소셜 데이터베이스에서 공간 단어의 효율적인 검색

양평우, 조현구, 남광우
군산대학교 컴퓨터정보공학과
e-mail : {manner7979, pseudojo, kwnam}@kunsan.ac.kr

Efficient Retrieval of Spatial Words in Social Database

Pyoung Woo Yang, Hyun Gu Joe, Kwang Woo Nam
Dept. of Computer and Information Engineering, Kunsan University

요 약

공간 웹 객체는 문서상에 지리정보를 포함하는 문서를 말한다. Twitter 나 FaceBook 같은 경우 문서가 생성된 위치를 문서 안에 포함하고 있다. 최근에는 공간 웹 객체와 같은 공간정보와 문자를 요구하는 검색이 많이 요구되고 있다. 본 논문에서는 공간 웹 객체를 검색하기 위한 효율적인 검색 기법을 제안한다. 이를 위하여 문서를 단어별로 나누고 각 단어와 문서의 위치정보를 포함하는 공간 객체를 만들어 공간객체를 검색하기 위한 QP-tree 를 제안한다.

1. 서론

최근 인터넷상의 문서에 지리 정보를 포함하거나 GPS 데이터를 포함하는 문서들이 많이 생성되고 있다. 예를 들어 문서의 내용에 '서울'이나 '군산'같은 지역명과 같이 지리정보를 포함하는 경우가 있고 Twitter 이나 FaceBook 같은 소셜 웹에서는 문서가 생성된 위치의 정보(GPS 정보)를 문서안에 포함하고 있다. 이와 같이 지리정보를 포함하는 문서를 공간 웹 객체(Spatial Web Object)라고 한다. 공간 웹 객체는 문서 정보와 공간 정보를 같이 가지고 있기 때문에 기존의 검색 기법 사용자가 원하는 정보를 검색하기는 어렵다. 예를 들어 "서울시에 맛있는 중국집을 찾아라"와 같은 질의는 "서울시", "맛있는", "중국집"과 같은 단어를 키워드로 하고 서울시에 해당하는 공간 좌표를 이용하여 검색을 해야 한다. 하지만 기존의 검색 기법은 단어만을 사용하거나 공간 정보만을 사용하는 방법을 쓰고 있어서 단어 정보와 공간정보를 같이 이용하는 방법이 필요하다. 따라서 본 논문에서는 공간 웹 객체를 위하여 단어 정보와 공간 정보를 같이 활용하여 검색을 할 수 있는 인덱스 기법을 제안한다.

2. Spatial Web Object

Spatial Web Object 는 다음과 같이 되어 있다고 볼 수 있다.

Doc = <id, words={$word_1$, $word_2$, … , $word_n$}, geo>

하나의 문서는 여러 단어(words)로 이루어져 있고, 각 문서는 id 와 지리정보(geo)를 포함한다. 따라서 각 단어들을 트리에 넣기 위하여 본 논문에서는 SpatialWord(sword)를 정의 하였다.

sword= <Doc_id, wordn, word_position, geo>

sword 는 단어가 포함된 문서의 id(Doc_id)와 단어(word), 단어의 문서상의 위치(word_position), 그리고 문서의 위경도 좌표에 해당하는 geo 를 포함하고 있다. 문서 하나를 읽게 되면, 문서안의 모든 단어들은 sword 로 변환하여 트리를 구축한다.

3. QP-tree : Spatial-First 검색

QP-tree 는 Quad-tree 와 Patricia trie 를 이용한 공간 웹 객체의 검색을 위하여 본 논문에서 제안하는 인덱스이다.

3.1 QP-tree 의 구조

QP-tree 는 공간정보를 우선으로 하는 tree 이다. QP-tree 는 우선 공간 웹 객체의 공간 정보를 이용하여 각 객체들을 인덱싱 한다. Quadratic tree 의 말단 노드에는 노드안에 있는 공간 웹 객체들의 단어에 대한 인덱싱을 위한 Patricia trie 가 존재하여 단어에 대한 인덱싱을 한다. 사용자가 공간 웹 객체에 대하여 검색을 할 경우 QP-tree 는 우선 공간 정보를 이용하여 Quadratic-tree 를 검색하고 검색된 말단노드에 있는 Patricia trie 를 이용하여 단어에 대한 인덱싱 정보를 검색한다.

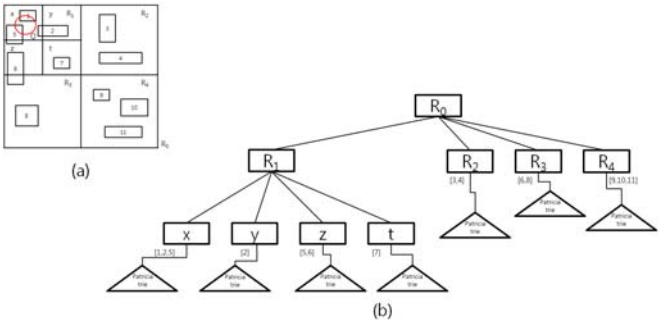

그림 1 QP-tree 의 구조

그림 1 은 QP-tree 의 구조를 보여준다. 그림 1 의

(a)는 QP-tree 가 객체의 수가 5 개가 넘는다고 했을 때의 분할된 모습을 보여준다. R_0가 4 개의 사각형 R_1, R_2, R_3, R_4로 분할이 되어 있고 다시 R_1은 또 x, y, z, t 로 분할이 되어 있다. 이 영역에 소셜 객체 1,2,3, … 11 이 들어가 있는 모습을 보여준다. 그림 1 의(b)는 실제 트리의 모습을 보여준다. x 노드에 있는 Patricia trie 에는 1, 2, 5 번 문서에 있는 단어에 대한 정보가 들어있고, z 노드에는 5,6 번 문서에 있는 단어에 대한 정보가 들어있다. 1 번 문서에 "서울의 맛있는 중국집" 이라는 글이 있다면, <"서울", "중국집"> 두개의 단어와 그림 1 의 (a)에 있는 Q 영역으로 질의를 하면 먼저 Q 영역을 포함하는 영역인 R_0 에서 R_1 을 검색하고 R_1 에서 x 영역을 검색한다. 마지막으로 x 에 있는 patricia trie 를 통하여 x 에 있는 문서들 중에 <"서울", "중국집">을 포함하고 있는 1 번 문서를 검색할 수 있다.

3.2 삽입

QP-tree 의 삽입(insert)과정은 다음과 같다. 삽입 과정에는 문서의 위치를 나타내는 geo(GPS 좌표), 문서의 아이디(doc_id), 문서정보가 필요하다. QP-tree 는 트리의 루트부터 삽입 연산을 시작한다. 일단 Node 가 자식 노드를 갖고 있는지 확인한다. 확인 후 자식 노드가 있다면 Node 의 자식노드들 중에서 문서의 위치를 포함하는 노드를 찾고, 찾은 노드를 이용하여 다시 삽입 연산을 실행한다. 자식 노드를 갖고 있지 않다면 현재 노드에 삽입하면된다. Quadratic tree 의 노드는 노드 안에 객체의 수를 가지고 분할을 결정하기 때문에 현재 노드의 객체 수를 비교하여 삽입이 가능하다면 단어 정보를 patricia trie 에 삽입한다. 현재 노드의 객체수가 가득 차있다면 노드는 분할을 해야할 경우, 노드의 분할을 알리는 isSplited 를 이용하여 노드가 분할하였음을 체크하고 노드를 4 개로 분할한다. 마지막으로 노드가 분할되였을 경우, 다시 객체 정보를 포함할 수 있는 노드를 검색하여 노드를 삽입한다.

3.3 검색

데이터 검색은 사용자가 입력한 위치정보를 이용하여 QP-tree 의 말단 노드를 찾는다. 말단 노드가 검색이 되면 해당 말단노드의 Patricia trie 를 이용하여 사용자가 검색하길 원하는 단어가 포함된 문서의 id 를 검색해온다. 사용자가 범위 검색을 할 경우에는 사용자가 입력한 범위가 포함되는 모든 말단 노드를 검색하고 검색된 말단 노드들에 있는 모든 Patricia trie 를 통하여 사용자가 찾고자 하는 단어가 포함된 모든 문서의 id 를 검색한다.

4. 결론

본 연구에서는 공간 웹 객체의 검색을 위한 QP-tree 와 PR-tree 를 제안하였다. 제안하는 기법을 사용하면 단어 뿐만 아니라 공간정보도 같이 활용을 하여 검색을 할 수 있다. 향후 연구로 공간 웹 객체의 집계질의를 위한 집계방법을 연구할 것이다.

Acknowledgment

이 논문은 2013 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 일반연구자지원사업의 결과물임(NRF-2013R1A1A4A01013416)

참고문헌

[1] B. S. Yoo, H. Y. Bae, "A Hybrid Index based on Aggregation R-tree for Spatio-Temporal Aggregation," Jornal of KIISE:Database, vol.33, no.5, pp.463-475, Oct. 2006 (in Korea)

[2] D. Padadias, P. kalnis, J. Zhang, and Y. Tao, "Efficient OLAP Operations in Spatial Data Warehouse." Technical Report: HKUST-CS01-01, University of Science & Technology, Hon Kong, 2001

[3] D. Papadias, Y. Tao, P. Kalnis, and J. Zhang. "Indexing Spatio-Temporal Data Warehouses," In Porc. of 18th IEEE International Conference on Data Engineering, pp. 166-175, 2002

[4] D. Wu, G. Cong, And C. S. Jensen, "A framework for efficient spatial web object retrieval", VLDB, 21(6), pp. 797-822, 2012.

[5] F. Rao, L. Zhang, X. L. Yu, Y. Li and Y. Chen, "Spatial Hierarchy and OLAP-Favored Search in Spatial Data Warehouse," In Proc. Data Warehousing and Knowledge Discovery, pp. 35-44, 2003.