(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0089036
(43) 공개일자 2021년07월15일

(51) 국제특허분류(Int. Cl.)
G06F 16/27 (2019.01) *G06F 16/28* (2019.01)
(52) CPC특허분류
G06F 16/278 (2019.01)
G06F 16/285 (2019.01)
(21) 출원번호 10-2020-0002253
(22) 출원일자 2020년01월07일
심사청구일자 2021년07월07일

(71) 출원인
한국전력공사
전라남도 나주시 전력로 55(빛가람동)
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
이규철
대전광역시 유성구 배울2로 133, 205동 203호 (용산동, 경남아너스빌 2단지 아파트)
김인아
대전광역시 유성구 대학로 131, 201호 (궁동)
이동재
서울특별시 노원구 석계로15길 25(월계동, 월계1차한일아파트)
(74) 대리인
특허법인 정안

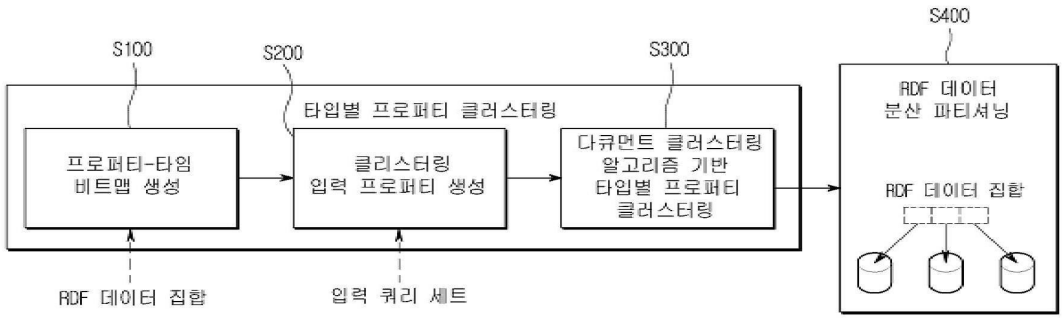
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **클러스터링 알고리즘을 사용한 RDF 데이터 분산 파티셔닝 방법 및 시스템**

(57) 요 약

본 발명의 RDF 데이터 분산 파티셔닝 방법은, RDF 데이터 집합을 분석하여 프로퍼티-타입 비트맵을 형성하는 단계; 입력 쿼리 세트에 상기 프로퍼티-타입 비트맵을 적용하여 타입-프로퍼티들을 생성하는 단계; 상기 타입-프로퍼티들을 클러스터링하는 단계; 및 클러스터링 결과에 따라 상기 RDF 데이터 집합를 분산 저장하는 단계를 포함할 수 있다. 여기서, 상기 RDF 데이터 집합은, 스마트 미터로부터 수집된 데이터로 타임스탬프를 포함하고 있는 미터 데이터와, 미터 데이터의 주체가 되는 사용자, 디바이스 등에 대한 기본 정보를 포함하고 있는 마스터 데이터와, 미터 데이터와 마스터 데이터 기반 온라인 배치 분석 결과를 나타내는 통계 데이터의 각 단위 데이터가 주어, 서술어, 목적어 3개의 트리플을 구성한 것의 집합일 수 있다.

대 표 도



명 세 서

청구범위

청구항 1

RDF 데이터 집합을 분석하여 프로퍼티-타입 비트맵을 형성하는 단계;

입력 쿼리 세트에 상기 프로퍼티-타입 비트맵을 적용하여 타입-프로퍼티들을 생성하는 단계;

상기 타입-프로퍼티들을 클러스터링하는 단계; 및

클러스터링 결과에 따라 상기 RDF 데이터 집합를 분산 저장하는 단계

를 포함하는 RDF 데이터 분산 파티셔닝 방법.

청구항 2

제1항에 있어서,

상기 RDF 데이터 집합은,

스마트 미터로부터 수집된 데이터로 타임스탬프를 포함하고 있는 미터 데이터와, 미터 데이터의 주체가 되는 사용자, 디바이스 등에 대한 기본 정보를 포함하고 있는 마스터 데이터와, 미터 데이터와 마스터 데이터 기반 온라인 배치 분석 결과를 나타내는 통계 데이터의 각 단위 데이터가 주어, 서술어, 목적어 3개의 트리플을 구성한 것의 집합인 RDF 데이터 분산 파티셔닝 방법.

청구항 3

제2항에 있어서,

상기 프로퍼티-타입 비트맵을 형성하는 단계에서는,

독출된 RDF 데이터 집합의 각 단위 트리플의 주어들을 타입들로 결정하고, 서술어들을 프로퍼티로 결정하는 RDF 데이터 분산 파티셔닝 방법.

청구항 4

제1항에 있어서,

상기 프로퍼티-타입 비트맵을 형성하는 단계에서는,

일축이 상기 추출/결정된 프로퍼티들이고 나머지 축이 상기 추출/결정된 타입들인 테이블을 형성하고, 각 교차점에 상기 RDF 데이터 집합에 각 교차점에 따른 프로퍼티-타입 페어의 존재여부를 표시하는 테이블을 작성하는 RDF 데이터 분산 파티셔닝 방법.

청구항 5

제1항에 있어서,

상기 타입-프로퍼티들을 생성하는 단계에서는,

특정 쿼리에 주어가 명확하지 않고, 상기 특정 쿼리가 속한 쿼리 그래프 세트에도 주어가 명확하지 않으면, 해당 쿼리 그래프 세트에 포함된 프로퍼티들이 상기 프로퍼티-타입 비트맵 상에서 공통적으로 가지고 있는 타입을 해당 타입으로 간주하는 RDF 데이터 분산 파티셔닝 방법.

청구항 6

제1항에 있어서,

상기 클러스터링 단계에서는,

다큐먼트 클러스터링(Document Clustering) 알고리즘을 이용하여, 쿼리 상에서 동시에 출현하는 빈도수가 높은 타입-프로퍼티 페어들을 클러스터링하는 RDF 데이터 분산 파티셔닝 방법.

청구항 7

RDF 데이터 집합이 저장된 RDF 데이터 저장부;

2개 이상의 저장 노드들로 이루어진 분산 저장부;

사용자의 입력 쿼리 세트가 저장된 쿼리 저장부;

RDF 데이터 집합를 분석하여 프로퍼티-타입 비트맵을 형성하는 비트맵 형성부;

입력 쿼리 세트에 상기 프로퍼티-타입 비트맵을 적용하여 타입-프로퍼티들을 생성하고, 상기 타입-프로퍼티들을 클러스터링하는 클러스터링부; 및

상기 타입-프로퍼티들의 클러스터링 결과에 따라 상기 RDF 데이터 집합를 상기 분산 저장부에 저장하는 분산부

를 포함하는 RDF 데이터 분산 파티셔닝 시스템.

청구항 8

제7항에 있어서,

상기 RDF 데이터 집합은,

스마트 미터로부터 수집된 데이터로 타임스탬프를 포함하고 있는 미터 데이터와, 미터 데이터의 주체가 되는 사용자, 디바이스 등에 대한 기본 정보를 포함하고 있는 마스터 데이터와, 미터 데이터와 마스터 데이터 기반 온라인 배치 분석 결과를 나타내는 통계 데이터의 각 단위 데이터가 주어, 서술어, 목적어 3개의 트리플을 구성한 것의 집합인 RDF 데이터 분산 파티셔닝 시스템.

청구항 9

제7항에 있어서,

상기 프로퍼티-타입 비트맵은,

일축이 상기 추출/결정된 프로퍼티들이고 나머지 축이 상기 추출/결정된 타입들인 테이블을 형성하고, 각 교차점에 상기 RDF 데이터 집합에 각 교차점에 따른 프로퍼티-타입 페어의 존재여부를 표시하는 테이블 형태로 작성되는 RDF 데이터 분산 파티셔닝 시스템.

청구항 10

제7항에 있어서,

상기 클러스터링부는,

특정 쿼리에 주어가 명확하지 않고, 상기 특정 쿼리가 속한 쿼리 그래프 세트에도 주어가 명확하지 않으면, 해당 쿼리 그래프 세트에 포함된 프로퍼티들이 상기 프로퍼티-타입 비트맵 상에서 공통적으로 가지고 있는 타입을

해당 타입으로 간주하는 RDF 데이터 분산 파티셔닝 시스템.

청구항 11

제7항에 있어서,

상기 클러스터링부는,

다큐먼트 클러스터링(Document Clustering) 알고리즘을 이용하여, 쿼리 상에서 동시에 출현하는 빈도수가 높은 타입-프로퍼티 페어들을 클러스터링하는 RDF 데이터 분산 파티셔닝 시스템.

청구항 12

제7항에 있어서,

상기 분산부는, 상기 RDF 데이터 집합을 상기 저장 노드들에 나누어 저장하되, 각 노드 저장소의 위치는 저장되는 클러스터에 속한 타입-프로퍼티 페어들에 대한 쿼리 처리 빈도가 높은데로 할당하는 RDF 데이터 분산 파티셔닝 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 RDF 클러스터링 데이터의 분산 저장 방법에 관한 것으로, 보다 구체적으로는 스마트 그리드 환경에서 데이터 사용자의 조회에 대한 대응 검색 속도를 높일 수 있도록 데이터들을 분산 파티셔닝을 적용하여 저장하는 RDF 클러스터링 데이터의 분산 저장 시스템 및 방법에 관한 것이다.

배 경 기 술

[0003] RDF 데이터는 엔티티(Entity; 주어(Subject), 목적어(Object)가 엔티티에 해당)와 엔티티를 연결하는 프로퍼티(서술어; Predicate; Property)로 구성된 트리플(Triple)을 가장 작은 단위의 구조로 가진다. 트리플이 서로 연결되어 더 큰 RDF 데이터를 구성하며, 이러한 트리플이 연결된 방식은 기존의 전통적인 데이터베이스인 관계형 데이터베이스(Relational Database) 저장 방식과 맞지 않아 많은 저장 방식 연구들이 진행되고 있다.

[0004] 상기 RDF 데이터를 이용한 데이터 저장 및 조회 기술은 최근 IOT 기술 등 신기술을 적용하여 그 범위를 확장하고 있는 스마트그리드 환경이나 이와 관련된 센서 네트워크 환경에 적용될 수 있다.

[0005] 다양한 신재생 에너지에 의한 발전이 가능해지고 소규모 발전사업자가 등장하며, 전력거래 시스템에 정보통신 기술을 융합한 스마트그리드 형태의 마이크로그리드가 도입되면서 소비자는 복수의 전력 판매자로부터 전력을 구매하는 것이 가능해지고 소비자가 직접 전력 판매자를 선택하는 것이 가능해졌다. 즉, 소규모 발전사업자가 태양광 발전 설비나 풍력 발전 설비를 운영하는 것이 이론적으로 가능하다.

[0006] 스마트 그리드의 데이터는 전력 발전, 설비제어, 운영, 서비스 등으로 구성되며, 다양한 도메인에 존재하는 정보와 연결을 통해 새로운 서비스 제공이 가능하다. 연결된 형태의 데이터는 여러 도메인의 자원을 연결하여 상위 레벨의 컨택스트 관리를 가능하게 하며, 특정 표준으로 정의된 온톨로지(Ontology)를 사용하여 표현이 가능하다.

[0007] 스마트 그리드 분야에서 온톨로지를 적용하는 사례로, 대표적으로 유럽전기통신표준협회 ETSI(European Telecommunications Standards Institute)의 ISG CIM 그룹과 TC SmartM2M 그룹이 있다. ISG CIM 그룹은 스마트 그리드를 위한 CIM(Common Information Model)의 OWL 프로파일, 온톨로지를 정의하고, 데이터의 상호 운용성을 위해 스마트 시티에 적용 가능한 공통 온톨로지를 정의하였다. 또한 TC SmartM2M 그룹은 스마트 시티에서 가전기기, 에너지, 환경, 빌딩 등을 위한 온톨로지를 설계하였으며, EU의 스마트 미터, 스마트 그리드와 같이 스마

트 시티의 주요 도메인을 위한 온톨로지 표준화를 진행하였다.

[0008] 이러한 온톨로지를 구성하기 위해 데이터는 RDF(Resource Description Framework) 표맷으로 작성된다. RDF 데이터는 엔티티(Entity; 주어(Subject), 목적어(Object)가 엔티티에 해당)와 엔티티를 연결하는 프로퍼티(서술어; Predicate; Property)로 구성된 트리플(Triple)을 가장 작은 단위의 구조로 가지며, 트리플이 서로 연결되어 더 큰 규모의 RDF 데이터를 구성한다.

[0009] 그러나, 이러한 데이터의 연결성은 분산 환경에서 높은 통신 비용을 발생시킨다. 스마트 그리드 데이터의 크기가 점차 증가하면서, 단일 저장소보다 몽고DB(MongoDB), HBase와 같은 분산된 저장소에 데이터를 나누어 저장하는 경우가 증가하고 있다. RDF 형식으로 저장된 데이터는 서로 연결된 형태를 가지며, 이로 인해 무작위로 데이터를 분산하는 것은 쿼리에 의한 데이터 검색 시 서로 다른 노드에서 데이터를 읽어와야 하기 때문에 결과적으로 높은 통신 비용이 발생하게 된다.

[0010] 이와 같이 분산된 노드에서의 통신 비용을 줄이기 위해서는 같은 쿼리에 의해 접근될 가능성이 높은 데이터를 같은 노드에 저장해야한다. RDF 데이터 분산 파티셔닝을 위해 데이터의 시맨틱 정보를 중심으로 분산 파티셔닝을 수행하는 여러 연구들이 제안되었다.

[0011] 분산 파티셔닝 연구들은 주로 RDF 데이터의 주어를 기준으로 파티셔닝을 수행하는데, 그 결과로 같은 주어를 가지는 연결된 데이터는 같은 노드에 위치하게 되지만, 쿼리의 워크로드를 고려하지 않은 파티셔닝으로 인해 다양한 쿼리에 최적화되지 않은 문제점을 가진다.

선행기술문헌

특허문헌

[0013] (특허문헌 0001) 대한민국 등록특허 10-1872414호

발명의 내용

해결하려는 과제

[0014] 본 발명은 스마트그리드 환경에서 검색 속도를 높이고 통신 비용을 절감할 수 있는 클러스터링 알고리즘을 사용한 RDF 데이터 분산 파티셔닝 방법 및/또는 시스템을 제공하고자 한다.

[0015] 구체적으로, 본 발명은 스마트그리드 환경에서 발생하는 다양한 쿼리 수행시 데이터가 분산 저장된 환경에서 통신 비용을 최소화하기 위해, 쿼리 워크로드가 고려된 클러스터링 알고리즘을 사용한 RDF 데이터 분산 파티셔닝 방법 및/또는 시스템을 제공하고자 한다.

과제의 해결 수단

[0017] 본 발명의 일 측면에 따른 RDF 데이터 분산 파티셔닝 방법은, RDF 데이터 집합을 분석하여 프로퍼티-타입 비트맵을 형성하는 단계; 입력 쿼리 세트에 상기 프로퍼티-타입 비트맵을 적용하여 타입-프로퍼티들을 생성하는 단계; 상기 타입-프로퍼티들을 클러스터링하는 단계; 및 클러스터링 결과에 따라 상기 RDF 데이터 집합를 분산 저장하는 단계를 포함할 수 있다.

[0018] 여기서, 상기 RDF 데이터 집합은, 스마트 미터로부터 수집된 데이터로 타임스탬프를 포함하고 있는 미터 데이터와, 미터 데이터의 주체가 되는 사용자, 디바이스 등에 대한 기본 정보를 포함하고 있는 마스터 데이터와, 미터 데이터와 마스터 데이터 기반 온라인 배치 분석 결과를 나타내는 통계 데이터의 각 단위 데이터가 주어, 서술어, 목적어 3개의 트리플을 구성한 것의 집합일 수 있다.

[0019] 여기서, 상기 프로퍼티-타입 비트맵을 형성하는 단계에서는, 독출된 RDF 데이터 집합의 각 단위 트리플의 주어들을 타입들로 결정하고, 서술어들을 프로퍼티로 결정할 수 있다.

[0020] 여기서, 상기 프로퍼티-타입 비트맵을 형성하는 단계에서는, 일축이 상기 추출/결정된 프로퍼티들이고 나머지

축이 상기 추출/결정된 타입들인 테이블을 형성하고, 각 교차점에 상기 RDF 데이터 집합에 각 교차점에 따른 프로퍼티-타입 페어의 존재여부를 표시하는 테이블을 작성할 수 있다.

[0021] 여기서, 상기 타입-프로퍼티들을 생성하는 단계에서는, 특정 쿼리에 주어가 명확하지 않고, 상기 특정 쿼리가 속한 쿼리 그래프 세트에도 주어가 명확하지 않으면, 해당 쿼리 그래프 세트에 포함된 프로퍼티들이 상기 프로퍼티-타입 비트맵 상에서 공통적으로 가지고 있는 타입을 해당 타입으로 간주할 수 있다.

[0022] 여기서, 상기 클러스터링 단계에서는, 다큐먼트 클러스터링(Document Clustering) 알고리즘을 이용하여, 쿼리 상에서 동시에 출현하는 빈도수가 높은 타입-프로퍼티 페어들을 클러스터링할 수 있다.

[0024] 본 발명의 다른 측면에 따른 RDF 데이터 분산 파티셔닝 시스템은, RDF 데이터 집합이 저장된 RDF 데이터 저장부; 2개 이상의 저장 노드들로 이루어진 분산 저장부; 사용자의 입력 쿼리 세트가 저장된 쿼리 저장부; RDF 데이터 집합를 분석하여 프로퍼티-타입 비트맵을 형성하는 비트맵 형성부; 입력 쿼리 세트에 상기 프로퍼티-타입 비트맵을 적용하여 타입-프로퍼티들을 생성하고, 상기 타입-프로퍼티들을 클러스터링하는 클러스터링부; 및 상기 타입-프로퍼티들의 클러스터링 결과에 따라 상기 RDF 데이터 집합를 상기 분산 저장부에 저장하는 분산부를 포함할 수 있다.

[0025] 여기서, 상기 RDF 데이터 집합은, 스마트 미터로부터 수집된 데이터로 타임스탬프를 포함하고 있는 미터 데이터와, 미터 데이터의 주체가 되는 사용자, 디바이스 등에 대한 기본 정보를 포함하고 있는 마스터 데이터와, 미터 데이터와 마스터 데이터 기반 온라인 배치 분석 결과를 나타내는 통계 데이터의 각 단위 데이터가 주어, 서술어, 목적어 3개의 트리플을 구성한 것의 집합일 수 있다.

[0026] 여기서, 상기 프로퍼티-타입 비트맵은, 일축이 상기 추출/결정된 프로퍼티들이고 나머지 축이 상기 추출/결정된 타입들인 테이블을 형성하고, 각 교차점에 상기 RDF 데이터 집합에 각 교차점에 따른 프로퍼티-타입 페어의 존재여부를 표시하는 테이블 형태로 작성될 수 있다.

[0027] 여기서, 상기 클러스터링부는, 특정 쿼리에 주어가 명확하지 않고, 상기 특정 쿼리가 속한 쿼리 그래프 세트에도 주어가 명확하지 않으면, 해당 쿼리 그래프 세트에 포함된 프로퍼티들이 상기 프로퍼티-타입 비트맵 상에서 공통적으로 가지고 있는 타입을 해당 타입으로 간주할 수 있다.

[0028] 여기서, 상기 클러스터링부는, 다큐먼트 클러스터링(Document Clustering) 알고리즘을 이용하여, 쿼리 상에서 동시에 출현하는 빈도수가 높은 타입-프로퍼티 페어들을 클러스터링할 수 있다.

[0029] 여기서, 상기 분산부는, 상기 RDF 데이터 집합을 상기 저장 노드들에 나누어 저장하되, 각 노드 저장소의 위치는 저장되는 클러스터에 속한 타입-프로퍼티 페어들에 대한 쿼리 처리 빈도가 높은데로 할당할 수 있다.

발명의 효과

[0031] 상술한 구성의 본 발명의 사상에 따른 클러스터링 알고리즘을 사용한 RDF 데이터 분산 파티셔닝 방법 및/또는 시스템을 실시하면, 스마트그리드 환경에서 검색 속도를 높이고 조회 활동에 대한 통신 비용을 절감할 수 있는 이점이 있다.

[0032] 구체적으로, 본 발명의 클러스터링 알고리즘을 사용한 RDF 데이터 분산 파티셔닝 방법 및/또는 시스템은, 쿼리에 의한 접근 빈도를 고려하고 다큐먼트 클러스터링 알고리즘을 사용하여 RDF 데이터를 분산 파티셔닝하고, 쿼리 워크로드를 고려하기 때문에, 스마트 그리드에서 발생하는 다양한 유형의 쿼리 수행 시 통신 비용을 최소화할 수 있는 이점이 있다.

[0033] 구체적으로, 본 발명의 클러스터링 알고리즘을 사용한 RDF 데이터 분산 파티셔닝 방법 및/또는 시스템은, 특정 타입이 유사 프로퍼티 셋을 가지는 RDF 포맷의 특징을 고려하여 각 데이터가 아닌 프로퍼티를 기준으로 클러스터링을 수행하고, 이를 통해 데이터가 증가해도 프로퍼티 수는 항상 제한되어있기 때문에 대량의 데이터에 대한 저장 알고리즘의 복잡성을 줄일 수 있는 이점이 있다.

도면의 간단한 설명

[0035] 도 1은 본 발명의 사상에 따른 RDF 데이터 분산 파티셔닝 방법에 대한 원 데이터가 될 수 있는 RDF(Resource Description Framework) 데이터 저장 구조의 일 실시예를 도시한 테이블.

도 2는 본 발명의 일 실시예에 따른 클러스터링 알고리즘을 사용한 RDF 데이터 분산 파티셔닝 방법을 도시한 흐름도.

도 3은 타입이 명확하지 않은 쿼리에 대한 타입-프로퍼티 페어 생성 과정을 나타내는 테이블들을 이용한 관계 개념도.

도 4는 타입-프로퍼티 페어들에 대한 클러스터링 과정을 나타내는 테이블들을 이용한 관계 개념도.

도 5는 도 2는 본 발명의 일 실시예에 따른 클러스터링 알고리즘을 사용한 RDF 데이터 분산 파티셔닝 시스템을 도시한 블록도.

발명을 실시하기 위한 구체적인 내용

[0036] 본 발명을 설명함에 있어서 제 1, 제 2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 구성요소들은 용어들에 의해 한정되지 않을 수 있다. 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제 1 구성요소는 제 2 구성요소로 명명될 수 있고, 유사하게 제 2 구성요소도 제 1 구성요소로 명명될 수 있다.

[0037] 어떤 구성요소가 다른 구성요소에 연결되어 있다거나 접속되어 있다고 언급되는 경우는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해될 수 있다.

[0038] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함할 수 있다.

[0039] 본 명세서에서, 포함하다 또는 구비하다 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것으로서, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해될 수 있다.

[0040] 또한, 도면에서의 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있다.

[0042] 스마트그리드의 전력데이터는 미터 데이터(Meter Data), 마스터 데이터(Master Data), 통계 데이터(Statistical Data)로 나눌 수 있다. 미터 데이터는 스마트 미터로부터 수집된 데이터로 타임스탬프를 포함하고 있는 데이터이며, 마스터 데이터는 미터 데이터의 주체가 되는 사용자, 디바이스 등에 대한 기본 정보를 포함하고 있는 데이터이다. 마지막으로 통계 데이터는 미터 데이터와 마스터 데이터 기반 온라인 배치 분석 결과를 나타낸다. 스마트 그리드 환경에서는 통계 데이터를 계산하기 위해, 미터 데이터와 마스터 데이터를 조회하는 다양한 형태의 분석 쿼리와 Ad-Hoc 쿼리가 발생한다.

[0043] 본 발명은 종래 기술인 주어 기반 RDF 데이터 분산 파티셔닝 방법이 쿼리 워크로드를 고려하지 않아 스마트 그리드와 같이 다양한 쿼리가 발생하는 환경에 적합하지 않은 한계점을 개선하기 위해, 쿼리에 의한 접근 빈도를 기반으로 클러스터링 알고리즘을 사용하여 스마트 그리드 환경에서의 RDF 데이터를 분산된 노드에 파티셔닝한다. 이에 따라, 본 발명은 다음 두 가지 특징을 가진다.

[0044] 제1 특징은 쿼리 워크로드 기반 클러스터링 알고리즘을 사용한 분산 파티셔닝이다. 본 발명은 다양한 쿼리 형태에 적응적으로 RDF 데이터를 분산 파티셔닝하기 위해, 쿼리에 의한 접근 빈도를 기반으로 클러스터링 알고리즘을 사용하여 같은 노드에 저장할 RDF 데이터를 그룹핑하여 분산 파티셔닝한다.

[0045] 제2 특징은 RDF의 타입(Type)을 고려한 프로퍼티 기반 분산 파티셔닝이다. RDF 포맷은 프로퍼티에 매우 의존적이며, 기본적으로 주어마다 특정 타입을 가지고 있으며, 타입별로 유사 프로퍼티 셋을 가진다. 예를 들어, 스마트 그리드 환경에서 RDF 데이터의 주어 <Account#1>은 <Account>라는 타입을 가지며, <Account> 타입은 주로 <hasID>, <hasContractStatusCode>, <hasCustomerName>, <hasRegionCode> 등의 프로퍼티를 가질 것이다. 따라서 RDF 포맷의 특징을 고려하였을 때, 같은 타입에 속하는 데이터들은 유사 프로퍼티 셋을 가질 가능성이 높기 때문에, 본 발명은 개별 데이터(e.g. Account#1)가 아닌 타입이 명시된 프로퍼티(e.g. Acoount-hasID,

Account-hasContactStatusCode)를 기준으로 클러스터링 알고리즘을 수행하여 RDF 데이터를 분산 파티셔닝한다.

[0047] 도 1은 본 발명의 사상에 따른 RDF 데이터 분산 파티셔닝 방법에 대한 원 데이터가 될 수 있는 RDF(Resource Description Framework) 데이터 저장 구조의 일 실시예를 도시한다.

[0048] 도시한 RDF 데이터 저장 구조는 단일의 저장 구조에 RDF 데이터를 저장하는 방식으로서, 대표적으로 모든 RDF 데이터를 주어, 서술어, 목적어 3개의 열(Column)로 구성된 하나의 테이블에 저장하는 단일 트리플 테이블(Naive Triple Table)이다. 이 외에도 같은 종류(Type)의 엔티티와 연결된 모든 프로퍼티들을 연결하여 하나의 테이블에 저장하는 프로퍼티 테이블(Property Table), 각 프로퍼티마다 수직 분할하여 같은 프로퍼티를 가진 엔티티들을 하나의 테이블에 저장하는 수직 분할 테이블(Vertical Partitioned Table) 저장 구조도 본 발명의 사상에 따른 RDF 데이터 분산 파티셔닝 방법에 대한 원 데이터가 될 수 있음은 물론이다.

[0049] 예컨대, 본 발명에서 분산 파티셔닝의 대상이 되는 원 데이터는, 스마트 미터로부터 수집된 데이터로 타임스탬프를 포함하고 있는 미터 데이터와, 미터 데이터의 주체가 되는 사용자, 디바이스 등에 대한 기본 정보를 포함하고 있는 마스터 데이터와, 미터 데이터와 마스터 데이터 기반 온라인 배치 분석 결과를 나타내는 통계 데이터의 각 단위 데이터가 주어, 서술어, 목적어 3개의 트리플을 구성한 것일 수 있다.

[0050]

[0051] 본 발명은 RDF 프로퍼티 클러스터링 과정; 및 그 결과로 생성되는 프로퍼티 그룹들을 기반으로 RDF 데이터를 분산된 노드에 파티셔닝 하는 과정으로 이루어지는데, 여기서, 'RDF 프로퍼티 클러스터링' 과정은, 1) 타입 비트맵 생성, 2) 타입-프로퍼티 생성, 3) 다큐먼트 클러스터링 알고리즘 기반 타입-프로퍼티 클러스터링의 3 단계로 수행될 수 있다.

[0052] 즉, 본 발명의 사상에 따른 RDF 데이터 분산 파티셔닝 방법은 도 2에 도시한 바와 같이, RDF 데이터 집합을 분석하여 프로퍼티-타입 비트맵을 형성하는 단계(S100); 입력 쿼리 세트에 상기 프로퍼티-타입 비트맵을 적용하여 타입-프로퍼티들을 생성하는 단계(S200); 상기 타입-프로퍼티들을 클러스터링하는 단계(S300); 및 클러스터링 결과에 따라 상기 RDF 데이터 집합를 분산 저장하는 단계(S400)를 포함할 수 있다.

[0053] 본 발명의 사상에 따른 RDF 데이터 분산 파티셔닝 방법에서는 RDF 데이터가 아닌 타입이 명시된 프로퍼티를 입력으로 클러스터링 알고리즘을 수행한다.

[0054] 상기 타입은 RDF 데이터 저장 구조에서 주어(Subject)에 적용되는 분류로 볼 수 있으며, 프로퍼티는 RDF 데이터 저장 구조에서 서술어(Predicate) 및/또는 주어에 대한 기술에 적용되는 분류로 볼 수 있다. 데이터 테이블 형태의 관점에서는 상기 타입은 개별 테이블을 가리키는 것으로, 상기 프로퍼티는 테이블의 컬럼을 가리키는 것으로 볼 수도 있다.

[0056] 상기 프로퍼티-타입 비트맵을 형성하는 단계(S100)에서는, 먼저, RDF 데이터 집합이 저장된 RDF 데이터 저장부 등으로부터 RDF 데이터 집합을 독출하고, 다양한 RDF 기준에 따른 데이터 저장 구조, 그래프 데이터 구조 등에 따른 데이터 세트를 분석하여, 프로퍼티들과 타입들을 추출/결정한다.

[0057] 예컨대, 가장 단순한 프로퍼티들 및 타입들 추출/결정 방법으로는, 독출된 RDF 데이터 집합의 각 단위 트리플의 주어들을 타입들로 결정하고, 서술어들을 프로퍼티로 결정할 수 있다. 다소 복잡한 구현에서는, 고정되어 사용되는 빈도가 매우 높은 서술어 - 목적어 페어를 별도의 프로퍼티로 추가할 수 있다.

[0058] 상기 프로퍼티-타입 비트맵을 형성하는 단계(S100)에서는, 다음, 일축이 상기 추출/결정된 프로퍼티들이고 나머지 축이 상기 추출/결정된 타입들인 테이블을 형성하고, 각 교차점에 상기 RDF 데이터 집합에 각 교차점에 따른 프로퍼티-타입 페어의 존재여부를 표시한다.(예: 존재시 1로 표시)

[0059] 다음, 상기 S100 단계에서의 타입 비트맵 생성과정을 구체적으로 예시하겠다.

[0060] 상기 프로퍼티-타입 비트맵을 형성하는 단계(S100)에서는, 클러스터링 입력이 되는 타입-프로퍼티 생성을 위해 필요한 자료구조로서 RDF 데이터를 읽고 타입 비트맵을 생성한다. 타입 비트맵은 이후 입력 쿼리에 포함된 프로퍼티의 타입을 추정하기 위한 비트맵이며, 프로퍼티가 어떤 타입에 속해있는지에 대한 정보를 저장한다. 프로퍼티가 특정 타입에 속할 경우 해당 셀의 비트(bit)를 '1'로 설정하며, 속하지 않은 경우는 해당 셀의 비트를 '

0'으로 설정한다.

[0061] 예컨대, 하기 표 1은 스마트 그리드의 RDF 데이터를 읽고 생성한 타입 비트맵을 예시한다.

표 1

Property \ Type	Meter	Account	AHE Server	Modem	...
hasContractClassCode	0	1	0	0	...
hasRegionCode	1	1	1	1	...
hasCustomerName	0	1	0	0	...
hasMeterTypeCode	1	0	0	0	...
hasMeterManufacturer Code	1	0	0	0	...
hasModemType	0	0	0	1	...
hasDCUID	1	0	0	1	...
...	...	...	...	...	...

[0062]

[0063] 프로퍼티 <hasContractClassCode>는 타입 <Account>에 속하므로 매칭되는 셀의 비트를 '1'로 설정하며, 타입 <Meter>에는 속하지 않으므로 '0' 으로 설정한다. 또한 프로퍼티 <hasRegionCode>는 모든 타입에 공통적으로 속하므로 해당 셀들의 비트들을 모두 '1'로 설정한다.

[0065] 상기 타입-프로퍼티들을 생성하는 단계(S200)에서의 타입-프로퍼티 페어는 쿼리에 대한 것임에, RDF 데이터에 대한 것인 상기 프로퍼티-타입 페어와 구분된다.

[0066] 상기 입력 쿼리 세트는 그래프 세트일 수 있으며, 사용자의 입력 쿼리 세트가 저장된 쿼리 저장부로부터 독출될 수 있다.

[0067] 상기 타입-프로퍼티들을 생성하는 단계(200)에서 각 쿼리에 주어가 되는 대상 및 서술어가 되는 조회 항목이 명확한 경우, 이를 각각 상기 프로퍼티-타입 비트맵 상에서의 타입 및 프로퍼티에 대한 것으로 대응할 수 있다.

[0068] 상기 타입-프로퍼티들을 생성하는 단계(200)에서 특정 쿼리에 주어가 명확하지 않은 경우, 타입을 특정하기 곤란한데, 이 경우, 상기 특정 쿼리가 속한 쿼리 그래프 세트에 주어가 존재하면, 이를 타입으로 대응시키면 된다. 그런데, 상기 쿼리 그래프 세트에도 주어가 불명확하면, 해당 쿼리 그래프 세트에 포함된 프로퍼티(서술어)들이 상기 프로퍼티-타입 비트맵 상에서 공통적으로 가지고 있는 타입을 해당 타입으로 간주할 수 있다.

[0069] 입력 쿼리에 타입-프로퍼티 페어가 명시된 경우에는 프로퍼티-타입 비트맵을 확인할 필요가 없으며, 쿼리에 명시된 페어를 클러스터링 입력으로 바로 사용할 수 있다.

[0071] 다음, 상기 S100 단계에서의 타입-프로퍼티 생성 과정을 구체적으로 예시하겠다.

[0072] 상기 타입-프로퍼티들을 생성하는 단계(200)에서는 쿼리 워크로드를 고려하여 과거에 입력된 쿼리셋에 의해 접근된 빈도를 기반으로 타입이 명시된 프로퍼티를 클러스터링한다. 이를 위해 쿼리셋에 포함되는 프로퍼티에 대해 프로퍼티가 속한 타입으로 명시하는, 즉 타입-프로퍼티를 생성하는 과정을 수행한다. 먼저 쿼리셋에 포함된 프로퍼티의 연결을 파악하기 위해, 쿼리셋의 각 쿼리를 주어 기반의 그래프 형태로 변환한다. 쿼리를 그래프 형태로 변환한 후 각 주어에 연결된 프로퍼티에 대한 타입을 파악하며, 만약 주어 타입이 쿼리에 있는 경우는 프로퍼티들을 해당 타입으로 명시하여 타입-프로퍼티를 생성한다. 하지만 주어 타입이 쿼리에 없는 경우 주어에 연결된 프로퍼티들을 통해 주어의 타입을 추정하는 과정이 필요하다. 이를 위해 프로퍼티들을 앞에서 생성한 타입-프로퍼티 비트맵에서 검색하며, 입력된 프로퍼티들이 모두 속하는 타입이 있는 경우 프로퍼티에 해당 타입을 명시함으로써 타입-프로퍼티를 생성한다.

[0073] 예컨대, 도 3은 타입이 명확하지 않은 쿼리에 대한 타입-프로퍼티 페어 생성 과정을 나타내는 테이블들을 이용한 관계 개념도이다. (a)의 입력 쿼리를 (b)와 같이 주어 'Subject' 기반의 쿼리 그래프 형태로 변환한다. (b)의 쿼리 그래프에는 'Subject'의 타입에 대한 정보가 없으므로, 연결된 프로퍼티 'hasCustomerName',

'hasRegionCode', 'hasContractClassCode'를 타입 비트맵에 검색하여 해당 프로퍼티들이 어떤 타입에 속하는지 추정해야한다. (c)는 타입 비트맵을 검색하여 프로퍼티들의 타입을 추정하는 과정을 나타낸 것이며, 프로퍼티들과 타입 <Account>의 매칭 비트가 모두 1이므로 해당 프로퍼티들은 타입 <Account>에 속하는 것으로 추정이 가능하다. 따라서 타입이 명시된 프로퍼티, 즉 타입-프로퍼티 'Account-hasContractClassCode', 'Account-hasRegionCode', 'Account-hasCustomerName'가 생성된다.

[0075] 상기 클러스터링 단계(S300)에서는, 공지된 다큐먼트 클러스터링(Document Clustering) 알고리즘 즉, 웹상에서의 문서에 대한 분류/저장에 사용되는 클러스터링 기술들 중 하나 이상을 적용할 수 있다. 이에 따라, 쿼리 상에서 동시에 출현하는 빈도수가 높은 타입-프로퍼티 페어들을 하나로 클러스터링할 수 있다.

[0076] 상기 다큐먼트 클러스터링 알고리즘 기반 타입-프로퍼티 클러스터링 단계(S300)를 구체적으로 예시하여 설명하겠다.

[0077] 상기 S300 단계에서는 다큐먼트 클러스터링(Document Clustering) 알고리즘을 사용하여 같은 쿼리에 의해 자주 접근되는 타입이 명시된 프로터피(타입-프로퍼티)를 클러스터링한다. 다큐먼트 클러스터링 알고리즘은 문서 분류에 활용되는 알고리즘으로, 각 문서에 포함된 단어의 출현 빈도를 기반으로 유사 문서를 분류한다. 상기 S300 단계에서는 다큐먼트 클러스터링 알고리즘에서 문서를 쿼리로, 문서에 포함된 단어를 쿼리에 포함된 타입-프로퍼티로 대응하여 유사 쿼리에 의한 관계성이 높은 타입-프로퍼티를 분류한다.

[0078] 상기 S300 단계에서는, 먼저 쿼리에 포함된 타입-프로퍼티의 출현 빈도를 계산하기 위해 빈도 점수 TF-IDF(Term Frequency-Inverse Document Frequency)값을 계산할 수 있다. TF-IDF는 본래 문서에 포함되는 단어의 빈도를 나타내는 점수이며, TF와 IDF를 곱한 값을 의미한다. TF는 단어가 특정 문서 내에 나오는 빈도 값이며, 동일 단어가 문서 집합 내에서 자주 사용되는 정도를 나타내는 DF(Document Frequency)의 역수를 IDF(Inverse Document Frequency)로 나타낸다. 본 발명은 문서를 쿼리로, 단어를 타입-프로퍼티로 대응하여 각 타입-프로퍼티의 특정 쿼리에 포함되는 빈도 점수 TF-IDF 값을 계산한다.

[0079] 각 쿼리와 타입-프로퍼티 간 TF-IDF 값은 벡터 스페이스 모델로 불리는 별도의 테이블에 저장되며, 벡터 스페이스 모델에 저장된 벡터, 즉 TF-IDF 값을 기반으로 프로퍼티 간 관계성(유사도; Similarity)을 계산한다. 계산된 관계성은 관계성 테이블로 불리는 별도의 테이블에 저장된다. 그 후 K-means 클러스터링 알고리즘을 사용하여 관계성 테이블을 기반으로 관계성이 높은 타입-프로퍼티를 클러스터링한다. 도 4는 전체적인 다큐먼트 클러스터링 알고리즘 기반 타입-프로퍼티 클러스터링 과정을 나타낸 예시이다. 이러한 클러스터링 과정은 클러스터된 타입-프로퍼티 그룹들을 생성한다.

[0081] 상기 분산 저장하는 단계(S400)에서는 상기 RDF 데이터 집합을 분산된 노드에 나누어 저장한다. 바람직하게는 각 노드 저장소의 위치는 저장되는 클러스터에 속한 타입-프로퍼티 페어들에 대한 쿼리 처리 빈도가 높은데로 할당할 수 있다.

[0083] 상기 S300 단계에서의 클러스터링 과정은 클러스터된 타입-프로퍼티 그룹들을 생성하는데, 타입-프로퍼티가 속한 클러스터는 물리적으로 분산된 노드를 의미하며, 같은 클러스터에 속한 타입-프로퍼티들은 같은 노드에 저장되는 것을 의미한다. 따라서 본 발명은 타입-프로퍼티별로 RDF 데이터를 그룹핑하여 테이블 형태로 저장하고, RDF 데이터를 타입-프로퍼티들이 배정된 분산된 노드에 분산 파티셔닝한다.

[0084] 상기 노드는 실제로는 각 클러스터링된 데이터들을 저장하는 특정 사이트에 위치한 서버나 데이터베이스 장치이지만, 노드라고 약칭하였다.

[0086] 도 5은 도 1에 도시한 RDF 데이터 분산 파티셔닝 방법이 수행될 수 있는 RDF 데이터 분산 파티셔닝 시스템의 일실시예를 도시한다.

[0087] 도시한 RDF 데이터 분산 파티셔닝 시스템은, RDF 데이터 집합이 저장된 RDF 데이터 저장부(10); 2개 이상의 저장 노드들로 이루어진 분산 저장부(200); 사용자의 입력 쿼리 세트가 저장된 쿼리 저장부(20); RDF 데이터 집합를 분석하여 프로퍼티-타입 비트맵을 형성하는 비트맵 형성부(110); 입력 쿼리 세트에 상기 프로퍼티-타입 비트

맵을 적용하여 타입-프로퍼티들을 생성하고, 상기 타입-프로퍼티들을 클러스터링하는 클러스터링부(120); 및 상기 타입-프로퍼티들의 클러스터링 결과에 따라 상기 RDF 데이터 집합를 상기 분산 저장부에 저장하는 분산부(140)를 포함할 수 있다.

[0089] 상기 비트맵 형성부(110), 상기 클러스터링부(120) 및 상기 분산부(140)는 하나의 독립된 서버 장치인 RDF 데이터 분산 파티셔닝 서버(100)를 구성할 수 있다.

[0090] 상기 RDF 데이터 저장부(10) 및 상기 쿼리 저장부(20)도 상기 RDF 데이터 분산 파티셔닝 서버(100)의 구성요소로 될 수도 있으나, 외부 DB 서버로 구현될 수도 있다. 특히, 상기 쿼리 저장부(20)는 사용자의 조회를 직접 받아서 처리하는 사용자 응용 서비스 제공 서버측에 구비되는 것이 유리하다.

[0091] 상기 비트맵 형성부(110)는 도 2에 도시한 상기 프로퍼티-타입 비트맵을 형성하는 단계(S100)를 수행하며, 상기 클러스터링부(120)는 상기 타입-프로퍼티들을 생성하는 단계(200) 및 상기 클러스터링 단계(S300)를 수행하고, 상기 분산부(140)는 상기 분산 저장하는 단계(S400)를 수행한다.

[0093] 본 발명이 속하는 기술 분야의 당업자는 본 발명이 그 기술적 사상이나 필수적 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있으므로, 이상에서 기술한 실시 예들은 모든 면에서 예시적인 것이며 한정적인 것이 아닌 것으로서 이해해야만 한다. 본 발명의 범위는 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 등가개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0095]

10 : RDF 데이터 저장부

20 : 쿼리 저장부

110 : 비트맵 형성부

120 : 클러스터링부

140 : 분산부

100 : RDF 데이터 분산 파티셔닝 서버

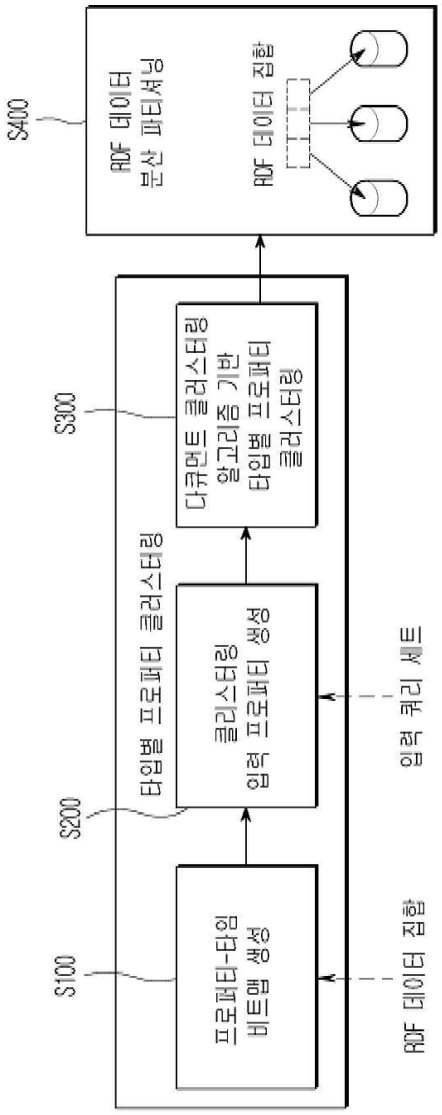
200 : 분산 저장부

도면

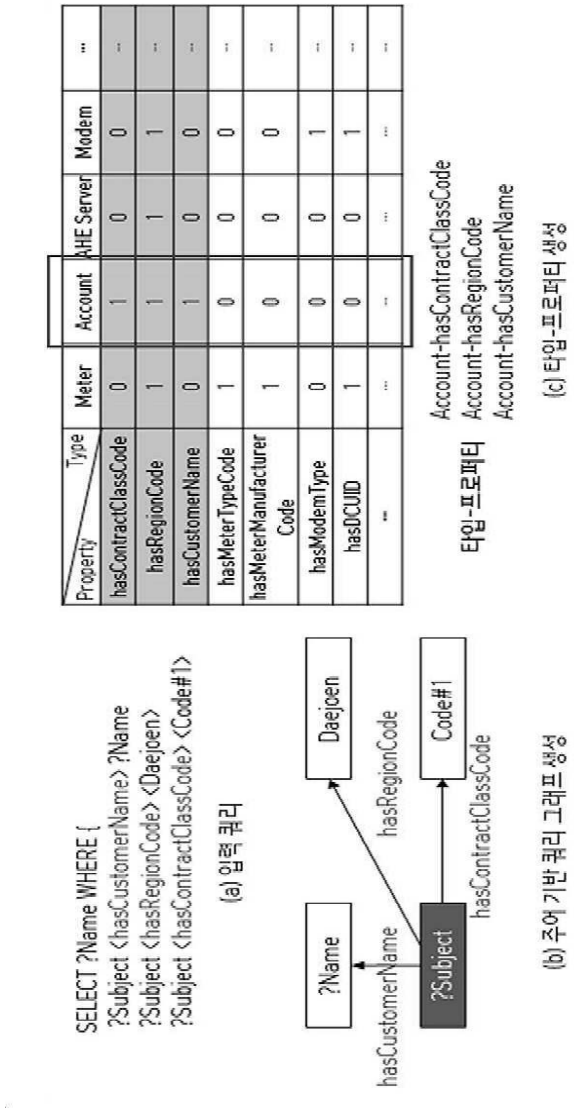
도면1

엔티티(주어)	프로퍼티(서술어)	엔티티(목적어)
ID1	Type	Booktype
ID1	Title	"XYZ"
ID1	Author	"Fox,Joe"
ID1	Copyright	"2001"
ID2	Type	CDType
ID2	Title	"ABC"
ID2	Artist	"Orr,Tim"
ID2	Copyright	"1985"
ID2	Language	"French"

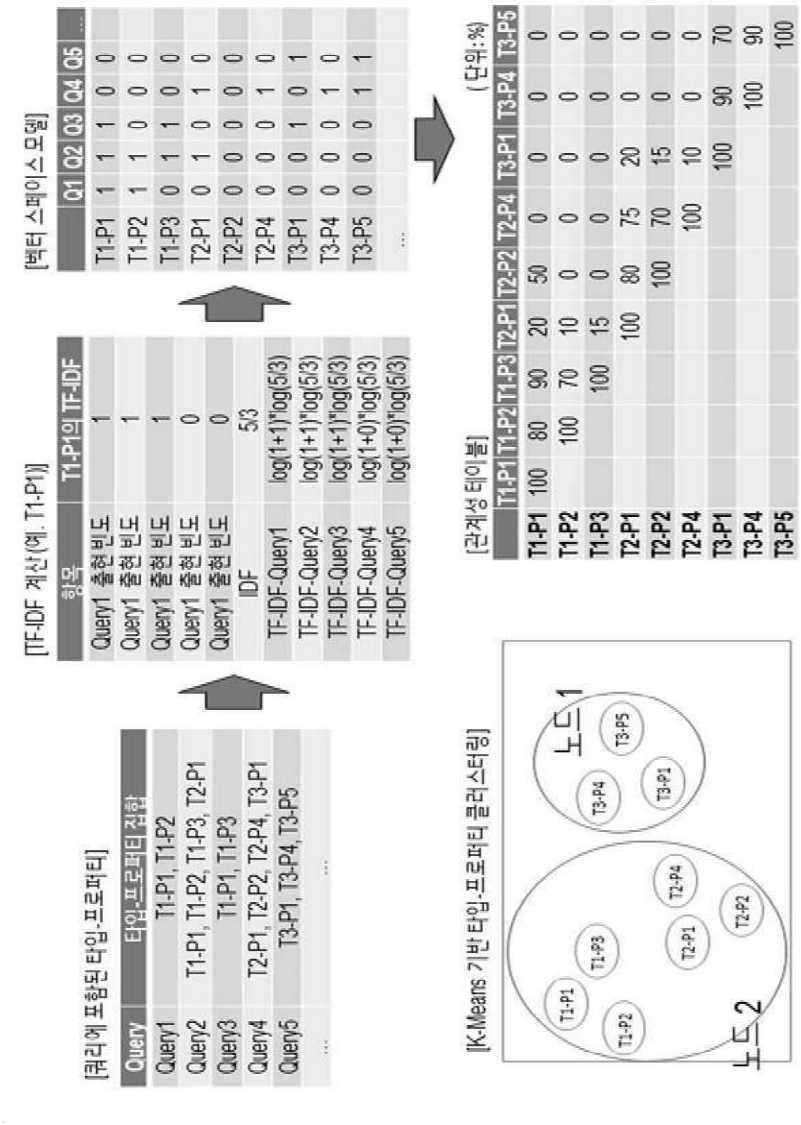
도면2


S400
RDF 데이터
분산 파티셔닝
RDF 데이터 집합
타입별 프로퍼티 클러스터링
S300
다큐먼트 클러스터링
알고리즘 기반
타입별 프로퍼티
클러스터링
S200
클러스터링
입력 프로퍼티 생성
입력 쿼리 세트
S100
프로퍼티-타입
비트맵 생성
RDF 데이터 집합

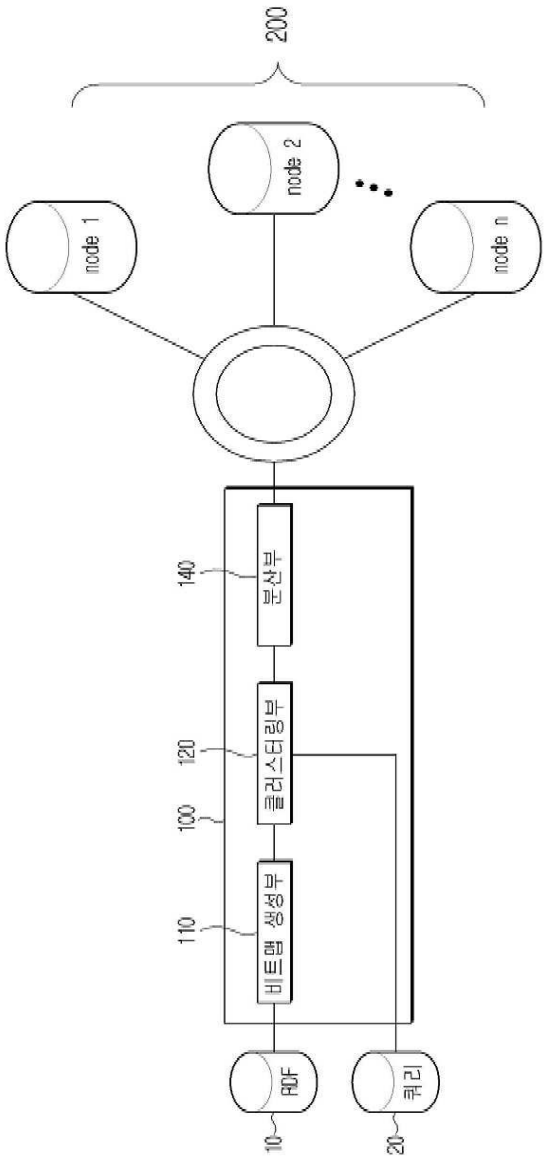
도면3


SELECT ?Name WHERE {
?Subject <hasCustomerName> ?Name
?Subject <hasRegionCode> <Daejoen>
?Subject <hasContractClassCode> <Code#1>
(a) 입력 쿼리
?Name
Daejoen
hasCustomerName
hasRegionCode
?Subject
Code#1
hasContractClassCode
(b) 주어 기반 쿼리 그래프 생성
Property \ Type
Meter
Account
AHE Server
Modem
hasContractClassCode
hasRegionCode
hasCustomerName
hasMeterTypeCode
hasMeterManufacturer Code
hasModemType
hasDCUID
타입-프로퍼티
Account-hasContractClassCode
Account-hasRegionCode
Account-hasCustomerName
(c) 타입-프로퍼티 생성

도면4


[쿼리에 포함된 타입-프로퍼티]
Query
타입-프로퍼티 집합
Query1
T1-P1, T1-P2
Query2
T1-P1, T1-P2, T1-P3, T2-P1
Query3
T1-P1, T1-P3
Query4
T2-P1, T2-P2, T2-P4, T3-P1
Query5
T3-P1, T3-P4, T3-P5
[TF-IDF 계산(예: T1-P1)]
항목
T1-P1의 TF-IDF
Query1 출현 빈도
1
Query1 출현 빈도
1
Query1 출현 빈도
1
Query1 출현 빈도
0
Query1 출현 빈도
0
IDF
5/3
TF-IDF-Query1
log(1+1)*log(5/3)
TF-IDF-Query2
log(1+1)*log(5/3)
TF-IDF-Query3
log(1+1)*log(5/3)
TF-IDF-Query4
log(1+0)*log(5/3)
TF-IDF-Query5
log(1+0)*log(5/3)
[벡터 스페이스 모델]
Q1 Q2 Q3 Q4 Q5
T1-P1 1 1 1 0 0
T1-P2 1 1 0 0 0
T1-P3 0 1 1 0 0
T2-P1 0 1 0 1 0
T2-P2 0 0 0 0 0
T2-P4 0 0 0 1 0
T3-P1 0 0 1 0 1
T3-P4 0 0 0 1 0
T3-P5 0 0 0 1 1
[K-Means 기반 타입-프로퍼티 클러스터링]
노드 1
T3-P5
T3-P4
T3-P1
노드 2
T1-P1
T1-P2
T1-P3
T2-P1
T2-P4
T2-P2
[관계성 테이블]
(단위:%)
T1-P1 T1-P2 T1-P3 T2-P1 T2-P2 T2-P4 T3-P1 T3-P4 T3-P5
T1-P1 100 80 90 20 50 0 0 0 0
T1-P2 100 70 10 0 0 0 0 0
T1-P3 100 15 0 0 0 0 0
T2-P1 100 80 75 20 0 0
T2-P2 100 70 15 0 0
T2-P4 100 10 0 0
T3-P1 100 90 70
T3-P4 100 90
T3-P5 100

도면5


200
node 1
node 2
node n
140
분산부
120
클러스터링부
100
110
비트맵 생성부
RDF
쿼리
10
20