다중 릴레이션에서의 Skyline 연산 방법 연구

임선영*, 박은영**, 박영호*
*숙명여자대학교 멀티미디어과학과
**협성대학교 시각디자인학과
e-mail: sunnyihm@sm.ac.kr, parkey@uhs.ac.kr, yhpark@sm.ac.kr

A Study on Skyline Operators over Multi-Relation

Sun-Young Ihm*, Eun-Young Park**, Young-Ho Park*
*Dept of Multimedia Science, Sookmyung Women's University
**Dept of Visual Design, HyupSung University

요 약

인터넷이 발달함에 따라 데이터가 방대해지고 더욱 이질적인 특성을 가지게 되었다. 따라서 이질적이고 대용량 데이터에서의 빠른 검색이 중요시되고 있다. 또한, 다중 릴레이션 환경에서의 사용자 질의가 증가함에 따라 다중 릴레이션에서의 Skyline 연산의 필요성이 증가하고 있다. 본 논문에서는 다중 릴레이션 환경에서의 Skyline 연산의 기존 연구들을 소개하며 문제점을 분석하고 해결책을 간략하게 제시한다.

1. 서론

최근 인터넷의 발달로 데이터가 이질적이고 방대해짐에 따라 사용자가 원하는 결과를 이질적이고 대용량의 데이터에서 빠르게 검색하는 질의 처리의 중요성이 높아지고 있다. Borzsonyi가 [5]에서 처음 제안한 Skyline 연산은 여러 속성을 가진 대용량의 데이터에서 사용자가 원하는 결과를 빠르게 검색하도록 도와주며, 이후로 Skyline 연산과 변형된 Skyline의 연구 [2, 3, 4, 5] 등에서 활발히 이루어져 왔다.

많은 데이터베이스 환경에서 사용자들이 다중 릴레이션 테이블에서의 검색의 필요성이 높아지고 있기 때문에 최근에는 다중 릴레이션에서의 Skyline 연산이 중요해지고 있다. 예를 들어, 데이터베이스에 학생 정보가 담긴 'Student' 테이블과 영어 성적 정보가 담긴 'English' 테이블과 수학 성적 정보가 담긴 'Math' 테이블이 있다고 가정 한다. 그림 1은 각 다중 릴레이션 테이블의 예를 나타낸 것이다. 'Student' 테이블은 학생들의 학번인 Id와 이름인 Name을 속성으로 가지고 있고, 'English'와 'Math' 테이블은 학번인 Sid와 점수인 Score를 속성으로 가지고 있다.

Students

Id	Name
12001	Alice
12002	Bella
12003	Dona
12004	James
12005	Thomas

English

Sid	Score
12001	90
12002	85
12003	70
12004	95
12005	90

Math

Sid	Score
12001	70
12002	75
12003	85
12004	90
12005	80

(그림 1) 다중 릴레이션 테이블의 예

이 때 영어와 수학 성적이 가장 높은 성적을 받은 학생을 검색하는 간단한 SQL 질의 예는 다음과 같다.

```
SELECT S.Id, S.Name, E.Score, M.Score
FROM Student S, English E, Math M
WHERE S.Id = E.Sid = M.Sid
SKYLINE OF E.Score MAX, M.Score MAX
```

이와 같이 다중 릴레이션 테이블에서 각 테이블의 정보를 함께 검색하는 경우에는 단일 릴레이션 테이블에서의 Skyline 연산과는 조금 다른 Skyline 연산이 필요하다. 본 논문에서는 다중 릴레이션 테이블에서의 Skyline 연산을 하는 기존의 연구들을 소개하고 문제점을 분석한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 대표 연구들을 소개한다. 3장에서는 문제점을 비교 분석한 후, 해결하는 방안에 대하여 간략하게 설명한다. 마지막으로 4장에서 결론을 맺는다.

2. 기존 연구

본 장에서는 다중 릴레이션 환경에서의 Skyline 연산에 대한 연구들을 소개하고 문제점을 도출한다. 다중 릴레이션 테이블에서 Skyline 연산을 하는 방법 중 가장 단순한 방법은 모든 조인 튜플들을 계산한 후에 Skyline을 구하는 방법이다. 이 방법은 Skyline에는 오직 일부 튜플들만 포함되지만, 모든 튜플들을 조인하기 때문에 불필요한 계산이 많다. 따라서, 다중 릴레이션 환경에서의 Skyline 연산에서는 불필요한 계산을 줄이기 위하여 결과에 속하지 않는 튜플을 미리 배제하는 것이 중요하다.

대표적인 연구로는 Jin이 제안한 알고리즘 [7]과 [8] 그

리고 Raghavan이 제안한 ProgX [6], 마지막으로 Vlachou가 제안한 SFSJ 알고리즘 [1]이 있다.

2.1 다중 릴레이션에서의 Skyline 연산 알고리즘

Jin은 [7]에서 다중 릴레이션에서의 Skyline 연산 알고리즘을 제안하였는데, 이 방법은 Skyline 계산 과정과 sort-merge 조인 알고리즘을 결합한 방법으로 다중 릴레이션 테이블을 조인하고 정렬하여 Skyline을 구한다.

Jin은 [8]에서 [7]을 개선하는 알고리즘을 제안하였는데, 이 방법은 nested-loop 알고리즘과 기존의 sort-merge 조인 알고리즘을 함께 이용하여 조인과 Skyline 연산을 하는 방법이다. [7]에 비하여 질의 처리 속도는 향상되었지만, 정확한 결과를 구하기까지 여전히 각 테이블에 불필요한 접근이 많다는 단점이 있다.

2.2 ProgXe의 Skyline 연산 알고리즘

ProgXe [6]은 다중 릴레이션 테이블을 사용하는 데이터베이스 환경에서의 검색을 지원하는 프레임웍이다. ProgXe의 Skyline 연산 알고리즘은 다차원의 grid 접근 방법을 이용하여 입력된 다중 릴레이션을 분할한다.

2.3 SFSJ 알고리즘

SFSJ 알고리즘 [1]은 기존의 방법들을 개선한 알고리즘으로 입력된 데이터 튜플들의 일부만 읽으면서 정확한 Skyline 연산을 하는 알고리즘이다. 빠른 종료를 할 수 있는 알고리즘을 제안하여 조인 연산을 통해 다중 릴레이션에서 Skyline 연산을 수행한다. 전체 튜플에 접근하지 않고 일부만 읽은 후에 배제되는 튜플을 계산하지 않기 때문에 질의 처리 속도가 향상되었다.

3. 문제점 비교 분석

본 장에서는 2장에서 살펴 본 다중 릴레이션 환경에서의 Skyline 연산 방법들의 문제점을 분석하고 해결 방안을 간략하게 소개한다.

Jin이 [7]과 [8]에서 제안한 Skyline 연산은 다중 릴레이션 테이블을 서로 조인하여 결과를 리턴하지만 정확한 Skyline을 구하기 위하여 각 테이블에 불필요하게 여러 번 접근해야 한다는 단점이 있다. ProgXe에서 사용하는 Skyline 연산은 오직 테이블을 기반으로 한 조인 연산에만 적용된다는 제한이 있고, 또 결과를 얻을 때 까지 빠른 종료를 지원하지 않는다. SFSJ 알고리즘은 정확한 Skyline을 구하기까지 여전히 결과에 포함되지 않는 튜플에 불필요한 접근을 한다는 단점이 있다.

이에 대한 해결 방법으로는 각 튜플 간의 배제 관계가 분석되어야 한다. 그리고 더 많은 튜플을 배제하도록 하는 연구가 필요하다.

4. 결론 및 기대효과

본 논문에서는 다중 릴레이션 환경에서의 Skyline 연산 방법들을 소개하고 문제점을 분석하였다. 단일 릴레이션에서의 Skyline 연산과 달리 다중 릴레이션에서는 조인 연산이 필요하기 때문에 많은 튜플들을 읽어야 한다. 따라서 입력된 데이터 튜플을 전부를 읽지 않고도 정확한 Skyline을 구하는 것이 중요하다. 이를 위하여 결과에 포함되지 않는 튜플들에 대한 불필요한 접근을 줄이고 빨리 배제하는 연구가 필요하다.

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2012-003797)

참고문헌

[1] A. Vlachou, C. Doulkeridis, N. Polyzotis "Skyline Query Processing over Joins" Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, pp.73-84, 2011

[2] C. Y. Chan, H. V. Jagadish, K. L. Tan, A. K. H. Tung, Z. Zhang "Finding k-Dominant Skylines in High Dimensional Space" The ACM SIGMOD International Conference on Management of Data, 2006

[3] M. D. Morse, J. M. Patel, H. V. Jagadish "Efficient Skyline Computation over Low-Cardinality Domains" The International Conference on Very Large Data Bases (VLDB), 2007

[4] P. Godfrey, R. Shipley, J. Gryz "Maximal Vector Computation in Large Data Sets" The International Conference on Very Large Data Bases (VLDB), 2005

[5] S. Borzsonyi, D. Kossmann, K. Stocker "The Skyline Operator" The IEEE International Conference on Data Engineering, 2001

[6] V. Raghavan, E. Rundensteiner "Progressive Result Generation for Multi-Criteria Decision Support Queries" IEEE 26th International Conference on Data Engineering (ICDE), 2010

[7] W. Jin, M. Ester, Z. Hu, J. Han "The Multi-Relational Skyline Operator" IEEE 23th International Conference on Data Engineering (ICDE), 2007

[8] W. Jin, M. Morse, J. Patel, M. Ester, Z. Hu "Evaluating Skylines in the Presence of Equijoins" IEEE 26th International Conference on Data Engineering (ICDE), 2010