최대 면적 차이 분할 방법을 이용한 선택률 추정

이 미 란[0], 황 환 규
강원대학교 컴퓨터 정보통신공학과
miran23@hanmail.net, wkwhang@cc.kangwon.ac.kr

Selectivity Estimation using Maximum Area Difference

Mi-Ran Lee[0], Whan-Kyu Whang
Dept. of Information & Telecommunication Engineering,
Kangwon National University

요 약

공간데이터베이스에서 공간 질의를 최적화하기 위해서는 질의 결과 크기를 계산하는 것이 필수적이다. 그러나 공간 데이터베이스의 크기는 매우 방대하여 질의 결과 크기를 계산하는데 비용이 많이 든다. 이러한 문제를 해결하기 위해서는 실제 공간 데이터의 분포와 특성에 근접하도록 공간 데이터의 분포를 요약하여 이를 토대로 질의 결과 크기를 추정하는 것이 효과적이라 할 수 있다. 공간 분할 방법에는 균등 분할 방법과 비균등 분할 방법이 있으며, 본 논문에서 제안한 방법은 1차원 데이터에 대한 선택률 추정 기법 중에서 그 성능이 가장 우수하다고 평가된 바 있는 최대 면적 차이 분할을 공간 데이터베이스에 적용하여 공간 분할하는 것이다. 공간 데이터베이스에서의 선택률 추정 방법은 공간 분할 방법에 따라 성능상의 차이가 있으며 본 논문은 기존의 방법과 제안한 방법을 실험을 통하여 선택률 추정의 정확성을 비교, 평가하여 제안한 방법이 우수함을 보였다.

1. 서 론

지리 정보 시스템(GIS: Geographic Information System)은 방대한 양의 공간 데이터를 유지 관리한다[1]. 공간 데이터에는 점, 선, 다각형 등이 있으며, 질의 최적자는 이러한 데이터를 대상으로 질의 결과를 찾고자 할 때 최소의 접근 비용과 효율성을 고려하여 질의 결과를 찾는다.

질의 결과를 구하는 방법은 질의 조건을 만족하는 데이터베이스 내 데이터들을 하나하나 세부적으로 탐색하여 찾아내는 방법과 원래 데이터에 대한 요약 데이터를 생성하여 그 요약된 정보를 이용하여 질의 결과를 추정하는 방법이 있다. 첫번째 방법은 데이터의 양이 매우 방대하므로 직접적인 접근으로 질의 결과를 찾는 것은 많은 시간과 처리 비용이 든다. 따라서 두 번째 방법과 같이 추정치를 통해 질의 결과를 얻어 내는 것이 효율적이라 볼 수 있다. 단, 이때 추정을 위한 요약된 데이터들은 원래 데이터에 대한 분포와 특성을 비교적 잘 반영할 수 있도록 만들어 내는 것이 무엇보다도 중요하다.

공간 데이터베이스에서 이러한 요약데이터를 얻어내기 위해서는 공간을 분할하여 분할된 영역 내에 존재하는 공간데이터의 개수를 요약 데이터로 유지해야 한다.

이때, 공간을 어떠한 방법으로 분할하느냐에 따라 그 정확성이 달라진다. 기존에 연구된 공간 분할 방법에는 균등 분할 방법과 비균등 분할 방법이 있다. 균등 분할 방법은 분할 영역의 크기를 모두 동일하게 분할하는 방법이며, 비균등 분할 방법은 공간데이터의 분포를 고려하여 편재되어 있는 영역을 더욱 세밀하게 나누는 방법을 말한다. 일반적으로 균일한 데이터 분포일 경우는 균등 분할 방법이 효과적이나 그렇지 않은 분포일 경우에는 비균등 분할 방법이 효과적이라 볼 수 있다.

본 논문은 1차원 데이터에 대한 선택률 추정을 위해 연구된 바 있는 분할 방법인 최대 면적 차이 분할 방법을 2차원 공간데이터베이스에 적용하여 공간 분할을 시도하였으며 기존의 방법과 선택률 추정의 정확성을 비교, 분석하여 제안한 방법의 우수함을 보였다[2].

본 논문의 구성은 다음과 같다. 2장은 본 논문이 제안한 분할 방법에 대해 논의하고, 3장은 실험을 통하여 위 3가지 방법에 대한 성능 평가를 한다. 마지막으로 4장에서 결론을 맺는다.

2. 제안한 공간 분할 방법

2.1 최대 면적 차이 분할 방법

1차원 데이터에 대한 선택률 추정기법으로 최대 면적 차이 분할 방법이 가장 효과적인 성능을 보이는 것으로 보고된 바 있다.[2]

최대 면적 차이 분할 방법이란 값의 분산 정도와 빈도수를 동시에 고려하여 분할하는 것이다. 즉, 데이터의 값 순으로 정렬한 한 후 면적간의 차이가 임계치 이상 넘어가는 곳을 분할 경계로 한다.

여기서 말하는 면적이란 존재하는 값간의 분산과 그 영역 내 MBR(최소 경계 사각형) 개수의 곱을 의미한다. 2차원 공간에서 최대 면적 차이 분할 방법은 <식1>을 토대로 구해진 면적에 따라 공간을 x축과 y축상으로 분할하는 것이다. x_i는 x 축상에 i번째 값을 의미하고, f_{MBR}은 $x_i \sim x_{i+1}$에 속한 MBR의 개수를 의미한다.(y 축에 대해

서도 동일함)

식1) $A_{x_i} = (x_{i+1} - x_i) \times f_{MBR}$, $A_{y_i} = (y_{i+1} - y_i) \times f_{MBR}$

면적 차이가 임계치 이상인 곳을 분할 경계로 하여 x축상의 모든 분할 경계가 결정되고 나면, y축에 대해서도 이와 동일한 방법으로 분할 경계를 찾아낸다. 최종적으로 각 축에 대한 분할 경계의 교차점에 의하여 2차원 공간의 분할 영역이 결정된다.

알고리즘을 설명하면 아래와 같다. 이 알고리즘은 x축, y축 각각에 대하여 동일하게 적용된다.

최대 면적 차이 공간 분할 알고리즘

단계1.x(또는 y) 축에 대하여, 전체 분할 영역 내 MBR의 개수가 임계치 이내가 될 때까지 이진 분할한다.

단계2.결과 분할 영역이 결정되어 단계1이 종료되면, 각 분할영역의 면적을 계산하여 구해진 면적간의 값 차이가 임계차이 이상인 지점을 분할 경계로 결정한다.

단계3.단계2가 종료되면, 각 x, y축상에 구해진 분할경계를 서로 교차시켜 최종적인 2차원 분할영역을 구한다.

2.2 적용 예

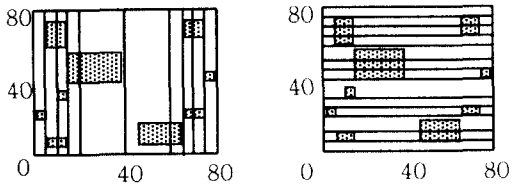
보다 이해를 쉽게 하기 위하여 <그림1>의 예를 살펴보자. <그림1>은 앞에서 설명한 분할 알고리즘의 단계1을 의미한다. 임계치를 1로 한다면, 1개 이상의 데이터가 존재하는 영역은 두개로 분할한다. 이와 같은 방법으로 x축과 y축상의 분할은 4차 분할까지 이루어진다(a)(b).

<그림1>과 같은 과정을 거친 후 값의 분산과 면적의 관계를 구하기 위해 마지막으로 분할된 영역과 그 영역에 속한 MBR의 개수를 히스토그램을 통하여 분석해보자.

<그림2>와 <그림3>는 알고리즘의 단계2에 해당한다. <그림2>는 x축의 분할 정보를 나타낸 것이다. (a)는 각 분할 영역과 MBR의 개수간의 관계를 히스토그램으로 나타낸 것이며 (b)는 이러한 정보를 토대로 <식1>에 근거하여 각 영역의 면적을 계산한 것이다. 면적간의 임계차이값을 5 라고 가정한다면 x축의 최종적인 분할 경계는 {5, 10, 15, 20, 60, 65, 75}가 된다.

<그림3>은 y축의 분할 정보를 나타낸 것이다. <그림3>의 (a)는 <그림2>의 (a)와 동일한 방법으로 진행되며, (b)는 <식1>을 이용하여 각 영역의 면적을 계산한 것이다. 면적간의 임계 차이는 마찬가지로 5로 한다면, 최종적인 y축의 분할 경계는 {5, 10, 20, 30, 45, 50, 60, 65, 70} 가 된다.

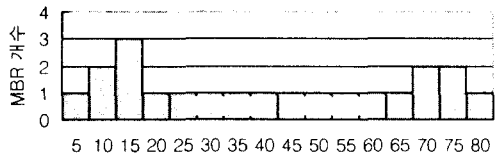
각 축에 대하여 분할경계가 결정되고 나면 이 분할 경계들이 교차하는 공간이 최종적인 2차원 분할 영역이 된다.



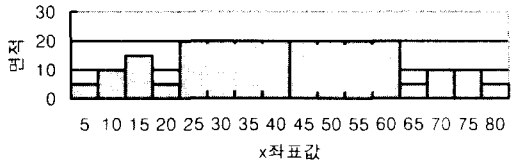
(a) x 축상의 분할

(b) y 축상의 분할

그림1. 2차원 공간을 차원별 분할

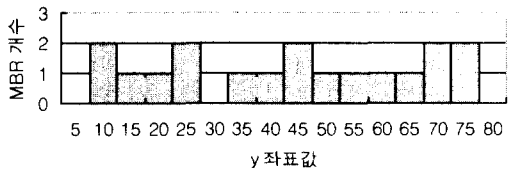


(a) 분할영역의 MBR수

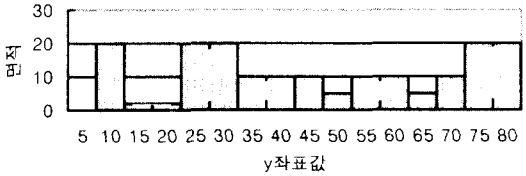


(b) 분할영역의 면적

그림2. x 축상의 분할 정보



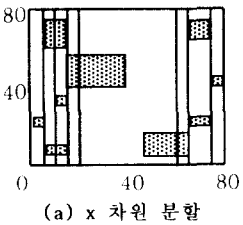
(a) 분할영역의 MBR수



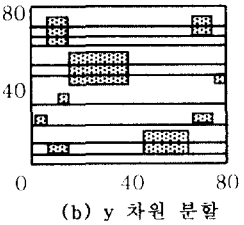
(b) 분할영역의 면적

그림3. y 축상의 분할 정보

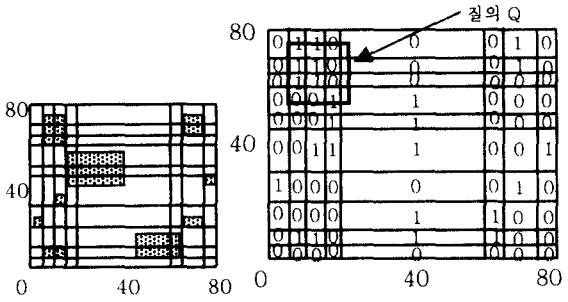
공간 질의에 대한 선택률 추정은 생성된 2차원 분할 영역 내 MBR의 개수를 이용하여 이루어질 수 있다. 최종적인 분할 영역을 나타내면 <그림4>과 같다. 각 축에 대하여 결정된 분할영역은 실제 공간에 (a)와(b)와 같고, 이 두 경계선이 교차하는 (c)가 2차원 분할영역이 되는 것이다. (d)는 각 분할 영역의 MBR개수를 나타낸 것이다. <그림4>의 (d)에서 질의Q가 주어졌을 때, 분할영역의 요약데이터를 토대로 <식1>을 이용하여 질의 결과를 계산하게 된다.



(a) x 차원 분할



(b) y 차원 분할



(c) 최종 분할 (d) 분할영역의 MBR개수

그림4. 최종 분할 결과

선택률을 추정하는 범위는 주어진 질의 영역과 교차하는 분할된 공간 영역이며, 질의 영역이 분할 영역과 교차하는 비율과 분할 영역내 MBR 개수를 곱한 것이 선택률 추정값(S)이 된다. 공식은 아래와 같다.

$$S = \sum_{i=1}^{k} n \times r_{overlap}$$

k : 질의 영역(Q)과 겹치는 분할영역의 개수($1 \le i \le k$)

n : i 번째 분할영역 내 MBR의 개수

$r_{overlap}$: i 번째 분할영역과 질의영역이 겹치는 비율

3. 실험

기존의 두 공간 분할 방법인 균등분할 방법(EP)과 비균등 분할 방법(QP), 그리고 본 논문이 제안한 최대 면적 차이 분할 방법(MP)을 실제 실험에 적용하여 성능을 비교하였다. 실험환경은 Sun Solaris Workstation에서 이루어 졌으며, C언어로 구현하였다. 성능 평가는 아래식에 의해서 구해진 오차율을 적용하여 이루어졌다.

$$\text{오차율}(\%) = \frac{|\text{실제선택률} - \text{계산된선택률}|}{\text{실제선택률}} \times 100$$

1) 실제 데이터에 대한 실험

실제 데이터는 long beach data와 mongocountry data를 사용하였으며, 실험 결과는 그림5와 같다.

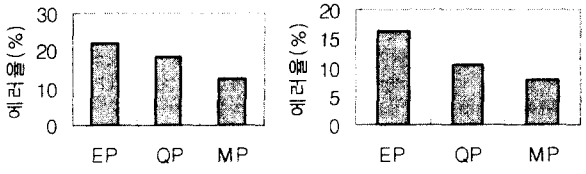


그림5. 실제 데이터에 대한 성능 비교

2) 데이터 크기에 의한 성능

40,000개의 공간 데이터를 균일하게 분포되도록 작성하였다. 전체 데이터 영역에 대한 데이터 크기를 10% ~ 40% 로 하여 실험하였다. 실험 결과는 <그림6>과 같다.

3) 질의 크기에 의한 성능

실제 데이터인 long beach data를 대상으로 하였으며, 질의 크기를 전체 데이터 영역의 5%~40%로 하여 실험하였다. 실험 결과는 <그림7>과 같다.

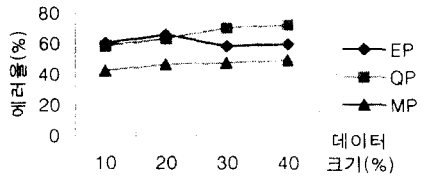


그림6. 데이터 크기에 대한 성능

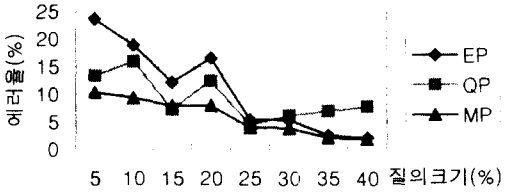


그림7. 질의 크기에 대한 성능 비교

4) 데이터 개수에 의한 성능

균일한 데이터 분포를 갖도록 데이터를 10000개~40000개로 생성하였으며, 임의의 질의를 수행하여 성능을 비교하였다. 실험 결과는 <그림8>과 같다.

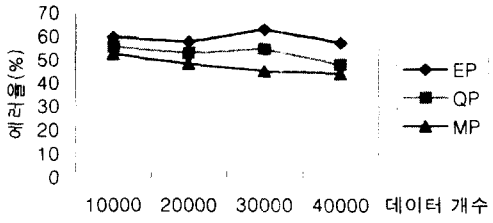


그림8. 데이터 개수에 의한 성능 비교

4. 결론

선택률 추정을 위해서는 공간데이터에 대한 요약된 정보가 필요하며, 이를 위해 다양한 접근 방법으로 공간을 분할한다. 공간 분할 방법에는 균등 분할 방법과 비균등 분할 방법이 있고, 본 논문은 최대 면적 차이 분할 방법을 제안하였다.

위 세가지 분할 방법을 실험을 통하여 성능을 비교해본 결과 비균등 분할방법이 균등분할 방법에 비해 월등히 좋은 성능을 보였으며, 최대 면적 차이 분할 방법이 비균등 분할 방법에 비해 좀 더 나은 성능을 보임을 알 수 있었다. 또한 최대 면적 차이 분할 방법은 x, y축상의 분할 경계를 결정하는 임계 면적 차이값에 따라 다른 성능을 보였다.

5. 참고문헌

[1] Swarup Acharya, Viswanath Poosala, Sridha Ramaswamy, " Selectivity Estimation in Spatial Databases" , SIGMOD 1999.

[2] Viswanath Poosala, Yannis E. Ioannidis, Peter J. Haas, Eugene J. Shekita, " Improved Histogram for Selectivity Estimation of Range Predicates" , SIGMOD 1996.