

GPS points { P_m, P_{m+1}, ... , P_n }에서 $\forall m < i \leq n$ 일 때, P_m부터 P_i까지의 거리가 거리 임계값 보다 작고, 이동하는데 걸린 시간이 시간 임계값 보다 클 때 포인트들을 병합하여 SP로 지정한다[4]. 본 논문에서는 병합된 포인트들의 평균값을 SP로 지정한다.

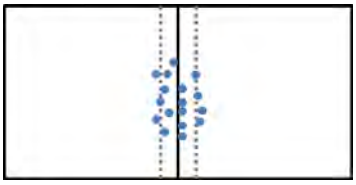
Map 단계에서는 GPS 데이터가 각 노드에 균등하게 할당 되도록 분할한다. 단, SP는 각 유저별로 추출하기 때문에 동일한 유저의 GPS 데이터는 같은 노드에 할당한다.

Reduce 단계에서는 할당 받은 유저들의 데이터에서 SP를 추출한다.

2) Local Clustering 모듈

SP는 각 노드별로 DBSCAN 알고리즘을 이용하여 클러스터링한다. DBSCAN은 Noise에 강건하고 다양한 모양과 크기로 군집이 형성되어 POI를 추출하는데 적합하다[5].

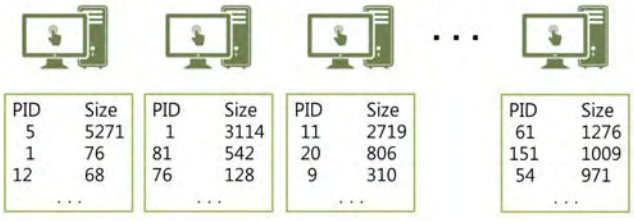
Map 단계에서는 SP 데이터의 최소, 최대 좌표 값을 구해서 넓이가 동일하도록 파티션을 분할하고, SP를 할당한다. MapReduce에서 DBSCAN을 사용하기 위해서는 각 파티션의 경계에 위치한 SP들을 고려해야 한다.



(그림 3) 데이터 분할의 문제점

각 파티션은 GPS좌표를 기준으로 나뉘기 때문에 그림 3과 같이 경계부분에 SP가 집중되어 있는 경우, 클러스터가 형성되지 않고 노이즈로 처리될 수 있다. 따라서 각 파티션은 DBSCAN의 파라미터인 eps 크기만큼의 외부 공간에 위치한 SP를 추가로 할당받는다. 이로 인해 중복된 SP 데이터는 Global Merging 모듈에서 제거한다.

파티셔닝 작업이 완료되면 각 파티션별로 할당된 SP의 개수를 구하고 SP의 개수에 따라 내림차순으로 정렬한다. 데이터는 로드밸런싱을 고려하여 노드별로 비슷한 양을 처리하도록 할당한다.



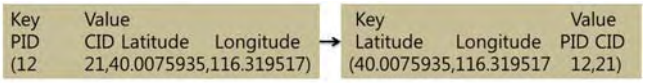
(그림 4) 데이터 분할 방법

Reduce 단계에서는 각 노드별로 DBSCAN 클러스터링을 수행한다. DBSCAN은 거리 임계값 eps과 밀도 임계값 MinPoints를 이용하여 군집과 잡음을 분류한다.

임의의 SP P의 eps 반경 이내에 MinPoints 이상의 SP가 존재하면 클러스터를 형성하고 P를 CorePoint로 지정한다. 형성된 클러스터 내에 있는 다른 SP에서 같은 방법을 사용하여 eps, MinPoints의 조건을 만족시키면 클러스터를 형성하고 병합한다. 이와 같은 방법을 반복하여 클러스터를 추출한다[6].

3) Global Merging 모듈

Global Merging 모듈은 각 노드에서 처리된 클러스터링의 결과 값을 수집하여 병합한다.



(그림 5) 데이터 구조 변경

Map 단계에서는 데이터의 구조를 그림 5와 같이 변경한다. MapReduce는 Key값을 기준으로 데이터가 정렬되기 때문에 중복된 SP의 좌표는 인접하여 출력되게 된다.

Reduce 단계에서는 중복된 SP가 속한 클러스터들을 병합해준다. 파티션의 경계에 위치한 SP들은 둘 이상의 파티션에 할당되게 되고 각각의 파티션에서 클러스터에 포함된 경우 현재의 단계에서 중복하여 출력된다. 이는 중복된 SP를 기준으로 각 클러스터의 연결이 가능하다는 것을 의미한다.

4. 결론

본 논문에서는 POI를 추출하기위해 MapReduce 환경에서 DBSCAN 클러스터링을 이용하였다. 또한 효율적인 데이터 분할방법을 제안하였다.

정확한 POI를 추출하기 위해서는 대용량의 위치 데이터가 필요하다. 하지만 단일 노드에서 데이터를 처리하기 위해서는 많은 시간이 소모된다. 때문에 본 논문에서는 MapReduce에서 DBSCAN을 수행하는데 적합한 데이터 분할 방법을 제안하여 알고리즘의 수행시간을 단축시켰다.

ACKNOWLEDGMENT

본 연구는 2015년도 정부 (교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2009-0093828)와 미래창조과학부 및 정보통신기술진흥센터의 ICT융합고급인력과정지원사업(IITP-2015-H8601-15-1006)의 연구결과로 수행되었음.

참고문헌

[1] Yu Zheng, Xing Xie, Wei-Ying Ma “GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory”, 2010, IEEE Data Eng. Bull.

[2] 정성현, 박영택, "스마트폰 사용자의 관심지점 추출을 위한 MapReduce기반의 Canopy 클러스터링 기법", 2013, 한국컴퓨터종합학술대회 논문집

[3] 김종환, 이석준, 김인철, “다음 장소 예측을 위한 맵리듀스 기반의 이동 패턴 마이닝 시스템 설계”, 2014, 정보처리학회논문지

[4] Matteo Zignani, Sabrina Gaito, "Extracting Human Mobility Patterns From GPS-based Traces", the 3rd IFIP Wireless Days Conference 2010.
[5] 윤애란, "DBSCAN 알고리즘을 이용한 유전자 발현 데이터 마이닝 시스템의 설계 및 구현", 이화여자대학교 과학기술대학원 2003학년도 석사학위 청구논문
[6] Hans-Peter Kriegel, Peer Kröger, Jörg Sander and Arthur Zimek1, "Density-based Clustering", WIREs Data Mining and Knowledge Discovery, Volume1, pp.231-240, 2011.3-5