Skyline 을 사용하는 Layer 기반 방법에 관한 조사

이지현*, 박영호*
*숙명여자대학교 멀티미디어과학과
e-mail : {jhlee7, yhpark}@sookmyung.ac.kr

An Survey on Layer-based Methods using Skylines

Ji-Hyeon Lee*, Young-Ho Park*
*Dept of Multimedia Science, Sookmyung Women's University

요 약

인터넷의 발달로 데이터가 이질적이고 방대해짐에 따라 사용자의 의도와 목적에 맞는 정보를 빠르고 정확하게 찾아내는 것이 어려워지고, 대용량의 데이터를 빠르게 검색 할 수 있는 효율적인 top k 질의 처리가 중요해 지고 있다. top k 질의 처리는 릴레이션에서 가장 높은 (또는 가장 낮은) 값을 가지는 k 개의 튜플을 반환하는 방법이며, 그 중 Layer 기반 방법은 객체가 가지는 d 개의 속성 값들을 d-차원의 공간상의 점 객체로 맵핑하여, layer 들의 list 를 생성 한다. 본 논문에서는 Layer 기반 방법 중 skyline 을 사용하여 layer 을 생성하고 인덱스를 구축하는 기존 연구에 대해서 조사한다. 그리고 대표적인 방법인 모든 객체를 순차적으로 비교하는 BNL 과 이의 비교 횟수를 감소시킨 SFS, 그리고 R-tree 를 사용한 NN 과 이의 계산 비용을 감소시킨 BBS 에 대해 설명한다.

1. 서론

우리는 특정 목적을 가지고 그 목적을 이루기 위해 검색을 한다. 그러나 정보화 사회에 접어 들며 인터넷의 발달로 데이터가 이질적이며 방대해짐에 따라 사용자의 목적에 맞는 정보를 빠르고 정확하게 찾아내는 것이 어려워지고 있다. 이에 따라 대용량의 데이터를 검색하여 빠르게 원하는 결과를 찾을 수 있는 효율적인 top k 질의 처리가 중요해 지고 있다.

top k 는 릴레이션에서 가장 높은 (또는 가장 낮은) 스코어를 가지는 k 개의 튜플을 반환하는 방법이다[3]. top k 질의 처리 연구는 인덱스 생성방법에 따라 Layer 기반 방법과 List 기반 방법으로 구분한다. 본 논문에서는 그 중에서 공간상에서 skyline 이나 convex hull 으로 layer 를 생성해서 인덱스를 구축하는 Layer 기반 방법의 기존 연구를 조사한다.

본 논문의 구성을 다음과 같다. 제 2 장에서는 top k 질의 처리 방법 중 Layer 기반 방법의 특징과 장, 단점에 대해서 설명한다. 제 3 장에서는 Layer 기반 방법 중 skyline 을 통해서 layer 을 생성하는 기존 연구에 대해 조사하고 제 4 장에서 제 3 장에서 조사한 방법들을 비교 분석한다. 마지막으로 제 5 장에서는 결론을 내린다.

2. Layer 기반 방법

Layer 기반 방법은 객체의 모든 속성들의 값을 이용하여 인덱스를 구성한다. 이 방법은 객체가 가지는 d 개의 속성 값들을 d-차원의 공간상의 점 객체로 맵핑하여, layer 들의 list(간단히, layer list)를 생성 한다 [3]. 여기서 i 번째 layer 는 top i 가 될 수 있는 객체들의 집합을 나타낸다. 따라서 최대 k 개의 layer 만 읽으면 top k 결과를 구할 수 있다.

Layer 기반 방법의 장점은 사용자 질의 벡터에 따른 질의 처리 성능 변화가 작으며, 모든 속성의 값들을 이용하여 layer 를 구성하기 때문에 여러 개의 속성을 고려하는 질의 처리 환경에서 유용하며 성능상의 손해가 작다. 그러나 layer 를 생성하는 비용이 크기 때문에 생성 및 갱신 비용이 크다는 단점이 있다.

3. Skyline 을 사용하는 Layer 기반 방법

본 장에서는 Skyline 을 사용하여 layer 을 생성하는 대표적인 방법인 BNL, SFS, NN, BBS 에 대해서 조사하고 각 방법에 대해 설명한다.

3.1 BNL

BNL (Block Nested Loops)은 전체 데이터를 순차적으로 읽으며 각 객체에 대해서 전체 객체와 비교하여 지배하는 객체가 없을 경우 skyline 결과에 포함되는 기법이다 [1]. BNL 은 먼저 파일로부터 데이터를 읽고 윈도우에 저장한다. 객체를 하나 읽었을 때, 윈도우에 있는 객체들과 비교하여 지배하는 객체가 있을 경우 삭제하고, 그렇지 않으면 윈도우에 삽입하고 윈도우에서 지배당하는 객체는 삭제하는 방식으로 진행된다 [5].

3.2 SFS

SFS (Sort-Filter-Skyline) [4]은 BNL 을 발전시킨 방법으로 객체의 엔트로피를 계산하여 엔트로피가 높은 포인트를 중심으로 정렬하여 skyline 을 계산하는 방법이다. 엔트로피 계산식은 식 (1)과 같다.

$$(1) \quad E(o) = \sum_{i=1}^{d} \ln(o'[i] + 1).$$

식 (1)의 결과 값을 통해 객체들이 정렬되며, 높은 엔트로피를 가지는 객체들이 다른 객체들을 지배할 가능성이 크다. 따라서 엔트로피 계산을 통해 가장 많은 객체를 지배하는 객체가 후보 리스트의 상위에 위치하게 하여 비교 횟수를 줄여 빠르게 Skyline 을 구할 수 있다. 하지만, 메인 메모리의 용량에 의존하는 문제점이 있다.

3.3 NN

NN (Nearest Neighbor) [2]은 R-tree 을 사용하여 데이터를 색인하고 최근접 질의(Nearest Neighbor Query)의 결과를 이용하여 검색 영역을 분할하는 방법이다. 첫 번째 최근접 질의의 결과 객체는 skyline 결과에 추가된다. 그리고 그 객체가 지배하는 영역은 다음 단계에서 제외되고, 분할된 영역에 대해서 다시 최근접 질의를 수행하여 분할된 영역에 객체가 없을 때까지 반복한다.

3.4 BBS

BBS (Branch and Bound Skyline) [1]은 NN 을 발전시킨 방법으로 중복된 노드의 접근을 제거하고, 입/출력 비용을 줄이기 위해 제안된 방법이다. 기존의 NN 기법과 같이 R-tree 로 색인하고, 노드의 중복 접근을 제거하기 위하여 노드 엔트리와 최단거리를 이용하여 힙(Heap)을 구성한다. BBS 는 처음 단계에서 루트노드의 모든 엔트리에 대한 midist 를 계산하여 힙을 구성한다. 힙의 상위 엔트리가 중간노드의 엔트리일 경우 skyline 결과 집합의 객체가 지배하지 않는 자식 엔트리만 힙에 추가하고, 단말노드의 엔트리일 경우 skyline 결과에 포함된 객체가 지배하지 않는 객체만 결과에 추가한다.

4. 비교 분석

BNL 은 skyline 을 계산하는 가장 간단한 접근 방법으로 각 객체에 대해서 전체 객체와 비교하는 방식이다. 그 결과 전체를 메인 메모리에 읽어서 비교하며 skyline 결과 집합에 객체의 삽입과 삭제가 반복하기에 계산 비용이 크다. 이러한 문제를 해결해 이동 가능성이 낮은 객체를 후보 리스트 상위에 위치하게 하여 계산 비용을 줄이는 SFS 가 연구되었다. 하지만 메인 메모리의 용량에 의존하는 문제가 여전히 존재한다.

NN 은 R-tree 를 사용하여 사전에 지배하는 영역에 포함되는 객체들을 읽지 않고 계산하게 되므로 메인 메모리의 의존하는 문제를 해결하였다. 그리고 BBS 가 연구되어 노드의 중복 접근이 감소되어 skyline 의 계산 시간이 감소하였다.

5. 결론

데이터가 이질적이며 방대해짐에 따라 효율적인 top k 질의 처리가 중요해 지고 있다. 본 논문에서는 top k 질의 처리를 위한 기존 연구 중 Layer 기반 방법에 관한 기존 연구에 대하여 조사하였다. 기존 연구로는 BNL, SFS, NN, BBS 를 소개하였으며 이들은 skyline 을 통해서 layer 를 생성하기 때문에 고차원 데이터에 대해서 계산 시간이 오래 걸리며 너무 많은 후보 객체가 결과로 나오기 때문에 layer 사이즈가 크다는 문제점이 있다. 이러한 문제점을 해결하기 위해 R-tree 를 이용하여 skyline 을 계산하기 전 후보 객체의 수를 줄이는 등 다양한 방법을 이용하여 효율적인 top k 질의 처리에 관한 연구가 필요하다.

이 논문은 0000 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2011-0002707)

참고문헌

[1] D. Papadias, Y. Tao, G. Fu and B Seeger. An optimal and progressive algorithm for skyline queries. Proceeding of the 2003 ACM SIGMOD international conference on Management of data, (2003)

[2] D. Kossmann, F. Ramsak and S. Rost. Shooting stars in the sky: an online algorithm for skyline queries. Proceeding of the 28th international conference on Very Large Data Bases (VLDB), (2002)

[3] J. Heo, K.Whang, M. Kim,Y. Kim and I. Song. The partitioned-layer index: Answering monotone top-k queries using the convex skyline and partitioning-merging technique. Information Science. Volume 179, Issue 19 (2009)

[4] J. Chomicki, P. Godfrey, J. Gryz and D. Liang. Skyline with Presorting. Proceedings of the 19th International Conference on Data Engineering (ICDE), (2003) March 5-8; Bangalore, India

[5] S. Borzsony, D. Kossmann and K. Stocker. The Skyline operator. Proceedings of the 17th International Conference on Data Engineering (ICDE), (2001) April 2-6; Heidelberg, Germany