평가함수에 따른 분기한정 스카이라인 질의 처리 기법의 성능 분석†

최우성*, 민종현*, 임태형*, 현경석*, 김민석*, 정순영*
*고려대학교 컴퓨터학과
{ws_choi, jh_m, th_lim, ks_hyun, rlaalstjr47, jsy}@korea.ac.kr

Performance Analysis of Branch and Bound Skyline Computation via Evaluation Function

Woo-Sung Choi*, Jong-Hyeon Min*, Tae-Hyung Lim*, Kyeong-Seok Hyun*,
Min-Seok Kim*, Soon-Young Jung*
*Dept of Computer Science, Korea University

요 약

스카이라인 질의는 '지배(dominate)'관계를 적용한 선호도 질의(preference query)의 한 종류로, 복수의 기준을 이용한 의사 결정 시 사용된다. 스카이라인 질의 결과는 다수의 선택지 중에서 사용자가 다른 객체에 비해 뒤처지지 않는 선택지를 제시함으로써 사용자가 검토해야하는 선택지의 수를 대폭 감소시키기 때문에 대용량 데이터 분석 시 매우 유용하게 활용될 수 있다. 본 논문에서는 기존에 제시된 BBS(Branch and Bound Skyline Computation)에서 사용되고 있는 평가함수를 설명하고, 스카이라인 계산을 위해 사용할 수 있는 대안 평가함수의 속성을 제시한다. 또한 다양한 대안 평가함수를 사용한 실험을 통해 성능을 분석했으며, 이를 통해 기존 기법의 성능보다 좋은 평가함수가 존재함을 보였다.

1. 서론

스카이라인 질의[1](Definition 2 참조)란 '지배(dominate)' 관계(Definition 1 참조)를 이용한 선호도 질의(preference query)이다. 사용자가 검토해야할 대상의 수를 축소시켜 줄 수 있는 스카이라인 질의는 빅데이터 환경에서 매우 유용하게 활용될 수 있다. 이에 스카이라인 질의와 관련된 다양한 계산 기법이 제안되어 왔다.

특히 R-Tree[2] 등의 다차원 데이터 색인 구조를 분기한정(Branch and Bound) 탐색함으로써 최소한의 노드 방문을 통해 스카이라인 객체를 탐색하는 기법인 BBS(Branch and Bound Skyline Computation)[3]이 개발되었으며 다양한 분야에서 활용되고 있다.

[3]에서는 BBS가 R-Tree 기반 스카이라인 질의들 중 최적의 I/O 비용을 갖는다는 사실을 증명했다. 그러나 최적 우선탐색 기준인 평가함수(evaluation function)에 따른 성능 분석은 이루어지지 않았다.

본 논문에서는 스카이라인 계산을 위해 사용할 수 있는 대안 평가함수의 속성을 정의 내렸으며 이를 기반으로 다양한 대안 평가함수를 도출했다. 또한 이를 통해 기존 기법의 성능보다 좋은 평가함수가 존재함을 밝힌다.

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) $\exists j \in \{1,2,....,d\}: p_j < q_j$

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.

2. BBS에서 사용될 수 있는 평가함수의 속성

BBS는 주어진 R-Tree를 최적우선탐색(Best Frist Search) 방식으로 순회한다. 최적우선탐색이란 매 순간 가장 유망한(promising) 노드를 방문/분기 하는 기법이다. 최적우선탐색에서는 노드의 유망성을 수치화한 평가함수의 값이 최적(최소 또는 최대)인 노드를 우선적으로 방문한다. BBS에서 사용되는 평가함수는 다음과 같다.

Definition 3. 평가함수 f_{BBS}
d-차원 데이터 집합에 대한 스카이라인 탐색을 위한 평가함수 f_{BBS}는 주어진 엔트리 N에 대해 다음과 같이 정의된다.

$$f_{BBS}(N)=\begin{cases}\sum_{i}^{d} minPoint(MBR\ of\ N)_i & (if\ N\ is\ a\ node\ of\ RTree)\\ \sum_{i}^{d} N_i & (if\ N\ is\ a\ data\ point)\end{cases}$$

(단, d-차원 점 p에 대해 p_i는 p의 i번 째 차원 값을 뜻하며 d차원 사각형 R에 대한 $minPoint(R)$은 원점으로부터 L_1 거리가 가장 가까운 꼭짓점 좌표를 뜻함)

† 이 논문은 2016년 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임[NRF-2016R1A2B1014013]

4. 스카이라인 계산을 위한 대안 평가함수

BBS의 엔트리 평가함수 f_{BBS}가 알고리즘의 정확성을 보장할 수 있는 이유는 다음과 같은 성질 때문이다.

Property 1. 임의의 데이터 p, N에 대해,
$N < p \rightarrow f_{BBS}(N) < f_{BBS}(p)$

아래 $f_{weightedBBS}$는 Property 1을 만족하는 대표적인 함수의 예시이다.

Definition 4. 대안평가함수 $f_{weightedBBS}$

d-차원 데이터 집합에 대한 스카이라인 탐색을 위한 평가함수 f_{BBS}는 주어진 엔트리 N과 각 차원의 값이 0 이상이며 0벡터가 아닌 임의의 벡터 w 대해 다음과 같이 정의된다.

$$f_{weightedBBS}(N,w) = \begin{cases} \sum_{i}^{d} w_i \times minPoint(MBR\ of\ N)_i & (if\ N\ is\ a\ node\ of\ RTree) \\ \sum_{i}^{d} w_i \times N_i & (if\ N\ is\ a\ data\ point) \end{cases}$$

[3]에 따르면 BBS가 최소의 노드 방문으로 스카이라인을 계산하며, f_{BBS} 대신 Property 1을 만족하는 대안 평가함수를 사용해도 정확성 보장에 논리적 결함이 발생하지 않으므로 $f_{weightedBBS}$ 또한 최소 노드 방문으로 스카이라인을 계산한다. 그러나 w가 달라질 경우 엔트리 방문 순서가 달라질 수 있으며, [4]에 따르면 엔트리 방문 순서는 알고리즘의 성능(response 시간)에 영향을 끼친다.

본 논문에서는 다양한 w에 따른 분기한정 스카이라인 계산 알고리즘의 성능을 실험해보았다. 본 논문에서 실험한 w는 다음과 같다.

Definition 5. 임의 벡터 rw

d-차원 데이터 집합에 대한 임의 벡터 rw는 각 차원의 값이 0 이상이며 0벡터가 아닌 임의(random) 벡터를 뜻한다.

Definition 6. Data Distribution-driven 벡터 ddw

d-차원 데이터 집합 및 해당 데이터 집합을 색인한 R-Tree R에 대한 Data Distribution-driven 벡터 ddw는 다음과 같이 정의된다.

$ddw = maxPoint(MBR\ of\ N) - minPoint(MBR\ of\ N)$

($minPoint(R)$은 원점으로부터 L_1 거리가 가장 가까운 꼭짓점 좌표를 뜻하며 $maxPoint(R)$은 원점으로부터 L_1 거리가 가장 먼 꼭짓점 좌표를 뜻함)

Definition 7. x 차원 사영 벡터 $projectiontoXw$

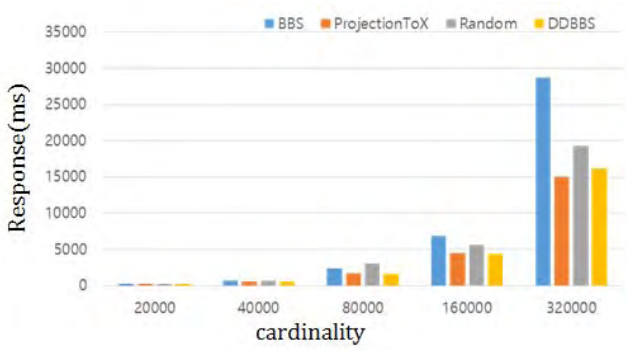
d-차원 데이터 집합에 x축 사영 벡터 $projectiontoXw$는

$$projectiontoXw = \begin{cases} projectiontoXw_1 = 1 \\ projectiontoXw_i = 0\ (i \neq 1) \end{cases}$$

를 만족한다.

5. 평가함수에 따른 실험 결과 분석

본 논문에서는 다양한 대안 평가함수 $f_{weightedBBS}$에 따른 분기한정 스카이라인 계산 알고리즘의 성능을 비교평가 했다. anti-correlated[5] 분포를 따르는 3차원 데이터 집합에 대해 실험했다. 실험에 사용된 알고리즘으로는 BBS[3]와 대안 평가함수 $projectiontoXw$, rw, ddw를 w로 사용하는 분기한정 스카이라인 계산 기법인 ProjectionToX, Random, DDBBS를 대상으로 response 시간을 측정했다.



(그림 2) 실험 결과

(그림 2)는 실험 결과를 요약하는 그래프이다. 로그스케일(log-scale)로 증가하는 데이터 수(cardinality)에 대해 Random을 제외한 각 알고리즘은 로그스케일로 성능이 증가하는 것으로 보아 확장가능(scalable)한 것으로 파악된다. 흥미로운 점은 ProjectionToX와 DDBBS는 일관적으로 BBS보다 좋은 성능을 보여주고 있다는 것이다. 이는 기존 f_{BBS}보다 좋은 성능을 보여줄 수 $f_{weightedBBS}$가 존재한다는 것을 의미한다.

6. 결론 및 제언

본 논문에서는 스카이라인 계산을 위해 사용할 수 있는 대안 평가함수의 속성을 정의 내렸으며 이를 기반으로 다양한 대안 평가함수를 도출했다. 또한 이를 통해 기존 기법의 성능보다 좋은 평가함수가 존재함을 밝혔다. 향후 연구로서 f_{BBS}보다 좋은 성능이 보장할 수 있는 최적의 $f_{weightedBBS}$의 w를 계산하는 기법에 대해 연구할 계획이다.

7. 참고문헌

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

[2] Guttman, Antonin. R-Trees: a dynamic index structure for spatial searching. Vol. 14. No. 2. ACM, 1984.

[3] Papadias, Dimitris, et al. "An optimal and progressive algorithm for skyline queries." Proceedings of the 2003 ACM SIGMOD international conference on Management of data. ACM, 2003.

[4] Lu, Ying, et al. "Efficient algorithms and cost models for reverse spatial-keyword k-nearest neighbor search." ACM Transactions on Database Systems (TODS) 39.2 (2014): 13.

[5] Haichuan Shang, et al. "Skyline Operator on Anti-correlated Distributions" VLDB. Vol. 6.9: pp.649-660. 2013.