Convex hull 을 사용하는 Top-k 질의처리 방법에 관한 분석

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

An Survey on Top-k Query Processing using Convex Hulls

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

요 약

최근 인터넷의 발달과 사용량의 증가로 데이터의 양이 급증함에 따라 대용량 데이터를 효율적으로 검색하는 top k 질의 처리가 중요시 되고 있다. Layer 기반 방법은 가장 잘 알려진 top k 질의처리 방법이며, 객체의 모든 속성의 값들을 이용하여 객체들을 layer 들의 리스트로 구성하는 방법이다. 본 논문에서는 그 중에서 convex hull 을 사용하여 layer list 를 생성하는 기존 연구를 조사하고 문제점을 파악한다.

1. 서론

최근 인터넷의 발달과 사용량의 증가로 데이터의 양이 급증함에 따라 대용량 데이터를 효율적으로 검색하는 top k 질의 처리가 중요시 되고 있다. top k 질의 처리는 릴레이션에서 가장 높은 (또는 가장 낮은) 스코어를 가지는 k 개의 튜플을 반환하는 방법[3]이며 Layer 기반 방법이 가장 잘 알려진 top k 질의처리 방법이다.

Convex hull 은 Layer 기반 방법 중 하나로 공간상에서 주어진 객체들을 감싸는 포인트들의 집합으로 이루어진 외곽선을 의미한다 [나 7]. Convex hull 은 컴퓨터 그래픽, 이미지 프로세싱, CAD/CAM, 데이터 마이닝 등 다양한 분야에서 사용되고 있다 [1]. 따라서 본 논문에서는 convex hull 을 사용하여 top k 질의를 하는 기존의 연구를 소개 및 분석하고 문제점을 파악한다.

본 논문의 구성은 다음과 같다. 제 2 장에서는 Layer 기반의 접근법 중 convex hull 을 사용하는 방법들을 설명하고, 제 3 장에서는 기존 방법의 문제점과 문제점을 해결하기 위한 방법에 대하여 간략하게 논의한다. 마지막으로 제 4 장에서는 결론을 내린다.

2. 관련연구

본 장에서는 top k 질의 처리 방법 중 Layer 기반 방법의 특징을 설명하고 convex hull 을 사용하는 방법에 대한 기존 연구들을 분석한다.

Layer 기반 방법은 객체의 모든 속성의 값들을 이용하여 객체들을 layer 들의 리스트로 구성하여 인덱스를 만드는 방법이다. 즉, layer 는 속성 값들의 모든 정보를 포함하고 있다. layer 들의 list 를 구성하면 적어도 k 개의 layer list 만 읽어서 top k 를 구할 수 있기에 top k 질의 처리에 효과적이다. 대표적인 convex hull 을 사용하는 layer 기반 방법 연구로 ONION[2]과 HL-index[4] 가 있다.

2.1 ONION

ONION[2]은 convex hull 의 정점들로 계층을 구성하는 방법으로 다차원 공간에서 객체들을 감싸는 convex hull 의 정점으로 이루어진 layer 를 구성하여 인덱스를 만든다. 즉, 전체 객체들의 집합에 대해 구한 convex hull 정점으로 첫 번째 layer 를 구성한다. 그 후에 남아있는 객체들의 집합에 대하여 convex hull 정점으로 이루어진 두 번째 layer 를 구하고, 같은 방법으로 계속해서 layer 를 구성한다. 반복하여 실행한 결과, 가장 바깥쪽 layer 는 기하학적으로 안쪽 layer 들을 둘러싸는 형태의 layer 들의 리스트가 만들어진다.

Chang 은 논문에서 optimally linearly ordered set 이라는 개념을 이용하여 convex hull 로 이루어진 layer 가 ONION 방법을 통해 바깥쪽부터 최대 k 개의 layer 만 읽고서 top k 질의에 답할 수 있음을 증명했다. 그러나 ONION 방법은 convex hull 을 구성 시에 새로 들어온 포인트가 convex hull 밖에 있는지 안에 있는지 domination relation 을 계산하는 비용이 커서 대용량 데이터를 처리하기에는 문제가 존재한다.

2.2 HL-index

HL-index[3]는 ONION 과 같은 방법으로 layer 리스트를 생성한 후, List 기반 방법인 TA [5]를 사용하여 각 layer 에 대해서 정렬한 리스트 생성하여 이를 색인으로 이용한다. 즉, convex hull 을 생성하고 layer 별로 TA 와 같은 방법으로 각 속성에 대하여 오름차순

으로 sorted list 를 구성한다. 이러한 과정을 반복하여 layer 와 리스트를 구성한 후 layer 에 속하는 일부 객체들만 읽어서 top k 결과를 구한다.

HL-index 는 데이터를 읽어들이는 비용은 작으나 고차원으로 갈수록 ONION 과 마찬가지로 convex hull 을 구성하는 비용이 크다는 문제가 여전히 존재한다.

3. 향후 top k 연구 방안

본 장에서는 2 장에서 설명한 기존 연구의 문제점에 대하여 향후 연구할 내용에 대해 간략하게 설명한다.

관련연구에서 살펴보았듯이 convex hull 은 고차원으로 갈수록 생성 비용이 높아진다는 단점이 있다. 그 이유는 먼저 convex hull 포인트 사이의 facet 정보에 의해 새로 들어온 포인트가 convex hull 밖에 있는지 안에 있는지 domination relation 을 일일이 계산하는 비용이 크다. 또, 이 계산을 위해 모든 포인트 사이의 facet 정보를 유지해야 하기 때문에 메모리를 많이 사용한다.

이에 대한 해결책으로 convex hull 에 비해 상대적으로 비용이 저렴한 skyline 을 이용하는 것을 제안한다. Skyline 은 다른 포인트로 인해서 지배될 수 없는 포인트 들을 연결한 선이며 convex hull 의 포인트를 모두 포함하게 된다. 따라서 convex hull 의 계산을 줄이기 위하여 skyline 계산을 우선으로 한다.

그러나 skyline 은 convex hull 에 비해 layer 사이즈가 크기 때문에 skyline 을 계산하기 전에 객체의 일부를 필터링하여 convex hull 사이즈로 줄이는 연구와 고차원으로 갈수록 대부분의 포인트가 skyline 포인트가 되는 것을 방지하기 위한 해결책이 필요하다.

4. 결론

본 논문에서는 기존의 Layer 기반 방법을 사용하여 top k 질의 처리를 하는 기존 연구를 소개 및 분석하였다. 기존 연구로는 ONION, HL-index 를 소개하였으며 이들은 convex hull 을 통해서 layer 를 생성하기 때문에 고차원 데이터에 대해서 layer 생성비용이 크다는 단점이 있다. 계산을 최소한으로 줄이기 위한 해결책으로 상대적으로 생성 비용이 작은 skyline 으로 layer 를 생성하는 방법을 제안한다. 그러나 skyline 은 convex hull 에 비해 읽어야 하는 객체의 수가 많으며 layer 사이즈가 크기 때문에 일부 객체를 계산 전에 제거하는 필터링에 관한 연구를 할 수 있을 것이다.

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

참고문헌

[1] C. Bohm and H.Kriegel. Determining the Convex Hull in Large Multidimensional Databases. Proceedings of the Data Warehousing and Knowledge Discovery (DaWaK), (2001) September; Munich, Germany

[2] Y. C. Chang, L. Bergman, V. Castelli, C. S. Li, M. L. Lo and J. R. Smith. The onion technique: indexing for linear optimization queries. Proceedings of the 2000 ACM SIGMOD international conference on Management of data, (2000)

[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] C.B. Barber, D. P. Dobkin and H. Huhdanpaa. The quickhull algorithm for convex hulls. ACM Transactionon Mathematical Software (TOMS). Volume 22, Issue 4 (1996)

[5] R. Fagin, A. Lotem, and M. Naor. Optimal Aggregation Algorithms for Middleware. Proceedings of the ACM Symposium on Principles of Database Systems(PODS), (2001) May, Santa Barbara, California