클러스터링 분석에 의한 공간데이터마이닝 방법*

손은정, 강인수, 김태완, 이기준
부산대학교 전자계산학과

A Spatial Data Mining Method by Clustering Analysis

Eun-Jeong Son, In-Soo Kang, Tae-Wan Kim and Ki-Joune Li
Department of Computer Science, Pusan National University
{ejson, iskang, twkim, lik}@spatios.cs.pusan.ac.kr

요약

지리정보시스템과 같이 방대한 양의 공간데이터를 다루는 응용시스템에서 공간데이터베이스로부터 규칙적인 특성이나, 혹은 관심있는 지식을 추출해내는 공간데이터마이닝의 역할은 매우 중요하다. 이를 위해 지금까지 이루어진 방법들에는 여러 가지가 있지만 그 중에서 대표적인 방법이 클러스터링으로 이는 단지 기하학적인 거리에 기반을 둔 공간적인 집중성과 분포도를 찾는 데에만 한정되어 있다. 그러나, 공간데이터마이닝을 위해서는 공간클러스터가 형성된 원인을 분석하는 것 또한 필요하다. 따라서 본 연구에서는 공간 클러스터링에서 얻어진 결과를 다른 공간적인 객체와의 연관성을 분석하여 공간적 집중성과 분포도를 유발하는 원인을 찾는 방법을 다룬다. 우선 몇 가지의 거리를 정의하는 것에 의해 클러스터와 공간객체사이의 연관성을 분석하는 방법을 제시하고, 생성된 공간 클러스터가 다수의 공간객체에 영향을 받을 경우, 그 공간클러스터를 각각 단위클러스터로 분리하는 방법을 제시한다.

1. 서론

지리정보시스템의 광범위한 활용으로, 공간데이터베이스에 저장되는 데이터의 양이 매우 커짐에 따라, 저장되어 있는 공간데이터의 분석은 지리정보시스템이나 공간데이터웨어하우스의 중요한 기능적 요구사항이 된다. 특히, 대량의 공간데이터베이스에 내재되어 있는 의미를 발견하는 작업은 매우 유용한 기능이다. 이를 위하여, 공간데이터마이닝에 대한 여러 연구가 진행되어 왔지만 아직은 초보적인 수준에 머무르고 있다.

지금까지 연구되어진 다수의 공간데이터마이닝의 방법은 주로 기하연산에 기반을 두고 있는 공간데이터의 집중성과 분포도에 한정된 클러스터링의 방법들에 대해서만 다루고 있다. 공간데이터마이닝의 궁극적인 목적 중 한가지는 공간데이터 사이의 공통적인 연관성을 찾는 것이다. 본 논문에서 제안하는 방법은 같은 레이어 상의 공간데이터를 클러스터링한 후, 생성된 클러스터와 다른 레이어 상의 공간객체와의 관계를 분석하여, 각각의 클러스터에 영향을 미치는 공간객체를 찾아내어 그 클러스터의 생성 원인을 밝히는 것이다. 예를 들면, 하나의 호수가 있을 때, 호수주위에는 여러 집들의 클러스터가 형성될 수 있다. 즉, 호수주위 집들의 클러스터는 호수라는 공간적인 객체에 의해 발생된 것이라고 볼 수 있다. 본 논문에서 제안하는 방법은 두 단계 즉, 공간데이터의 분포를 알아내기 위한 클러스터링의 단계와, 각 클러스터가 발생하게 된 원인을 찾기 위해 앞 단계에서 생성된 클러스터와 공간객체 사이의 관계를 분석하는 단계로 이루어져 있다. 본 논문에서 제시하는 방법을 적용하여 얻는 최종 결과는 각각의 클러스터와 그 클러스터에 영향을 주는 공간객체와의 쌍이다. 이러한 공간데이터마이닝은 범죄나 교통사고의 발생지역에 따른 원인분석이나, 환경오염의 원인 분석 등, 특정한 사건이 발생되는 지역의 원인을 분석하는데 매우 유용한 도구로 사용될 수 있을 것이다.

본 논문은 다음과 같이 구성된다. 먼저, 지금까지의 공간데이터마이닝에 대한 연구의 고찰과 본 논문의 목적에 대하여 2 장에서 살펴본다. 3 장에서는 각 클러스터와 공간객체사이의 관계를 분석하는 방법을 제안하고 4 장에서는 특별히 발생하는 복합클러스터를 분석하는 방법을 제시한다. 그리고 5 장에서는 이 방법들을 몇가지 실험 데이터에 적용한 실험결과를 보여주고 마지막으로 결론을 맺는다.

* 본 연구는 과학기술처의 국가 지리 정보 시스템 구축과제 DB Tool 과제의 공간 객체 저장 시스템 개발 세부과제의 위탁과제로 수행하였음

2. 관련연구

공간데이터마이닝을 위한 방법은 여러가지가 있지만 그 중에서 대표적인 방법이 클러스터링으로서, 특별한 속성정보나 어떠한 배경지식을 필요로 하지 않고 데이터로부터 직접 얻을 수 있다는 이점을 가지며, 이를 클러스터 분석이라 한다. 이 클러스터 분석은 공간객체의 분포를 알아내어 필요한 정보를 시각화하여 사용자에게 제공하거나, 또는 공간데이터의 집중성과 분포도를 알아내는데 매우 중요한 단서가 된다.

지금까지 알려진 대표적인 방법들에는 PAM[7], CLARA[10], CLARANS[10], BIRCH[11], DBSCAN[9], 그리고 SMTIN[4]등이 있다. 이러한 방법들은 모두 공간객체간의 거리를 기반으로 클러스터링을 한다. PAM 은 n 개의 공간객체에 대하여 k 개의 클러스터를 대표할 수 있는 *k-medoids* 를 찾는 방법으로 $O(k(n-k)^2)$의 시간복잡도를 갖는다. 이 단점을 극복하기 위해 개발된 CLARA 는 *medoid* 를 찾기위해서 샘플 데이터 집합을 선정하여 PAM 의 방법을 적용한 것이지만 결과가 샘플 데이터의 선정에 의존하므로 결과에 대한 신뢰성이 떨어진다. 이를 극복하기 위한 방법이 CLARANS 로써, 임의 탐색 알고리듬을 이용한 것으로 시간복잡도는 $O(k(n-k))$이다. 이는 전통적인 방법에 비해서는 우수한 성능을 보이지만, 데이터 집합의 패턴이나 분포도가 복잡해질 경우 충분한 정보를 제공하지 못하며, 계층적인 구조의 표현과 최적의 클러스터링 결과를 보장할 수 없다. BIRCH 는 공간객체의 통계적인 연관관계를 이용한 것으로 CF-vector(Clustering Feature Vector)를 이용한 방법이다. 이는 객체들간의 거리뿐만 아니라, 클러스터간의 거리도 고려하여 클러스터링을 하며 시간복잡도 또한 $O(n)$인 이유로 수행 성능면에서는 뛰어나지만, 클러스터링 자체에 중점을 두고 있기 때문에 클러스터의 형태에 대한 분석이 쉽지 않으며 데이터의 입력 순서에 민감한 단점을 갖고 있다. 이에 비해, SMTIN 은 먼저 각 객체들에 대하여 딜러니삼각화 과정을 통하여 TIN 데이터를 생성한 뒤, 이 TIN 데이터를 이용하여 클러스터링을 하는 방법으로 시간복잡도는 $O(n\log n)$이다. 이 방법은, BIRCH 와는 달리 클러스터링 이외에도 클러스터의 모양에 대한 정보도 제공하고 있으며, 데이터의 입력순서에 무관하게 클러스터링을 수행한다.[4] 이러한 이유로, 본 연구에서는 클러스터링을 위한 방법으로 SMTIN 을 적용한다.

본 연구에서 제시하는 클러스터링의 결과를 이용하여 클러스터와 다른 공간객체와의 상관관계를 밝혀내는 것은 공간데이터를 분석하는데 많은 도움을 줄 수 있다. 예를 들면, 환경 오염이 심각한 지역의 클러스터와 주변의 공장지대와는 매우 강한 관계가 있다는 것을 예상

할 수 있다. 물론 이와 같은 작업은 공간 클러스터를 시각화하여 사용자가 직접 확인할 수도 있지만, 공간객체의 수가 많아지고 클러스터가 많아지면, 단순한 시각화를 통한 분석은 한계가 있다.

본 논문에서 제안하는 방법은 클러스터와 공간객체사이의 위치적 관계를 분석하여 연관성이 가장 많은 공간객체를 찾아내는 것이다. 여기서, 단순히 가장 가까운 객체를 찾는 것만으로는 충분한 정보를 제공하지 못한다. 즉, 공간객체의 모양과 클러스터의 모양이 서로 일치할 때만 의미 있는 연관성을 발견할 수 있다. 예를 들면, 그림 2-1 에서 클러스터 A 는 단순히 가장 가까운 거리에 있는 객체 p 보다는 객체 q 가 더욱 밀접한 관계를 가지고 있다고 할 수 있다. 또한, 그림 2-2 에서 클러스터 B 와 연관성이 있는 공간객체는 r, s 가 있을 때, Reg.1 지역의 점들은 r 보다는 s 와 더 관련이 있으며, Reg.2 지역의 점들은 s 보다는 r 과 더 많은 관련이 있다. 그러므로, 이는 공간적 객체 r 과 s 에 의해 각기 형성된 공간클러스터가 겹쳐진 것일 가능성이 높기 때문에, 이를 위해서는 두개의 분리된 클러스터로 나누어 각각의 공간객체와 연관성을 찾아야 의미 있는 결과를 얻을 수 있다. 이와 같이 여러 개의 공간객체에 영향을 받은 클러스터를 복합클러스터라고 정의한다.

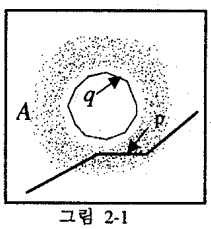

그림 2-1

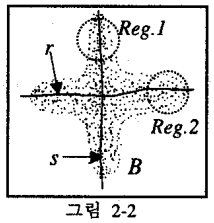

그림 2-2

따라서, 공간객체와 클러스터간의 연관성을 분석하기 위해서 해결해야 할 사항을 세가지로 요약할 수 있다. 첫째, 어떠한 공간 클러스터링 방법을 사용하는 것인가이다. 본 방법을 위해 사용하는 클러스터링 방법은 단순히 집중성만을 찾아내는 것이 아니라 클러스터의 모양, 즉 공간객체의 분포형태를 찾아낼 수 있어야 한다. 여기서, SMTIN 방법을 적용한다. 둘째, 공간클러스터와 공간객체의 거리를 적절하게 정의해야 한다. 그림 2-1 에서 공간객체 p 와 q 중 클러스터 A 와 더 가까운 것을 어느 것으로 정의하느냐에 따라, 연관성이 강한 객체를 찾을 수 있다. 셋째, 그림 2-2 와 같이 복합클러스터가 발생할 때 이를 단위클러스터로 분리하여, 연관성이 있는 객체와의 관계를 어떤 방법으로 찾아내는가의 문제이다. 다음 장에서는 이러한 문제에 대한 해결방안을 제시한다.

3. 클러스터와 공간객체의 연관성 분석

본 논문에서 제안하는 클러스터 분석과정은 다음의 세가지 단계로 이루어 진다.

단계 1 : 공간클러스터의 계산
단계 2 : 공간클러스터와 상관관계가 가장 큰 공간객체의 선택
단계 3 : 복합클러스터의 분석

단계 1 은 공간데이터의 클러스터링을 위한 방법으로 SMTIN 을 적용하며, 얻어진 결과는 여러 개의 공간 클러스터의 집합이다. 단계 2 에서 최소거리, 평균거리의 두 종류의 거리에 대한 정의를 하고, 이를 적용해서, 각 클러스터와 영향을 주는 공간객체의 쌍을 선택한다. 공간클러스터와 가장 강한 연관성을 가지고 있는 공간객체의 후보들은, 비교적 거리상으로 가까운 곳에 위치한다. 하지만, 클러스터와 공간객체의 거리를 단정적으로 정의하는 것은 그리 간단하지 않다. 여기서, 여러 용도에 따른 융통성있는 정의가 필요하므로, 본 논문에서는 최소거리 d_{min} 과 평균거리 d_{avg} 를 정의한다.

정의 1. 클러스터와 공간객체간 최소거리

$d_{min}(C, x) = 0$ when a point in x is in C
a minimum distance between C and x, otherwise.

C: 클러스터 영역 [1], x: 공간 객체.

[1] SMTIN 의 결과에 의해 생성된 클러스터의 경계선을 나타내는 polygon 이다.

정의 2. 클러스터와 공간객체간 평균거리

$$d_{avg}(C, x) = \frac{1}{n}\sum_{i=0}^{n} d(p_i, x) \qquad (p_i \in C)$$

C: 클러스터 영역, x: 공간 객체.

위의 두 종류의 거리에 대한 정의는 아래 그림과 같은 차이를 가지고 있다.

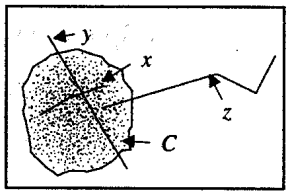

그림 3-1

먼저, 최소거리를 적용하면, 공간객체와 클러스터 영역이 교차하면 거리는 0 이 된다. 따라서, 그림 3-1 과 같이, 객체 x, y, z 모두 클러스터영역 C 와의 거리는 0 이 된다. 그러나, 평균거리를 적용하면 세 개의 공간객체중 x 가 클러스터영역과 가장 가까운 객체로 선정된다. 이 두가지 종류의 거리에 대한 유용성은 경우에 따라 다르게 결정된다.

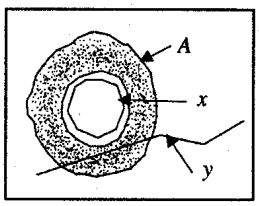

그림 3-2

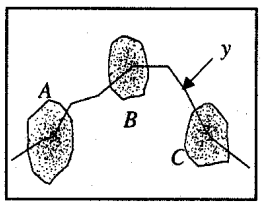

그림 3-3

그림 3-2 에서, 집들의 클러스터 A, 호수 x, A 를 통과하는 가스관 y 가 있을 때, 집들의 클러스터는 호수와의 최소거리, $d_{min}(A, x)$ 는 0 보다 크고, 가스관과의 최소거리, $d_{min}(A, y)$는 0 이다. 하지만, 평균거리를 적용했을 때, $d_{avg}(A, x) < d_{avg}(A, y)$가 되어 집들의 클러스터의 생성원인은 호수라고 할 수 있다. 반면에, 그림 3-3 의 경우, 최소거리를 적용했을 때, 어떤 한 도로 y 는 세 군락의 집들의 클러스터 A, B, C 에 영향을 미친다고 할 수 있다. 위의 예에서 보듯이, 최소거리와 평균거리는 모두 각각 다른 관점에서 의미 있게 이용될 수 있다. 결국 하나의 거리만 아니라 두가지 모두를 각각 적용하여 영향을 주는 공간객체 후보를 선택하는 것이 바람직하다.

4. 복합클러스터의 분석

본 장에서는 2 장에서 언급한 복합클러스터로부터 단위클러스터들을 추출해내는 방법을 제시한다. 이 방법은 두단계 즉, 클러스터가 단위클러스터인지 복합클러스터인지를 결정하는 단계와 복합클러스터인 경우 단위클러스터들로 분리하는 단계로 이루어 진다.

일단 공간클러스터와 공간객체의 연관성이 분석되면 많은 정보를 사용자에게 제공할 수 있다. 그러나, 경우에 따라 모호한 상황이 발생되는데, 그 중 한가지가 복합클러스터의 발생이다. 따라서, 클러스터가 복합클러스터인지 단위클러스터인지를 구분해야 하며, 만일 복합클러스터인 경우, 이를 그림 4-2 와 같이 여러 개의 단위클러스터로 분리할 필요가 있다. 이 처리과정은 다음과 같다.

Procedure 1. ProcessComplexCluster(C, X) // C: 클러스터, X: 객체집합.
1. $d(C, x) < \epsilon, x \in X$ 를 만족하는 X'를 선택
2. 만일 $n(X') = 1$ 이면, C 는 단위클러스터로 판정 종료.
3. 만일 $n(X') > 1$ 이면 C 는 복합클러스터로 판정
   3-1. 모든 $x \in X'$에 대하여,
   가장 적합한 단위클러스터 C_i를 선택

위의 과정을 통해서 얻어지는 결과는 하나의 공간객체와 단위클러스터

의 쌍의 집합 $\{(x, C_i) \mid x \in X \text{ and } C_i \subset C\}$이 된다. 먼저, 주어진 클러스터에 영향을 주는 공간객체의 후보는 단순한 방법으로 구한다. 아래와 같이, 주어진 δ값보다 가까운 거리에 있는 모든 객체를 후보객체로 선택한다. 그 다음, 후보중에서 가장 적합한 공간객체와 이에 대응되는 단순 복합객체를 찾는 3-1 과정은 상당히 복잡하다. 복합클러스에서 단위클러스터를 분리하는 과정은 주어진 공간객체가 영향을 줄 수 있는 영역을 찾아내는 것이다. 물론 영향을 받은 클러스터는 주어진 공간객체와 일정 범위내가 될 것이다. 그러나, 문제는 어떻게 이 거리를 찾아내는가이다. 아래 그림 4-1 은 복합클러스터의 예이며, 그림 4-2 는 복합클러스터에서 추출된 두개의 단위클러스터들의 예를 보인다. 이때, 그림 4-1 에서 그림 4-2 로 가는 과정은 아래에 서술한 알고리듬과 그림 4-3, 그림 4-4, 그림 4-5 에서 설명한다. 결국 이 문제는 어떠한 방법으로 복합클러스터를 분리하여 단위클러스터로 만드는 가의 문제가 된다.

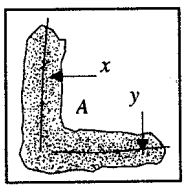

그림 4-1

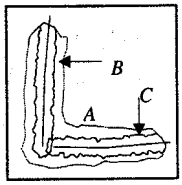

그림 4-2

이를 위해서, 본 논문에서 제안하는 방법은 다음과 같다.

Procedure 2. ExtractSimpleCluster(C, x) // C: 복합클러스터, x: 후보객체

1. C와 x의 거리평균과 분산을 계산하여 (m, σ)를 구한다.
2. $C' \leftarrow C$의 객체중 $[m-3\sigma, m+3\sigma]$의 거리 안에 들어오는 객체들
3. C'와 x의 거리평균과 분산 계산하여 (m', σ')를 구한다.
4. 만일 $|m' - m| > \delta$ 이면 step 2 와 step 3 을 반복
5. C'를 단위클러스터로 지정.

만일 공간클러스터에 영향을 주는 공간객체와 클러스터안의 객체사이의 거리가 정규분포를 따른다고 가정하면, C'는 단위클러스터의 99.7% 의 객체들을 포함한다고 할 수 있다. 즉, 그림 4-1 과 같이 주어진 공간객체와 클러스터를 구성하는 객체들 사이의 거리 분포에서 $[m-3\sigma, m+3\sigma]$ 거리 안에 들어오는 객체를 선택하여 그림 4-4 와 같이 새로운 평균과 분산을 구하고 더 이상 평균의 변화가 생기지 않을 때 까지 반복하여, 그림 4-5 와 같은 단위클러스터를 얻을 수 있다. 이와같은 과정을 모든 후보공간객체에 대하여 반복하면 그림 4-2 같이 하나의 복합클러스터가 여러 개의 단위클러스터로 분리된다. 물론, 클러스터내의 하나의 객체는 여러 개의 단위클러스터와 쌍이 될 수도 있다.

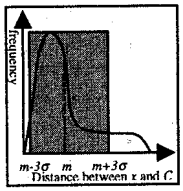

그림 4-3

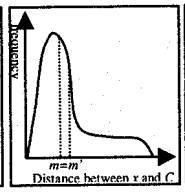

그림 4-4

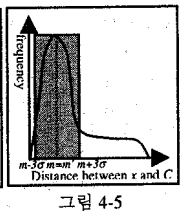

그림 4-5

본 장에서는 복합클러스터에 영향을 줄 수 있는 공간객체를 고려하여 여러 개의 단위클러스터로 분리하는 방법을 제시하였다. 이는 다른 원인에 의해 발생된 여러 개의 클러스터가 우연히 같은 지역에 중복되었을 경우 이를 분리하는 목적에 이용될 수 있다.

5. 실험결과

본 장에서는 본 논문에서 제시한 방법을 하나의 예에 적용하여 나오는 결과를 보인다. 본 예에서 취한 예는 집들로 이루어져 있는 공간 클러스터와 도로라는 공간객체의 연관성을 분석한 것이다.

그림 5-1 에서와 같이, 편의상 집들은 점으로 도로는 다각선으로 단순화하였다. 단계 1 을 통해 계산된 클러스터는 그림 5-2 의 경우와 같이 여러가지 형태의 클러스터가 발견된다. 이를 이용하여 단계 2 에서 영향을 주는 공간객체를 선택하면 그림 5-3 과 그림 5-4 와 같이 여러 개의 후보공간객체가 선택된다. 이때 5-3 은 최소거리로 선택한 경우이고, 5-4 는 평균거리로 선택한 경우이다. 그림 5-2 에서처럼, 몇몇의 클러스터는 두개 이상의 후보객체에 의해 영향을 받으므로 이를 단계 3 의 복합클러스터 분석을 통해서 단위클러스터들로 분리된다. 결국 각각의 클러스터와 연관된 공간객체의 쌍을 찾을 수 있다.

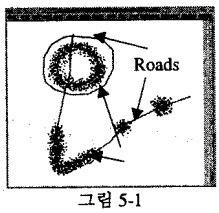

그림 5-1

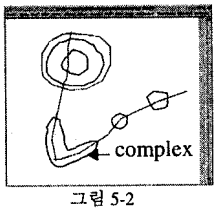

그림 5-2

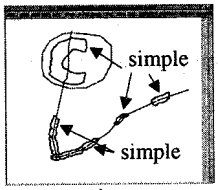

그림 5-3

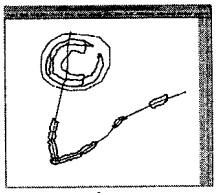
그림 5-4

6. 결론

본 논문에서 제시하는 공간데이터마이닝 방법은 생성된 클러스터와 공간객체사이의 관계를 찾아내어 클러스터의 생성 원인을 밝혀낸다. 클러스터링 방법으로는 SMTIN 방법을 적용했으며, 각 클러스터에 영향을 주는 가장 적합한 공간객체를 찾아내기 위해, 두 종류의 거리를 정의하였으며, 또한, 여러 개의 공간객체에 의해 영향을 받은 복합클러스터를 여러 개의 단위클러스터들로 추출해 내는 방법을 제안하였다. 최종적으로, 각 단위클러스터와 그것에 영향을 미치는 공간객체의 쌍을 찾아낸다. 향후 연구과제로, 클러스터와 공간객체의 관계뿐만 아니라 서로 성질이 다른 클러스터와 클러스터간의 관계 분석을 통해 클러스터의 생성원인을 찾아내는 방법 또한 개발 중에 있다.

7. 참조문헌

[1] A.C. Cressie, *Statistics for spatial Data, John Wiley & Sons,* 1991

[2] B.W. Silverman, *Density Estimation for Statistics and Data Analysis, Champman and Hall Ltd,* 1986

[3] F.P. Preparata. and M.I. Shamos, *Computational Geometry: An Introduction, Springer-Verlag,* 1985

[4] I-S. Kang, T-W. Kim, and K-J. Li, "A Spatial Data Mining Method by Delaunay Triangulation," 5th ACM-GIS, 1997, pp. 35-39

[5] J. A. Hartigan, *Clustering Algorithms, John Wiley & Sons,* 1975

[6] J. R. Schewchuk, "Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Trianglultor," First Workshop on Applied Computational Geometry, ACM, 1996

[7] L. Kaufman, and P.J. Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons,* 1990

[8] M. Ester, H-P. Kriegel, and X. Xu, "A Database Interface for Clustering in Large Spatial Databases," Proc. 1st Int'l Conf. On knowledge Discovery and Data Mining, 1995

[9] M. Ester, H.P.Kriegel, J. Sander, and X. Xu, "A Density-based algorithm for Discovering Clusters in Large Spatial Databases with Noise,"Proc. Of th 2nd Int.Conf.on KDD-96, 1996, pp.226-231

[10] T. Ng. Raymond, and J.W. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994, pp.144-155

[11] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH:An Efficient Data Clustering Methods for Very Large Databases," Proceedings of ACM SIGMOD International Conf. of Management of Data, 1996, pp.104-114