PDA에서 공간데이터를 저장하기 위한 혼성색인구조

김 은영°, 전 봉기, 서 영덕, 홍 봉희

부산대학교 컴퓨터공학과

A Hybrid Indexing Structure for Storing Spatial Data in PDA

Eun-Young Kim°, Bong-Gi Jun , Young-Duk Soe, Bong-Hee Hong

Dept. of Computer Engineering, Pusan National University

요 약

컴퓨터와 통신의 기술 발달에 따라 개인용 휴대 기기의 보급이 확산되고 있다. 휴대 단말기에서의 다양한 지도서비스의 요구 또한 나날이 높아지고 있다. 그런데, 휴대용 기기들은 기존 지도서비스 환경이었던 서버나 PC와 비교하여 연산처리속도가 낮고 저장용량이 적다. 그래서 기존 환경에서 적용된 저장 및 색인구조가 휴대용 기기에 그대로 적용될 수 없다. 또한, 휴대용 기기의 이동성을 고려할 때, 질의 수행 시 저장되지 않은 지도 데이터는 무선 통신을 이용하여 서버로부터 전송되어져야 한다. 새롭게 전송된 지도 데이터는 기존에 저장된 지도 데이터에 삽입되기 위해서 색인 재구축 비용이 발생한다. 즉, 서버와 무선통신을 하는 휴대용 단말기에서 지도서비스를 하기 위해서는 휴대 기기에 적합한 데이터 저장 및 색인구조가 필요하다.

이 논문에서는 무선단말기, 특히 PDA 환경에서의 벡터지도시스템을 위해 공간 데이터의 최적화된 저장구조와, 비연속적인 다양한 지역에 대한 데이터를 효율적으로 통합·관리하는 색인구조를 제시한다.

1. 서론

IMT2000 및 UMTS(Universal Mobile Telecommunication System)으로의 초고속 무선통신의 발전과 무선단말기 하드웨어의 발전에 따라 휴대용 무선단말기의 보급이 빠르게 확산되고 있다. 특히, PDA는 전자수첩보다 더 다양하고 강력한 기능을 지원할 뿐 아니라 노트북이나 PC보다 저렴하고 이동성이 뛰어나다는 장점들로 인해 사용이 더욱 급속히 늘어나고 있다.

그리고 휴대용 단말기를 가진 사용자들은 이동 중에 무선단말기를 통해 다양한 지도서비스를 받기를 원한다. 즉, PDA에서 전자지도 및 지역정보 검색 기능, 실시간 교통정보 서비스 뿐 아니라 GPS 칩셋을 장착한 이동객체의 위치 기반 차량항법 및 물류 관제 등의 필요성이 대두되고 있다.

그러나 기존의 벡터지도시스템은 벡터지도의 높은 연산비용과 많은 데이터량으로 인해서 대규모 저장공간/메모리, 높은 CPU의 연산능력을 가진 컴퓨팅 환경을 대상으로 하고 있다. 따라서, 기존의 벡터지도데이터의 저장기법이나 색인을, PDA와 같이 비교적 자원이 적고 이동성을 가지며 무선통신을 하는 컴퓨팅 시스템에서 그대로 적용하기에는 비효율적이다. 그러므로 새로운 컴퓨팅 환경에 적합한 벡터지도데이터의 저장 및 색인 구조를 제시할 필요가 있다.

또한 상대적으로 많은 공간데이타를 적은 용량의 휴대용단말기에 저장하기위한 압축저장기법이 필요하다.

이 논문에서는 이동 중에 있는 단말기에서 지도를 저장, 검색, 서버와의 통신을 통한 지도데이터 다운로드를 지원하기 위한 효율적인 공간 데이터의 저장기법과 색인 구조를 제시한다.

이 논문의 구성은 다음과 같다. 먼저 2장에서는 관련 연구를 기술하고 3장에서는 혼성색인기법에 대해서 설명한다. 4장에서는 공간데이터를 실제 저장하기위한 압축기법에 대해서 서술하고, 마지막으로 5장에서는 결론 및 향후 연구를 기술한다..

2. 관련연구

공간 데이터를 위한 기존의 색인으로는 그리드파일로 대표되는 영역기준 분할방식의 색인과 객체 기준 분할방식의 R-tree 계열 색인이 있다. 그리드파일은 해쉬 기반의 색인으로서 색인구축과 검색 속도가 빠른 장점이 있으며, R-tree 계열의 색인은 삽입되는 객체에 따라서 도메인 공간(Domain space)이 가변적이다.

이동 중인 휴대 단말기에서 다양한 영역의 상대적으로 많은 데이터를 서버로부터 전송받아 저장하게 될 경우, 휴대 기기에 저장되는 공간 데이터의 도메인 공간은 질의영역에 따라 가변적이다. 또한 전송된 데이터 역시 소수의 객체가 아니라 상대적으로 많은 객체를 한 번에 삽입(Bulk-Insertion)하게 된다.

[3]에서는 기 구축된 R-tree에 많은 데이터를 삽입하는 Bulk-Insertion의 방법으로서 STLT(Small-Tree-Large-Tree)를 제시하였다. 그러나 이 방법 또한 작은 트리를 만들기 위한 비용이 필요하며, 같은 데이터에 대해 그리드파일을 구축하는 비용에 비해 훨씬 크다. 또한 그리드파일은 고정된 도메인 공간을 대상으로 색인을 구축함으로 도메인 영역이 바뀌는 데이타의 삽입은 전체 색인 재 구축의 비용이 든다. 따라서, 도메인 공간이 가변적이면서 Bulk-Insertion에 효율적인 새로운 색인구조의 제시가 필요하다.

기존의 압축방법은 주로 텍스트, 이미지, 동영상, 사운드 데이터의 중복정보를 제거하기 위해 부호화하는 방식을 사용했으며, 원본 데이터로 복원이 가능한가에 따라 손실 압축과 무손실 압축으로 분류할 수 있다. 그러나 부호화 방식의 압축은 압축을 풀어야만 실제 데이터를

사용할 수 있다는 단점이 있으므로 저장공간과 메모리의 용량이 적은 PDA에서 압축하기와 풀기를 해야 하는 기존의 압축방법은 효율적이지 못하다.

3. 혼성색인구조

3.1 대상환경과 문제정의

이 논문은 사용자가 GIS 기능을 가진 휴대 기기(예를 들어, PDA)를 가지고 이동 중에 필요한 지도 데이터를 무선통신을 통해서 서버로부터 전송받아서 저장하는 환경을 대상으로 한다.(그림 1)

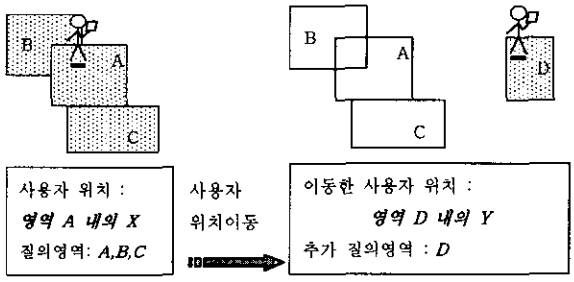

그림 1 사용자의 위치 변화에 따른 질의영역의 변화

이 경우 휴대용 단말기의 입장에서 다음과 같은 시나리오를 가진다.

사용자는 X위치에서 질의 영역 A를 가지고 있으며, 해당영역에 대한 질의를 수행한다.

사용자는 A영역에 인접한 B,C 영역의 데이터를 서버에게 요청하여 전송받아서 PDA에 저장하고 사용자는 저장된 영역(A B C)에 대해서 질의 수행한다. 이때 A와 B는 $A \not\subset B \& B \not\subset A \& (A \cap B) \neq \emptyset$ 의 조건을 만족하므로 B영역에 대한 질의를 할 때 무선통신의 비용을 줄이기 위해서 서버로부터 전송되는 질의영역은 (B-A) 이다.

사용자가 위치 X와 동떨어진 D영역 내의 위치 Y로 이동한다.

이동한 위치 Y에서 D영역에 대한 데이터를 무선통신으로 서버로부터 전송받아서 저장한다.

시나리오 에서 사용자가 질의하는 영역이 PDA에 저장된 영역 내에 완전히 포함되지 않을 경우 추가적으로 필요한 영역의 데이터를 서버에 질의하여 전송받아야 하는데, 사용자가 질의한 영역을 q라고 하면 다음 번에 질의할 영역은 q와 인접할 확률이 높다. 매 질의마다 상대적으로 작은 영역인 q의 데이터를 전송받는 것보다, 그림 2와 같이 q를 포함하는 상대적으로 넓고 잠정적으로 사용자가 질의할 가능성이 높은 영역 Q로 확장한 영역의 데이터를 한번에 전송하는 것이 통신 비용 면에서 효율적이다.

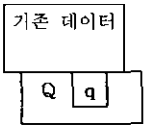

그림 2 사용자 질의영역(q)과 확장된 서버 질의영역(Q)

그리드파일은 검색속도가 비교적 일정하고 빠르며, 초기 색인 구축 시 한 번의 데이터 스캔으로 색인을 구축할 수 있다는 장점이 있다[4]. 그러나 위의 예시와 같은 경우에, 서버로부터 질의영역(Q)의 데이터를 전송받아 저장할 때마다 그리드파일의 도메인 공간이 아래 그림 3과 같이 A, MBR(A B), MBR(A B C), MBR(A B C D)로 확장된다. 따라서 그 때마다 전체 색인을 다시 재구축 해야 하는 문제가 발생하게 된다.

그리드파일과는 달리 R*-tree는 도메인 공간이 미리 고정된 것이 아니라 삽입되는 객체에 따라 달라질 수 있다. 그러나 R*-tree 많은 객체를 한번에 삽입(Bulk-Insertion)하는 비용은 그리드파일에 비해 크기 때문에[3] 사용자의 질의에 대해 빠른 응답시간을 보장할 수 없다.

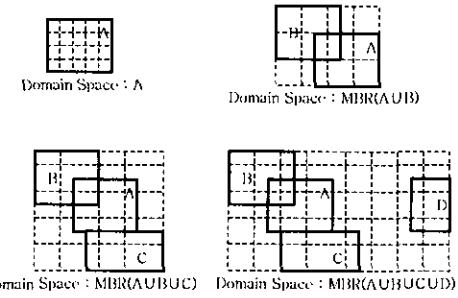

그림 3 그리드파일의 도메인 공간의 확장

3.2 혼성색인구조

이 논문에서 제시하는 혼성색인구조는 아래 그림 4와 같다. 아래 그림에서 혼성색인구조는 (1)도메인 색인(Domain-Index), (2)질의영역정보(Query Region Information), (3)질의영역 색인(QR-Index, Query Region-Index) 으로 구성된다.

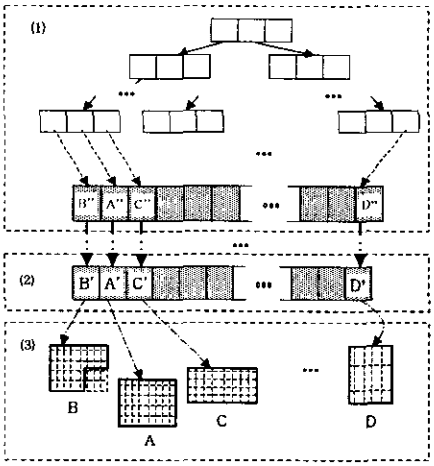

그림 4 혼성색인구조

(1) *도메인 색인*은 질의영역들에 대한 색인으로서, 도메인 공간의 증감이 자유로운 트리 계열의 색인을 사용한다.

(2) *질의영역정보*의 각 필드에는 질의영역의 실제 경계에 대한 정보(다각형 형태)와 하위의 (3)질의영역색인에 대한 정보를 포함한다.

(3) *질의영역 색인*은 각 질의영역(Q)에 포함되는 객체들에 대한 색인으로서, 고정된 도메인 공간에 대해서 빠른 구축과 검색을 지원하는 색인을 사용한다.

예를 들어, 도메인 색인으로서 R*-tree를 사용할 수가 있고, 질의영역 색인으로서 고정그리드파일을 사용할 수가 있다. 이 방법은 도메인 공간이 자유롭게 증감할 수 있다는 R*-tree의 장점과 색인구축 비용이 적다는 고정그리드파일의 장점을 모두 취한다. 이러한 가정 하에 그림 1의 질의영역에 대해서 그림 4와 같이 혼성색인이 구축된다. 이 때, (3)질의영역 색인인 고정그리드는 질의영역(A,B,C,D)에 대하여 각각 구축되고, (2)질의영역정보에서는 해당되는 고정그리드의 정보(scale)와 다각형인 질의영역 형태(A,C,D는 사각형, B는

육각형)를 저장하는 질의영역정보 필드(A',B',C',D')가 저장되고, (1)도메인 색인인 R*-tree에서는 각 질의영역의 MBR 정보와 질의영역정보필드에 대한 참조를 가진다(A'',B'',C'',D'').

3.3 혼성색인구조를 위한 알고리즘

삽입 알고리즘

이 논문의 대상이 되는 환경에서 혼성색인구조에 대한 삽입은 객체 한 개의 삽입이 아니라 질의영역에 포함되는 모든 객체의 삽입, 즉 Bulk-Insertion이 된다. 혼성색인구조에 대한 삽입은 다음과 같이 3단계로 거쳐서 수행된다.

질의영역 색인 구축 – 질의영역에 포함되는 객체들에 대해서 적절하게 질의영역 색인을 구축하고 정보(질의영역 색인 정보와 질의영역형태(다각형) 정보)를 수집한다.

질의영역정보 삽입 – 에서 수집한 정보를 질의영역정보에 하나의 필드로서 삽입하고 이에 대한 참조를 구한다.

도메인 색인 삽입 – 에서 구한 질의영역정보필드의 참조를 도메인 색인에 삽입한다.

검색 알고리즘

혼성색인구조에서 검색은 크게 세 단계의 여과 과정과 한 단계의 정제 과정을 거쳐서 수행된다.

여과 과정1 – 도메인 색인에서의 여과 과정으로서, 검색조건과 질의영역의 MBR만을 비교하여 1개 이상의 질의영역정보 필드에 대한 참조를 추출한다.

여과 과정2 – 질의영역정보에서의 여과 과정으로서, 여과 과정1을 거쳐서 추출된 질의영역정보 중에서 질의영역의 MBR이 아닌 다각형 형태의 요약정보를 이용하여 검색조건과 좀 더 일치하는 질의영역정보 필드를 추출한다.

여과 과정3 – 질의영역 색인에서의 여과 과정으로서, 여과 과정2를 거친 질의영역정보필드와 매핑되는 질의영역 색인에서 여과하여 후보 객체 집합을 선별한다.

정제 과정 – 여과 과정3을 거쳐서 선별된 객체 집합에 대해서 검색조건과 완전히 일치하는 객체만을 선택한다.

예를 들어, *위 그림 1의* 예시에서 서버에 질의한 영역이 A,B,C,D가 되면 혼성색인은 그림 4와 같이 된다. 저장된 영역에 대해서 사용자가 아래 그림 5의 빗금부분(S)과 같이 질의를 하면, 여과 과정1에서 A'',B''가 추출되고, 여과 과정2에서 A'가 추출된다; 서버에 B 영역을 질의할 때 사각형 B가 아니라 (B-A)에 해당하는 육각형으로 질의하여 데이터를 전송받아 저장하므로 질의영역 B에 대한 질의영역 정보필드에는 사각형 B가 아닌 육각형 B의 정보가 포함된다. 따라서 여과 과정2에서 이 육각형 정보를 사용하므로 B는 제외된다. 여과 과정3에서 검색영역(S)와 겹쳐지는 그리드 셀을 추출하고, 마지막 정제 과정에서 추출된 그리드 셀 내의 객체에서 검색영역(S)과 동떨어지지 않은(not disjoint) 객체를 추출한다.

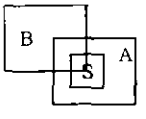

그림 5 사용자 질의(빗금 영역)

4. 공간데이타의 압축기법

디스크에 저장되는 공간데이터의 전체 크기에 가장 큰 영향을 미치는 것은 선, 면 데이터이다. *이 장에서 설명하는 공간데이터의* 압축기법은 혼성색인구조의 최하위인 질의영역 색인의 data bucket에 적용되어 디스크에 저장될 선, 면 공간 객체의 크기를 감소시킨다.

선과 면 객체는 n개의 점으로 구성된 객체이며, 저장할 때는 MBR, 구성원인 점의 개수를 저장하는 정수형 변수 n과 점 데이터를 저장한다고 가정한다. 즉, 선(면) 객체 L을

$$L = \{P_i(x_i, y_i) \mid x_a \le x_i \le x_b, y_a \le y_i \le y_b, i = 1,2,\cdots,n\}$$
$$MBR(L) = ((Xa, Ya), (Xb, Yb))$$

라고 정의하면 서버에서 선(면) 객체 L을 저장하기 위한 공간의 크기 SS(L)은 다음과 같다.

$$SS(L) = 4 * B + N + n * 2 * B$$

n = *선(면) 객체를 구성하는 점의 개수*

B = *점의 X좌표를 표현하기 위한 byte 수*

N = *서버의 정수형 변수의 byte수*

MBR을 위해서 2*2*B byte, 변수 n을 위해서 N byte, 구성원인 n개의 점을 저장하기 위해 n*2*B byte 필요하다.

객체의 MBR 정보와 구성원인 점들 간의 difference를 이용하여 각 객체 별로 압축을 한다. 압축은 선(면) 객체의 구성원인 점에 대해서만 압축한다; MBR과 점의 개수를 저장하는 변수 n은 압축 전과 동일하다. 선(면) 객체를 이루고 있는 점들 중 첫 번째 점 $P_1(X_1,Y_1)$은 변환하지 않은 값을 그대로 사용하고, 두 번째 이하의 점 Pi(Xi,Yi)은 앞점(Pi-1)과의 difference만으로 표현한다. 따라서 선(면) 객체 L은 다음과 같이 표현될 수 있다.

$$i = 1, P_i(X_i, Y_i) = (X_1, Y_1)$$
$$i = 2,3,\cdots,n \text{ 일 때, } P_i = (X_i - X_{i-1}, Y_i - Y_{i-1})$$

의 경우에는 2*B byte 가 필요하고, 에서는 Xi의 값이 0 Xi' 2^{8*K}) (단, 0<K B) 이므로, K byte 필요하다. 그러므로 선(면) 객체 L을 압축하여 저장하기 위해서는 4*B+ N + 2*B+(n-1)*2*K byte 가 소요된다. 이때의 압축률은 압축전 : 후가 n*B : B+(n-1)*K 의 비율로서 K가 B에 비해 작은 값일수록 높은 압축효과를 기대할 수 있다.

5. 결론 및 향후 연구

이 논문에서는 메모리와 저장공간이 비교적 적으면서 이동성을 가지고 무선통신을 하는 휴대용 단말기에 대해 최적화된 공간데이터의 혼성색인구조 및 압축기법을 제시하였다.

혼성색인구조는 비연속적인 여러 개의 질의영역에 대한 객체를 저장할 때 적은 비용으로 전체 영역에 대한 색인을 구축할 수 있다. 또한, 사용자가 질의를 했을 때, 질의 영역이 현재 저장된 영역과 중복이 되는지 빠르게 검사하여 중복되지 않은 영역의 확장영역을 질의함으로써 대역폭이 좁은 무선망을 통해 전송비용을 줄일 수 있다.

또한, 선,면 데이터를 압축하는 기법은 상대적으로 적은 저장공간을 가진 휴대 기기에 더 많은 지도 데이터를 저장할 수 있도록 하며, 무선통신을 할 경우 통신 패킷의 크기를 줄이는데 크게 기여한다.

향후 연구로서 제한적 저장구조를 효율적으로 사용하기 위해 캐쉬 교체 전략을 고려한 색인 기법과 성능평가가 필요하다.

6. 참고 문헌

[1] Norbert Beckmann, Hans-Peter Kriegel, "The R*-tree:An Efficient and Robust Access Method for Points and Rectangles", *Proceedings of the 1990* ACM SIGMOD, pages 332-331, 1990.

[2] Christos Faloutsos, Timos Sellis, Nick Roussopoulos ,"Analysis of Object Oriented Spatial Access Methods", Proceedings of the 1987 ACM TODS, pages 426-439

[3] Li Chen, Rupesh Choubey and Elke A. Rundensteiner, "Bulk-Insertions into R-Trees", ACM-GIS 1998: 161-162

[4] J. Nievergelt, Hans Hinterberger and Kenneth C. Sevcik,"The Grid File: An Adaptable, Symmetric Multikey File Structure ",ACM TODs, pages 38-71,1984