해시 색인 군집화 기반 스카이라인 질의

최종혁† · 나스리디노프 아지즈†
† 충북대학교 컴퓨터과학과

Clustered Hash Index-based Skyline Query

Jong-Hyeok Choi† · Aziz Nasridinov†
† Dept. of Computer Science, Chungbuk National University

요 약

스카이라인 질의는 지배라는 개념을 활용, 주어진 데이터로부터 데이터를 대표할 수 있는 데이터들을 탐색하기 때문에 사용자의 요청에 부합하는 최적의 결과를 탐색하거나 기업에서 의사결정을 이루기 위해 사용되는 등 넓은 활용을 보이고 있다. 하지만 스카이라인 질의는 데이터의 차원이 증가하는 경우 전체적인 성능의 감소와 함께 스카이라인으로 선택되는 데이터의 수가 급증하여 사용자에게 유용한 결과를 반환하지 못하게 된다. 이러한 문제를 해결하기 위해 최근에는 Top-k 질의 기반의 방식이나 군집화 기반의 기법을 적용한 방식의 스카이라인 질의들이 새롭게 제안되고 있지만 이들은 데이터의 편향이나 사용자로부터 입력된 k에 큰 영향을 받는 등 해당 질의 결과가 데이터들을 충분히 대표하거나 다양성을 만족시키지 못했다. 이러한 문제를 해결하기 위해 본 논문에서는 해시 색인 기법과 군집화 기법인 DBSCAN을 통해 주어진 데이터들을 충분히 대표함과 동시에 다양성을 만족할 수 있는 새로운 방식의 스카이라인인 *CHI-SQ*의 이론적 배경을 제안하고자 한다.

1. 서 론

스카이라인 질의는 다른 데이터에 의해 지배되지 않은 데이터들의 최소 집합을 탐색하는 방법이다[1-5]. 이때 지배되지 않는다는 것은 다른 데이터들과 비교하여 모든 차원에서 같은 값을 지니거나 최소 하나 이상의 차원에서 다른 데이터들보다 더욱 좋은 값을 갖는다는 것을 이야기한다[1-2]. 따라서 주어진 데이터로부터 스카이라인으로 탐색된 데이터는 주어진 데이터들을 대표할 수 있는 데이터들로 구성되었다고 할 수 있다. 이러한 특징으로 인해 스카이라인 질의는 사용자에 요청에 부합하는 최적의 결과를 탐색하거나 기업 등에서 다양한 기준을 바탕으로 최적의 의사 결정을 선정하기 위해 사용되는 등 다양한 분야에서 그 가치를 보이고 있다. 하지만 스카이라인 질의는 지배 유무의 판단이라는 단순함과 반대로 데이터의 차원이 증가하는 경우에 전체적인 성능과 함께 탐색된 스카이라인의 의미가 크게 감소하는 문제가 발생한다. 이는 차원의 수가 증가할수록 지배 유무를 확인하기 위한 비교 연산의 비용이 증가하며, 동시에 스카이라인으로 선택되는 데이터의 수가 매우 크게 증가하기 때문이다. 이러한 문제는 스카이라인 질의를 통해 얻어진 스카이라인이 주어진 데이터들을 충분히 대표하지 못하게 하며, 동시에 수많은 스카이라인으로부터 최적의 선택을 내리기 위해 사용자의 개입이 필요해지는 등 스카이라인 질의의 목적을 크게 희석시키게 된다.

이러한 문제로 인해 최근의 몇몇 스카이라인 질의 기법들은 단순히 스카이라인을 빠르게 탐색하기 위한 기존의 연구 방향에서 벗어나 사용자에게 보다 유의미하면서도 실제로 활용 가능한 수의 스카이라인을 전달하기 위해 다양한 방법들을 시도하였다. 이들 연구 중 가장 대표적인 연구는 Top-k 질의를 활용한 연구들로써, 해당 연구들은 사용자가 활용할 수 있는 k개의 결과만을 스카이라인으로부터 선택하여 전달하는 기법들이다[3-4]. 하지만 이들 기법들은 사용자에게 전달된 k개의 스카이라인 데이터가 특정 데이터들로 편향되는 등 주어진 데이터들을 충분히 대표할 수 없는 경우가 자주 발생했다.

이러한 문제를 해결하기 위해 최근에는 대표적인 군집화 기법인 k-평균 기법을 통해 스카이라인을 군집화하고 각 군집으로부터 해당 군집을 대표하는 스카이라인만을 선택하여 사용자에게 전달하는 기법이 연구되었다[5]. 이렇게 선택된 결과는 Top-k 기반의 기법들이 목표한 실제 활용 가능한 수의 스카이라인 결과라는 목적을 만족하며 동시에 스카이라인의 군집으로부터 스카이라인을 선택하기 때문에 보다 주어진 데이터들을 대표할 수 있는 결과가 사용자에게 전달되었다. 하지만 이러한 k-평균 기반의 질의는 k-평균이 갖는 기존의 문제점처럼 사용자로부터 선택된 k가 주어진 데이터들을 군집화하기에 적합치 않은 값인 경우, 군집의 다양성(diversity)을 충분히 반영하지 못하는 경우가 잦았으며 동시에 스카이라인을 탐색하기 위해 모든 데이터를 활용해야만 하기 때문에 탐색에 많은 시간을 요구하였다.

본 논문에서는 이러한 문제들을 해결하기 위해 격자 분할 기반의 해시 인덱스와 k-평균 기법의 문제점을 효과적으로 해결할 수 있는 대표적인 군집화 기법인 DBSCAN에 기초한 새로운 방식의 스카이라인 질의 기법인 *CHI-SQ(Clustered Hash Index-based Skyline Query)*을 제안한다. *CHI-SQ*는 격자 분할 기법과 이들을 효과적으로 관리하기 위한 해시 인덱스를 통해 데이터들을 관리하며 동시에 격자 분할된 공간이 갖는 특징을 통해 스카이라인이 발생할 수 없는 공간을 조기에 제거함으로써 탐색에 불필요한 데이터들을 조기에 제거한다[6]. 이후 분할된 영역과 실제 스카이라인을 활용하여 효과적으로 DBSCAN을 수행, 대표성을 띄면서도 동시에 다양한 형태의 군집을 모두 나타낼 수 있는 결과를 최종 결과로써 사용자에게 전달한다[7].

이러한 *CHI-SQ*를 제안하기 위한 본 논문의 구성은 다음과 같다. 제 2장에서는 k-평균 기반의 기존 연구를 소개한다. 제 3장에서는 *CHI-SQ*의 스카이라인 질의 처리를 제안한다. 제 4장에서는 결론을 내리며 본 논문을 마친다.

2. 관련 연구

스카이라인은 데이터의 차원이 증가할수록 스카이라인 질의를 통해 탐색되는 스카이라인의 수는 크게 증가하는 특징이 있다. 이는 차원이 증가할수록 데이터가 지배되지 않을 확률이 증가하기 때문으로, 이로 인해 고차원의 데이터로부터 탐색된 스카이라인은 주어진 데이터를 대표할 수 있으나 충분히 요약하고 있지는 못하기 때문에 결과 활용의 측면에 있어 사용자의 개입을 크게 요구하게 된다. 이러한 문제를 해결하기 위해 기존에 주로 사용된 방법들은 Top-k 기반의 방법들로써, 스카이라인 질의를 통해 얻어진 결과로부터 일정한 평가 기준에 부합하는 상위 k개의 결과를 사용자에게 전달하는 방식이다[3-4]. 하지만 이러한 방식은 종종 탐색된 스카이라인이 데이터 공간에서 편향되어 있거나 스카이라인 평가 기준이 바르지 못한 경우, 다양한 영역을 대표할 수 있는 스카이라인이 아닌 일부 영역만을 대표할 수 있는 스카이라인만을 결과로써 사용자에게 전달하는 등 스카이라인이 갖는 대표성을 크게 감소시켰으며 동시에 사용자에게 편향된 선택만을 제안하는 문제가 발생했다.

이러한 문제를 해결하기 위해 최근에는 군집화 기법 기반의 접근을 통해 탐색된 스카이라인을 군집화하고, 이들 군집으로부터 대표적인 스카이라인을 선택함으로써 스카이라인이 갖는 대표성을 유지하며 동시에 다양성을 보장할 수 있는 방법들이 연구되고 있다. 이들 연구들 중 가장 대표적인 연구는 k-평균 기반의 접근을 수행한 연구로써 해당 연구에서는 탐색된 스카이라인을 k-평균을 이용하여 k개의 군집으로 군집화하고 각각의 군집으로부터 가장 대표성을 띄는 스카이라인들을 선택하여 사용자에게 전달하였다[5]. 이러한 방법은 유사한 형태를 갖는 스카이라인들이 동일한 군집으로 분류하며 대표성을 유지할 수 있었으며 동시에 군집으로부터 고르게 스카이라인을 선택하여 사용자에게 전달하기 때문에 편향되지 않고 군집의 수만큼 다양한 결과를 사용자에게 제안할 수 있었기 때문에 사용자 측면에서 매우 유의미한 결과들을 전달할 수 있었다.

하지만 k-평균 기반의 접근법은 기존의 k-평균 기법이 갖고 있는 문제점을 그대로 계승하고 있기 때문에 사용자로부터 선택된 k가 주어진 데이터들을 군집화하기에 이상적이지 못한 값인 경우, 생성된 군집이 갖는 대표성이 크게 감소하며 동시에 생성된 k개만큼 생성된 결과 또한 사용자에게 다양한 결과를 전달하기에는 매우 불충분하게 된다.

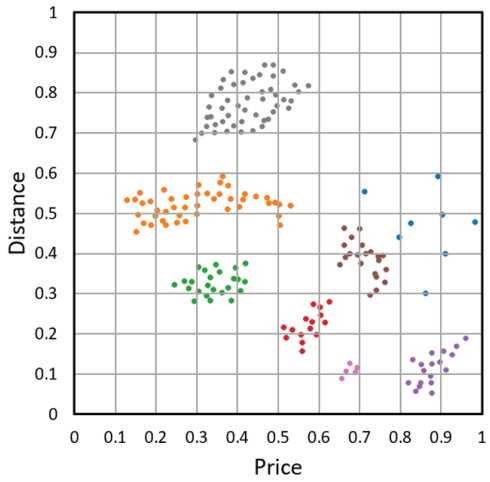
따라서 이러한 k로 인한 문제점을 해결할 수 있는 새로운 방법들의 연구가 크게 필요하다 할 수 있다.

1. 제안 기법

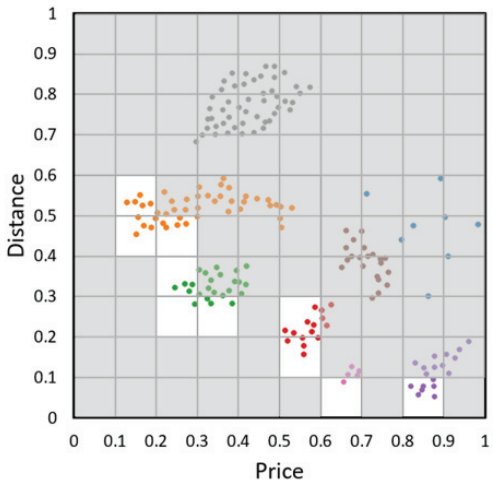
본 장에서는 *CHI-SQ(Clustered Hash Index-based Skyline Query)*의 이론적 배경과 함께 그림을 통해 *CHI-SQ*의 질의 처리 과정을 실제로 보인다.

*CHI-SQ*는 기존의 방법들이 스카이라인을 탐색하기 위해 주어진 모든 데이터를 비교해야만 하는 문제점을 해결하기 위해 주어진 데이터들을 해시 색인으로 관리하며 이후 질의가 요청되었을 때 해당 색인을 기반으로 스카이라인을 탐색한다. 이를 위해 *CHI-SQ*에서는 주어진 데이터들을 격자 분할 할 수 있는 해시 함수를 기반으로 [그림 1]과 같이 주어진 데이터들을 격자 분할한 후, 동일한 격자 영역에 속하는 데이터를 동일한 색인에 할당하여 관리한다.

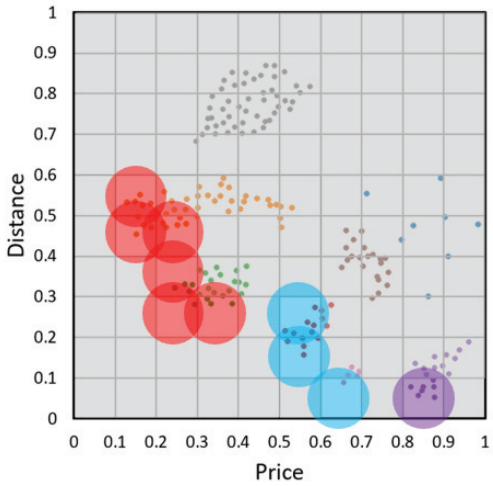
이러한 격자 형태의 해시 색인은 스카이라인을 탐색하는 과정에 있어 색인 사이의 비교만으로도 특정 색



[그림 1] 격자 분할된 데이터 공간



[그림 2] 데이터 공간 제거 단계 후의 모습

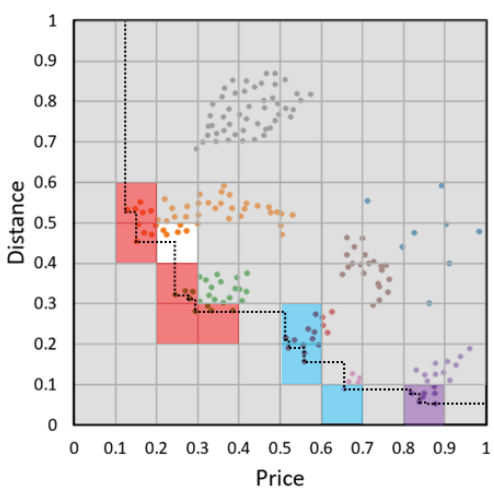


[그림 3] 근사 군집화 단계를 통해 얻어진 군집

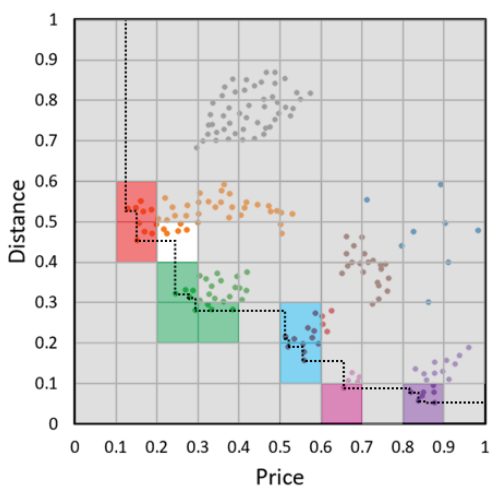
인이 지배 가능한 색인, 즉 격자 공간을 검출할 수 있도록 한다. 따라서 *CHI-SQ*의 해시 색인은 스카이라인이 될 수 없는 데이터들에 대한 불필요한 접근을 색인 사이의 비교를 통해 제거시키며 이러한 단계를 데이터 공간 제거 단계라 칭한다. 따라서 [그림 1]과 같은 데이터가 앞선 데이터 공간 제거 단계를 거치게 되면 [그림 2]와 같이 스카이라인이 절대 탐색될 수 없는 다수의 색인과 함께 해당 색인에 포함된 데이터들이 조기에 제거되게 된다.

이렇게 첫 번째 단계를 통해 제거되지 않은 색인들을 이용하여 *CHI-SQ*는 두 번째 단계인 근사 군집화(approximate clustering) 단계를 수행한다. 해당 단계에서는 해시 색인만을 이용하여 DBSCAN을 수행하며 이러한 근사 군집화 단계를 통해 얻어진 근사 군집은 스카이라인 탐색 결과를 활용하여 군집화를 진행하는 마지막 단계에서 군집화 연산에서 활용되어야할 데이터를 동일한 근사 군집의 데이터로 한정시킴으로써 실제 군집의 빠른 탐색이 가능토록 한다. [그림 3]은 이와 같은 근사 군집화 단계를 [그림 2]에 대해 수행한 결과로써 총 3개의 근사 군집이 얻어졌으며, 이후 마지막 군집화 단계에서는 근사 군집화 단계에서 얻어진 동일 군집의 데이터만을 고려하여 최종적인 군집을 탐색한다.

이어지는 세 번째 단계에서는 제거되지 않은 색인들에 속한 실제 데이터들을 이용하여 스카이라인을 탐색하는 스카이라인 탐색 단계를 수행한다. 이때 앞선 첫 번째 단계에서 제거되지 않은 일부의 해시 색인은 데이터 사이의 비교를 통해서 스카이라인 발생하지 않음



[그림 4] 스카이라인 탐색 단계의 수행 결과



[그림 5] 실제 군집화 단계의 수행 결과

이 최종적으로 확인되기도 한다. 따라서 스카이라인이 발생하지 않는 색인이 발생하는 경우 앞선 근사 군집에서 해당 색인을 제거한다. [그림 4]는 [그림 3]에 대해 실제 스카이라인을 탐색한 결과로써 기존에 적색으로 표기된 군집의 일부 색인에서 스카이라인이 발생하지 않음으로써 근사 군집의 일부 형태가 변화한 것을 알 수 있다.

이후 *CHI-SQ*는 최종적인 단계로써 탐색된 스카이라인 결과를 바탕으로 실제 군집을 탐색하는 실제 군집화(actual clustering) 단계를 수행한다. 해당 단계에서는 앞선 근사 군집화 단계와 마찬가지로 DBSCAN을 이용하여 군집화를 진행한다. 이때 DBSCAN의 경우 특정 데이터가 일정한 거리 안에 속한 데이터인지를 연산해야할 필요가 있다. 하지만 *CHI-SQ*는 앞선 근사 군집화 단계를 통해 연산에 대상이 되는 공간을 근사적으로 한정하였기 때문에 해당 단계에서는 동일한 근사 군집에 속한 데이터만을 확인함으로써 불필요한 연산을 최소화한다. 따라서 적색으로 구성된 영역의 경우 적색에 속한 데이터만을, 청색에 속한 영역의 데이터는 청색에 속한 데이터만을 고려하여 실제적인 군집을 탐색한다. 이러한 과정에서 실제 데이터들 사이의 군집이 발생하지 않는 경우, 이들을 새로운 군집으로써 명명하여 지속적인 탐색을 진행한다. 따라서 기존 [그림 3]에서 적색으로 근사 군집화 되었던 영역의 일부는 녹색의 영역으로, 청색으로 근사 군집화 되었던 영역의 일부는 분홍색의 영역으로써 실제 군집을 이루게 되며 실제 군집화를 통해 얻어진 군집의 결과는 [그림 5]와 같이 총 5개의 실제 군집이 얻어지게 되며 이 군집들을 바탕으로 사용자에게 질의 결과를 순차적으로 반환한다.

이러한 *CHI-SQ*는 스카이라인을 탐색하기 위해서 모든 데이터를 비교해야한다는 기존의 문제를 해시 색인을 통해 해결함과 동시에 두 단계에 거친 군집화 단계를 통해 군집화의 성능을 개선할 수 있으며 k-평균의 k와 같이 사용자로부터 입력된 부정확한 수의 군집이 아닌 최적의 군집들을 탐색하게 함으로써 탐색 결과의 다양성을 충분히 확보할 수 있으며 이들 군집으로부터 선택된 스카이라인들이 동일한 군집을 충분히 대표하는 것을 가능하게 한다.

2. 결론

본 논문에서는 스카이라인 질의를 요청한 사용자에게 보다 유용하면서도 보다 다양성을 지닌 결과를 빠르게 전달할 수 있도록 새로운 스카이라인 질의 기법인 *CHI-SQ*의 이론적 배경을 제안하였다. 이러한 *CHI-SQ*는 기존의 Top-k나 k-평균 군집을 활용ㅎ한 연구들이 갖는 스카이라인의 편향으로 인한 대표성의 감소나 다양성 감소의 문제를 효과적으로 해결할 수 있음을 이론적 제안을 통해 선보였으며, 동시에 이들 결과를 보다 빠르게 탐색하여 사용자에게 전달할 수 있는 처리 방법 또한 동시에 제안하였다. 차후 지속될 연구에서는 이와 같은 *CHI-SQ*의 이론적 배경을 실제 구현화하고 기존 기법들과 비교 실험을 수행하는 등 실증적인 방법들을 통해 해당 기법의 우수성을 비교 판단할 예정이다.

Acknowledgement

이 논문은 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2017R1D1A3B03035729).

참 고 문 헌

[1] Borzsony, S., Kossmann, D., & Stocker, K. (2001). The skyline operator. In *Proceedings of the International Conference on Data Engineering*, 421-430.

[2] Afrati, F. N., Koutris, P., Suciu, D., & Ullman, J. D. (2015). Parallel skyline queries. *Theory of Computing Systems, 57*(4), 1008-1037.

[3] Lin, X., Yuan, Y., Zhang, Q., & Zhang, Y. (2007). Selecting stars: The k most representative skyline operator. In *Proceedings of the International Conference on Data Engineering*, 86-95.

[4] Ilyas, I. F., Beskales, G., & Soliman, M. A. (2008). A survey of top-k query processing techniques in relational database systems. *ACM Computing Surveys*, *40*(4), 11.

[5] Huang, Z., Xiang, Y., Zhang, B., & Liu, X. (2011). A clustering based approach for skyline diversity. *Expert Systems with Applications*, *38*(7), 7984-7993.

[6] 최종혁・류관희・나스리디노프 아지즈 (2017). 다차원 데이터 및 동적 이용자 선호도를 위한 색인 구조의 연구. **예술인문사회융합멀티미디어논문지**, **7**(7), 925-934.

[7] Gan, J., & Tao, Y. (2015). DBSCAN revisited: mis-claim, un-fixability, and approximation. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 519-530.