하둡 맵리듀스를 이용한 웹 스케일 수준의 공간 지식 추출기 설계

이석준, 김인철
경기대학교 컴퓨터과학과
e-mail:{20151101162, kic}@kgu.ac.kr

Design of a Web-Scale Spatial Knowledge Extractor Using Hadoop MapReduce

Seokjun Lee, Incheol Kim
Dept of Computer Science, Kyonggi University

요 약

최근 들어 공간 지식을 활용한 다양한 서비스들이 개발됨에 따라, 공간 객체들 간의 정성적 공간 관계를 표현한 정성 공간 지식의 수요가 크게 늘어나고 있다. 공간 객체 각각의 세부 정보를 담은 대용량의 공간 데이터들은 개방화가 점차 확대되고 있으나, 공간 객체들 간의 정성적 관계를 표현한 정성 공간 지식은 상대적으로 확보하기 어려운 실정이다. 본 논문에서는 하둡 맵리듀스 병렬 분산 컴퓨터 환경을 이용해, 대용량의 공간 데이터로부터 공간 객체들 간의 위상 관계와 방향 관계를 나타내는 정성 공간 지식을 자동으로 추출하는 공간 지식 추출기를 제안한다. 본 논문에서 제안하는 대용량의 공간 지식 추출기는 맵리듀스 프레임워크를 기반으로 R-트리 색인과 범위 질의들을 효과적으로 이용함으로써, 웹 스케일 수준의 정성 공간 지식을 매우 효율적으로 추출해낸다. Open Street Map (OSM) 공개 데이터를 이용한 성능 분석 실험을 통해, 본 논문에서 제안하는 대용량 공간 지식 추출기의 높은 성능을 확인할 수 있었다.

1. 서론

최근 들어 Open Street Map(OSM), USGS, OS Open Data 과 같은 대용량의 공간 데이터들의 개방화가 가속화됨에 따라, 이들을 활용한 다양한 형태의 공간 정보 서비스들도 함께 늘어나고 있는 추세이다. 최근에 개발되는 많은 공간 정보 서비스들은 공간 객체 각각에 관한 구체적인 정보를 담은 공간 데이터(spatial data)들을 필요로 하는 경우도 있지만, 공간 객체들 간의 관계를 보다 직관적이고 함축적으로 표현한 정성 공간 지식(qualitative spatial knowledge)을 요구하는 경우도 많다. 하지만, 일반적으로 대용량의 공간 지식베이스 구축 작업은 숙련된 지식 공학자(knowledge engineer)들의 수작업이 반드시 필요한 고수준의 작업이기 때문에 자동화하기 어렵고, 이러한 이유에서 공간 정보 서비스 개발에 이용할 양질의 공간 지식베이스를 확보하기는 쉽지 않은 실정이다.

이러한 정성 공간 지식의 결핍 문제를 해결하기 위해 시도해볼 수 있는 대표적인 방법으로는 정성 공간 추론(qualitative spatial reasoning)을 이용해 기존의 공간 지식베이스를 확장하는 방법[5]과 기계 학습(machine learning)과 자연어 처리(natural language processing) 기술을 이용해 웹 문서로부터 공간 지식을 추출해내는 방법[6] 등이 있다. 하지만 정성 공간 추론을 통한 지식 베이스 확장 방법은 어느 정도 충분한 양의 초기 공간 지식을 확보하고 있을 때만 가능하다는 한계점이 있다. 또한, 기계 학습과 자연어 처리 기법을 이용한 공간 지식 추출 방법은 현재 기술 수준으로는 정확도와 신뢰도가 충분치 않아 아직 실용화하기 어려운 실정이다.

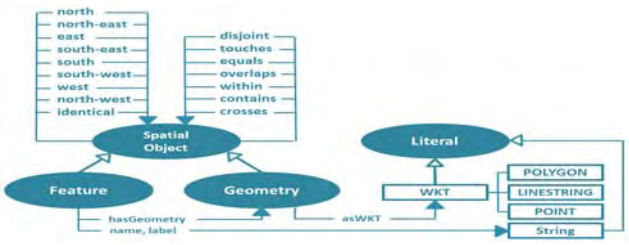
한편, 선행 연구[1]에서는 정성 공간 지식 결핍 문제를 해결하기 위한 새로운 시도로서, 표준 모델[3]을 따르는 공간 객체 각각의 기하학적 데이터로부터 두 공간 객체들 간의 공간 관계를 나타내는 정성 공간 지식을 추출해내는 방법을 제안하였다. 연구[1]에서 제시한 공간 지식 추출 방법은 단일 컴퓨터 환경에서 수행하는 것을 가정하였으므로, 웹 스케일 수준의 대용량 공간 지식 추출기로는 한계가 있다. 이러한 한계점을 극복하고자 본 논문에서는 하둡 맵리듀스(Hadoop MapReduce) 병렬 분산 컴퓨터 환경을 이용해, 대용량의 정성 공간 지식을 자동으로 추출해내는 공간 지식 추출기를 제안한다. 본 논문에서 제안하는 대용량의 공간 지식 추출기는 맵리듀스 프레임워크를 기반으로 R-트리 색인과 범위 질의들을 효과적으로 이용함으로써, 웹 스케일 수준의 정성 공간 지식을 매우 효율적으로 추출해낸다. 본 논문에서 제안하는 하둡 맵 리듀스 기반의 공간 지식 추출기의 성능을 분석하기 위해, Open Street Map(OSM) 공개 데이터를 이용한 실험들을 수행하고 그 결과를 소개한다.

※ 본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음.
[10044494, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발]

2. 정성 공간 지식 추출

2.1 공간 지식 표현

정성 공간 지식 추출 방법을 설명하기에 앞서 공간 객체의 기하학적 데이터와 공간 지식이 어떻게 표현되는지, 즉 공간 지식 표현 체계를 먼저 정의할 필요가 있다. 본 논문에서 공간 지식은 시맨틱 웹 표준 온톨로지(ontology) 언어인 RDF/OWL로 표현되고, 특히 표준 공간 질의 언어인 OGC GeoSPARQL[3]에서 정의한 클래스(class)들과 서술자(property)들을 사용하여 공간 객체들의 성질과 관계들이 표현되는 것으로 가정한다. 하지만 GeoSPARQL에서는 두 공간 객체들 간의 위상 관계 서술자(topological property)들은 정의되어 있으나, 방향 관계 서술자(directional property)들은 정의되어 있지 않다. 따라서 본 논문에서는 GeoSPARQL의 핵심 온톨로지에 CSD(Cone-Shaped Directional relations)-9 이론에서 정의한 9가지 방향 관계 서술자들을 추가하여 (그림 1)과 같은 공간 지식 표현 체계를 가정하였다.



(그림 1) 공간 지식 표현 체계

(그림 1)에서 공간 객체(spatial object)는 모든 공간 객

체들을 나타내는 최상위 클래스(class)이다. 두 공간 객체들 사이의 경계 및 포함 관계는 분리(disjoint), 맞닿음(touches) 등 총 7가지 위상 관계 서술자(topological property)들로 표현 가능하며, 두 공간 객체들 사이의 방향 관계는 북(north), 북동(north-east) 등 총 9가지 방향 관계 서술자(directional property)들로 표현할 수 있다. 한편, 공간 객체의 하위 클래스로 피처(feature)와 지오메트리(geometry)가 있으며, 피처는 실세계에서 도시, 도로, 건물과 같은 특정한 장소를 의미하고 반면에 지오메트리는 점, 선, 면과 같은 피처의 기하학적 데이터를 나타낸다. 그리고 피처는 문자열(string) 형태의 리터럴로 표현되고 반면에 지오메트리의 기하학적 데이터는 WKT(well-known text) 형태의 리터럴(literal)로 표현된다.

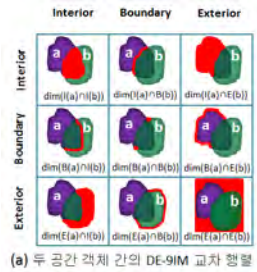
```
georesource:2297418
        a               geo:Feature   ;
        name            "Seoul"       ;
        geo:hasGeometry georesource:geom_2297418    .

georesource:200227274
        a               geo:Feature   ;
        name            "HanGang"     ;
        geo:hasGeometry georesource:geom_200227274  .

georesource:geom_2297418
        a               geo:Geometry ;
        geo:asWKT       "<http://www.opengis.net/def/crs/OGC/1.3/CRS84>
                        POLYGON((126.7639878 37.5549376, ...))"
                        ^^<http://www.opengis.net/ont/sf#wktLiteral>   .

georesource:geom_200227274
        a               geo:Geometry ;
        geo:asWKT       "<http://www.opengis.net/def/crs/OGC/1.3/CRS84>
                        LineString(127.3059215 37.5160799, ...))"
                        ^^<http://www.opengis.net/ont/sf#wktLiteral>   .

georesource:200227274
        geo:crosses     georesource:2297418
```

(그림 2) turtle 형식의 공간 지식 예시

(그림 2)는 공간 지식 표현 체계에 따라 기술된 공간 지식의 한 예를 나타낸다. 이 예에서 서울과 한강은 각각 별도의 지오메트리를 통해 자신의 기하학적 데이터를 가지고 있다. 즉, 서울의 지오메트리는 2차원 좌표들로 구성된 면 형태(geom_2297418 asWKT POLYGON)의 리터럴로 표현되어 있고, 한강의 지오메트리 역시 2차원 좌표들로 구성된 선 형태(geom_200227274 asWKT LineString)의 리터럴로 표현되어 있다. 한편, (그림 2)는 "한강은 서울을 가로지른다(Hangang crosses Seoul)"라는 정성 공간 지식도 포함하고 있다. 본 논문에서는 이와 같은 서울과 한강의 기하학적 데이터로부터 둘 간의 위상 관계와 방향 관계를 분석해내고 이들을 앞서 정의한 관계 서술자들로 표현한 정성 공간 지식을 자동 생성하는 방법을 제시한다.

2.2 R-트리 기반의 위상 관계 추출

이 절에서는 3x3 교차 행렬(3x3 intersection matrix)을 이용하는 DE-9IM(dimensionally extended nine-intersection model)모델을 토대로, R-트리를 이용하여 특정 범위 내에 존재하는 다수의 공간 객체들을 대상으로, 효율적으로 위상 관계 지식을 추출해내는 방법도 소개한다.



(a) 두 공간 객체 간의 DE-9IM 교차 행렬

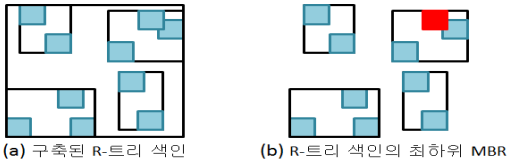
Topological Relationship	Applies to Geometry Types	DE-9IM Intersection Pattern
disjoint	All	(FF*FF****)
touches	All except P/P	(FT*******, F**T*****, F***T****)
equals	All	(TFFFTFFFT)
overlaps	A/A, P/P, L/L	(T*T***T**) for A/A, P/P; (1*T***T**) for L/L
within	All	(T*F**F***)
contains	All	(T*****FF*)
crosses	P/L, P/A, L/A, L/L	(T*T***T**) for P/L, P/A, L/A; (0********) for L/L

(b) 위상 관계 결정 표

(그림 3) DE-9IM 교차 행렬 예시와 위상 관계 판정표

두 공간 객체 각각의 기하학적 데이터가 주어졌을 때, 두 공간 객체 간에 만족하는 7 가지 위상 관계(disjoint, touches, equals, overlaps, within, contains, crosses) 중 하나를 판별하는 대표적인 종래의 방법은 DE-9IM 교차 행렬을 이용하는 것이다. (그림 3) (a)는 면 형태의 두 기하학적 데이터 A와 B의 DE-9IM 교차 행렬을 나타낸다. 행렬의 각 가로행과 세로열은 기하학적 데이터의 내부(interior), 경계선(boundary), 외부(exterior)를 의미하며 각 행렬 값은 가로행과 세로열의 교집합(∩)의 차원(dimension)을 의미한다. 예컨대, (그림 3) (a)의 객체 a의 내부와 객체 b의 내부의 교집합은 면이므로 차원은 2(dim[I(a)∩I(b)]=2)이고 a의 내부와 b의 경계선의 교집합은 선이므로 차원은 1(dim[I(a)∩B(b)]=1)이다. 만약 교집합이 공집합(∅)이면 차원은 -1이 된다. 행렬의 결과는 각 차원을 위에서 아래, 왼쪽에서 오른쪽 순으로 나열한다. 따라서 행렬의 결과는 212101212 또는 불 방식(boolean)으로서 TTTTTTTTT로 표현할 수 있다. DE-9IM은 비교적 쉬운 교집합을 이용한 정형화된 방법으로 모든 점, 선, 면 형태의 두 기하학적 데이터가 가지는 위상 관계를 분류할 수 있다는 장점이 있다. DE-9IM 교차 행렬이 계산되면, (그림 3) (b)의 판정표에 따라 두 공간 객체 간의 위상 관계를 판별해낼 수 있다.

(그림 3) (b)는 DE-9IM 교차 유형(intersection pattern)별로 각각 대응하는 위상 관계 서술자들을 나타낸다. 따라서 두 공간 객체의 기하학적 데이터로부터 DE-9IM 교차 유형이 결정되면, 이것을 토대로 (그림 3) (b)에 따라 두 공간 객체들 사이에 만족되는 위상 관계와 서술자를 결정할 수 있다. 예컨대, DE-9IM(Seoul, Suwon)의 결과가 FFTFFTTTT이면, "서울은 수원과 서로 떨어져 있다(Seoul disjoint Suwon)"는 위상 관계 지식을 추출해낼 수 있다. 하지만 위상 관계 서술자는 서술자에 따라 적용 가능한 기하학적 데이터 형태(applies to geometry types)에 제약이 있다. (그림 3) (b)의 적용 가능한 기하학적 데이터 종류에서 P는 점, L은 선, A는 면, All은 이 세 가지 모두 적용 가능함을 의미한다. 예컨대, 공간 서술자 중 분리(disjoint)는 두 공간 객체의 기하학적 데이터 형태가 점, 선, 면 모두 제약 없이 추출 가능하며 중복(overlaps)은 두 공간 객체의 기하학적 데이터 형태가 면과 면(A/A), 점과 점(P/P), 선과 선(L/L)일 때만 추출이 가능하다. 한편, 본 연구에서는 이와 같은 DE-9IM 기반의 위상 관계 판별법을 기초로, 다수의 공간 객체들에 대한 R-트리 색인을 구축하고 이를 효과적으로 이용함으로써, 대용량의 정성적 공간 지식을 매우 효율적으로 추출해내는 방법을 제시한다.



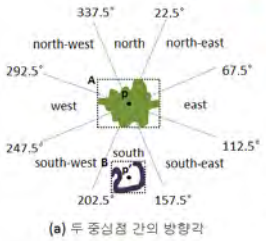
(a) 구축된 R-트리 색인 (b) R-트리 색인의 최하위 MBR

(그림 4) 공간 객체들의 MBR과 R-트리 색인 구축

(그림 4) (a)는 음영으로 표기된 사각형과 같이 위상 관계 지식을 얻고자 하는 다수의 공간 객체들로부터 구해진 MBR과, 이들을 기초로 구축한 R-트리 색인을 나타낸다. (그림 4) (b)는 구축된 R-트리 색인에서, 4개의 최하위 MBR들을 나타낸다. 최하위 MBR들은 각 MBR에 포함된 공간 객체들의 MBR과 디스크 상에 저장되어 있는 공간 객체들의 포인터 정보를 가지고 있다. 한편, R-트리 색인의 MBR들은 기준 객체를 포함하고 있는 MBR과 그렇지 않은 MBR로 구분될 수 있으며, (그림 4) (b)의 진하게 표기된 사각형과 같이 모든 공간 객체는 순차적으로 한 번씩 기준 객체로 선정되어 자신을 포함한 나머지 모든 공간 객체들과 위상 관계가 구해진다. 본 논문에서 제안하는 이 방법은 (그림 4) (b)와 같이 R-트리 색인을 이용해, 기준 객체를 포함하는 MBR과 그렇지 않은 MBR들을 쉽게 구별해내고, 기준 객체의 MBR과 겹치지 않는 MBR들 내에 속한 다수의 모든 공간 객체들은 DE-9IM 교차 행렬 계산을 생략하고 모두 기준 객체와 서로 떨어져 있다(disjoint)는 위상 관계로 쉽게 판별해낸다. 반면, 기준 객체의 MBR과 겹치는 MBR 내 소수의 객체들만 DE-9IM 교차 행렬을 구하여 위상 관계를 판별함으로써 계산의 효율성을 매우 높였다.

2.3 R-트리 기반의 방향 관계 추출

이 절에서는 공간 객체를 둘러싸는 MBR을 구하고, 이들의 중심점들이 놓인 방향각에 따라 임의의 두 공간 객체들 사이의 방향 관계 지식을 추출하는 방법을 소개한다. 또한, R-트리를 이용하여 다수의 공간 객체들을 대상으로, 효율적으로 방향 관계 지식을 추출해내는 방법도 소개한다.



(a) 두 중심점 간의 방향각

Directional Relationship	Applies to Geometry Types	Angle
north	All	[0°~22.5°), [337.5°~360°)
north-east	All	[22.5°~67.5°)
east	All	[67.5°~112.5°)
south-east	All	[112.5°~157.5°)
south	All	[157.5°~202.5°)
south-west	All	[202.5°~247.5°)
west	All	[247.5°~292.5°)
north-west	All	[292.5°~337.5°)
identical	All	P = P' (no angle)

(b) 방향 관계 결정 표

(그림 5) 방향각 예시와 방향 관계 판정표

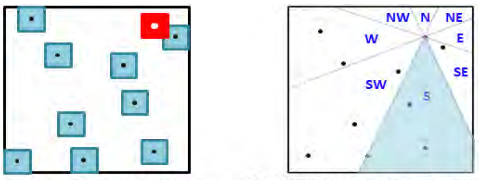
(그림 5) (a)는 점선으로 표기된 두 공간 객체 A, B의 MBR과 그 중심점 P, P′ 및 P를 기준으로 형성되는 콘(cone) 모양의 8가지 방향각을 나타낸다. 두 공간 객체 각

각의 가하학적 데이터가 주어졌을 때, 두 공간 객체 간에 만족하는 9 가지 방향 관계(north, north-east, east, south-east, south, south-west, west, north-west, identical)중 하나를 판별하는 대표적인 종래의 방법은 (그림 8)와 같이 두 공간 객체 각각을 둘러싸는 MBR을 구한 뒤, 두 MBR의 중심점들이 이루는 방향각을 계산하는 것이다. 동일 방향(identical)을 제외한 총 8가지 방향 관계 서술자 중 하나를 선정하기 위해서, (그림 5) (a)과 같이 방향각의 범위를 45°씩 360°를 8등분으로 분할하였다.

$$ANGLE(p, p') = \arctan\left(\frac{p'_y - p_y}{p'_x - p_x}\right) * 180/\pi \qquad (1)$$

두 MBR의 중심점이 이루는 방향각은 식 (1)과 같이 두 점 사이의 기울기의 역-탄젠트(arctan) 값을 이용하여 계산한다. MBR과 그 중심점을 이용하면 (그림 5)의 객체 B와 같이 불규칙적인 형태의 공간 객체들에 대해서도 쉽게 방향각을 구할 수 있다는 장점이 있다. 두 MBR 중심점이 이루는 방향각이 계산되면, (그림 5) (b)의 판정표에 따라 두 공간 객체 간의 방향 관계를 판별해낼 수 있다.

(그림 5) (b)는 방향 관계 서술자별로 각각 대응하는 MBR의 중심점들의 방향각을 나타낸다. 따라서 두 공간 객체의 기하학적 데이터로부터 각 공간 객체를 둘러싸는 MBR의 중심점들을 구하고 이들 간의 방향각을 계산하면, 이것을 토대로 (그림 5) (b)에 따라 두 공간 객체 사이에 만족되는 방향 관계와 서술자를 결정할 수 있다. 예컨대, 서울을 둘러싸는 MBR의 중심점과 수원의 MBR의 중심점이 이루는 방향각이 180°이면, "서울은 수원의 북쪽에 위치하고 있다(seoul north suwon)"는 방향 관계 지식을 추출해낼 수 있다. 한편, 본 연구에서는 이와 같은 MBR 기반의 방향 관계 판별법을 기초로, 다수의 공간 객체들에 대한 R-트리 색인을 구축하고 이를 효과적으로 이용함으로써, 대용량의 정성적 공간 지식을 매우 효율적으로 추출해내는 방법을 제시한다.



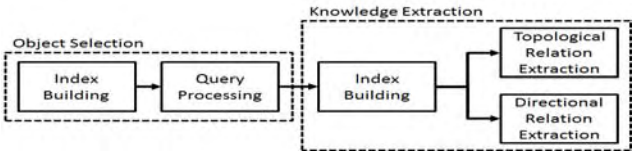
(a) R-트리 색인의 최상위 MBR (b) 방향각의 영역들과 범위 질의

(그림 6) 최상위 MBR과 공간 객체의 MBR들의 중심점

(그림 6) (a)는 구축된 R-트리 색인에서, 의미적으로 모든 공간 객체를 포함하는 최상위 MBR(root MBR)을 나타내고 (그림 6) (b)는 최상위 MBR의 내부에서 기준 객체 MBR의 중심점을 기준으로 형성되는 방향각의 영역(area)들(N, NE, E, SE, S, SW, W, NW)과 나머지 공간 객체 MBR들의 중심점을 나타낸다. 기준이 되는 공간 객체의 MBR의 중심점은 순차적으로 한 번씩 선정되며, 자신을 포함한 나머지 모든 공간 객체들의 MBR의 중심점과 방향 관계가 구해진다. 또한, 최상위 MBR 내의 방향각 영역은 최상위 MBR의 4개의 변과 방향각을 나타내는 4개의 방사선을 2차원 평면상의 1차 그래프로 가정하여 1차 그래프 간의 접점을 구하고, 각 접점을 토대로 다각형 모양의 영역을 구해낸다. 본 논문에서는 기준 객체를 중심으로 개별 객체의 방향각을 따로 따로 계산하는 방식 대신, (그림 6) (b)와 같이 공간 객체들의 R-트리 색인을 토대로, 9가지 방향 관계별로 기각의 영역을 이용한 범위 질의(range query)를 각각 수행함으로써, 질의 결과에 속한 객체들에 대해서는 한꺼번에 동일한 방향 관계를 효율적으로 판별해낸다. 예컨대, (그림 6) (b)에서 음영으로 표기된 남쪽(s) 영역에 포함되는 두 개의 중심점을 남쪽 영역의 범위 질의를 이용하여 한꺼번에 얻어 낼 수 있고 두 개의 중심점을 가지는 공간 객체는 기준 객체와의 방향각 계산을 생략하고 모두 기준 객체의 "남쪽에 있다"고 쉽게 판별해낸다. 한편, (그림 6) (b)의 남쪽 영역과 같이 방향각 영역의 경계선(boundary) 상에 존재하는 중심점이 존재할 수가 있는데, 이 경우에는 실제로 방향각을 계산해야만 정확한 방향 관계를 알 수 있다. 하지만 이런 경우는 매우 드물어 계산 시간에 큰 영향을 주지 않는다.

3. 맵리듀스 기반의 공간 지식 추출기 설계

본 논문에서는 웹 스케일 수준의 정성 공간 지식을 추출하기 위해, 하둡 맵리듀스 기반의 대용량 공간 지식 추출기를 설계하였다. 대용량 공간 지식 추출기의 작업 흐름도는 (그림 7)과 같이 크게 객체 선택(object selection)과 지식 추출(knowledge extraction) 단계로 나뉜다.

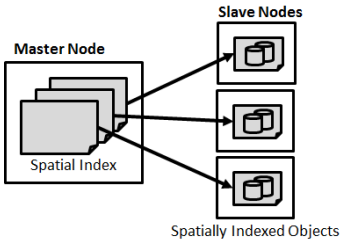


(그림 7) 작업 흐름도

공간 지식을 추출하기에 앞서, 기준 객체(base object)와의 공간 지식을 추출하고자 하는 대상 객체(target object)들을 선택해야한다. 먼저, 객체 선택 단계에서는 전체 공간 객체에 대해 색인을 구축(index building)하고 이를 토대로 범위 질의나 k-최근접 이웃과 같은 공간 질의(spatial query)를 수행하여 대상 객체들을 선택한다. 대상 객체들을 선택하면, 지식 추출 단계에서는 선택한 대상 객체들에 대해 색인을 구축(index building)하고 색인을 이용하여 효율적으로 위상 관계 지식(topological relation extraction)과 방향 관계 지식을 추출(direction relation extraction)한다.

3.1 R-트리 색인

본 논문에서 제안하는 대용량 공간 지식 추출기는 R-트리를 효과적으로 이용함으로써, 정성 공간 지식을 효율적으로 추출해낸다. (그림 8)은 SpatialHadoop[3]에서 지원하는 하둡 맵리듀스 기반의 R-트리 색인을 나타낸다.

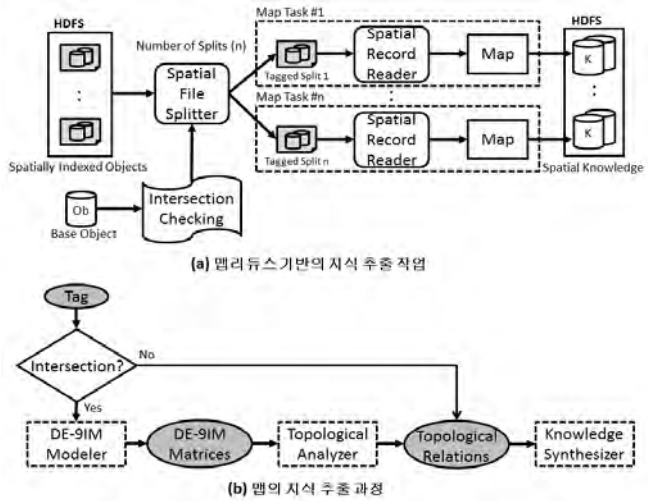


(그림 8) 하둡 맵리듀스 기반의 R-트리 색인

(그림 8)과 같이 색인(spatial index)에 대한 정보는 마스터 노드(master nonde)에 저장되고 색인된 공간 객체들(spatially indexed objects)은 슬레이브 노드들(slave nodes)에 분산되어 저장된다. 특히, 각 슬레이브 노드들에 위치한 공간 객체들은 각각의 노드들에 분산되기 위해 하둡 블록 크기(block size)만큼 할당되어 저장된다. 즉, R-트리의 최하위 MBR들은 하둡 블록 크기 단위로 구성된다.

3.2 위상 관계 추출 작업

본 논문에서 제안하는 하둡 맵리듀스 기반의 위상 관계 추출 작업은 (그림 9)와 같다.



(그림 9) 위상 관계 추출 작업

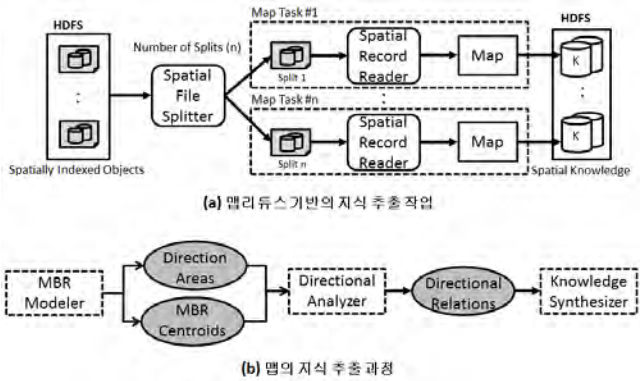
(그림 9) (a)는 하둡 맵리듀스 기반의 위상 관계 추출 작업을 나타낸다. 먼저, 공간 파일 분열기(spatial file splitter)는 앞서 구축된 R-트리 색인을 참조하여, HDFS 상에서 하둡 블록 크기만큼 할당되어 색인된 객체들을 각각의 맵 작업(map task)에 전달한다. 특히, 공간 파일 분열기는 색인된 객체들의 MBR이 기준 객체와 교차되어

있는가를 판별하여 꼬리표(tag)를 붙인다. 따라서, 꼬리표가 붙은 색인된 객체들(tagged splits)이 각각의 맵(map)에 전달되어 동시다발적으로 위상 관계 추출 작업이 진행된다. (그림 9)의 (b)는 각각의 맵(map)에서 수행되는 위상 관계 지식 추출 과정을 나타낸다. 맵은 우선적으로 꼬리표를 확인하여 기준 객체와 교차되어 있는 경우와 그렇지 않은 경우로 나누어 지식을 추출한다. 먼저, "기준 객체와 교차되어 있다"는 꼬리표의 경우에는, 기준 객체와 대상 객체 간의 DE-9IM 행렬을 계산하여 위상 관계를 판별하고, "기준 객체와 교차되어 있지 않다"는 꼬리표의 경우에는 DE-9IM 행렬 계산을 생략하고 모두 기준 객체와 떨어져 있다고 판별한다.

이러한 분산된 위상 관계 추출 작업은 사전에 지역성이 보장되는 R-트리를 토대로, 색인된 대상 객체들의 MBR이 기준 객체와 교차되어 있는가를 확인하기 때문에 교차되지 않은 MBR의 대상 객체들에 대해서는 계산을 생략하고 매우 효율적으로 위상 관계를 판별할 수 있다. 또한, HDFS 블록 크기 단위로 R-트리 색인을 구축했기 때문에 색인된 공간 객체들이 각각 부하 분산되어 동시다발적으로 위상 관계를 계산하므로 뛰어난 확장성을 가진다.

3.3 방향 관계 추출 작업

본 논문에서 제안하는 하둡 맵리듀스 기반의 방향 관계 추출 작업은 (그림 10)과 같다.



(그림 10) 방향 관계 추출 작업

(그림 10) (a)는 하둡 맵리듀스 기반의 분산된 방향 관계 추출 작업을 나타낸다. 먼저, 공간 파일 분열기(spatial file splitter)는 앞서 구축된 R-트리 색인을 참조하여, HDFS 상에서 하둡 블록 크기만큼 할당되어 색인된 객체들을 각각의 맵 작업(map task)에 전달한다. 따라서, 전달된 객체들이 각각의 맵에 전달되어 동시다발적으로 방향 관계 추출 작업이 진행된다. (그림 10) (b)는 각각의 맵(map)에서 수행되는 방향 관계 지식 추출 과정을 나타낸다. 먼저, MBR 모델러(MBR modeler)는 방향각 영역들(direction areas)과 공간 객체의 MBR의 중심점들(MBR centroids)을 계산하고, 방향각의 영역들을 토대로 MBR의 중심점들에 대해 범위 질의를 수행하여 방향 관계를 판별한다.

이러한 분산된 방향 관계 추출 작업은 사전에 R-트리의 최상위 MBR을 토대로 방향각의 영역을 구하고, 범위 질의를 이용하여 일괄적으로 방향 관계를 판별하기 때문에 방향각을 직접 계산하는 것보다 매우 효율적이다. 또한, HDFS 블록 사이즈 단위로 R-트리 색인을 구축했기 때문에 색인된 공간 객체들이 각각 부하 분산되어 동시다발적으로 방향 관계를 계산하므로 뛰어난 확장성을 가진다.

4. 성능 실험

본 논문에서는 Open Street Map(OSM) 공개 데이터를 이용하여 공간 지식 추출기의 성능 분석 실험을 수행하였다.

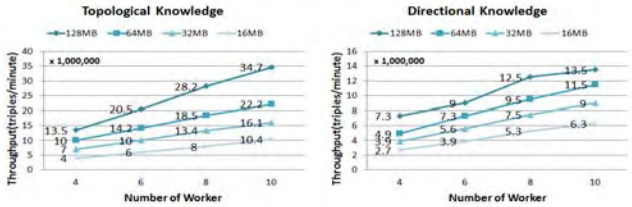
<표 1> 실험 데이터

Dataset	Capacity	Point(n)	Linestring(n)	Polygon(n)	Total(n)
OSM roads	24.1GB	3,529	2,111,1934	70,224,463	91,339,926

실험에 이용한 데이터는 <표 1>과 같다. Open Street Map은 전 세계 범위의 행정구역, 도로, 하천 등의 대규모 공간 데이터들을 포함하고 있으나, 본 실험에서는 전 세계 범위의 도로와 관련된 데이터들만 주로 이용하였다. 본 논문에서는 크게 하둡 맵리듀스 환경에서, 방법에 따른 공간 지식 추출기의 효율성과 작업 노드 수 증가에 따른 공간 지식 추출기의 확장성을 분석해보는 실험들을 수행하였다.

(그림 11) 방법에 따른 처리량 비교

(그림 11)은 방법에 따른 공간 지식 추출기의 처리량(throughput)을 기준 객체의 수(number of base object)를 증가하면서 비교하였다. 위상 관계 지식 추출의 경우 R-트리를 이용하여 색인된 객체들의 MBR과 기준 객체 간의 교차 검사(intersection checking)를 이용한 방법이 이 방법을 이용하지 않았을 때(no checking)보다 더 많은 처리량을 보였다. 교차 검사를 이용한 방법은 분당 평균 2천 7백만의 지식을 추출했고, 교차 검사를 이용하지 않은 방법은 분당 평균 1천 9백 6십만의 지식을 추출했다. 마찬가지로 방향 관계 지식 추출도 교차 검사를 이용한 방법이 더 많은 처리량을 보였다. 교차 검사를 이용한 방법은 분당 평균 2천 1백 1십만의 지식을 추출했고, 교차 검사를 이용하지 않은 방법은 분당 1천 2백만의 지식을 추출했다. (그림 11)의 실험 결과를 통해, 본 논문에서 제안하는 분산 공간 지식 추출기의 높은 효율성을 확인할 수 있었다.



(그림 12) 작업 노드 수에 따른 처리량 비교

(그림 12)는 HDFS 블록 크기 단위 별로, 작업 노드의 수를 증가하면서 공간 지식 추출기의 처리량을 비교하였다. 위상 관계 추출 시간과 방향 관계 추출 시간 모두 HDFS 블록 크기가 커질수록, 작업 노드 수가 많아질수록, 더 좋은 성능을 보였다. (그림 12)의 실험 결과를 통해 본 논문에서 제안하는 분산 공간 지식 추출기의 높은 확장성을 확인할 수 있었다.

6. 결론

본 논문에서는 하둡 맵리듀스 병렬 분산 컴퓨팅 환경을 이용해, 대용량의 공간 데이터로부터 공간 객체들 간의 위상 관계와 방향 관계를 나타내는 정성 공간 지식을 자동으로 추출하는 공간 지식 추출기를 제안하였다. 본 논문에서 제안한 대용량의 공간 지식 추출기는 맵리듀스 프레임워크를 기반으로 R-트리 색인과 범위 질의들을 효과적으로 이용함으로써, 웹 스케일 수준의 정성 공간 지식을 매우 효율적으로 추출해낸다. 향후 연구로는 아파치 스파크(Apache Spark)와 같은 인-메모리 병렬 프로그램 개발 환경을 이용해 좀 더 효율성이 높은 공간 지식 추출기를 구현해보는 연구를 계획하고 있다.

참고문헌

[1] S. J. Lee, et al. "Spatial Knowledge Extraction from Geometric Data", Proc. of KIISE Winter Conference, pp.1337-1339, 2014.
[2] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on large Clusters", Communications of the ACM, Vol.51, No.1, pp.107-113, 2008.
[3] A. Eldawy and M. F. Mokbel, "SpatialHadoop: A MapReduce Framework for Spatial Data", in ICDE, 2015.
[4] R. Battle and D. Kolas, "Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL", Semantic Web Journal, Vol.3, No.4, pp.355-370, 2012.
[5] M. Egenhofer, "Qualitative spatial-relation reasoning for design", Studying Visual and Spatial Reasoning for Design Creativity, pp.153-175, 2015.
[6] L. Dong, et al. "Extraction of Spatial Information from Chinese Story Text for Animation Generation", Journal of Computational Information Systems, Vol.10, No.23, pp.10283-10292, 2014.