셀 상태 값을 이용한 연속 스카이라인 질의

최현식° 강상원 정하림 구원교 정승희 황종선 이상근
고려대학교 컴퓨터학과
{hyunsik, swkang, harim, lkhkwk, shchung, hwang}@disys.korea.ac.kr yalphy@korea.ac.kr

Continuous Skyline Queries Using Cell State

Hyunsik Choi°, Sang-Won Kang, Ha-Rim Jung, Won-Kyo Ku, Seung-Hee Chung
Chong-Sun Hwang, SangKeun Lee

Department of Computer Science and Engineering, Korea University

스카이라인 질의(Skyline Query)는 다중속성을 갖는 튜플(Tuple)들의 집합에서 다른 어떤 튜플들 에게도 지배(domination) 되지 않는 튜플을 찾는 문제이다. 스카이라인에서 지배의 의미는 중요한데 지배의 정의는 다음과 같다. 튜플 r과 r'이 있고 r과 r'의 모든 속성을 각각 비교했을 때 r의 모든 속성들이 r'의 모든 속성들보다 작거나[1] 같고 r이 r' 보다 최소한 한 속성이 반드시 작다면 r이 r'를 지배한다고 한다.

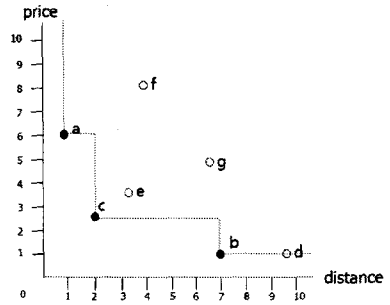

그림 1. 거리와 가격에 대한 호텔들의 스카이라인

스카이라인 질의를 설명하기 위해 그림 1과 같이 호텔을 호텔의 두 가지 속성인 '투숙비' 와 '해변으로부터의 거리' 를 x,y 좌표값으로 하여 2차원 좌표평면 위에 점 {a,b,c,d,e,f,g}로 표현했다. 이때 호텔 a는 해변으로부터의 거리와 투숙비 모두 호텔 f 보다 작기 때문에 a 가 f를 지배 한다고 한다. c 또한 {e,f,g} 보다 모든 속성 값이 작기 때문에 {e,f,g}를 지배한다. b와 d의 관계는 약간 다르다. b가 d보다 거리는 짧지만 투숙비는 같다. 이런 경우에도 앞서 정의한 대로 지배한다고 한다. 반면 a와 b의 경우 a 는 b 보다 거리는 짧고, b가 a 보다 투숙비는 싸다. 이런 경우는 비교불가(incomparable)라고 한다. 결과적으로는 그림 1의 튜플들 중에서 스카이라인 튜플은 어떤 점에게도 지배되지 않는 {a, b, c}가 된다.

위와 같은 특징을 갖는 스카이라인은 다중속성을 고려한 질의로써 데이터베이스 분야에서 집중적으로 연구 되어 왔다. 그러나 기존 스카이라인 연구들은 최근 발전하는 모바일(Mobile) 및 스트림 데이터

1) 응용에 따라 큰 것이 작은 것을 지배한다고 할 수 도 있다. 그러나 대부분의 스카이라인 논문들에서와 같이 본 논문에서도 작은 값이 큰 값을 지배한다고 하겠다.

(Stream Data)와 같이 데이터가 실시간으로 변하는 동적인 환경에는 적합하지 않다.

기존 정적 환경에서는 한 묶음의 정적 데이터에 대한 한번의 질의 처리를 요구하는데 반해 동적인 환경은 실시간으로 변하는 데이터가 주어지며, 장기간 동안 변하는 데이터에 대한 계속적인 질의처리를 요구한다. 따라서 정적 데이터에 대한 질의만을 고려했던 기존 연구들은 데이터가 실시간으로 변하는 동적인 환경에서는 적합하지 않다.

이와 같은 문제를 해결하기 위해 Tao et al.는 스트림 환경에서의 연속적인 스카이라인 질의처리 기법을 제안했다. 그러나 R-tree 를 기반으로 한 색인을 사용했기 때문에 빈번한 데이터 추가, 삭제 및 데이터 값 변경 시 색인 갱신에 대한 과비용(overhead) 문제가 발생한다. 또한 매번 튜플 추가 시, 해당 튜플의 스카이라인 여부를 판단하기 위해 범위질의(Range Query) 수행으로 인한 과비용 문제가 유발된다. Li Tian et al.은 의 문제를 해결하기 위해 격자 셀을 색인으로 하는 연속적인 스카이라인 질의처리(Continuous Monitoring Skyline Query, CMSQ)를 제안했다. CMSQ는 영향 영역(Influence Region) 에 삽입된 튜플에 한해서만 추가적인 연산을 수행함으로써 연산 비용을 감소시켰다. 그러나 CMSQ는 다음과 같은 문제를 갖고 있다. (i) 영향 영역을 연결 리스트(Linked List) 로 유지하여 매번 튜플이 결과에 영향을 주는지 여부를 확인하기 위해 $O(n)$의 비용이 든다. (ii) 어떤 튜플의 스카이라인 여부를 알기 위해 시스템이 유지하는 모든 스카이라인과 순차적으로 비교하므로 큰 비용이 발생한다. 더군다나 차원이 높아짐에 따라 스카이라인의 수는 더욱 증가하기 때문에 차원 증가에 따라 더 큰 비용이 요구된다. (iii) 셀 방문 시 큐(Queue)에 현재 방문 셀의 인접 셀들을 넣어 하나씩 꺼내면서 방문하는데 이때 고차원에서 큐의 크기가 너무 증가한다. 또한 인접 셀 방문 알고리즘이 갖는 특성 때문에 중복 방문 문제가 발생하므로 중복방문을 회피하기 위한 추가 비용이 요구된다.

위의 문제를 해결하기위해 본 논문에서는 격자 셀에 상태값을 설정하고 이렇게 설정된 상태값을 활용한 연속적인 스카이라인 질의(CSQUCS)를 제안한다. 제안 기법은 다음과 같은 기여를 한다. (i)셀의 상태 값을 셀에 직접 유지하여 해당 셀에 추가된 튜플이 결과에 영향을 주는지 확인하는 비용을 $O(1)$로 감소 시킨다. (ii) 인접 셀의 상태값을 확인함으로써 방문해야 할 셀의 범위를 최소로 설정하여 질의처리 시 방문해야 하는 셀의 수를 감소 시킨다. (iii) 각 셀이 지역 스카이라인(하나의 셀 안에서의 스카이라인)을 별도로 유지하게 하여 셀들 간 튜플 비교 시 지역 스카이라인들 끼리만 비교하게 함으로써 비교 연산 횟수를 줄인다.

제안 기법의 연산비용 성능평가를 위해 기존 기법인 CSMQ와의 비교평가를 시뮬레이션을 통해 수행하였다. 비교평가는 튜플 개수에 따른 단위시간당 질의 처리 비용과 격자 셀의 입도(Granularity)의 변화에 따른 질의 처리 비용을 수행했다. 시뮬레이션 결과는 제안 기법이 기존 연구에 비해 질의처리 비용을 현저하게 감소시킨다는 것을 보여준다.