Vol.6, No.1 pp.1~8 pISSN: 2287-5905 https://doi.org/10.3745/KTSDE.2017.6.1.1

An Efficient Angular Space Partitioning Based Skyline Query Processing Using Sampling-Based Pruning

Woosung Choi† · Minseok Kim†† · Gromyko Diana††† · Jaehwa Chung†††† · Soonyong Jung†††††

ABSTRACT

Given a multi-dimensional dataset of tuples, a skyline query returns a subset of tuples which are not 'dominated' by any other tuples. Skyline query is very useful in Big data analysis since it filters out uninteresting items. Much interest was devoted to the MapReduce-based parallel processing of skyline queries in large-scale distributed environment. There are three requirements to improve parallelism in MapReduced-based algorithms: (1) workload should be well balanced (2) avoid redundant computations (3) Optimize network communication cost. In this paper, we introduce MR-SEAP (MapReduce sample Skyline object Equality Angular Partitioning), an efficient angular space partitioning based skyline query processing using sampling-based pruning, which satisfies requirements above. We conduct an extensive experiment to evaluate MR-SEAP.

Keywords : Skyline Computation, MapReduce, Pruning, Data Sampling

데이터 샘플링 기반 프루닝 기법을 도입한 효율적인 각도 기반 공간 분할 병렬 스카이라인 질의 처리 기법

최 우 성† · 김 민 석†† · Gromyko Diana††† · 정 재 화†††† · 정 순 영†††††

요 약

다기준 의사결정 시 활용할 수 있는 스카이라인 질의는 다수의 선택지 중에서 사용자가 '선호하지 않을 만한'(uninteresting) 선택지를 제거함으로써 사용자가 검토해야 하는 선택지의 수를 대폭 감소시키기 때문에 대용량 데이터 분석 시 매우 유용하게 활용될 수 있다. 이러한 배경에서 대용량 데이터에 대한 스카이라인 질의를 분산·병렬 처리하는 기법이 각광을 받고 있으며, 특히 맵리듀스(MapReduce) 기반의 분산·병렬 처리 기법 연구가 활발히 진행되어 왔다. 맵리듀스 기반 알고리즘의 병렬성 제고를 위해서는 부하 불균등 문제·중복 계산 문제·과다한 네트워크 비용 발생 문제를 해소해야 한다. 본 논문에서는 부하 불균등 문제와 중복 계산 문제를 해소하면서도 데이터 샘플링 기반 프루닝을 통해 네트워크 비용 절감 시킬 수 있는 맵리듀스 기반 병렬 스카이라인 질의 처리 기법인 MR-SEAP(MapReduce sample Skyline object Equality Angular Partitioning)을 소개한다. 또한 다양한 관점에서의 실험 평가함으로써 제안 기법의 효용성을 다방면으로 검증했다.

키워드 : 스카이라인 질의, 맵리듀스, 프루닝, 데이터 샘플링

1. 서 론

데이터 분석 기법은 정보량이 폭증하고 있는 오늘날 더욱 중요해지고 있다. 대표적인 데이터 분석 기법인 스카이라인 질의(skyline query)는 사용자가 검토해야 할 데이터를 대폭 줄여주는 질의로, 다기준 의사결정 문제에 활용된다. 다기준 의사결정 문제란 주어진 대안 중 하나를 선택할 때 고려해야 할 기준이 하나 이상인 의사결정 문제를 뜻한다.

다기준 의사결정 문제의 예인 Fig. 1은 호텔을 선택하는 기준이 두 가지인 의사결정 문제를 나타낸다. 고객은 Fig. 1에서 점으로 표현된 10개의 호텔 중, 호텔과 해변의 거리(x축)와 숙박비(y축)를 종합적으로 고려하여 투숙할 호텔을 선택해야 한다. 이때 스카이라인 질의를 활용하면 고객은 4개(t_1, t_2, t_3, t_4)의 호텔만 검토해도 충분히 합리적인 의사결정을 할 수 있다.

※ 이 논문은 2013년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2013R1A1A2010616).
※ 이 논문은 2016년도 한국정보처리학회 춘계학술발표대회에서 '효율적인 각 기반 공간 분할 병렬 스카이라인 질의 처리를 위한 데이터 샘플링 기반 프루닝 기법'의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 고려대학교 컴퓨터학과 석·박사통합과정
†† 비 회 원 : 고려대학교 컴퓨터학과 학사과정
††† 준 회 원 : 고려대학교 컴퓨터학과 박사과정
†††† 종신회원 : 한국방송통신대학교 컴퓨터과학과 조교수
††††† 종신회원 : 고려대학교 컴퓨터학과 교수
Manuscript Received : July 4, 2016
First Revision : September 20, 2016
Accepted : September 29, 2016
* Corresponding Author : Soonyoung Jung(jsy@korea.ac.kr)

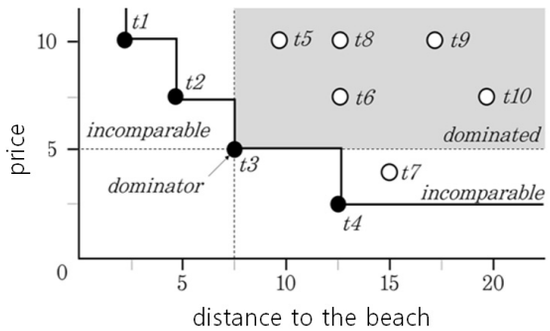


Fig. 1. Skyline of Hotels

6개(t_5, t_6, t_7, t_8, t_9, t_{10})의 호텔을 고려하지 않아도 되는 이유는 해당 호텔들보다 자명하게 선호될 호텔이 존재하기 때문이다. 이를테면 모든 기준에서 t_5보다 우위인 t_3가 존재하기 때문에 t_3는 t_5보다 자명하게 선호된다. 이러한 경우 t_3가 t_5를 지배한다(t_3 dominates t_5)고 표현한다.

호텔선택 예제 외에도 다양한 다기준 의사결정 문제에서 선택지 간의 지배관계가 발생한다. 고려해야할 기준이 d개인 의사결정 문제에서 각 선택지가 $d-$차원 공간의 한 점으로 표현된다면, 두 선택지 간의 지배관계는 Definition 1과 같이 정의된다(단, 수식의 단순화를 위해 각 차원(기준)에서 낮은 값이 더 선호된다고 가정함).

Definition 1. 지배(dominate) 관계
d-차원 공간에 속하는 두 점 $p=(p_1,p_2,...,p_d)$와 $q=(q_1,q_2,...,q_d)$에 대해 아래 (1), (2)가 성립할 경우 p가 q를 지배한다고 정의하며, 이를 $p \prec q$로 표현한다.
(1) $\forall i \in \{1,2,...,d\}: p_i \leqq q_i$
(2) $\exists j \in \{1,2,....,d\}: p_j < q_j$

Fig. 1과 같이 다른 선택지에 의해 지배되는 선택지를 제외시키는 스카이라인 질의는 Definition 2와 같이 정의된다.

Definition 2. 스카이라인 질의 (skyline query)
d-차원 공간에 속하는 점들의 집합 P에 대한 스카이라인 질의는 $\{p_i \in P \mid \nexists p_x \in P: p_x \prec p_i\}$ 를 반환한다.

위와 같이 정의된 스카이라인 질의는 검토해야하는 선택지의 수를 대폭 감소시키기 때문에 데이터가 폭증하는 환경에서 매우 유용하게 활용될 수 있다. 특히 분산·병렬 스카이라인 질의 처리 기법은 대용량 데이터에 대한 스카이라인 질의 처리가 가능하기 때문에 각광을 받고 있으며, 그 중에서도 최근 급부상한 맵리듀스(MapReduce) 프레임워크 기반의 스카이라인 분산·병렬 처리 기법 연구[2-4, 6]가 활발히 연구되고 있다. 맵리듀스 기반 스카이라인 질의 처리 알고리즘의 병렬성 제고를 위해서는 부하 불균등 문제·중복 계산 문제·과다한 네트워크 비용 발생 문제를 해소해야 한다(이는 §2에서 자세히 다룸). 현재까지 연구된 기법들 또한 이러한 문제를 해소시키는 방향으로 발전해 왔다.

최근 부하 불균등 문제를 해소한 각 기반 공간 분할 기반의 기법[4]이 제안되었으나 [4]는 네트워크 비용 관점에서 최적화되어 있지 않다. 본 논문에서는 맵리듀스 프레임워크 상에서 세 가지 문제를 극복할 수 있는 각 기반 공간 분할을 사용하는 기법인 MR-SEAP (MapReduce sample Skyline object Equality Angular Partitioning)[6]을 소개하고, 다양한 관점의 실험을 통해 해당 기법의 효용성을 검증한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 맵리듀스 기반 스카이라인 질의 처리와 관련된 기존 연구를 다룬다. 제 3장에서는 스카이라인 객체 추출을 통한 각 기반 공간 분할 병렬 스카이라인 질의처리 기법[4]에 대해 설명한다. 제 4장에서는 부하 불균등 문제와 중복 계산 문제를 해소하면서도 프루닝을 통해 네트워크 비용 절감 시킬 수 있는 맵리듀스 기반 병렬 스카이라인 질의 처리 기법[6]에 대해 설명한다. 제 5장에서는 [6]의 효용성을 검증하기 위한 다양한 관점에서의 실험 수행 결과에 대해 요약 및 해석한다. 제 6장은 결론 및 제언을 다룬다.

2. 관련 연구

2장에서는 본 연구와 직접적인 관련된 맵리듀스 기반 스카이라인 질의 처리 기법에 대해 소개한다.

현재까지 제안된 맵리듀스 기반 스카이라인 질의 처리 기법은 일반적으로 두 단계에 걸쳐 스카이라인 질의를 계산하며 각 단계는 독립적인 맵리듀스 작업으로 이루어진다. 첫 번째 맵리듀스 작업인 '로컬 스카이라인 계산' 단계에서는 주어진 d-차원 데이터 집합을 특정 기준에 따라 k개로 분할한다(map 과정). 그렇게 만들어진 k개의 하위 데이터 집합 각각에 대해 스카이라인 질의를 수행하며, 질의 결과를 통합하여 아웃풋을 산출한다(reduce 과정). 두 번째 맵리듀스 작업인 '글로벌 스카이라인 계산' 단계에서는 로컬 스카이라인 계산 결과에 포함되어 있는 모든 데이터에 대해 스카이라인 질의를 수행하여 최종 아웃풋을 산출한다.

기존 연구는 로컬 스카이라인 계산 과정에서 사용하는 데이터 분할 방법에 따라 그리드 기반(grid-based) 공간 분할 기반 기법군(2.1)과 각 기반(angle-based) 공간 분할 기반 기법군(2.2)으로 분류할 수 있다. 두 연구군은 데이터를 분할하는 방법에서의 차이가 있으나 부하 불균등 문제·중복 계산 문제·과다한 네트워크 비용 발생 문제를 해소하는 방향으로 발전해왔다는 공통점이 있다.

부하 불균등 문제란 리듀스 과정에서 특정 리듀서에 부하가 집중되어 병렬성이 저해되는 문제이며, 중복 계산 문제란 단순 그리드 기반 공간 분할 등으로 인한 중복된 연산이 발생하는 문제를 뜻한다. 과다한 네트워크 비용 발생 문제란 네트워크 자원 사용 전에 비스카이라인(non-skyline) 객

체를 빠르게 프루닝(pruning)하는 기법 등의 부재로 인해 과도한 네트워크 비용이 발생하게 되는 문제를 뜻한다.

2.1와 2.2에서는 세 가지 문제를 해결하기 위해 제안되어 온 기존연구에 대해 자세히 살펴본다.

2.1 그리드 기반 공간 분할 기반 기법

그리드 기반 공간 분할 방식은 주어진 차원을 수직으로 분할하는 초평면(hyperplane)을 삽입하여 공간을 분할하는 방식을 말한다. 대표적인 그리드 기반 공간 분할 기법인 MR-BNL[2]은 각 차원을 균등하게 수직 이등분하는 초평면을 삽입하여 공간을 분할한다($k=2^d$). 공간 분할에 필요한 연산이 단순하기 때문에 신속한 데이터 매핑(mapping)이 가능하다는 장점이 있다.

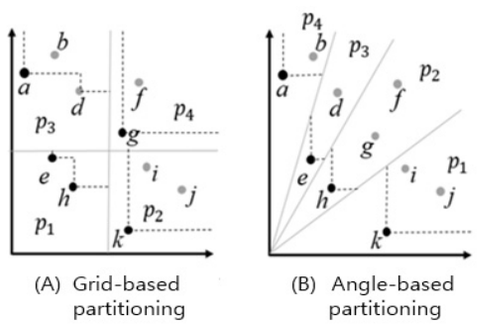


Fig. 2. Comparison of Space Partition Methods

MR-BNL의 로컬 스카이라인 계산 단계에서는 주어진 데이터 집합을 2^d개로 분할한 한 뒤, 2^d개의 하위 데이터 집합 각각에 대해 스카이라인 질의를 수행한다. 이때 Fig. 2(A)에서의 영역 p_1과 p_4영역에서 볼 수 있듯 하위 영역 간의 지배 관계가 발생하는데, 이 경우 중복 계산이 많아진다는 단점이 있다. 예를 들어 p_4에 속하는 대부분의 데이터는 p_1에 속하는 데이터에 지배당할 확률이 매우 높다. p_4에서 수행된 스카이라인 질의 처리 결과는 추후 글로벌 스카이라인 계산 단계의 결과 집합에서 배제될 가능성이 매우 높음에도 불구하고 스카이라인 계산에서 고려되어야 한다.

그리드 기반 공간 분할 기법 사용 시 이러한 중복 계산 문제를 극복하기 위해 [3]에서는 이른 프루닝(early pruning)을 도입했다. [3]에서는 데이터 샘플링을 통해 구축한 Quad-Tree를 기반으로 공간을 세밀하게 분할한 후 하위 영역간의 지배 관계를 파악함으로써 지배당하는 하위 영역에 대한 스카이라인 중복 계산을 사전에 방지한다. 또한 데이터 샘플링을 통해 전체 데이터 분포를 추정하고 이를 기반으로 Quad-Tree를 사용해 하위 영역을 분할하므로 각 리듀서에 전달되는 부하가 비교적 균일해진다는 장점이 있다.

2.2 각 기반 공간 분할 기반 기법

각 기반 공간 분할 기법은 원점에서 각 데이터에 이르는 각도를 기준으로 공간을 분할한다. Fig. 2(B)는 각 기반 공간 분할 기법에 대한 예시 그림이다. 각 기반 공간 분할 기법은 그리드 기반 공간 분할 기법에 비해 분할된 하위 영역 간의 지배 관계가 발생하지 않기 때문에 맵리듀스뿐만 아닌 다른 병렬 컴퓨팅 환경에서의 스카이라인 질의 처리 기법에서 자주 사용되어온 기법이다.

Fig. 2(A)에서의 p_1과 같이, 그리드 기반 분할 방식에서는 원점 및 원점과 가까운 영역이 소수의 특정 하위 영역에 귀속되며 해당 영역은 다른 하위 영역을 지배한다. 이에 따라 글로벌 스카이라인 결과의 대다수는 소수의 특정 영역에서 나타난다. 반면 각 분할 방식에서는 모든 하위 영역이 원점과 가까운 영역을 포함하기 때문에 다른 하위 영역을 지배하는 하위 영역은 존재하지 않는다. 따라서 글로벌 스카이라인의 결과는 모든 하위 영역에서 고르게 나타나며 중복되는 계산이 방지된다.

최근 각 기반 공간분할 기법을 사용하여 부하 불균등 문제와 중복 계산 문제를 해소하는 맵리듀스 기반 스카이라인 질의 처리 기법인 MR-DEAP[4](MapReduce Data Equality Angular Partitioning) 제안되었으나 해당 기법은 네트워크 비용 관점에서 최적화되어 있지 않다. MR-SEAP[6]은 MR-DEAP을 네트워크 비용 관점에서 최적화했다. 두 기법의 차이점은 로컬 스카이라인 단계에서 발생하며, 로컬 스카이라인 단계의 결과를 하나의 리듀서로 전송한 후 스카이라인 질의를 처리하여 최종 아웃풋을 산출하는 글로벌 스카이라인 단계는 두 기법이 같다.

3장에서는 MR-DEAP에 대해 4장에서는 MR-SEAP에 대해 자세히 다룬다. 이때 두 기법은 글로벌 스카이라인 단계에서의 차이점이 없으므로, 로컬 스카이라인 단계를 위주로 두 기법의 차이점을 설명한다.

3. MR-DEAP: 데이터 샘플링을 통한 각 기반 공간 분할 병렬 스카이라인 질의처리 기법

3장에서는 MR-DEAP에 대해 다룬다. MR-DEAP의 로컬 스카이라인 map 단계에서는 데이터를 초구면 좌표계(hyperspherical coordinates)로 변환(§3.1)시키며, 초구면 좌표를 기준으로 데이터 집합을 균등 분할한다. 단, 전수조사를 통한 균등 분할 대신, 데이터 샘플링을 기반으로 근사(approximate) 분할(§3.2)한다. 이후 reduce 과정에서 k개로 분할된 각 하위 집합 대해 로컬 스카이라인을 계산한다.

3.1 초구면 좌표계 변환

MR-DEAP의 로컬-스카이라인 map 단계에서는 데이터 분할을 위해 주어진 직교 좌표(rectangular coordinates)를 초구면 좌표로 변환한 뒤, 반지름 차원을 제외한 나머지 차원 값을 기반으로 공간 분할을 적용한다. d-차원 직교 좌표

계의 한 점 $x=(x_1,x_2,x_3,...,x_d)$를 d-차원 초구면 좌표계의 한 점 $y=(r,\theta_1,...,\theta_{d-1})$로 변환하는 방법은 다음과 같다.

$$r=\sqrt{x_1^2+x_2^2+...+x_d^2}$$
$$\theta_1=\arctan\left(\sqrt{x_2^2+x_3^2+...+x_d^2}/x_1\right)$$
$$...$$
$$\theta_{d-2}=\arctan\left(\sqrt{x_{d-1}^2+x_d^2}/x_{d-2}\right)$$
$$\theta_{d-1}=\arctan\left(\sqrt{x_d^2}/x_{d-1}\right)$$

이때, 반지름 값을 제외시키는 이유는 데이터의 실제 위치가 아닌 오직 각도만 사용하여 분할하고자하기 위함이다.

3.2 샘플링을 이용한 분할

부하 불균등 문제를 방지하기 위해 MR-DEAP의 로컬 스카이라인 map 과정에서는 데이터 집합을 균등하게 분할한다. 초구면 좌표계로 변환된 데이터 집합을 균일한 크기로 분할하기 위해 map 과정에서는 공간 분할 기법인 kd-tree가 사용된다. (반지름 차원 값을 제외한) 초구면 좌표계로 변환된 데이터 집합에 대해 kd-tree를 구축한다면, 데이터를 k개로 균등하게 분할가능하다.

그러나 전체 데이터 집합에 대한 kd-tree 구축에는 데이터 전수조사가 필요하며 이는 대용량 데이터 환경에서 적합하지 않다. 이 경우 전수조사 방법에 대한 차선책으로 데이터 샘플링 기반의 방법이 대안이 될 수 있다. MR-DEAP에서는 대량의 데이터에서 균일하게 데이터 표본을 추출하기 위해 저수지 샘플링(reservoir sampling)[5] 기법을 사용하여 소량의 데이터를 추출한 후, 추출된 데이터만을 이용하여 공간을 근사(approximate)하게 균등 분할한다.

kd-tree 구축을 통한 근사 균등 분할은 top-down 방식으로 이루어지며 구체적인 방법은 다음과 같다. 추출된 샘플 데이터 전체를 초구면 좌표계로 변환하여 첫 번째 각도 차원(즉, θ_1)을 기준으로 샘플 데이터를 정렬한 후, 해당 각도 차원 값의 중앙값(즉, 샘플 데이터 집합을 양분하는 각도 값)을 루트 노드로 설정한다. 루트 노드는 생성될 때 기준이 된 각도 차원과 해당 차원을 양분하는 각도 값을 저장하고 있다. 루트 노드의 왼쪽 자식노드는 부모 노드의 각도 값보다 작은 각도 값을 가지는 데이터 집합, 오른쪽 자식노드는 해당 차원에서 부모 노드의 각도 값보다 큰 각도 값을 가지는 데이터 집합에 대한 정보를 담고 있다.

자식 노드들은 루트 노드와 같은 방식으로 재귀적으로 구축된다. 이를테면 θ_1 차원 값이 루트 노드의 각도 값보다 작은 샘플 데이터의 집합을 대상으로 구축되는 왼쪽 자식 노드의 경우, 데이터는 두 번째 각도 차원인 θ_2를 기준으로 정렬되며 θ_2 차원 값의 중앙값을 왼쪽 자식 노드의 값으로 저장한다. MR-DEAP의 kd-tree는 위와 같이 매 level마다 정렬 기준 축을 round-robin 방식으로 바꾼다. 본 논문에서는 이러한 kd-tree를 DEAP-kd-tree라고 명명한다.

DEAP-kd-tree는 top-down 방식으로 구축되며, 충분한 수의 하위 영역을 얻은 시점(즉, 2×'자식노드의 개수가 1개 이하인 노드의 수'≧k 인 시점)에서 구축을 중단한다.

Fig. 3(A)는 소량의 2차원 데이터를 대상으로 하는 MR-DEAP의 로컬-스카이라인 구동 예시로, (1)단계에서는 저수지 샘플링을 함으로써 일부 데이터(검은색)를 전체 데이터(검은색+하얀색)에서 추려낸다. 이렇게 추출된 샘플 데이터를 초구면 좌표계로 변환시킨 후 각 하위공간에 균등한 개수의 데이터가 들어가게끔 분할하며, 이에 대한 각도 정보를 DEAP-kd-tree 형태로 구축한다.

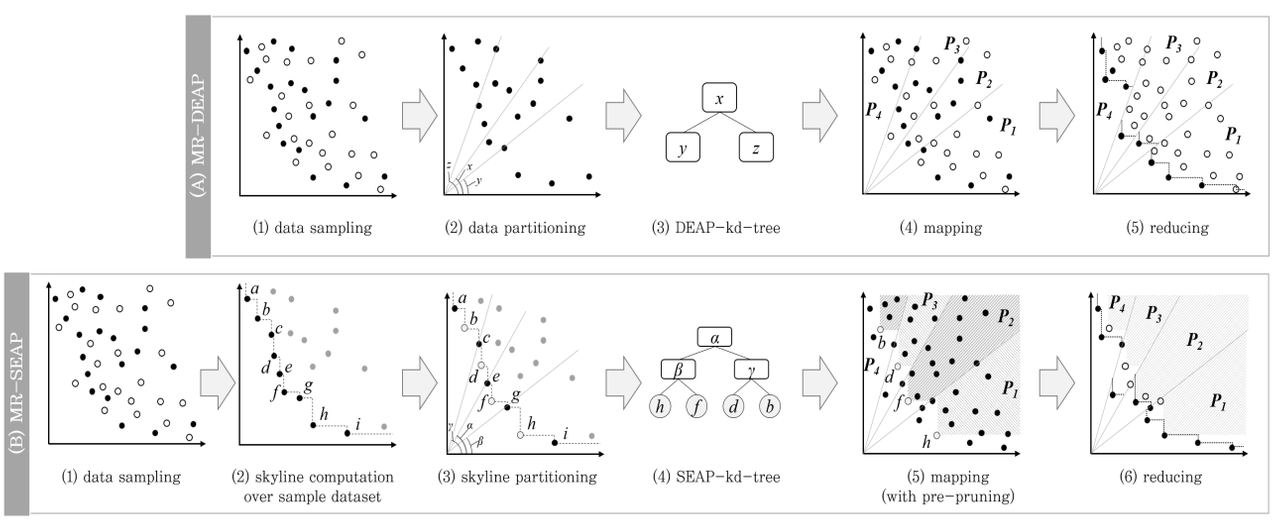


Fig. 3. Local Skyline Computation of MR-DEAP and MR-SEAP

3.3 공간 분할을 통한 로컬 스카이라인 계산

DEAP-kd-tree는 로컬 스카이라인 맵 단계에서 모든 매퍼에 브로드캐스팅된다. 매퍼는 브로드캐스팅된 tree를 이용하여 각 데이터가 속하는 하위 영역을 판정한다. 이후 하위 영역의 식별자를 key값으로, 데이터의 직교좌표를 value로 만들어 (key, value)쌍을 송출(emit)한다.

DEAP-kd-tree는 로컬 스카이라인 리듀스 단계에서는 같은 key값을 가지는 데이터들을 대상으로 스카이라인을 계산하며, 이중 스카이라인 객체만 송출한다.

4. MR-SEAP: 샘플 스카이라인 추출을 통한 각 기반 공간 분할 병렬 스카이라인 질의처리 기법

MR-DEAP은 데이터 샘플링을 기반으로 근사 균등 분할 방법을 도입함으로써 중복 계산 문제를 해소했다. 또한 각 기반 공간 분할을 사용함으로써 중복 계산 문제를 방지한다. 그러나 MR-DEAP은 별도의 데이터 프루닝 방법의 부재로 인해 네트워크 비용 면에서 최적화되어 있지 않다.

이 장에서는 부하 불균등 문제와 중복 계산 문제를 해소하면서도 프루닝을 통해 네트워크 비용을 절감시킬 수 있는 질의 처리 기법인 MR-SEAP을 소개한다. MR-DEAP은 MR-SEAP과 유사하지만 데이터 프루닝을 위한 과정(§4.1)이 추가되었다는 점이 다르다.

4.1 샘플 스카이라인 추출을 통한 이른 데이터 프루닝

MR-DEAP과 마찬가지로 MR-SEAP에서는 저수지 샘플링을 통해 데이터를 임의 추출하며(Fig. 3(B)의 (1)), 이를 기반으로 kd-tree를 구축하여 데이터를 분할한다는 점이 같다. 그러나 MR-DEAP이 샘플 데이터 전체를 근사 균등 분할하는 반면, MR-SEAP에서는 추출된 데이터 샘플에 대한 스카이라인 객체 집합을 균등하게 분할한다는 점이 다르다.

Fig. 3(B)는 소량의 2차원 데이터를 대상으로 하는 MR-SEAP의 로컬-스카이라인 구동 예시이다. 데이터를 샘플링하는 Fig. 3(B)의 (1)단계의 결과는 Fig. 3(A)의 (1)과 같다. Fig. 3(B)의 (2)에서 추출된 샘플 스카이라인 객체는 9개로, 해당 객체 집합은 $\{a,b,c,d,e,f,g,h,i\}$이다. 이후 Fig. 3(B)의 (3)과 같이 추출된 샘플 스카이라인 객체 집합을 대상으로 kd-tree를 구축한다. 본 논문에서는 이렇게 구축된 kd-tree는 DEAP-kd-tree와 구분하기 위해 SEAP-kd-tree라고 명명한다.

SEAP-kd-tree는 DEAP-kd-tree와 구축방법이 유사하나 2×'자식 노드의 개수가 1개 이하인 노드의 수'가 k개일 때까지 구축되는 DEAP-kd-tree와는 달리 말단노드의 개수가 k개일 때 까지 재귀적으로 구축된다는 차이점이 있다. 또한 말단 노드(leaf node) 구조가 다르다. SEAP-kd-tree의 말단 노드에는 데이터를 양분하는 각과 함께 해당 데이터의 직교좌표를 함께 저장한다. 저장된 직교 좌표는 향후 로컬 스카이라인 단계의 map 과정에서 활용된다.

하위 영역에서 특정 각도 차원을 기준으로 데이터를 양분하는 샘플 스카이라인 객체는 Fig. 3(B)의 (5)에서 볼 수 있듯 하위 영역의 많은 영역을 지배한다. 이를테면 객체 h는 13개의 객체가 존재하는 하위 영역에서 8개에 이르는 객체를 프루닝한다. 이렇게 프루닝된 객체는 맵 단계에서 네트워크 전송을 하지 않아도 되기 때문에(즉, 로컬 스카이라인 결과에 포함되지 않을 것임을 알기 때문에) 네트워크 전송 비용이 절감된다.

이후 MR-SEAP에서는 map 과정에서 프루닝되지 않은 데이터에 한하여 데이터를 송출하며, 리듀서는 같은 하위 영역에 속하는 데이터를 수합하여 로컬 스카이라인을 계산한다. 이렇게 계산된 로컬 스카이라인 결과를 재수합하여 글로벌 스카이라인을 계산하는 또 한 번의 글로벌 스카이라인 계산 결과를 거치면 MR-SEAP은 종료된다.

5. 실험 결과

본 장에서는 MR-SEAP의 성능을 검증하기 위한 실험 내용과 그 결과에 대해 설명한다. 실험을 위해 4대의 IBM 서버(각각 E3-1270V2 x 4 3.5GHz 인텔 제온 CPU와 4GB 메인 메모리 탑재)를 사용했으며, 각 서버의 운영체제로는 Ubuntu 14.04를 사용했다. 맵리듀스 프레임워크로는 오픈소스 맵리듀스 프레임워크인 Hadoop-2.6.0을 사용했다. 샘플링 비율은 입력 데이터 수의 0.05%이다.

5.1 데이터 개수에 따른 효율성 비교 실험

본 절에서는 데이터 분포 변화(anti-correlated, uniform) 및 데이터 수 변화(1백만, 2백만, 3백만, 4백만, 5백만)에 따른 MR-SEAP의 성능을 실험한 결과에 대해 다룬다. 평가 척도는 초 단위 응답시간(response)이며 대조 알고리즘으로는 MR-DEAP을 사용했다. 2차원 데이터를 기준으로 실험 평가 했을 때, anti-correlated 분포와 uniform 분포에서의 실험 결과에 대한 요약 그래프는 Fig. 4, 5와 같다. 모든 데이터 규모에서 비교해 보았을 때 MR-SEAP이 MR-DEAP보다 응답시간이 수초 이상 빨랐다.

이러한 성능 차이는 이른 프루닝 도입을 통한 네트워크 비용 절감 기법의 유무에 있다. MR-SEAP의 로컬 스카이라인 계산 과정에서는 샘플 스카이라인을 계산해 이들 중 일부를 브로드캐스팅해 활용한다.

샘플 스카이라인은 다른 객체에 의해 지배되지 않는 객체들의 집합이다. 이 객체들은 비록 소량이지만 비스카이라인 객체에 비해 다른 객체를 지배할 확률이 높다. 이에 따라서 다량의 데이터가 프루닝되며 네트워크 비용을 절감되기 때문에 수행 시간이 감소된다.

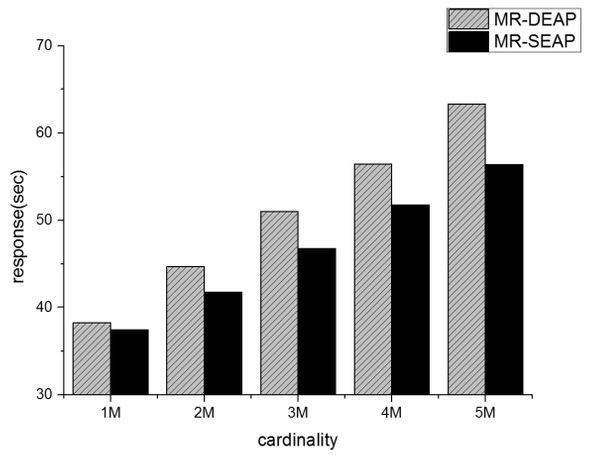


Fig. 4. Response Time Over Varying Cardinality: Anti-Correlated Distribution

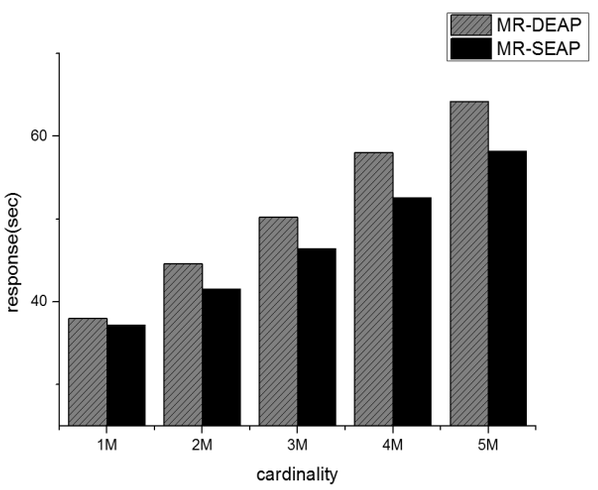


Fig. 5. Response Time Over Varying Cardinality: Uniform Distribution

5.2 차원 증가에 따른 MR-SEAP의 프루닝률

본 절에서는 차원이 증가함에 따라 MR-SEAP의 프루닝률이 어떻게 변화하는지를 알아보기 위해 각 분포에서 데이터 개수를 고정하되 차원을 늘려감에 따라 프루닝되는 데이터의 비율을 조사하는 실험을 수행에 대해 설명한다. 데이터의 경우 3백만(3M)개로 고정시켰으며, 리듀서의 개수는 주어진 차원 d에 대해 2^d개로 설정했다. 각 분포에 대한 실험 결과는 Fig. 6, 7과 같다.

anti-correlated 분포의 경우 차원이 증가함에 따라 프루닝률이 급격하게 감소했다. 6차원에 이르면 프루닝의 효과가 10% 이하로 떨어지는 것으로 확인됐다. 이러한 이유에 대해서는 두 가지로 해석할 수 있다.

첫째로, 본 논문에서 사용한 anti-correlated 데이터 분포는 [7]의 데이터 모델을 이용하여 생성했다. 그러나 [7]의 모델 특성상, [7]의 가장 정도가 덜한(즉, [7]의 c값이 1인) 분포에서조차 초평면 근처에 다수의 점이 분포하게 됨으로 인해서 다른 anti-correlated 분포 모델에 비해 anti-correlated의 정도가 매우 강하다. 만약 실제 데이터 집합(real dataset)을 대상으로 실험을 한다면 Fig. 4와 같이 극단적으로 프루닝 효과가 떨어지는 현상이 일어나지 않을 것으로 판단된다.

둘째로, 프루닝 효과가 미비해지는 현상은 차원의 저주(The curse of dimensionality)로 해석이 가능하다. 차원이 증가함에 따라 MR-SEAP이 MR-DEAP에 비해 비슷한 수준의 성능을 낼 수 있으며, 심지어는 프루닝을 위한 절차에서 소요되는 오버헤드로 인해 MR-DEAP 보다 성능이 떨어질 수 있다. 차원의 저주 효과를 경감시키기 위해 리듀서의 개수, 즉 분할 공간 수를 증가시키는 방법을 시도해보았으며, 리듀서 분할 공간 수 증가로 인해 프루닝률이 올라가는 효과를 확인했다. 그러나 프루닝 효과를 유지하기 위해서는 2^d개 이상 분할 공간 수를 증가시켜야하기 때문에 유의미한 결과라고 볼 수 없다.

반면 uniform 분포의 경우 상대적으로 차원의 저주 효과가 적었다. 6차원에서도 프루닝률이 약 50%였으며, 이는 네트워크 전송 비용이 절반으로 감소한다는 뜻이다.

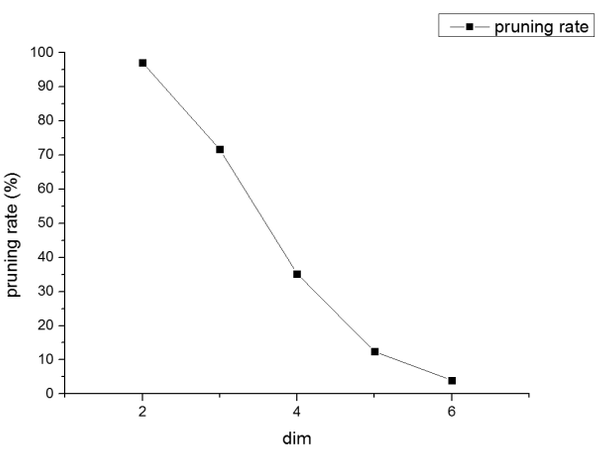


Fig. 6. Pruning Rate Over Varying Dimensionality: Anti-Correlated

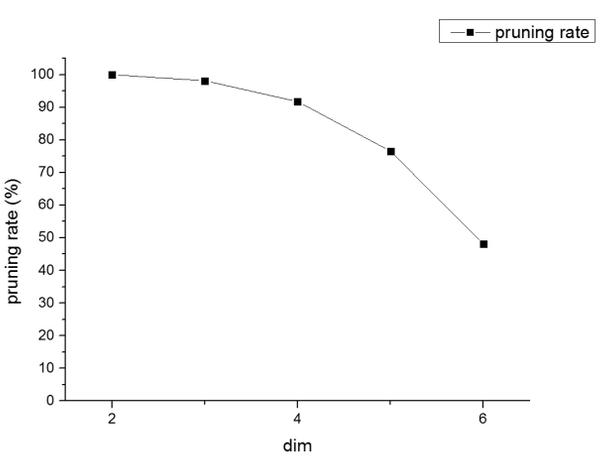


Fig. 7. Pruning Rate Over Varying Dimensionality: Uniform

6. 결론 및 제언

본 논문에서는 부하 불균등 문제, 중복 계산 문제, 네트워크 비용 증가 문제를 해결한 각 기반 공간 분할을 사용하는 맵리듀스 기반 스카이라인 질의 처리 기법인 MR-SEAP에 대해 소개하고 다양한 관점의 실험을 통해 해당 기법의 효용성을 검증했다.

본 논문에서는 저차원에서는 데이터 규모가 늘어남에 따라 MR-SEAP이 스카이라인을 계산하는 데에 걸린 응답 시간이 MR-DEAP보다 수초 이상 빠르다는 것을 밝혔다. 저차원일수록, uniform 분포일수록 MR-SEAP은 MR-DEAP보다 성능이 우수했다. 또한 본 논문에서는 고차원일수록, anti-correlated되는 경향이 심해질수록 MR-SEAP의 프루닝률이 떨어짐에 따라 그 성능은 MR-DEAP과 비슷해진다는 것을 밝혔다.

실험을 통해 얻은 결과를 종합하자면, 5차원 이하이며 anti-correlated 분포가 아닌 대용량 데이터를 검토해야하는 다기준 의사결정 문제에는 MR-SEAP을 활용할 경우 신속하게 스카이라인 객체를 구함으로써 합리적인 의사 결정을 지원할 수 있다. 반면 6차원 이상이거나 anti-correlated 분포인 데이터 집합의 경우 프루닝률이 떨어지기 때문에 MR-DEAP 보다 높은 성능을 보이기 위해서는 추가적인 연구가 필요할 것으로 보인다.

References

[1] Borzsony, Stephan, Donald Kossmann, and Konrad Stocker, “The skyline operator,” *Data Engineering, 2001, Proceedings, 17th International Conference on. IEEE*, 2001.

[2] Zhang, Boliang, Shuigeng Zhou, and Jihong Guan, “Adapting skyline computation to the mapreduce framework: Algorithms and experiments,” *International Conference on Database Systems for Advanced Applications*, Springer Berlin Heidelberg, 2011.

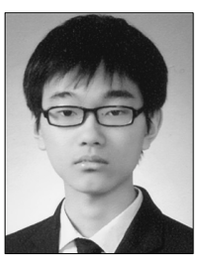
[3] Park, Yoonjae, Jun-Ki Min, and Kyuseok Shim, “Parallel computation of skyline and reverse skyline queries using mapreduce,” *Proceedings of the VLDB Endowment*, Vol.6, No.14, pp.2002-2013, 2013.

[4] Jaehwa Chung, “Data Samping-based Angular Space Partitioning for Parallel Skyline Query Processing,” *The Korean Association Computer Education*, Vol.18, No.5, pp.63-70, 2015.

[5] J. S. Vitter, “Random sampling with a reservoir,” *ACM Transactions on Mathematical Software (TOMS)*, Vol.11, No.1, pp.37-57, 1985.

[6] Woo-Sung Choi, Jong-Hyeon Min, Jaehwa Chung, and Soon-Young Jung, “A Sampling based Pruning Approach for Efficient Angular Space Partitioning based Skyline Query Processing,” *2016 KIPS Spring Conference*, Vol.23, No.1, pp.55-58, 2016.

[7] Shang, Haichuan and Masaru Kitsuregawa, “Skyline operator on anti-correlated distributions,” *Proceedings of the VLDB Endowment*, Vol.6, No.9, pp.649-660, 2013.



최 우 성

e-mail : ws_choi@korea.ac.kr
2013년 고려대학교 컴퓨터교육학과 (이학사)
2013년～현 재 고려대학교 컴퓨터학과 석・박사통합과정
관심분야 : 데이터베이스, 빅데이터, 인공지능, 시공간 데이터

김 민 석

e-mail : rlaalstjr46@korea.ac.kr
2014년～현 재 고려대학교 컴퓨터학과 학사과정
관심분야 : 데이터베이스, 인공지능

Gromyko Diana

e-mail : dialen@korea.ac.kr
2001년 우수리스크 국립사범대학교 컴퓨터/수학과(학사)
2011년 고려대학교 컴퓨터교육학과(석사)
2011년～현 재 고려대학교 컴퓨터학과 박사과정
관심분야 : Trajectory Pattern Mining, Spatio-Temporal Query Processing

정 재 화

e-mail : jaehwachung@knou.ac.kr
1999년 고려대학교 컴퓨터교육과(이학사)
2011년 고려대학교 컴퓨터교육과
(이학석・박사)
2012년~현 재 한국방송통신대학교
컴퓨터과학과 조교수
관심분야 : 공간질의처리 및 인덱싱, 분산 컴퓨팅 플랫폼
(Mapreduce, Spark), 모바일 데이터 관리, RFID,
무선 센서 네트워크

정 순 영

e-mail : jsy@korea.ac.kr
1990년 고려대학교 전산학(이학사)
1992년 고려대학교 전산학(이학석사)
1997년 고려대학교 전산학(이학박사)
2008년~현 재 고려대학교 컴퓨터학과
교수
관심분야 : 데이터베이스, 빅데이터, 시공간 데이터, 분산 시스템