공간 데이터 분포와 질의 크기를 고려한 선택률 추정

문현수 이미란[O] 황환규
강원대학교 컴퓨터·정보통신공학과
leo@alink.co.kr kiki@dblab1.kangwon.ac.kr wkwhang@cc.kangwon.ac.kr

Selectivity Estimation for Spacial Data Distribution and Query Size

Hyun-Soo Moon Mi-ran Lee[O] Whan-Kyu Whang
Dept. of Computer, Information & Communication Eng.
Kangwon Chuncheon, Kangwon National Univ.

요 약

공간 데이터베이스에서의 질의에 대한 선택률 추정에 대해서는 많은 연구가 있었지만 공간 데이터베이스에서의 공간 질의에 대한 선택률 추정이 매우 중요함에도 불구하고 이에 대한 연구는 아직 미흡한 상태이다. 이 논문에서는 공간 검색 조건의 정확한 선택률 추정을 위해 공간 데이터 분포를 통계 데이터로 저장하고 이를 이용하여 선택률을 추정하는 방법을 제안하고 구현하였다. 공간 질의에 대한 선택률 추정을 위해서 기존의 통계 데이터를 작성하는 방법으로 균등 분할 방법과 비균등 분할 방법이 사용되고 있지만 보다 정확한 선택률을 추정하기 위해서 본 논문에서는 새로운 통계 데이터 작성 방법인 크기별 분할 방법을 제안하였다. 각 방법의 성능은 다양한 파라미터에 대한 선택률 오차를 산출하여 평가하였다.

1. 서론

지리 정보 시스템(Geographic Information Systems: GISs)은 일반적으로 점(points), 선(lines), 다각형(polygons), 표면(surface)과 같은 공간 객체를 가지고 있으며, 이러한 각 객체는 공간 데이터베이스에 저장되어 관리된다[Ege88].

데이터베이스에 저장되어 있는 공간 데이터 중에서 질의 검색 조건을 만족하는 데이터의 수는 질의 조건에 따라 제각기 다르다. 질의 처리를 위한 연산 비용은 검색 조건을 만족하는 데이터의 개수에 의해 결정된다. 일반적으로 전체 데이터 중 검색 조건을 만족하는 데이터 수의 비율을 그 검색 조건의 선택률이라 정의한다. 질의에 대한 선택률 추정은 가장 효율적인 질의 수행을 위해 질의 최적화에 사용되기도 하며[SAC+79] 실제 질의가 수행되기 전에 질의의 실행 시간을 사용자에게 알려주기 위해 데이터베이스 시스템에 의해서 사용되기도 한다. 질의 결과의 크기를 계산하기 위해 전체 질의를 실행하는 것은 비실용적이므로 대부분의 상용 시스템은 데이터를 기반으로 대략적인 추정을 계산하기 위해 여러 형태의 통계를 사용하며, 이러한 통계를 기반으로 결과의 크기를 추정하게 된다.

공간 데이터에 대한 선택률 추정의 문제점은 데이터베이스 논문에서 광범위하게 연구되어 왔던 관계 데이터베이스 선택률 추정 문제와는 많은 차이가 있다. 대부분의 이전 작업들은 하나의 애트리뷰트에 대한 분포를 근사적으로 추정하는 데에 초점을 맞춰 왔다. 다차원 데이터에 대한 연구[PI97][MPS99] 역시 공간에서의 점에 대한 분포의 근사화에 초점을 두어 왔다. 이러한 기법들은 심하게 치우친 분포를 지니는 데이터에 잘 적용될 수 있다.

본 논문에서는 공간 데이터의 영역 선택 속성(predicate)에 대한 선택률 추정의 문제를 다루고자 하며, 특히 여기서는 이차원 사각형 데이터에 초점을 두게 된다. 이는 공간 데이터베이스 시스템에서 공간 데이터를 근사화시키기 위해 최소 경계 사각형(MBR)을 사용하며 질의 처리 수행 역시 MBR을 사용하는 것이 가장 일반적인 방법이기 때문이다. 본 논문에서는 이러한 사각형 데이터에 초점을 두고 있지만 제안된 기법은 점과 선 데이터에 역시 적용 가능하다.

본 논문의 구성은 다음과 같다. 2절에서는 본 논문에서 제안한 크기별 분할 방법의 통계 데이터 작성 및 관리에 초점을 두며, 3절에서는 선택률 추정 공식을 제시하고 이를 이용하여 선택률을 계산해 주는 방법을 다루게 된다. 4절에서는 제안한 크기별 분할 방법을 실험에 적용하여 선택률에 대한 성능을 평가하게 된다. 5절에서는 결론을 내리고 향후 연구과제를 제시하며, 마지막 6절에서는 참고문헌을 제시한다.

2. 통계 데이터 작성 및 관리

본 논문에서 제안한 크기별 분할 방법은 공간 데이터의 분포와 질의의 크기를 모두 고려한 방법으로 대부분의 형태의 질의에 대해 적은 오차율로 선택률을 추정할 수 있다. 이 방법은 전체 데이터를 스캔하여 전체 공간을 가장 작은 데이터가 포함될 수 있을 때까지 분할의 수를 증가시키면서 균등 분할을 반복하여 각 분할 결과를 파일로 저장하는 것이다. 즉, 전체 데이터 집합을 하나의 단계적 파일로 표현될 수 있도록 하는 것이다. 공간 데이터 집합 내에 데이터 객체가 10개 있을 때의 단계적 파일 분할 방법이 그림 1에 나타나 있다. 먼저 공간을 4개로 분할하여 2개 이상의 영역에 속하는 데이터가 있는 지를 살핀다. 이때 공간 데이터 2, 7, 8, 10이 2개 이상의 영역에

속하므로 분할하기 이전의 단계 File 1을 형성하게 된다. 전체 데이터는 10개이므로 File1의 통계 데이터는 10이 된다. 이렇게 모든 데이터가 둘 이상의 영역에 속하지 않을 때까지 통계 데이터를 단계적으로 작성하게 되며, 이 예제에서는 총 4개의 파일이 생성되었으며, 이 4개의 통계 데이터로 선택률 추정을 하게 된다.

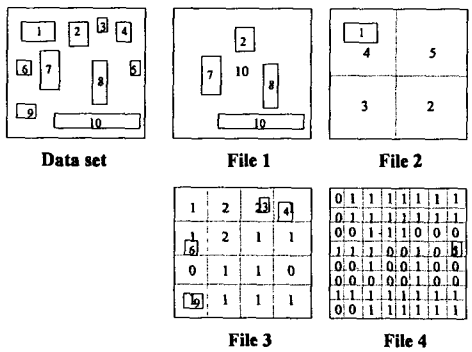


그림 1. 크기별 분할 방법

3. 통계 데이터를 이용한 선택률 추정

본 논문에서 사용하는 미리 작성된 통계 데이터에 질의를 수행시킴으로써 선택률을 추정하게 된다. 선택률을 추정하는 범위는 질의가 속하는 통계 데이터의 분할된 영역이며, 질의가 속하는 각 영역에 대해, 질의가 속한 영역의 카운트에 질의가 그 영역을 차지하는 비율을 곱한 결과를 합한 결과가 선택률 추정 값이 된다. 본 논문에서 사용되는 선택률 추정 공식은 다음과 같다[KJD97].

선택률 추정 공식: $V(Q) = \{P(1) * C(1)\} + \cdots + \{P(N) * C(N)\}$

Q: 공간 데이터에 대한 질의
V: 질의 Q에 대한 선택률 값
N: 통계 데이터에서 Q가 속한 영역의 수
P: 질의가 통계 데이터 분할 영역 내에서 차지하는 비율($0 \le P \le 1$)
C: 통계 데이터 분할 영역의 카운트

제안된 크기별 분할 방법의 선택률 계산 알고리즘을 간단하게 정리하면 다음과 같다.

질의 Q와 overlap 관계에 있는 데이터에 대한 선택률 추정

Step 1: 질의 영역에 해당되는 부분의 데이터 존재 유무 확인을 위해 가장 마지막에 생성된 통계 데이터 파일 조사
- 질의 영역이 속하는 분할 영역의 카운트가 모두 0이면 선택률은 0
- 그렇지 않으면 Step 2

Step 2: 통계 데이터를 선택하여 선택률 추정 공식을 적용하여 선택률 계산
- 질의 영역의 면적을 구함
- 각 통계 데이터의 분할 영역 중 질의 영역의 면적과 가장 유사한 면적을 지니는 파일을 선택하여 질의 적용한 후 선택률 계산

그림 2는 그림 1의 데이터 집합과 여기서 생성된 통계 데이터를 사용하여 질의가 크기별로 분할된 통계 데이터에 질의 Q와 overlap하는 공간 데이터의 수를 추정하는 경우에 대한 예이다.

그림 2의 경우, 먼저 질의 Q를 마지막에 생성된 통계 데이터 File 4에 수행시켜 본다. 질의 영역이 속한 분할 영역의 카운트가 0이 아니므로 질의 Q의 면적을 구하게 된다. 질의 Q의 면적은 이미 구성된 통계 데이터 파일 중 File 2의 통계 데이터의 한 영역의 면적과 같으므로 File 2가 질의가 수행될 통계 데이터로 선택되며, 선택률 추정 공식을 적용한 선택률 계산 결과는 $V(Q) = 1*4 = 4$이다.

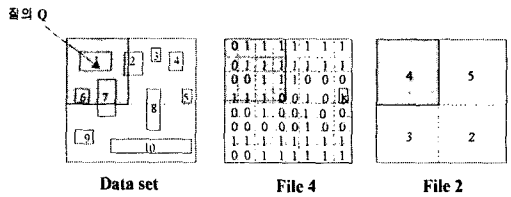


그림 2. 크기별 분할 방법에의 영역 질의 수행

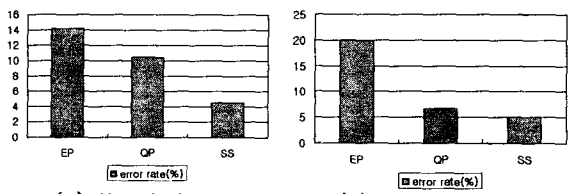
4. 실험

공간 질의는 영역 질의를 사용하고 있으며, 공간 관계 overlap을 적용시켜 실험하였다. 실험 방법은 먼저 주어진 데이터에 제안한 분할 방법을 적용하여 통계 데이터를 작성한 후, 주어진 질의에 대한 선택률 값을 추정하게 된다. 먼저, 작성된 통계 데이터에 질의를 적용시킨 후 선택률 공식을 사용하여 선택률을 계산한다. 그리고 실제 공간 관계 overlap을 만족하는 데이터의 수를 구하여 각 통계 데이터로부터 계산된 선택률의 오차율을 구한다. 본 실험에서는 최종적으로 구해진 오차율을 성능 평가의 척도로 삼았으며, 균등 분할 방법(EP)과 비균등 분할 방법(QP)를 실험 대상으로 하여 본 논문에서 제안한 크기별 분할 방법(SS)과 성능을 비교하였다. 선택률에 대한 오차율을 구하기 위해 다음 공식을 사용한다.

$$\text{오차율} = \frac{|\text{계산된선택률} - \text{실제선택률}|}{\text{실제선택률}} * 100$$

1) 실제 데이터 적용

실제 데이터는 long beach data와 mongocounty data를 사용하였으며, 실험 결과는 그림 3과 같다.

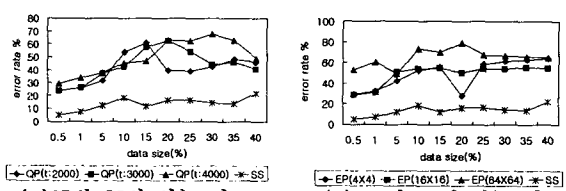


(a) lbeach data (b) mgcounty data

그림 3. 실제 데이터에 대한 성능 비교

2) 데이터 크기에 의한 성능

공간 데이터의 크기를 두 부류로 나누어 총 40,000개의 데이터를 전체 데이터 영역에 대해 작은 데이터 크기는 0.5~15%, 큰 데이터 크기는 20~40%로 전체 데이터 영역에 거의 균일하게 분포되도록 작성하였다. 실험 결과는 그림 4와 같다.

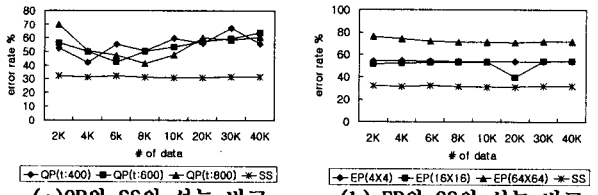


(a)QP와 SS의 성능 비교 (b) EP와 SS의 성능 비교

그림 4. 데이터 크기에 의한 성능 비교

3) 데이터 개수에 의한 성능

데이터의 개수를 작은 데이터 수(2,000~8,000)부터 큰 데이터 수(10,000~40,000)까지 분류하여 성능을 비교하였으며 데이터는 전체 데이터 영역에 거의 균일하게 분포되도록 작성하였다. 실험 결과는 그림 5와 같다.

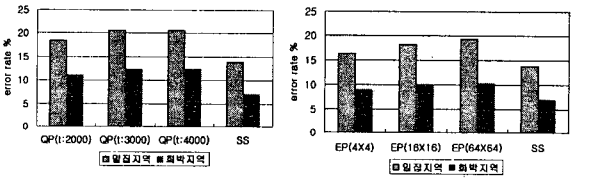


(a)QP와 SS의 성능 비교 (b) EP와 SS의 성능 비교

그림 5. 데이터 수에 의한 성능 비교

4) 데이터 분포에 의한 성능

여기에서는 lbeach data 영역을 데이터 밀집 지역과 데이터 희박 지역으로 나누어 각 부분에 대한 선택률의 오차를 평가하였다. 실험 결과는 그림 6과 같다. 각 그림의 왼쪽 막대는 밀집 지역, 오른쪽 막대는 희박 지역에 대한 오차율을 나타내는 것이다.

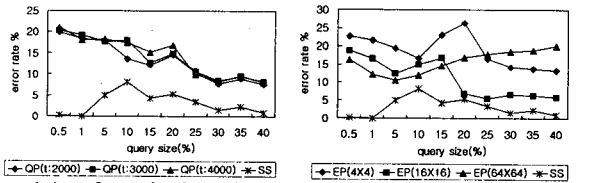


(a)QP와 SS의 성능 비교 (b) EP와 SS의 성능 비교

그림 6. 데이터 분포에 의한 성능 비교

5) 질의 크기에 의한 성능

공간 데이터에 대한 질의는 점 질의와 영역 질의가 있을 수 있는데, 본 논문에서는 영역 질의를 사용하여 질의를 처리하고 이에 대한 선택률을 추정하였다. 영역 질의의 크기를 두 부류로 나누어 전체 데이터 영역에 대해 작은 질의 크기는 0.5~15%, 큰 질의 크기는 20~40%로 하여 실험하였다. 실험에 쓰인 데이터는 실제 데이터인 lbeach data이며 실험 결과는 그림 7과 같다.



(a)QP와 SS의 성능 비교 (b) EP와 SS의 성능 비교

그림 7. 질의 크기에 의한 성능 비교

5. 결론

관계 데이터베이스에서 질의 수행 결과를 위한 선택률 추정이 중요한 것과 같이 공간 데이터베이스에서 역시 선택률 추정은 중요한 문제로 여겨지고 있다. 선택률을 추정의 대상이 되는 데이터는 하나의 애트리뷰트에 기반하는 것부터 다차원 데이터에 이르기까지 다양하지만 기존의 연구는 하나의 애트리뷰트에 대한 분포를 근사적으로 추정하는 데에 초점을 두어 왔으므로, 다차원 공간 데이터에 대해 오차율을 최소로 하는 새로운 선택률 추정 기법이 필요하다.

공간 데이터베이스에 대한 선택률 추정은 전통적인 선택률 추정과 두 가지 중요한 측면에서 차이가 있는데, 하나는 각 공간 엔티티의 모양과 크기가 다르다는 것이며, 다른 하나는 입력 도메인의 분포 빈도는 공간 데이터에 대해서는 다양하지 않은 반면 값은 공간상에 비균일하게 분포되어 있다는 것이다. 그러므로 공간 데이터의 근사치를 추정하는 문제는 값의 분포를 정확하게 근사시킬 수 있는 기법을 요구하게 된다.

이상의 이유로 인해 본 논문에서는 공간 데이터의 점과 영역 선택 속성에 대한 선택률 추정의 문제를 이차원 사각형 데이터에 초점을 두어 다루었으며 기존의 균등 분할 방법과 비균등 분할 방법으로 통계 데이터를 사용하여 선택률을 추정한 방법의 단점을 극복할 수 있는 새로운 통계 데이터 작성 방법인 크기별 분할 방법을 제안하였다. 새로 제안된 크기별 분할 방법은 데이터의 크기별로 통계 데이터를 여러 단계의 파일로 작성하였으며 공간 데이터에 수행되는 질의에 대해 필터링을 적용하고, 질의의 크기를 고려하여 여러 통계 데이터 파일 중 하나를 선택하는 방법을 사용하였다. 이렇게 함으로써 선택률 계산 결과의 오차율을 줄이고 선택률 계산 수행 시간을 단축시킬 수 있도록 하였다.

본 논문에서는 공간 데이터의 분포 상태를 통계 데이터로 저장하고 이를 이용하여 선택률을 계산하는 방법을 제안한 후 실험적으로 성능을 평가하였다. 통계 데이터 작성 방법 중 제안한 크기별 분할 방법을 사용한 것이 가장 좋은 성능을 보였다. 통계 데이터를 이용한 선택률 계산 방법의 성능 평가는 다음과 같다. 데이터가 희박한 곳보다 데이터가 밀집되어 있는 곳에서 선택률의 오차율이 증가하며, 데이터의 크기는 성능에 그다지 큰 영향을 미치지 않는다. 그리고 질의 크기가 중간일 경우 가장 낮은 성능을 보였다.

향후 연구 과제는 통계 데이터 작성시 데이터가 위치한 분할 영역의 카운트를 증가시키는데 있어서 보다 오차율을 줄일 수 있는 방법을 적용시킬 필요성이 있으므로 이에 대해 계속 연구할 것이다.

6. 참고 문헌

[ACH99] S. Acharya, V. Poosala, S. Ramaswamy. Selectivity Estimation in Spatial Databases.

[BF95] A. Belussi and C. Faloutsos. Estimating the selectivity of spatial queries using the correlation's fractal dimension. In Proceedings of the 21st International conference on Very Large Data Bases, Zurich, 1995.

[Ege88] M. J. Egenhofer, A. Frank. Towards a spatial query language: User Interface consideration, International Conference on Very Large Data Bases, pages 124-133, 1988.

[KJD97] J. D. Kim, J. B. Seon, B. H. Hong, J. S. Kim. Estimation Methods of Selectivity for Spatial Query Optimization. 1997.

[MPS99] S. Muthukrishnan, Viswanath Poosala, and Torsten Suel. On rectangular partitionings in two dimensions: Algorithms, complexity, and applications. 7th International Conference on Database Theory, January, 1999

[PI97] Viswanath Poosala and Yannis Ioannidis. Selectivity estimation without the attribute value independence assumption. Proc. of the 23rd Int. Conf. on Very Large Data Bases, August 1997.

[SAC+79] P. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, and T. G. Price. Access Path Selection in a Relational Databases Management System. In Proceedings of the 1979 ACM SIGMOD International Conference on management of Data, pages 23-34, Boston, Massachusetts, June 1979.