데이터 샘플링을 통한 각 기반 공간 분할 병렬 스카이라인 질의처리 기법

정재화†

요 약

상호 연관되는 복잡한 데이터 조건이 존재하는 환경에서 스카이라인 질의는 의사결정 시스템 등 폭넓은 애플리케이션 활용 가능성으로 다양한 분야에서 연구되어 왔다. 중앙집중식 환경에서 스카이라인 질의처리 기법이 초기에 제안되었으며 최근 대량의 다차원 데이터에 대해 데이터 공간을 분할하여 맵/리듀스 플랫폼 상에서 병렬적으로 처리하는 기법이 제안되었다. 그러나 현재까지의 기법이 비균등적 실행과 높은 중복 작업으로 효율성이 저하된다는 문제점을 배경으로 본 논문에서는 랜덤 샘플링을 통해 데이터 분포를 추정하여 비균등 분할 문제를 해결하고 각 기반의 데이터 공간을 분할하여 스카이라인 처리 과정에서 중복 작업을 최소화한 새로운 기법 MR-DEAP를 제안한다. 마지막으로 다양한 환경에서의 실험결과 제안된 기법이 다른 각 기반 분할과 그리드 분할 기법보다 우수한 것을 입증하였다.

주제어 : 각 기반 공간 분할, 스카이라인, 맵/리듀스, k-d tree, 랜덤 샘플링

Data Sampling-based Angular Space Partitioning for Parallel Skyline Query Processing

Jaehwa Chung†

ABSTRACT

In the environment that the complex conditions need to be satisfied, skyline query have been applied to various field. To processing a skyline query in centralized scheme, several techniques have been suggested and recently map/reduce platform based approaches has been proposed which divides data space into multiple partitions for the vast volume of multidimensional data. However, the performances of these approaches are fluctuated due to the uneven data loading between servers and redundant tasks. Motivated by these issues, this paper suggests a novel technique called MR-DEAP which solves the uneven data loading using the random sampling. The experimental result gains the proposed MR-DEAP outperforms MR-Angular and MR-BNL scheme.

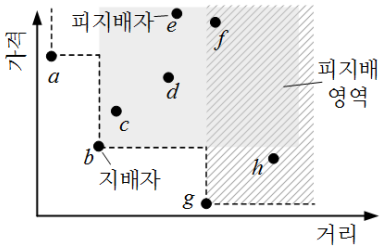
Keywords : Angular-based Space Partitioning, Skyline, Map/Reduce, k-d Tree, Random Sampling

† 종신회원: 한국방송통신대학교 컴퓨터과학과 조교수
논문접수: 2015년 8월 7일, 심사완료: 2015년 8월 10일, 게재확정: 2015년 8월 11일
* 본 논문은 2014학년도 후기 한국방송통신학교 학술연구비 지원으로 수행되었음

1. 서론

상호 연관되는 복잡한 데이터 조건이 존재하는 환경의 데이터베이스에서 스카이라인 질의는 의사결정 시스템[1] 등의 폭넓은 응용 가능성으로 많은 관심을 받고 있다. 데이터베이스 학계에서 또한 다양한 환경(중앙집중, 분산, 병렬처리 등)에서 스카이라인 질의를 효율적으로 처리하기 위한 다양한 연구가 진행되어 왔다. 스카이라인 질의는 주어진 다차원 데이터 집합에 대해 다른 데이터에 의해 지배(dominate)되지 않는 데이터들의 집합을 반환한다. 형식적으로 d차원 공간에서 주어진 데이터 $p_1, p_2, ..., p_n \in R$에 대하여 스카이라인 질의의 결과 $Skyline(R)$는 $\forall t_i \in Skyline(R)$과 $p_j \in R - Skyline(R)$에 대해 $\exists a \in \{1, 2, ..., d\}$, $t_i[a] \geq p_j[a]$이다.

예를 들어, <그림 1>의 x축은 호텔에서 해변까지의 거리이고, y축은 호텔의 숙박비라고 할 때, 원점에 가까운 호텔이 가장 선호된다. 이때 호텔 c는 b에 비해 가격도 높고 해변과 호텔의 거리도 멀기 때문에 c는 b보다 선호도가 떨어진다. 이를 'c는 b에 의해 지배된다' 라고 한다. 반면 호텔 a는 b에 비해 가격이 높지만 해변에서 b보다 가깝기 때문에 지배되지 않는다. 그림에서 회색 영역과 사선 영역에 존재하는 모든 호텔들은 각각 b와 g보다 가격이 높고 해변과 호텔의 거리가 멀기 때문에 b와 g에 의하여 지배당하는 영역을 나타낸다. 따라서 결과적으로 지배되지 않은 객체 집합, 즉 스카이라인은 ={a, b, g}가 된다.



<그림 1> 스카이라인 질의 예

스카이라인은 여러 조건이 독립적이지 않은 상황에서 데이터 간 우열 관계를 파악할 수 있는 기능으로 응용 가능성이 높아 많은 연구가 진행되어 왔다. 스카이라인 질의를 처리하기 위한 접근방법으로 다차원 인덱스를 생성, 분할 후 정복, 최근접 객체 검색, 분기한정 등의 기법이 제안되었지만 차원의 수나 데이터의 개수가 증가할 경우 상호작용이 필요한 시스템에서 적합하지 않을 정도로 성능이 저하되는 문제점이 있을 뿐만 아니라, 질의 처리의 성능이 서버의 성능에 종속되는 중앙집중식은 빅데이터 환경과 같은 대량의 데이터에 대한 질의처리에 구조적인 한계가 있다.

최근 이러한 문제점을 해결하기 위해 다수의 시스템으로 구성된 서버 클러스터에 맵/리듀스 플랫폼을 도입하여 병렬적으로 스카이라인 질의처리를 할 수 있는 기법이 제안되었다. 맵/리듀스 플랫폼은 데이터를 배분하는 매퍼(mapper)와 배분된 데이터를 처리하는 리듀서(reducer)로 구성되는데 맵/리듀스 환경에서 스카이라인 질의 처리는 d차원의 공간 데이터를 여러 개의 하위공간으로 분할하고 매퍼가 각각의 하위공간에 해당하는 데이터를 리듀서에 할당하여 하위공간별 개별적으로 스카이라인 질의를 처리(로컬 단계)하게 한 후 임시 스카이라인 결과를 최종적으로 병합(글로벌 단계)하여 질의 결과를 생성하는 방식으로 진행된다.

이러한 데이터 공간 분할 방식은 분할 방법에 따라 처리 알고리즘과 성능이 크게 좌우되며, 최근까지 그리드 분할 방식(MR-BNL)과 각 기반 분할 방식(MR-Angular)으로 구분된다.

MR-BNL 알고리즘[13]은 각 차원을 균등하게 수직 분할하는 초평면(hyperplane)을 삽입하여 공간을 분할하고 각 하위공간에서 스카이라인을 찾는다. MR-BNL의 분할 방식은 매핑 단계가 신속하게 처리되는 장점이 있는 반면 몇몇 하위공간에 대해서는 불필요한 질의처리가 수행되고 글로벌 단계에서 중복적인 질의처리 발생하기 때문에 비효율적이다.

MR-Angular 알고리즘[14]은 질의 결과가 원점과 차원 축 근처에서 나타나는 스카이라인 질의 특성을 이용하여 원점을 기준으로 균등한 각도로 공간을 분할하는 각 기반 분할 방식(Angle-Based Space Partitioning)[3]을 사용한 맵/리듀스 기반 스카이라인 질의 처리 방법을 고안했다. 각 기반 분할 방식은 불필요하게 질의처리가 수행되는 하

위공간이 적고 글로벌 단계가 단순하여 그리드 분할 방식에 비해 질의처리 과정이 효율적이다. 그러나 MR-Angular는 데이터 분포가 비균일(Skewed)할 경우 매핑 과정에서 데이터 균등 분배(로드 밸런싱)이 적절하게 이루어지지 않기 때문에 전체적인 질의처리 과정이 비효율적이다.

이러한 두 분할 방식의 문제점을 배경으로 본 논문에서는 데이터 샘플링을 통해 비균등한 공간 분할 문제를 해결하고 각 기반의 데이터 공간 분할을 이용하여 스카이라인 결과를 생성하는 MR-DEAP(Map/Reduce Data Equality Angular Partitioning) 기법을 제안한다. MR-DEAP 기법은 랜덤 샘플링으로 추출된 샘플 데이터에 대해 k-d tree로 데이터 분포를 추정하고 분할 공간(파티션) 간 데이터양이 균일하도록 각의 범위를 조절한다. 분할 공간에 대해 맵/리듀스 프레임워크에서 지배 관계를 판단하여 스카이라인 질의 결과를 최대한 균등하게 처리한다.

이후의 논문 구성은 다음과 같다. 2장은 중앙집중, 분산 및 병렬 방식의 스카이라인 질의 처리 기법을 제시한다. 3장에서는 각 기반 공간 분할하고 이를 k-d tree에서 관리하는 기법을 소개하고 4장에서는 맵/리듀스 환경에서 병렬적으로 스카이라인 질의를 처리하는 알고리즘을 제시한다. 5장에서는 제안된 기법의 우수성을 입증하고 마지막으로 6장에서 정리한다.

2. 관련연구

이번 장에서는 폭넓은 활용 가능성으로 많은 분야에서 활용되는 스카이라인 질의연구 동향과 제안된 기법의 특징에 대하여 살펴본다.

2.1 중앙 집중 처리 방식

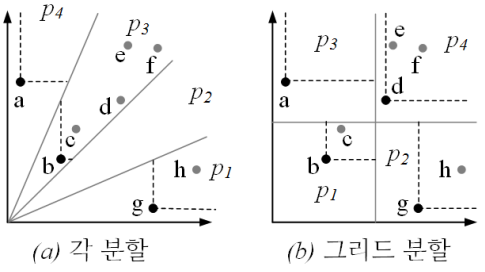
중첩 반복 스카이라인 질의는 객체가 스카이라인의 결과에 속하는지 판별하기 위해 나머지 객체를 비교하는 무작위 대입(brute force) 접근방식이다. 중첩 반복문의 복잡도를 개선한 블록 중첩 반복문 스카이라인 질의(BNL)[2]는 객체를 블록 단위로 읽어 들인 후 해당 지점까지의 스카이라인 후보 객체 집합을 계산하여 메모리에 저장한다. i번째 블록 B_i를 읽었을 때 스카이라인 후보 객체가 l개라면, 그 블록내의 각각의 객체와 l번만 비교하여 스카이라인 여부를 판단한다.

BNL을 개선하여 모든 객체를 특정 값으로 정렬하는 기법을 도입하여 후보 간의 비교 횟수를 감소시킨 SFS[5]가 제안되었으며, SFS의 외부 정렬 비용을 줄이기 위해 후보 객체만 정렬하여 성능을 향상시킨 LESS[6]가 제안되었다. 또한 anti-correlated 분포에 특화된 BNL 기반 스카이라인 질의 처리 방법[7]이 최근에 제안되었다.

스카이라인 질의 처리를 위한 다른 접근 방법으로 인덱스를 사용하여 데이터 탐색 비용을 줄이는 기법이 있다. 스카이라인 질의는 다차원 공간으로 확장될 수 있기 때문에 공간 인덱스인 R-tree를 사용하는 연구가 제안되었다. [3]은 R-tree를 사용하여 원점에서 가장 가까운 객체(NN) 검색을 재귀적으로 실행하여 스카이라인 질의를 처리하는 기법을 제안했다. [4]는 원점과의 맨해튼 거리로 R-tree를 분기한정 탐색하는 방식의 BBS를 제안하였다. 또 동적 스카이라인 질의 처리를 위한 분기한정 동적 처리 기법 등 다양한 기법이 제안되었다.

2.2 맵/리듀스 기반 분산 처리 방식

센서 및 모바일 기기에서 대량으로 수집되는 빅데이터를 처리할 수 있는 애플리케이션에 대한 요구가 늘어남에 따라 스카이라인과 같은 데이터 집중적 질의를 중앙 집중적으로 처리하는 것에 대한 한계가 지적되었다. 따라서 서버 클러스터에서 병렬처리가 가능한 프로그래밍 모델인 맵/리듀스 프레임워크를 스카이라인 질의 처리에 응용한 MR-BNL과 MR-Angular가 제안되었다.



<그림 2> 데이터 공간 분할 방식

MR-BNL 방식은 <그림 2>의 (b)와 같이 각 차원을 균등하게 분할하는 초평면(hyperplane)을 삽입하여 공간을 분할하는 방식을 말한다. 예를 들어 이 그림에서 각각의 차원 축 x, y에 수직인 초평면이 삽입되었으며, 각 분할 공간 p_i에 대해 $p_1 = \{b, c\}$, $p_2 = \{g, h\}$, $p_3 = \{a\}$, $p_3 = \{d, f, e\}$ 이 된다. 두 단계에 걸쳐 맵/리듀스 과정이 진행되는 MR-BNL 방식은 로컬 단계에서 맵핑 과정에서 데이터가 분할되고 리듀싱 과정에서 각 파티션별 로컬 스카이라인을 판단한다. 따라서 로컬 스카이라인은 각각 $\{b\}$, $\{g\}$, $\{a\}$, $\{d\}$가 된다. 맵/리듀스 과정이 재실행되는 글로벌 단계에서는 매핑 과정에서 이전 단계에서 생산된 로컬 스카이라인이 하나의 파일로 통합되고 리듀싱 과정에서 최종 질의처리 결과인 $\{a, b, g\}$가 생성된다.

MR-Angular 알고리즘은 원점을 기준으로 균등한 각도로 공간을 분할한다. <그림 2>의 (a)에서와 같이 초구면 좌표계로 표현된 데이터 공간은 균등한 각으로 분할된다. 로컬 단계에서는 데이터 공간과 데이터가 극좌표계로 변한되고 매핑 과정에서 분할 된 공간별로 데이터가 분할되어 매퍼에게 할당된다. 리듀싱 과정에서 분할 공간별로 로컬 스카이라인 $\{g\}$, $\{b\}$, $\{a\}$가 생성된다. 그리고 글로벌 단계에서는 매핑 과정에서 하나의 데이터로 통합되어 리듀싱에서 최종 결과 $\{a, b, g\}$가 만들어진다.

그러나 두 방식 모두 서로 다른 파티션에 속한 데이터가 질의처리에 영향을 미치지 않도록 공간 분할에 공헌이 있지만 서버 간 로드 밸런싱에 대해서는 고려가 되지 않았다. 맵/리듀스 프레임워크에서 매퍼와 리듀서는 서로 간 통신 없이 완전하게 독자적으로 동작하기 때문에 분할된 시스템 사이에 균형된 로드 밸런싱 이루어지지 않을 경우 시스템의 성능이 저하된다. 즉, 특정 서버에 네트워크 I/O, CPU 시간 및 저장 공간 등의 과부하가 걸리지 않도록 균형된 작업 배분이 요구된다.

따라서 본 논문에서는 샘플링으로 데이터 분포를 추정하고 사용하여 양적으로 데이터를 균등하게 배분할 수 있도록 공간을 분할하여 로드 밸런싱이 가능한 MR-DEAP 기법을 제안하고자 한다.

3. 데이터 균등 각 기반 공간 분할

이번 장에서는 기존의 각 기반 공간 분할 방식의 비균등 공간 분할 문제를 해결하기 위한 MR-DEAP 기법에서 사용되는 데이터 표현과 샘플링 기법에 대하여 소개한다.

3.1 초구면 좌표 데이터 표현

각도 기반으로 공간을 분할하기 위해서는 위치 좌표값이 데카르트 좌표에서 표현하는 것이 아닌 원점에서의 거리와 각도로 표현되는 초구면 좌표계(hyperspherical coordinate system)로 표현되어야 한다. 예를 들어, 2차원 데카르트 좌표계에서의 한 점 $x = [x_1, x_2]$는 <그림 3>의 (a)와 같이 초구면 좌표의 $x = (r, \theta)$로 표현된다. 이때, $r = \sqrt{x_1^2 + x_2^2}$이고 $\theta = \arctan(x_2/x_1)$이다. 이를 일반화하면, d차원 공간에서 한 점 x는 거리 r과 $d-1$개의 각 $\theta_1, \theta_2, ..., \theta_{d-1}$으로 표현된다. r과 θ는 다음의 식에 의해 계산된다.

$$r = \sqrt{x_1^2 + x_2^2 + ... + x_d^2}$$

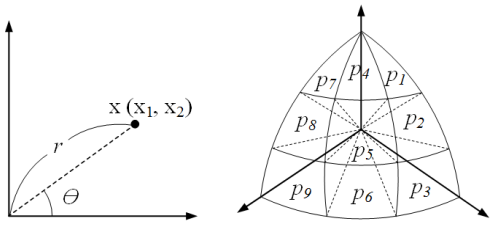
$$\theta_1 = \arctan\left(\sqrt{x_2 + x_3 + ... + x_d}/x_1\right)$$

$$...$$

$$\theta_{d-2} = \arctan\left(\sqrt{x_{d-1} + x_d}/x_{d-2}\right)$$

$$\theta_{d-1} = \arctan(x_d/x_{d-1})$$

<그림 3>의 (b)는 3차원 공간에서 데이터 표현 및 공간 분할을 나타낸다.

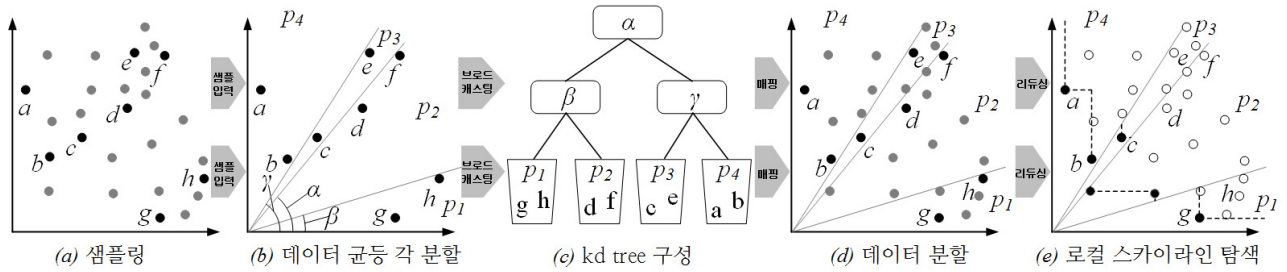


(a) 데이터 표현 (b) 3차원 공간의 분할

<그림 3> 초구면좌표 공간

3.2 샘플링을 이용한 분할

맵/리듀스 플랫폼에서의 처리 성능은 로드 밸런싱에 많은 영향을 받는다. 따라서 초구면 좌표계



<그림 4> 스카이라인 질의처리(로컬 단계)

에서 표현된 데이터에 대해 균일한 개수로 공간이 분할되도록 각의 범위를 결정하기 위해서는 전체적인 데이터 분포가 필요하다. 선형 탐색이 필요한 데이터 분포 파악은 데이터의 양이 거대한 빅데이터 환경에서 불필요하기 때문에 <그림 4>의 (a)와 같이 데이터의 일부분을 표본으로 추출하여 데이터 분포를 유추하고 데이터 균등 분할 범위를 결정하는 데 사용된다. 본 논문에서는 대량의 데이터에서 랜덤하게 데이터 표본을 추출하는 저수지 샘플링(reservoir sampling)[12] 알고리즘 사용한다.

초구면 좌표로 표현된 샘플링 데이터의 θ만을 고려하여 1차원 k-d tree를 생성한다. k-d tree는 <그림 4>의 (b)와 같이 샘플링 데이터가 절반이 되는 순간 초평면을 삽입하여 공간을 분할하고 하위공간에 포함된 데이터의 개수가 다시 절반이 되는 순간 초평면을 삽입하는 재귀적 방식으로 동작하여 리듀서의 개수보다 분할 공간의 개수가 같거나 커질 때까지 계속 반복된다. 이 작업으로 <그림 4>의 (c)와 같이 데이터 파일의 크기가 균등한 공간으로 분할 할 수 있다.

4. 병렬 스카이라인 질의 처리 기법

이번 장에서는 본 논문에서 제안하는 맵/리듀스 프레임워크 상에서 MR-DEAP의 실행 과정을 제시한다.

MR-DEAP은 알고리즘 1에서 데이터 집합 D, 추출할 표본의 크기 k, 분할 공간 수 n을 입력받아 랜덤 샘플 추출 작업을 완료한 이후 두 번의 맵/리듀스 작업을 거쳐 스카이라인 객체를 반환한다. 먼저 데이터 분포를 파악하고 데이터가 양적으로 균등하게 분할 될 수 있도록 저수지 샘플링을 이용하여 k개만큼의 표본 집합 샘플을 추출(라인 1)한다. 샘플에 속하는 각 데이터의 데카르트 좌표계를 초구면 좌표계로 사상(라인 2)한다. 단, 초구면 좌표계에서 거리는 각과 무관하므로 제외된다. 즉, 2차원의 경우 (x_1, x_2)는 (r,θ)로 변형되지만 r을 제외하고 θ만 사용한다.

초구면 좌표계로 사상된 랜덤 샘플 angle_sample을 기준으로 n개로 분할된 하위공간을 생성하기 위한 sky-kdtree(k-d tree)를 구축(라인 3)하고 이를 모든 작업 노드(매퍼)에 브로드캐스트(라인 4)한다(<그림 4>의 (c) 참조).

알고리즘 1: MR-DEAP

Input: a data set R, sample size k,
number of partition n
Output: the skyline of S
1. *sample* = ReservoirSampling(R, k)//표본추출
2. *angle_sample* = toPolar_set(*sample*)
3. sky-kdtree = KDTREE(*angle_sample*, n)
4. Broadcast sky-kdtree
5. LocalSkyline = Runmap/reduce(Local-MR)
6. Return Runmap/reduce(Global-MR)

랜덤 샘플과 k-d tree를 기반으로 하위 공간 생성을 마친 후 MR-DEAP은 총 두 번의 맵/리듀스 단계를 통해 최종 스카이라인 결과를 반환한다.

알고리즘 2는 첫 번째 로컬 단계의 맵/리듀스 작업을 설명한다. 우선 각 하위 공간 별 스카이라인 객체를 생성한다. 매핑에서 각 객체의 좌표를 초구면 좌표계로 변환한 후, sky-kdtree를 이용해 해당하는 하위 공간을 담당하는 리듀서에게 해당

객체를 전송한다(라인 3). 그리고 리듀싱에서 중앙 집중 환경에서의 스카이라인 질의처리 기법인 BNL을 사용하여 각 하위 공간에서의 스카이라인 객체(로컬 스카이라인)를 생성한다.

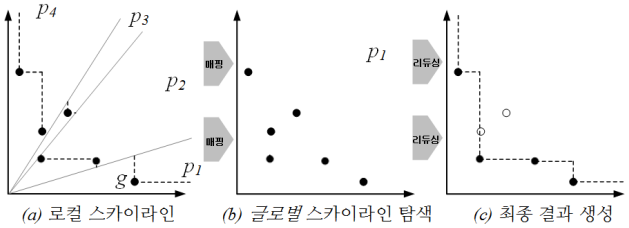
알고리즘 2: Local-MR

```
//Map Procedure
Input: a <K, V> pair //offset K, Object V
1. o_angle = toPolar(V) //초구면 좌표 변환
2. sky-kdtree = LoadTree();
3. output(sky-kdtree.get_partition(o_angle), o)
//Reduce Procedure
Input: a <K, L> pair
//// partition offset K, list of objects L
1. result = compute_Skyline(L)
2. For each o ∈ result
3.       output(NULL, o)
```

알고리즘 3: Global-MR

```
//Map Procedure
Input: a <K, V> pair //offset K, Object V
1. output(NULL, o) //하나의 리듀서에 보냄
//Reduce Procedure
알고리즘 2의 Reduce Procedure와 동일
```

알고리즘 3은 두 번째 글로벌 단계의 맵/리듀스 작업을 설명한다. 매핑 과정에서 각 객체의 키를 NULL 통일하여 <그림 5>의 (b)와 같이 하나의 리듀서에게 로컬 스카이라인 데이터를 전송한다(라인 1). 그리고 이후 단계는 알고리즘 2의 리듀싱 단계와 동일하다. 즉, 로컬 스카이라인을 하나로 병합한 후 최종 글로벌 스카이라인을 구한다(<그림 5>의 (c) 참조).



<그림 5> 스카이라인 질의처리(글로벌 단계)

5. 성능 평가

이번 장에서는 제안된 MR-DEAP 기법의 효율성 입증하기 위해 알고리즘을 구현하고 다양한 실험 환경에서 성능을 비교 분석한다.

5.1 실험 환경

MR-DEAP의 효율성을 평가하기 위해 실험 환경은 데이터 개수, 차원, 데이터 분포, 샘플링 비율을 변화시켜 실행시간 관점에서 MR-Angular와 MR-BNL 알고리즘과의 성능을 비교하였다. 구체적인 실험 환경 설정은 <표 1>과 같다. 밑줄은 기본 값을 나타낸다.

<표 1> 실험 환경 설정

항목	범위
data cardinality(n)	10M, 30M, 50M, 70M, 100M
data distribution(d)	uniform, anti-correlated
sampling ratio(%)	1, 2, 3
# of servers	4

맵/리듀스 환경은 4대의 IBM 서버를 사용하였으며, 각각의 서버는 리눅스 운영체제와 E3-1270 V2 x 4 3.5GHz 인텔 제온 CPU, 4GB의 메인 메모리로 구성된다. 또한 본문에서 제시된 모든 알고리즘은 Javac 1.7로 컴파일 되었으며 Hadoop 2.6.0을 사용하였다.

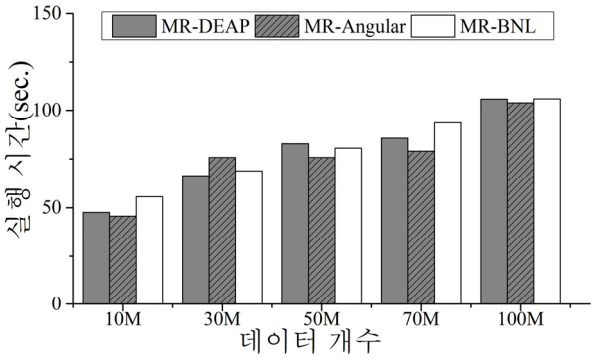
5.2 실험 결과

실험 결과는 데이터 개수, 데이터 분포, 데이터 차원, 샘플링 비율 관점에서 측정되었으며, 각 실험은 10번을 수행한 실행 시간의 평균값을 사용하여 도표를 작성하였다.

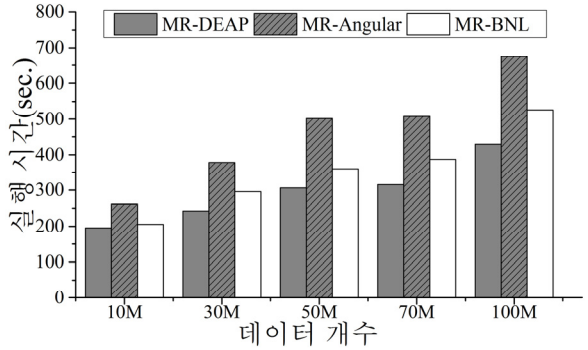
데이터 개수를 10^7개에서 10^8까지 단계적으로 증가시켰으며 [그림 6], [그림 7]과 같이 수행되는 시간을 측정하였다. MR-DEAP의 샘플링 비율은 2%이다.

MR-DEAP이 MR-BNL에 비하여 전반적인 구간에서 우수한 성능을 보였다. 예상했던 바와 같이 MR-DEAP는 불필요한 구간에서 스카이라인

검색을 수행하고 불필요한 후보를 탐색하여 처리 시간과 네트워크 전송시간 때문으로 판단된다. 그러나 MR-Angular에 비하여 구분되는 성능을 보이지는 못하였다. 이는 데이터가 균일(uniform)하게 분포되어 랜덤 샘플링을 통한 공간 분할의 영향력이 없어지고 두 기법 모두 유사하게 공간을 분할하여 스카이라인을 처리한 것으로 생각된다.



<그림 6> 데이터 개수 영향(d = uniform)

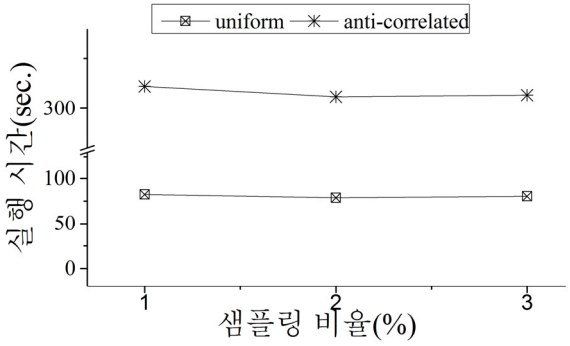


<그림 7> 데이터 개수의 영향(d = anti-correl.)

데이터 분포에 대해서 크게 uniform, anti-correlated 두 가지 분포를 사용하여 실험을 진행하였다. 이미 앞서 언급한 바와 같이 균일한 데이터 분포에서는 MR-DEAP의 성능은 크게 우수하게 보이지 않았다. 그러나 <그림 7>과 같이 데이터 분포가 한쪽으로 치우친, anti-correlated 분포에서는 MR-DEAP가 대조 알고리즘에 비하여 보다 빠른 수행 성능을 보이고 있다.

<그림 8>은 샘플링 비율을 1 ~ 3%에 따른 MR-DEAP의 수행시간을 나타낸다. 데이터 개수는 50M개로 고정되었다. 그림에서와 같이 anit-correlated 분포의 경우 샘플링 비율이 높아짐에 따라 실행 시간 또한 0.63 ~ 0.99% 정도 향상되는 것을 볼 수 있다. 그러나 uniform 분포에서는 1%에서는 82.482초, 2%에서는 78.783초, 3%에서는 80.513초로 실행성능이 등락을 보여 샘플링 기법으로 인한 성능향상이 이루어 졌다고 보기 어렵다.

이러한 원인은 anti-correlated의 경우 샘플링 비율을 증가함에 따라 데이터 분포의 추정이 정확해 짐에 따라 데이터 분할이 보다 균등하고 이루어져 실행 시간이 감소한 반면, uniform 분포에서는 다른 실험과 유사하게 분할 각이 거의 균일하게 이루어져 데이터 균등 분할의 효과가 미미했기 때문이라고 생각된다.



<그림 8> 샘플링 비율의 영향(n = 50M)

6. 결론

본 논문은 다양한 애플리케이션에서 응용되고 있는 스카이라인 질의를 맵/리듀스 환경에서 효율적으로 처리할 수 있는 기법을 제안하였다. 기존 MR-Angular와 MR-BNL 알고리즘에 중복적인 작업과 분포가 균일하지 않은 데이터에 대해 비효율적인 처리 문제 제기하고 이를 해결하기 위해 본 논문에서 제안한 MR-DEAP 기법은 랜덤 샘플링 기법인 저수지 샘플링 기법으로 데이터 분포를 추측하고, k-d tree를 이용하여 초구면 좌표계 상에서 데이터양이 균일하게 각 기반 분할한다. 마지막으로 데이터 개수, 데이터 분포, 차원 등의 설정을 변화시켜 MR-Angular 및 MR-BNL과 제안한 MR-DEAP과 비교 분석하고 실효성을 입증하였다.

참 고 문 헌

[1] 조성경, 김동은, 김응모. "효율적인 프렌차이즈 지점 선택을 위한 맵리듀스를 이용한 스카이라인 질의 처리 기법." 한국정보과학회 2014 한국컴퓨터종합학술대회 논문집 (2014): 1683-1685.

[2] Borzsony, S., Kossmann, D., & Stocker, K. (2001). The skyline operator. In Data Engineering, 2001. Proceedings. 17th International Conference on (pp. 421-430). IEEE.

[3] Vlachou, A., Doulkeridis, C., & Kotidis, Y. (2008, June). Angle-based space partitioning for efficient parallel skyline computation. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 227-238). ACM.

[4] Zhang, S., Mamoulis, N., & Cheung, D. W. (2009, June). Scalable skyline computation using object-based space partitioning. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (pp. 483-494). ACM.

[5] Chandler, P., & Sweller, J. (1991). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10*(1), 151-170.

[6] Kriegel, H. P., & Zimek, A. (2008, August). Angle-based outlier detection in high-dimensional data. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 444-452). ACM.

[8] Samet, H. (2006). Foundations of multidimensional and metric data structures. Morgan Kaufmann.

[9] Park, Y., Min, J. K., & Shim, K. (2013). Parallel computation of skyline and reverse skyline queries using mapreduce. Proceedings of the VLDB Endowment, 6(14), 2002-2013.

[10] Kossmann, D., Ramsak, F., & Rost, S. (2002, August). Shooting stars in the sky: An online algorithm for skyline queries. In Proceedings of the 28th international conference on Very Large Data Bases (pp. 275-286). VLDB Endowment.Dean, J., & Ghemawat, S. (2008).

[11] mapreduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

[12] Vitter, J. S. (1985). Random sampling with a reservoir. ACM Transactions on Mathematical Software (TOMS), 11(1), 37-57.

[13] B. Zhang, S. Zhou, and J. Guan. (2011). Adapting skyline computation to the mapreduce framework: Algorithms and experiments. In DASFAA Workshops (pp. 403-414).

[14] Chen, L., Hwang, K., & Wu, J. (2012). mapreduce skyline query processing with a new angular partitioning approach. In Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International (pp. 2262-2270). IEEE.

정 재 화

1999 고려대학교 컴퓨터교육과 (이학사)
2011 고려대학교 컴퓨터교육과 (이학석·박사)
2012~현재 한국방송통신대학교 컴퓨터과학과 조교수
관심분야: 공간질의처리 및 인덱싱, 분산 컴퓨팅 플랫폼 (mapreduce, spark), 모바일 데이터 관리, RFID, 무선 센서 네트워크

E-Mail: jaehwachung@knou.ac.kr