Spark Streaming 기반의 그리드 색인을 이용하는 이동객체를 위한 연속 질의 기법

Continuos Query Method for Moving Objects using Grid Index based on Spark Streaming

최 도 진, 송 석 일
한국교통대학교

Do-jin Choi, Seokil Song
Korea National University of Transportation

요약

이 논문에서는 Spark Stream의 Discretized Streams 모델을 기반의 그리드 인덱스를 제안하고, 이를 이용한 이동객체를 위한 연속질의 기법을 제안한다. 제안하는 연속질의 처리 방법은 Spark 의 RDD 모델을 이용하여 그리드 색인 및 연속질의 목록을 구현하여, 시스템 고장 시에도 빠르게 복구할 수 있는 내 장애성을 확보 하였다.

Ⅰ. 서론

최근 스마트폰, 태블릿 같은 모바일 기기의 대중화가 가속화 되고 있고, 사용자들에게 제공 되는 응용 서비스들의 범위 또한 증가하고 있다. 그에 맞추어 SNS, 교육, 의료/헬스, 각종 IOT(Internet of Things) 서비스, 클라우드 서비스와 같은 삶의 편의를 한층 높이는 모바일 서비스가 운용되고 있다. 이들 중 LBS (Location-Based Service)는 모바일 기기 사용자 (이동 객체)의 위치 변경을 실시간으로 추적하고 위치에 따라 특정 정보를 제공하는 서비스를 의미한다.

이동객체에게 주위의 건물, 차량, 이벤트와 같은 정보를 제공해주기 위해서는 반복적인 위치 데이터 및 이벤트 데이터에 대한 접근이 필요하다. 이를 위해서는 이동객체의 현재 위치를 저장하고, 이동객체의 현재 위치 주변의 정보 검색을 반복적으로 수행해야 한다. 이동객체 수가 매우 많을 경우에 이와 같은 반복적인 데이터 접근은 운영 서버에 많은 부하를 주게 된다. 이와 같은 반복적인 데이터 접근을 피하기 위해서 이동객체의 관심 영역이 같을 경우 관심영역에 대한 질의를 공유하는 연속질의 처리기법[1]이 제안된 바 있다. 서로 다른 이동객체의 관심 영역이 같을 경우 연산이 이동 객체 수만큼 발생하지 않고, 한 개의 연산을 공유하는 형태로 변경 되어 많은 수행시간을 절약 시킬 수 있다.

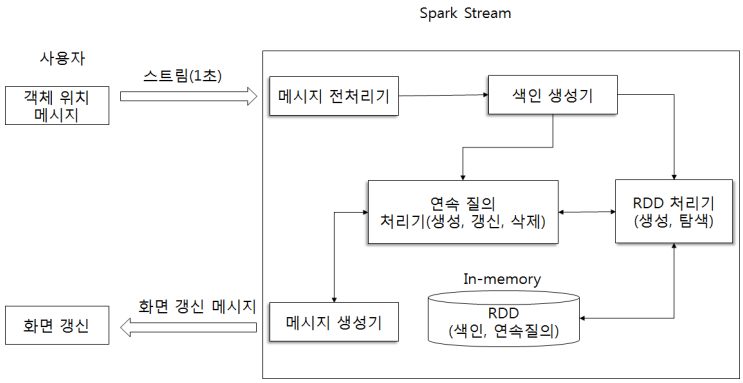
본 논문에서 사용되는 연속 질의 기법은 Spark Stream을 기반으로 한다. Spark Stream은 실시간으로 들어오는 메시지 혹은 로그, 텍스트 파일을 미리 정의 된 시간(초 단위)만큼 입력 받은 후 처리를 하는 인 메모리 스트림 처리 시스템이다. Spark Stream에서는 RDD (Resilient Distributed Datasets)[3]모델의 추상 데이터 집합을 사용하고 있으며, 미리 정의 된 시간만큼 모여진 RDD들을 D-Streams (Discretized streams)[4] 라는 모델을 통해 정의하고 있다. RDD는 RDD가 변형되는 과정을 저장하여 인 메모리의 데이터 손실을 빠르게 복구할수 있는 기능을 제공한다.

본 논문에서는 효율적인 처리 및 복구를 위해 Spark Stream의 RDD를 기반으로 하는 연속질의 처리 기법을 제안한다. 연속질의 처리 속도를 높이기 위해 [2]에서 제안하는 분산 이동객체에 대한 그리드 색인기법을 Spark Stream 에 맞도록 변형하여 이용한다.

Ⅱ. 제안하는 연속 질의 기법

제안하는 연속 질의 기법을 수행하기 위한 시스템 구조도는 그림 1과 같다. 사용자와 서버간의 통신을 위해서 메시지 처리기가 필요하다. 메시지는 실시간으로 사용자에게 받을 수 있고, 혹은 클러스터 서버에 저장되어 있는 파일을 읽어서 처리를 할 수 있다. 색인 생성기는 전 처리 된 메시지를 [2]에서 제안하는 그리드 색인 기법을 통하여 생성한다. 연속질의 처리기는 본 논문에서 제안하는 기법이다. 색인 된 정보와 생성, 수정 된 연속 질의 들은 메모리 내에 RDD 형태로 저장된다.

제안하는 연속 질의 기법은 가장 최신의 D-Stream에 대해서 연속 질의를 생성한다. 기존의 연속 질의가 연속 질의를 참조하는 객체가 1이면 질의의 범위만 수정을 하고 아니라면 새로운 연속 질의를 생성한다. 그림 2, 3에서는 연속질의 처리 절차를 보여준다. 질의 생성 후에는 삭제, 갱신 연산을 수행 한다. 참조 하는 객체가 없는 질



▶▶ 그림 1. 연속 질의 처리를 위한 시스템 구조도

의인 경우에는 삭제하고 그렇지 않은 경우에는 질의에 포함되는 이동 객체를 검색하고, 이를 참조 하는 객체에 갱신 메시지를 보낸다.

```
createCQ(object)
{
 oldCQ = object.CQ
 if(oldCQ.reference == 1)
  oldCQ.MBR = object.MBR
 else
  newCQ = new CQ(object.MBR)
  newCQ.ReferenceObjs += object
}
```

▶▶ 그림 2. 연속 질의 생성

```
deleteNupdateCQ(CQ, objects)
{
 if(CQ.reference == 0)
  delete(CQ)
 else
  containCheck(CQ.MBR, objects)
  refreshMessage(CQ.ReferenceObjs)
}
```

▶▶ 그림 3. 연속 질의 삭제 및 갱신

연속 질의가 저장 하는 정보는 표 1과 같다. MBR (Minimum Bounding Rectangle)은 연속 질의가 실제 질의하는 관심 영역을 의미한다. ContainsObjs는 관심 영역에 포함되는 객체들을 리스트 형태로 저장한다. ReferenceObjs는 질의를 참조하는 객체들을 리스트 형태로 저장한다. 연속 질의의 범위 수정은 MBR을 수정하고, 새로운 질의를 생성할때는 MBR을 기준으로 생성한다. reference는 질의를 참조하는 객체 수를 의미한다. containCheck는 ContainObjs에 포함 될 객체들을 탐색하는 것이다. refreshMessage는 연속 질의를 참조하는 모든 객체에게 갱신 메시지를 보내는 것이다. 모든 객체의 탐색은 [2]에서 제안된 인-메모리 분산 그리드 색인을 기반으로 한다. 마지막으로, 수정, 생성, 삭제 된 연속 질의들은 Spark Stream에서 RDD형태로 메모리에 저장된다. 이는 분산 환경에서 데이터 손실이 생길 경우를 대비하여 효율적인 복구를 수행하기 위하여 사용된다.

표 1. 연속 질의 정보

Name	Description
MBR	질의 영역
Contain Objs	질의에 포함되는 객체 리스트
Reference Objs	질의를 참조하는 객체 리스트

Ⅲ. 결론

제안하는 연속 질의 기법은 Spark Stream 인-메모리 분산 그리드 색인을 기반으로 한다. 제안 하는 방법은 Spark Stream에서 이동 객체의 LBS 컨텐츠를 연속 질의 기법을 통해 제공한다. 질의의 생성, 삭제, 수정은 질의 영역, 질의 포함 객체, 질의 참조 객체를 이용한다. 향후 연구로는 가상의 LBS를 구축하여 성능평가를 수행 할 예정이다.

▌참 고 문 헌▐

[1] 서기언, 이주일, 이원석 “데이터 스트림에서 다중 조인 연속질의의 효과적인 처리를 위한 전처리 기법”, Journal of Korean Socieity Internet Information 2012. Apr: 13(3): 93-105

[2] 이윤수, 송석일 “Spark 기반의 인 메모리 분산 이동객체 색인 기법”, 한국콘텐츠학회 추계종합학술대회, pp 35-36, 2014

[3] Zaharia Matei, et al. “Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing”, NSDI'12, 2012

[4] Zaharia Matei, et al. “Discretized Streams: An Efficient and Fault-Tolerant Model for Stream Processing on Large Clusters”, 4th USENIX conference, 2012