무선 방송 환경에서 스카이라인 질의 처리 기법*

하종우, 박상현, 류병걸, 이상근
고려대학교 정보통신대학 컴퓨터통신공학부
e-mail :{okcomputer, newtypus, smart123, yalphy}@korea.ac.kr

Skyline Query Processing Method in Wireless Broadcast Environments

Jong-Woo Ha, Sang-Hyun Park, Byung-Gul Ryu, SangKeun Lee
Division of Computer and Communication Engineering, Korea University

요 약

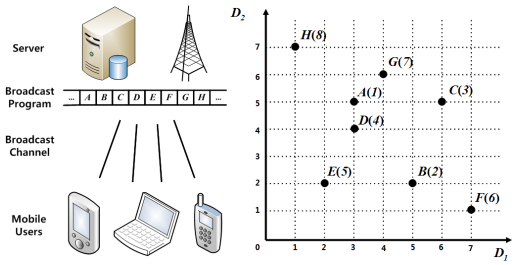
본 논문에서 무선 방송 환경에서 효율적으로 스카이라인 질의를 처리하는 기법을 제안한다. 무선 방송 환경의 순차적 데이터 접근 특성 때문에 기존의 기법을 적용할 경우 접근시간 및 튜닝시간에 큰 제약이 있다. 이를 해결하고 모바일 사용자가 에너지 효율적으로 스카이라인 질의를 처리하기 위하여 DSI(Distributed Spatial Index) 구조에 기반한 SOA(Skyline On Air) 알고리즘을 제안하였다. 제안된 기법은 접근시간이 한 주기의 방송 프로그램 길이를 넘지 않도록 한다. 또한 성능 평가를 통하여 제안 된 기법이 접근시간 및 튜닝시간 측면에서 효율적임을 확인하였다.

1. 서론

모바일 기기와 무선 통신 기술의 지속적인 발달로 최근 데이터 방송 서비스가 현실화 되었으며 그 사용자가 급속하게 증가하고 있다[1][2]. T-DMB 의 경우, 뉴스, 날씨, 주식, 교통 정보 등의 데이터가 방송 채널을 매개로 모바일 사용자에게 전달되고 있으며 서비스 개시 2 년 반 만에 1000 만 명의 사용자가 서비스를 이용하고 있다[2]. 이러한 무선 방송 환경에서 에너지 효율적으로 데이터를 전송하고 받는 기법과 더불어 사용자가 적극적으로 특정 질의를 처리하는 기법이 지속적으로 연구되고 있다[6][8][12][13][14].

본 논문에서는 데이터 방송 서비스가 제공되는 무선 방송 환경에서 스카이라인 질의를 효율적으로 처리하기 위한 문제를 다룬다. 스카이라인은 다차원공간에서 다른 데이터들로부터 지배되지 않는 데이터의 집합이다[3]. 전체 데이터 집합에서 스카이라인을 계산하는 기법은 복수 개의 기준으로 의사를 결정하는 응용 분야에서 중요하기 때문에 데이터베이스 분야에서 많은 연구가 이루어져왔다[3][7][11]. T-DMB 의 경우, 방송되는 주식 데이터를 입력으로 하여 스카이라인 질의를 처리하여 사용자의 의사결정에 도움을 줄 수 있다.

현재 서비스되고 있는 데이터 방송에서 서버 측의 방송 제작 기법을 수정하지 않아도 모바일 클라이언트는 스카이라인 질의를 처리할 수 있다. 하지만 이 경우 질의 처리의 정확성을 보장하기 위하여 방송 프로그램에 있는 모든 데이터를 튜닝해야 하기 때문에 최악의 튜닝시간[6] 성능을 가진다. 그림 1 은 데이터 방송 서비스가 제공되는 환경에서 2 차원의 데이터 집합이 방송되는 예제이다. 클라이언트가 데이터 A 를 튜닝하여 스카이라인 질의 처리를 시작한다고 가정하였을 때, 정확한 스카이라인을 얻기 위하여 한 주기의 방송 프로그램을 전부 튜닝하며 연속적으로 지배 검사(Dominance test)를 수행해야 한다. 따라서 다음 주기의 데이터 A 를 받음으로써 질의 처리가 종료된다.



(그림 1) 데이터 방송 서비스의 예

한편, 스카이라인 처리를 위하여 많은 연구가 진행되고 있지만 기존 연구는 서버 측에서 실행하는 질의 처리 기법에만 집중 되었으며, 무선 방송 환경의 순차적 데이터 접근 특성을 고려한 연구는 이루어 지지 않은 실정이다[7][11]. 스카이라인 질의 처리 기법의 최신 연구에서와 같이 트리 구조에 데이터를 색인하여 방송 프로그램을 제작하고 랜덤하게 데이터를 접근하는 알고리즘으로 질의를 처리하였을 경우에도 스카이라인을 계산할 수 있다. 하지만 질의 처리가 몇 주기의 방송 프로그램에 완료되는지에 대한 보장이

* 이 연구에 참여한 연구자는 '2 단계 BK21'의 지원비를 받았음

없다는 관점에서 접근시간[6]의 성능에 제약이 있다.

이러한 문제점을 해결하기 위하여 본 논문에서는 방송되는 데이터를 효율적으로 가지치기(Pruning)하여 튜닝시간 성능을 높임과 동시에 접근시간은 한 주기의 방송 프로그램으로 보장하는 기법을 제시한다. 제안하는 기법은 무선 방송 환경의 순차적 데이터 접근 특성에 특화된 색인 구조인 DSI(Distributed Spatial Index)를 기반하여 데이터를 SWEEP 순서[10]로 색인한 후 방송 프로그램을 제작한다. 또한 방송되는 데이터를 효율적으로 가지치기하여 튜닝시간 성능을 높인다. 더욱이 SWEEP 순서의 특성을 활용하여 질의 처리에 필요한 지배 검사 수를 크게 줄일 수 있다.

본 논문의 구성은 다음과 같다. 2 장에서는 스카이라인 질의 처리의 관련 최신 연구와 무선 방송 환경에서 효율적인 색인 구조에 관한 기존 연구를 살펴본다. 3 장에서는 무선 방송 환경에서 스카이라인 질의를 처리하는 효율적인 기법을 제안한다. 4 장에서는 제안하는 기법의 성능 평가 결과를 제시하며, 5 장에서는 향후 연구 방향을 포함하여 본 논문의 결론을 내린다.

2. 관련 연구

스카이라인 질의 처리를 위한 최신 기법은 데이터 집합을 트리 구조에 색인한 후 효율적인 분지한계(Branch-and-bound) 알고리즘을 개발하는 방향으로 연구되고 있다[7][11]. 이것은 특정 데이터가 지배하는 영역(Dominant region)을 기준으로 많은 데이터를 동시에 가지치기를 수행함으로써 비용이 높은 지배 검사의 수를 줄이기 위함이다. 이러한 아이디어에 기반한 연구에서는 다차원 데이터를 색인하기 위하여 기존의 R-tree 를 활용하여, ZB-tree 와 같이 스카이라인 질의 처리에 특화된 새로운 트리 구조를 제시하였다[7][11]. 하지만 무선 방송 환경의 순차적 데이터 접근 특성으로 인하여 이러한 기존 기법을 무선 방송 환경에 적용하는 것은 Branch-and-Bound 문제[8][13]를 일으켜 접근시간 성능이 크게 떨어지기 때문에 적합하지 않다.

무선 방송 환경에서 순차적 데이터 접근 특성을 고려하여 위치 종속적인 질의(Location-dependent query)와 같이 복잡한 질의를 효율적으로 처리하기 위한 연구가 진행되어 왔다 [8][13][14]. 이러한 연구의 대표적인 것으로 무선 방송 환경에 특화된 색인 구조인 DSI(Distributed Spatial Index)가 있다[8]. DSI 에서는 무선 방송 환경에서 데이터 전송의 기본 단위인 각각의 버켓(Bucket)에 색인 구조를 분산하여 할당함으로써, 기존의 트리 색인 구조가 무선 방송 환경에서 가지는 제약을 해소하였다. DSI 구조와 더불어 위치 종속적인 질의인 윈도우 질의와 *kNN*(k-Nearest Neighbor) 질의 처리 기법이 제안되었다.

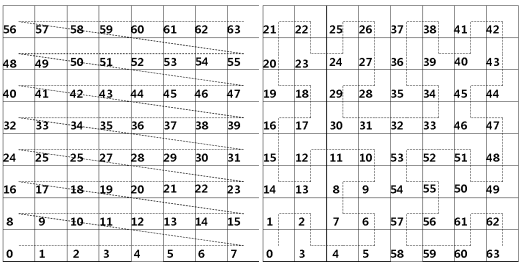
3. 제안 기법

3.1. 방송 프로그램 제작 기법

서버 측에서는 DSI 구조를 활용하여 방송 프로그램을 제작한다. 더불어 각각의 데이터는 잘 알려진 공간 채움 순서인 SWEEP 순서에 기반하여 색인된다. SWEEP 순서에서는 아래의 식 (1)과 같이 간단하게 인코딩이 가능하다 [10].

$$\sum_{i=1}^{n_D}\left(v_{d_i}\times b^{i-1}\right) \quad (1)$$

식 (1)에서 n_D 는 차원의 수, v_{di} 는 i 번째 차원의 속성, b 는 차원의 공간 너비를 의미한다. 1 차원 색인 값을 다차원 데이터로 디코딩 하는 과정은 인코딩 과정의 역을 취한다.



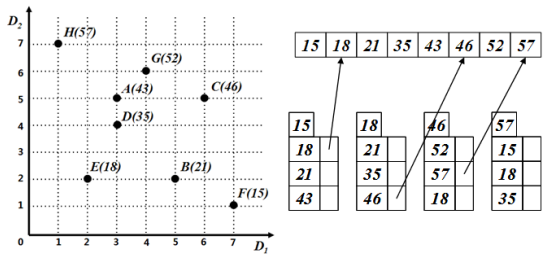
(그림 2) SWEEP 순서 색인과 힐버트 커브 순서 색인

기존의 DSI 에서 데이터 색인을 위하여 활용한 힐버트 커브(Hilbert curve) 순서[5]를 적용하지 않고 새로운 색인 기법을 제안한 이유는 크게 두 가지이다. 첫째, 힐버트 커브 순서와 달리 SWEEP 순서는 스카이라인 질의 처리에서 지배 연산의 수를 줄이는데 활용될 수 있기 때문이다. 그림 2 의 좌측과 같이 SWEEP 순서로 색인된 데이터는 색인 값이 큰 데이터가 색인 값이 작은 데이터를 지배할 수 없다는 특성이 있다. 따라서 특정 데이터보다 색인 값이 더 큰 데이터에 대한 지배 검사를 수행하지 않아도 그 데이터의 지배 여부를 정확하게 판단할 수 있다. 하지만 힐버트 커브 순서에서는 이러한 특성이 존재하지 않는다. 그림 2 의 우측 힐버트 커브 순서 색인의 경우, 색인 값이 낮은 11 이 색이 값이 더 작은 13 에게 지배 된다. 둘째, SWEEP 커브의 인코딩과 디코딩 작업은 식 (1)에서 알 수 있듯이 몇 번의 사칙연산만으로 쉽게 가능하다. 하지만 힐버트 커브 순서는 색인을 디코딩 하여 다차원 공간에 맵핑하고 다차원 값을 인코딩 하여 색인을 얻는 작업이 느리다. 특히 차원의 수가 높아질수록 인코딩과 디코딩 작업의 비용이 비약적으로 증가한다 [4][9].

3.2 스카이라인 질의 처리 기법

아래의 그림 3 은 모바일 사용자가 본 논문에서 제안하는 SOA(Skyline On Air) 알고리즘으로 스카이라인 질의를 처리하는 예를 보여준다. 그림 3 에서 좌측 그림은 그림 1 에서의 예제 데이터가 SWEEP 순서로 인코딩 된 것을 보여주며, 우측 상단과 하단의 그림은 각각 방송 채널 상에 존재하는 방송 프로그램과 버켓을 튜닝함으로써 얻을 수 있는 색인 구조를 나타낸다. 색인 구조에는 현재 튜닝한 데이터로부터 1, 2, 4 만큼

떨어진 데이터의 색인 값과 포인터가 포함되어 있다. 일반적으로 DSI 구조에서 각각의 버켓은 $base^0$, $base^1$, $base^2$, ..., $base^i$ 만큼 떨어진 데이터의 색인 값과 포인터를 포함하며 i 는 n 이 방송 프로그램의 길이(데이터 집합의 원소 개수)일 때 $(log_{base}(n) - 1)$보다 크지 않은 정수로 할당된다 [9]. 그림 3 은 각각 $base$ 를 2 로 설정하고, i 에 2 가 할당된 예제이다.



(그림 3) 스카이라인 질의 처리 예

먼저, 사용자가 버켓 15 를 튜닝함으로써 스카이라인 질의 처리를 시작하였다고 가정하자. 버켓 15 의 인덱스 구조를 해석하면 아직 방송 프로그램의 길이를 알 수 없지만 방송 프로그램 상에서 데이터가 (15, 18, 21, ?, 43, ...)와 같이 존재한다는 것을 알 수 있다. 발견된 데이터를 기준으로 지배 검사를 수행하면 데이터 15 와 데이터 18 이 현재의 스카이라인 후보(Skyline candidate)가 됨을 알 수 있다. 현재 튜닝한 데이터가 스카이라인 후보에 포함이 되므로 저장하여 두고, 다음 번으로 튜닝할 버켓을 결정한다. 데이터 18 은 스카이라인 후보이기 때문에 다음 번 튜닝할 버켓은 버켓 18 이 된다. 데이터 21 과 데이터 35 는 데이터 18 일 지배되기 때문에 튜닝할 필요가 없다.

다음으로 버켓 18 일 튜닝하고 색인 구조를 분석하면 방송 프로그램이 (15, 18, 21, 35, 43, 46, ...)으로 편성되어 있다는 것을 알 수 있다. 현재의 스카이라인 후보인 15 와 18 을 새롭게 발견된 데이터인 35 와 46 에 지배 검사를 수행하여 35 와 46 을 튜닝할 필요가 없다는 것을 알게 된다. 따라서 현재 발견된 데이터들 중에서 읽을 필요가 있는 데이터가 없지만 방송 프로그램 상에 더 많은 데이터가 존재할 수 있으므로 데이터 46 을 튜닝하여 다른 데이터의 존재를 확인한다.

이러한 과정이 방송 프로그램 상의 전체 데이터 집합을 발견하고, 지배 검사를 수행하여 스카이라인 후보를 모두 튜닝하였을 때 스카이라인 질의를 종료한다. 이러한 기법을 적용하면 반드시 방송 프로그램의 한 주기 안에 질의 처리가 종료된다. 또한 새롭게 발견된 모든 데이터에 대하여 지배 검사를 수행하여 매 순간 스카이라인 후보를 추출하기 때문에 스카이라인 질의의 정확성이 보장된다.

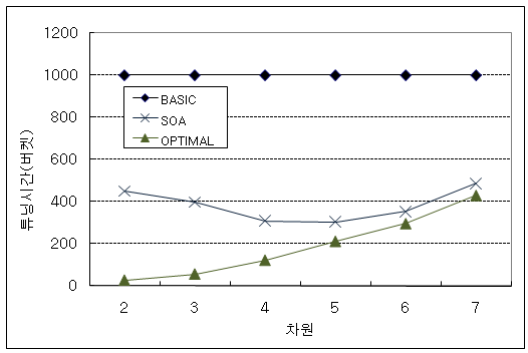
더불어 스카이라인 후보를 추출하기 위하여 지배 검사를 수행할 때, SWEEP 순서의 이점을 활용하면 하나의 스카이라인 질의를 처리하기 위하여 수행하는 지배 검사의 횟수를 줄일 수 있다. SWEEP 순서로 인코딩된 데이터는 자신 보다 높은 순서의 데이터에게 지배되지 않기 때문에 이러한 지배 검사를 수행하지 않아도 질의 처리의 정확성이 보장된다.

4. 성능 평가

본 논문에서 제안한 기법의 성능을 평가하기 위하여 데이터 방송 서비스에서의 서버와 방송 채널, 모바일 사용자를 시뮬레이터 상에서 구현하였다. 시뮬레이터는 Java SDK 1.6.0 버전으로 작성되었다. 총 1000 개의 원소로 이루어진 데이터 집합을 DSI 색인 구조에서 SWEEP 순서로 인코딩하여 방송 프로그램을 제작하였다.

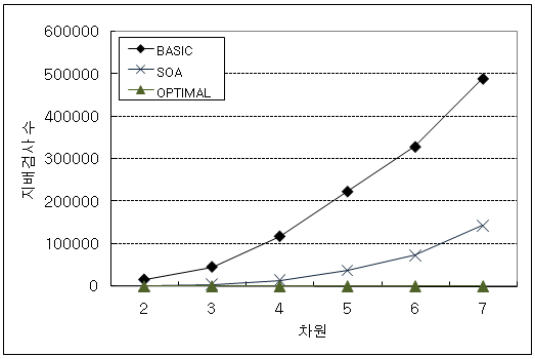
무선 방송 환경에서 잘 알려진 지연시간과 튜닝시간을 측정하여 질의 처리의 효율성을 평가하였다[6]. 또한 스카이라인 질의를 처리하기 위하여 모바일 사용자가 계산을 위하여 소모하는 에너지를 간접적으로 측정하기 위하여 제안된 기법이 수행하는 지배 검사의 횟수를 측정하였다.

제안하는 기법은 성능 평가 그림 4, 5, 6 의 결과에서 SOA 로 표시하였다. 제안 기법의 성능을 상대적으로 평가하기 위하여 BASIC 알고리즘과 OPTIMAL 알고리즘을 추가 구현하였다. BASIC 알고리즘은 방송 프로그램 상의 모든 버켓을 읽고 스카이라인 질의를 처리하는 기법으로서, 인덱스가 주어지지 않은 상황에서 모바일 클라이언트가 채택할 수 있는 유일한 처리 기법이다. OPTIMAL 알고리즘은 방송 프로그램 상의 스카이라인을 미리 알고 스카이라인 데이터만을 에너지 효율적으로 튜닝하는 이상적인 알고리즘으로서, 무선 방송 환경에서 가능한 최선의 지연시간 및 튜닝시간 성능을 알려준다.



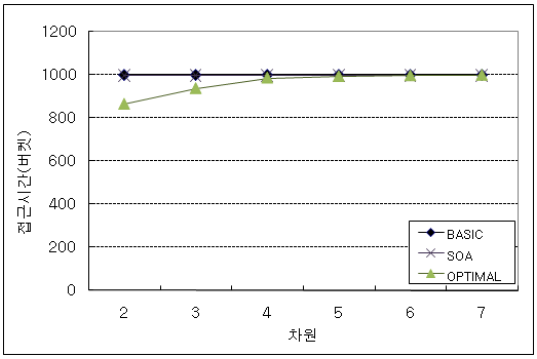
(그림 4) 튜닝시간 성능

그림 4 는 데이터 집합의 차원을 달리하였을 때 제안하는 기법의 튜닝시간 성능을 나타낸 것이다. 낮은 차원에서는 데이터 집합에서 스카이라인 포인트의 수가 작다. 따라서 낮은 차원에서 튜닝시간이 큰 이유는 스카이라인 후보가 아니지만 질의 처리의 정확성을 보장하기 위하여 튜닝하는 버켓의 개수가 많기 때문이다. 이러한 현상은 차원이 올라가고 스카이라인의 수가 많아짐에 따라 줄어들어서, 5 차원을 넘기면서 OPTIMAL 알고리즘과의 성능 차이가 낮게 유지된다.



(그림 5) 스카이라인 질의 처리 시 지배 검사 수

그림 5 는 모바일 사용자가 스카이라인 질의 처리 시, 계산에 소모되는 에너지를 간접적으로 측정하기 위하여 지배 검사의 수를 측정한 결과이다. BASIC 알고리즘에 비교하여 제안하는 기법은 지배 검사의 횟수를 60~70% 정도를 줄인다는 것을 알 수 있다. OPTIMAL 알고리즘은 질의 처리 전에 스카이라인을 이미 알고 있기 때문에 지배 검사를 수행한다.



(그림 6) 지연시간 성능

그림 6 은 스카이라인 질의 처리의 지연시간 성능을 나타낸 것이다. 제안된 기법은 BASIC 알고리즘에 비교하여 약간의 성능 향상을 보인다. 이것은 질의 처리의 정확성이 보장되면 바로 질의 처리를 종료할 수 있기 때문이다. 더욱이 4 차원 이상의 데이터 집합에서 지연시간은 OPTIMAL 알고리즘과 거의 유사한 성능을 보임을 알 수 있다. 이것은 차원이 늘어날수록 스카이라인의 개수가 많아지기 때문이다.

5. 결론 및 향후 연구

본 논문에서는 데이터 방송 서비스가 제공되는 무선 방송 환경에서 모바일 사용자가 효율적으로 스카이라인 질의를 처리하는 기법을 제시하였다. 제안된 기법은 튜닝시간 성능의 측면에서 이득이 있기 때문에 모바일 사용자가 에너지 효율적으로 스카이라인 질의를 처리하는 것을 가능하게 한다. 동시에 한 주기의 방송 프로그램 안에 질의 처리가 반드시 종료되는 특성을 가진다.

향후 동일한 환경에서 제약 부공간 스카이라인 (Constrained subspace skyline) 등 다양한 종류의 스카이라인 질의 처리 기법에 대하여 다루고자 한다. 또한 데이터의 업데이트와 통신 에러가 빈번한 환경에서 효율적인 스카이라인 처리 기법을 연구할 계획이다.

참고문헌

[1] 1seg. http://www.dpa.or.jp/english/.

[2] Terrestrial digital multimedia broadcasting (T-DMB). http://www.t-dmb.org/.

[3] S. Borzsonyi, K. Stocker, and D. Kossmann. The skyline operator. In *Proceedings of the International Conference on Data Engineering*, pages 421-430, 2001.

[4] L. Chenyang, Z. Hong, and W. Nengchao. Fast *n*-dimensional hilbert mapping algorithm. In *Proceedings of the IEEE International conference on Computational Sciences and Its Applications*, pages 507-513, 2008.

[5] C. Gotsman and M. Lindenbaum. On the metric properties of discrete space-flling curves. *IEEE Transactions on Image Processing*, 5(5):794-797, 1996.

[6] T. Imielinski, S. Viswanathan, and B. Badrinath. Data on air: Organization and access. *IEEE Transactions on Knowledge and Data Engineering*, 9(3):353-372, 1997.

[7] K. C. K. Lee, B. Zheng, H. Li, and W.-C. Lee. Approaching the skyline in z order. In *Proceedings of the International Conference on Very Large Data Bases*, pages 270- 290, 2007.

[8] W.-C. Lee and B. Zheng. DSI: A fully distributed spatial index for location-based wireless broadcast services. In *Proceedings of the IEEE International conference on Distributed Computing Systems*, pages 349-358, 2005.

[9] X. Liu and G. Schrack. Encoding and decoding the Hilbert order. *Software-Practice and Experience*, 26(12):1335-1346, 1996.

[10] M. F. Mokbel, W. G. Aref, and I. Kamel. Analysis of multi-dimensional space-flling curves. *Geoinformatica*, 7(3):179-209, 2003.

[11] D. Papadias, Y. Tao, G. Fu, and B. Seeger. An optimal and progressive algorithm for skyline queries. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 467-478, 2003.

[12] J. Xu, W.-C. Lee, and X. Tang. Exponential index: A parameterized distributed indexing scheme for data on air. In *Proceedings of the International conference on Mobile Systems, Applications, and Services*, pages 153-164, 2004.

[13] J. Xu, B. Zheng, W.-C. Lee, and D. L. Lee. Energy effcient index for querying location-dependent data in mobile broadcast environments. In *Proceedings of the IEEE International Conference on Data Engineering*, pages 239-250, 2003.

[14] B. Zheng, W.-C. Lee, and D. L. Lee. Spatial index on air. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications*, pages 297-304, 2003.