(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0122563
(43) 공개일자 2022년09월02일

(51) 국제특허분류(Int. Cl.)
G06F 16/174 (2019.01) *G06F 16/17* (2019.01)
G06F 3/06 (2006.01)
(52) CPC특허분류
G06F 16/1744 (2019.01)
G06F 16/1727 (2019.01)
(21) 출원번호 10-2022-0026411
(22) 출원일자 2022년02월28일
심사청구일자 2022년02월28일
(30) 우선권주장
1020210026749 2021년02월26일 대한민국(KR)

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
이영구
경기도 수원시 영통구 청명북로 33, 431동 1704호(영통동, 청명마을삼성아파트)
탄지나 술타나
경기도 수원시 영통구 매영로415번길 42-27, 203호(영통동)
(74) 대리인
두호특허법인

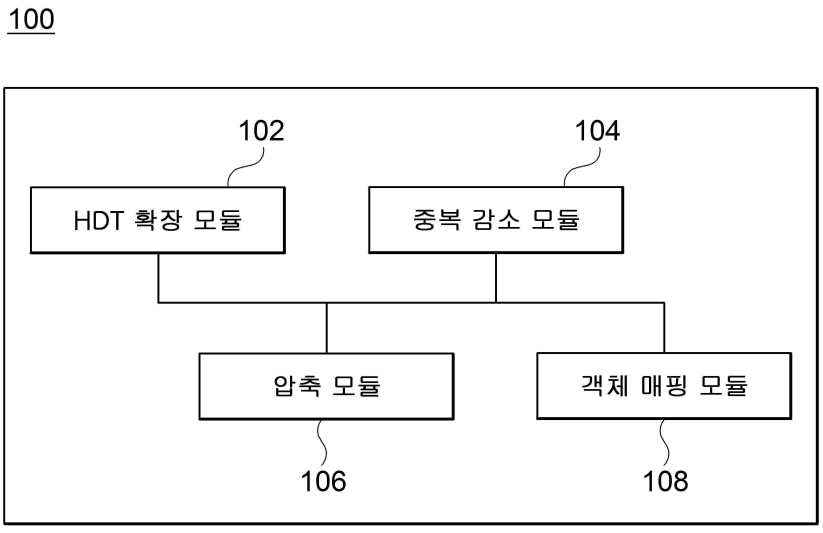
전체 청구항 수 : 총 23 항

(54) 발명의 명칭 **RDF 데이터 압축 방법 및 장치**

(57) 요 약

RDF 데이터 압축 방법 및 장치가 개시된다. 개시되는 일 실시예에 따른 RDF 데이터 압축 방법은, 하나 이상의 프로세서들, 및 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서, 주제(Subject), 술어(Predicate), 및 객체(Object)로 구성된 복수 개의 트리플(Triple)을 포함하는 RDF(Resource Description Framework) 데이터를 직렬화 하는 단계, 직렬화 된 트리플들을 재배열하여 직렬화된 트리플들에서 중복되는 부분을 감소시키는 단계, 및 재배열된 트리플들을 압축시키는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 3/0608 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711111586
과제번호 2018R1A2A2A05023669
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)(R&D)
연구과제명 대용량 RDF그래프 요약 및 압축 기반 고속 RDF질의 처리 기술
기 여 율 1/2
과제수행기관명 경희대학교(국제캠퍼스)
연구기간 2020.03.01 ~ 2021.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711125817
과제번호 2015-0-00742-007
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성(R&D)
연구과제명 라이프 컴패니온쉽 경험을 위한 지능형 인터랙션 융합 연구
기 여 율 1/2
과제수행기관명 성균관대학교 산학협력단
연구기간 2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

하나 이상의 프로세서들, 및

상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서,

주제(Subject), 술어(Predicate), 및 객체(Object)로 구성된 복수 개의 트리플(Triple)을 포함하는 RDF(Resource Description Framework) 데이터를 직렬화 하는 단계;

상기 직렬화 된 트리플들을 재배열하여 상기 직렬화된 트리플들에서 중복되는 부분을 감소시키는 단계; 및

상기 재배열된 트리플들을 압축시키는 단계를 포함하는, RDF 데이터 압축 방법.

청구항 2

청구항 1에 있어서,

상기 중복되는 부분을 감소시키는 단계는,

상기 직렬화 된 트리플들을 동일한 주제를 기준으로 그룹화 하여 상기 트리플들을 트리 형태로 배열하는 단계;

상기 트리 형태의 배열에서 동일한 술어 세트를 그룹화 하고, 그룹화 된 상기 술어 세트에 ID를 부여하는 단계; 및

상기 트리 형태의 배열에서 상기 술어 세트의 ID가 동일한 주제끼리 그룹화하여 재배열하는 단계를 포함하는, RDF 데이터 압축 방법.

청구항 3

청구항 2에 있어서,

상기 중복되는 부분을 감소시키는 단계는,

상기 재배열된 각 주제의 ID를 순차적으로 배열하여 주제 매핑 시퀀스를 생성하는 단계를 더 포함하는, RDF 데이터 압축 방법.

청구항 4

청구항 2에 있어서,

상기 중복되는 부분을 감소시키는 단계는,

상기 재배열된 주제들을 상기 술어 세트의 ID가 동일한 그룹 별로 각 주제에 대해 새로운 ID를 부여하는 단계; 및

상기 그룹 별로 각 주제에 대해 원본 ID와 상기 새로운 ID를 매핑하여 저장하는 단계를 더 포함하는, RDF 데이터 압축 방법.

청구항 5

청구항 4에 있어서,

상기 압축시키는 단계는,

상기 재배열된 트리플들에서 각 주제 별 트리에 따른 트리플 값을 추출하여 순차적으로 배열하는 단계;

상기 각 주제 별 트리의 트리플들에 대한 합계 값(Summed Value)을 산출하는 단계; 및

상기 산출한 합계 값에 기반하여 각 주제 별 트리의 트리플들을 기 설정된 형태의 데이터로 재구성하는 단계를 포함하는, RDF 데이터 압축 방법.

청구항 6

청구항 5에 있어서,

상기 합계 값은,

해당 주제의 새로 부여된 ID 값과 해당 주제의 트리플들 중 객체의 ID 값을 합산하여 산출하는, RDF 데이터 압축 방법.

청구항 7

청구항 5에 있어서,

상기 재구성하는 단계는,

상기 각 주제 별 트리의 트리플들을 <합계 값, {주제(Subject), 객체1($Object_1$), 객체2($Object_2$),..., 객체n($Object_n$)}>의 집합으로 재구성하는, RDF 데이터 압축 방법.

청구항 8

청구항 7에 있어서,

상기 압축시키는 단계는,

상기 각 주제 별 트리의 트리플들에 대한 재구성 데이터에서 동일한 합계 값이 존재하는 경우, 중복되는 객체의 값을 빈 문자열로 대체하는 단계를 더 포함하는, RDF 데이터 압축 방법.

청구항 9

청구항 2에 있어서,

상기 RDF 데이터 압축 방법은,

상기 술어 세트의 그룹별로 해당 그룹에 속하는 술어들에 대응하도록 상기 트리플들의 객체를 1차 재배열하는 단계;

상기 술어 세트를 구성하는 각 술어 별로 순차적으로 그에 대응하는 객체를 2차 재배열하는 단계; 및

상기 2차 재배열된 객체의 값들을 원래의 값보다 작은 범위의 값으로 정규화하여 매핑시키는 단계를 더 포함하는, RDF 데이터 압축 방법.

청구항 10

청구항 1에 있어서,

상기 RDF 데이터는, 복수 개의 지명 그래프를 포함하고,

상기 RDF 데이터 압축 방법은,

상기 RDF 데이터를 직렬화 할 때, HDT 사전(HDT Dictionary)에 상기 지명 그래프들의 그래프 이름을 나타내기 위한 섹션을 추가하는 단계; 및

상기 지명 그래프들에 특정 트리플이 있는지 여부를 나타내기 위한 그래프 정보를 생성하는 단계를 더 포함하는, RDF 데이터 압축 방법.

청구항 11

청구항 10에 있어서,

상기 그래프 정보를 생성하는 단계는,

트리플 인터프리터를 통해 각 트리플을 기준으로 해당 트리플이 존재하는 지명 그래프를 비트맵으로 생성하는 단계; 및

그래프 인터프리터를 통해 각 지명 그래프를 기준으로 각 지명 그래프가 어떤 트리플을 가지는지를 비트맵으로 생성하는 단계를 포함하는, RDF 데이터 압축 방법.

청구항 12

주제(Subject), 술어(Predicate), 및 객체(Object)로 구성된 복수 개의 트리플(Triple)을 포함하는 RDF(Resource Description Framework) 데이터를 직렬화 하는 HDT 확장 모듈;

상기 직렬화 된 트리플들을 재배열하여 상기 직렬화된 트리플들에서 중복되는 부분을 감소시키는 중복 감소 모듈; 및

상기 재배열된 트리플들을 압축시키는 압축 모듈을 포함하는, RDF 데이터 압축 장치.

청구항 13

청구항 12에 있어서,

상기 중복 감소 모듈은,

상기 직렬화 된 트리플들을 동일한 주제를 기준으로 그룹화 하여 상기 트리플들을 트리 형태로 배열하고, 상기 트리 형태의 배열에서 동일한 술어 세트를 그룹화 하고 그룹화 된 상기 술어 세트에 ID를 부여하며, 상기 트리 형태의 배열에서 상기 술어 세트의 ID가 동일한 주제끼리 그룹화하여 재배열하는, RDF 데이터 압축 장치.

청구항 14

청구항 13에 있어서,

상기 중복 감소 모듈은,

상기 재배열된 각 주제의 ID를 순차적으로 배열하여 주제 매핑 시퀀스를 생성하는, RDF 데이터 압축 장치.

청구항 15

청구항 13에 있어서,

상기 중복 감소 모듈은,

상기 재배열된 주제들을 상기 술어 세트의 ID가 동일한 그룹 별로 각 주제에 대해 새로운 ID를 부여하고, 상기 그룹 별로 각 주제에 대해 원본 ID와 상기 새로운 ID를 매핑하여 저장하는, RDF 데이터 압축 장치.

청구항 16

청구항 15에 있어서,

상기 압축 모듈은,

상기 재배열된 트리플들에서 각 주제 별 트리에 따른 트리플 값을 추출하여 순차적으로 배열하고, 상기 각 주제 별 트리의 트리플들에 대한 합계 값(Summed Value)을 산출하며, 상기 산출한 합계 값에 기반하여 각 주제 별 트리의 트리플들을 기 설정된 형태의 데이터로 재구성하는, RDF 데이터 압축 장치.

청구항 17

청구항 16에 있어서,

상기 합계 값은,

해당 주제의 새로 부여된 ID 값과 해당 주제의 트리플들 중 객체의 ID 값을 합산하여 산출하는, RDF 데이터 압축 장치.

청구항 18

청구항 16에 있어서,

상기 압축 모듈은,

상기 각 주제 별 트리의 트리플들을 <합계 값, {주제(Subject), 객체1($Object_1$), 객체2($Object_2$),..., 객체n($Object_n$)}>의 집합으로 재구성하는, RDF 데이터 압축 장치.

청구항 19

청구항 18에 있어서,

상기 압축 모듈은,

상기 각 주제 별 트리의 트리플들에 대한 재구성 데이터에서 동일한 합계 값이 존재하는 경우, 중복되는 객체의 값을 빈 문자열로 대체하는, RDF 데이터 압축 장치.

청구항 20

청구항 13에 있어서,

상기 RDF 데이터 압축 장치는,

상기 술어 세트의 그룹별로 해당 그룹에 속하는 술어들에 대응하도록 상기 트리플들의 객체를 1차 재배열하고, 상기 술어 세트를 구성하는 각 술어 별로 순차적으로 그에 대응하는 객체를 2차 재배열하며, 상기 2차 재배열된 객체의 값들을 원래의 값보다 작은 범위의 값으로 정규화하여 매핑시키는 객체 매핑 모듈을 더 포함하는, RDF 데이터 압축 장치.

청구항 21

청구항 20에 있어서,

상기 RDF 데이터는, 복수 개의 지명 그래프를 포함하고,

상기 HDT 확장 모듈은,

상기 RDF 데이터를 직렬화 할 때, HDT 사전(HDT Dictionary)에 상기 지명 그래프들의 그래프 이름을 나타내기 위한 섹션을 추가하고, 상기 지명 그래프들에 특정 트리플이 있는지 여부를 나타내기 위한 그래프 정보를 생성하는, RDF 데이터 압축 장치.

청구항 22

청구항 21에 있어서,

상기 HDT 확장 모듈은,

상기 그래프 정보 생성 시, 트리플 인터프리터를 통해 각 트리플을 기준으로 해당 트리플이 존재하는 지명 그래프를 비트맵으로 생성하고, 그래프 인터프리터를 통해 각 지명 그래프를 기준으로 각 지명 그래프가 어떤 트리플을 가지는지를 비트맵으로 생성하는, RDF 데이터 압축 장치.

청구항 23

비일시적 컴퓨터 판독 가능한 저장 매체(non-transitory computer readable storage medium)에 저장된 컴퓨터 프로그램으로서,

상기 컴퓨터 프로그램은 하나 이상의 명령어들을 포함하고, 상기 명령어들은 하나 이상의 프로세서들을 갖는 컴퓨팅 장치에 의해 실행될 때, 상기 컴퓨팅 장치로 하여금,

주제(Subject), 술어(Predicate), 및 객체(Object)로 구성된 복수 개의 트리플(Triple)을 포함하는 RDF(Resource Description Framework) 데이터를 직렬화 하는 단계;

상기 직렬화 된 트리플들을 재배열하여 상기 직렬화된 트리플들에서 중복되는 부분을 감소시키는 단계; 및

상기 재배열된 트리플들을 압축시키는 단계를 수행하도록 하는, 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 개시되는 실시예는 RDF 데이터 압축 기술과 관련된다.

배 경 기 술

[0003] 최근, RDF(Resource Description Framework)와 같은 그래프 모델을 이용하여 시맨틱 웹에서 데이터를 표현하고 교환하고 있으며, RDF 데이터의 규모가 기하 급수적으로 증가하고 있다. RDF 데이터는 Subject(S : 주어부 또는 주제), Predicate(P : 서술부 또는 술어), 및 Object(O : 목적부 또는 객체)로 이루어지는 트리플(Triple)을 기본 단위로 표현된다.

[0004] RDF 데이터는 저장 및 교환을 위해 효율적으로 직렬화되어야 한다. 이때 JSON-LD, RDF/XML 또는 터틀과 같은 RDF 직렬화 형식을 통해 RDF 그래프를 일반 텍스트처럼 직렬화하는데, 그 결과 직렬화된 파일은 그 형식의 장황함으로 인해 데이터 볼륨이 커지게 된다. 따라서, RDF 데이터를 저장하기 위한 저장 공간을 줄이기 위해 RDF 데이터를 압축할 필요가 있다.

[0005] 한편, RDF 데이터를 직렬화 하는 효율적인 대안 중 하나는 HDT(Header-Dictionary-Triples)인데, HDT는 RDF 트리플의 각 용어(S, P, O)를 숫자 식별자로 매핑하게 된다. 여기서, HDT는 동일한 그래프에 존재하는 모든 트리

플을 고려하기 때문에 지명 그래프(Named Graphs)가 있는 RDF 데이터 셋을 관리할 수 없으나, RDF 데이터 셋은 [Subject, Predicate, Object, Graphname]으로 알려진 RDF 트리플의 확장인 다양한 지명 그래프를 포함하므로, 기존의 HDT로는 지명 그래프가 있는 RDF 데이터 셋을 직렬화 하여 관리하는데 한계가 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 공개특허공보 제10-2020-0094673호(2020.08.07)

발명의 내용

해결하려는 과제

[0008] 본 발명의 실시예는 지명 그래프를 포함하는 RDF 데이터 셋을 직렬화 하여 관리할 수 있는 RDF 데이터 압축 방법 및 장치를 제공하기 위한 것이다.

[0009] 본 발명의 실시예는 압축 효율을 높일 수 있는 RDF 데이터 압축 방법 및 장치를 제공하기 위한 것이다.

과제의 해결 수단

[0011] 개시되는 일 실시예에 따른 RDF 데이터 압축 방법은, 하나 이상의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서, 주제(Subject), 술어(Predicate), 및 객체(Object)로 구성된 복수 개의 트리플(Triple)을 포함하는 RDF(Resource Description Framework) 데이터를 직렬화 하는 단계; 상기 직렬화 된 트리플들을 재배열하여 상기 직렬화된 트리플들에서 중복되는 부분을 감소시키는 단계; 및 상기 재배열된 트리플들을 압축시키는 단계를 포함한다.

[0012] 상기 중복되는 부분을 감소시키는 단계는, 상기 직렬화 된 트리플들을 동일한 주제를 기준으로 그룹화 하여 상기 트리플들을 트리 형태로 배열하는 단계; 상기 트리 형태의 배열에서 동일한 술어 세트를 그룹화 하고, 그룹화 된 상기 술어 세트에 ID를 부여하는 단계; 및 상기 트리 형태의 배열에서 상기 술어 세트의 ID가 동일한 주제끼리 그룹화하여 재배열하는 단계를 포함할 수 있다.

[0013] 상기 중복되는 부분을 감소시키는 단계는, 상기 재배열된 각 주제의 ID를 순차적으로 배열하여 주제 매핑 시퀀스를 생성하는 단계를 더 포함할 수 있다.

[0014] 상기 중복되는 부분을 감소시키는 단계는, 상기 재배열된 주제들을 상기 술어 세트의 ID가 동일한 그룹 별로 각 주제에 대해 새로운 ID를 부여하는 단계; 및 상기 그룹 별로 각 주제에 대해 원본 ID와 상기 새로운 ID를 매핑하여 저장하는 단계를 더 포함할 수 있다.

[0015] 상기 압축시키는 단계는, 상기 재배열된 트리플들에서 각 주제 별 트리에 따른 트리플 값을 추출하여 순차적으로 배열하는 단계; 상기 각 주제 별 트리의 트리플들에 대한 합계 값(Summed Value)을 산출하는 단계; 및 상기 산출한 합계 값에 기반하여 각 주제 별 트리의 트리플들을 기 설정된 형태의 데이터로 재구성하는 단계를 포함할 수 있다.

[0016] 상기 합계 값은, 해당 주제의 새로 부여된 ID 값과 해당 주제의 트리플들 중 객체의 ID 값을 합산하여 산출할 수 있다.

[0017] 상기 재구성하는 단계는, 상기 각 주제 별 트리의 트리플들을 <합계 값, {주제(Subject), 객체1($Object_1$), 객체2($Object_2$),..., 객체n($Object_n$)}>의 집합으로 재구성할 수 있다.

[0018] 상기 압축시키는 단계는, 상기 각 주제 별 트리의 트리플들에 대한 재구성 데이터에서 동일한 합계 값이 존재하는 경우, 중복되는 객체의 값을 빈 문자열로 대체하는 단계를 더 포함할 수 있다.

[0019] 상기 RDF 데이터 압축 방법은, 상기 술어 세트의 그룹별로 해당 그룹에 속하는 술어들에 대응하도록 상기 트리플들의 객체를 1차 재배열하는 단계; 상기 술어 세트를 구성하는 각 술어 별로 순차적으로 그에 대응하는 객체를 2차 재배열하는 단계; 및 상기 2차 재배열된 객체의 값들을 원래의 값보다 작은 범위의 값으로 정규화하여 매핑시키는 단계를 더 포함할 수 있다.

[0020] 상기 RDF 데이터는, 복수 개의 지명 그래프를 포함하고, 상기 RDF 데이터 압축 방법은, 상기 RDF 데이터를 직렬화 할 때, HDT 사전(HDT Dictionary)에 상기 지명 그래프들의 그래프 이름을 나타내기 위한 섹션을 추가하는 단계; 및 상기 지명 그래프들에 특정 트리플이 있는지 여부를 나타내기 위한 그래프 정보를 생성하는 단계를 더 포함할 수 있다.

[0021] 상기 그래프 정보를 생성하는 단계는, 트리플 인터프리터를 통해 각 트리플을 기준으로 해당 트리플이 존재하는 지명 그래프를 비트맵으로 생성하는 단계; 및 그래프 인터프리터를 통해 각 지명 그래프를 기준으로 각 지명 그래프가 어떤 트리플을 가지는지를 비트맵으로 생성하는 단계를 포함할 수 있다.

[0022] 개시되는 일 실시예에 따른 RDF 데이터 압축 장치는, 주제(Subject), 술어(Predicate), 및 객체(Object)로 구성된 복수 개의 트리플(Triple)을 포함하는 RDF(Resource Description Framework) 데이터를 직렬화 하는 HDT 확장 모듈; 상기 직렬화 된 트리플들을 재배열하여 상기 직렬화된 트리플들에서 중복되는 부분을 감소시키는 중복 감소 모듈; 및 상기 재배열된 트리플들을 압축시키는 압축 모듈을 포함한다.

[0023] 상기 중복 감소 모듈은, 상기 직렬화 된 트리플들을 동일한 주제를 기준으로 그룹화 하여 상기 트리플들을 트리 형태로 배열하고, 상기 트리 형태의 배열에서 동일한 술어 세트를 그룹화 하고 그룹화 된 상기 술어 세트에 ID를 부여하며, 상기 트리 형태의 배열에서 상기 술어 세트의 ID가 동일한 주제끼리 그룹화하여 재배열할 수 있다.

[0024] 상기 중복 감소 모듈은, 상기 재배열된 각 주제의 ID를 순차적으로 배열하여 주제 매핑 시퀀스를 생성할 수 있다.

[0025] 상기 중복 감소 모듈은, 상기 재배열된 주제들을 상기 술어 세트의 ID가 동일한 그룹 별로 각 주제에 대해 새로운 ID를 부여하고, 상기 그룹 별로 각 주제에 대해 원본 ID와 상기 새로운 ID를 매핑하여 저장할 수 있다.

[0026] 상기 압축 모듈은, 상기 재배열된 트리플들에서 각 주제 별 트리에 따른 트리플 값을 추출하여 순차적으로 배열하고, 상기 각 주제 별 트리의 트리플들에 대한 합계 값(Summed Value)을 산출하며, 상기 산출한 합계 값에 기반하여 각 주제 별 트리의 트리플들을 기 설정된 형태의 데이터로 재구성할 수 있다.

[0027] 상기 합계 값은, 해당 주제의 새로 부여된 ID 값과 해당 주제의 트리플들 중 객체의 ID 값을 합산하여 산출할 수 있다.

[0028] 상기 압축 모듈은, 상기 각 주제 별 트리의 트리플들을 <합계 값, {주제(Subject), 객체1($Object_1$), 객체2($Object_2$),..., 객체n($Object_n$)}>의 집합으로 재구성할 수 있다.

[0029] 상기 압축 모듈은, 상기 각 주제 별 트리의 트리플들에 대한 재구성 데이터에서 동일한 합계 값이 존재하는 경우, 중복되는 객체의 값을 빈 문자열로 대체할 수 있다.

[0030] 상기 RDF 데이터 압축 장치는, 상기 술어 세트의 그룹별로 해당 그룹에 속하는 술어들에 대응하도록 상기 트리플들의 객체를 1차 재배열하고, 상기 술어 세트를 구성하는 각 술어 별로 순차적으로 그에 대응하는 객체를 2차 재배열하며, 상기 2차 재배열된 객체의 값들을 원래의 값보다 작은 범위의 값으로 정규화하여 매핑시키는 객체 매핑 모듈을 더 포함할 수 있다.

[0031] 상기 RDF 데이터는, 복수 개의 지명 그래프를 포함하고, 상기 HDT 확장 모듈은, 상기 RDF 데이터를 직렬화 할 때, HDT 사전(HDT Dictionary)에 상기 지명 그래프들의 그래프 이름을 나타내기 위한 섹션을 추가하고, 상기 지명 그래프들에 특정 트리플이 있는지 여부를 나타내기 위한 그래프 정보를 생성할 수 있다.

[0032] 상기 HDT 확장 모듈은, 상기 그래프 정보 생성 시, 트리플 인터프리터를 통해 각 트리플을 기준으로 해당 트리플이 존재하는 지명 그래프를 비트맵으로 생성하고, 그래프 인터프리터를 통해 각 지명 그래프를 기준으로 각 지명 그래프가 어떤 트리플을 가지는지를 비트맵으로 생성할 수 있다.

발명의 효과

[0034] 개시되는 실시예에 의하면, HDT 확장 모듈을 통해 각 지명 그래프의 그래프 이름을 나타내기 HDT 사전에 별도의 섹션을 추가하고 그래프 정보를 트리플 인터프리터 및 그래프 인터프리터를 통해 나타냄으로써, RDF 데이터 셋에서 지명 그래프도 압축 및 관리를 할 수 있으며, 각 그래프 별 트리플의 포함 여부를 효율적으로 검색할 수 있게 된다.

[0035] 또한, 중복 감소 모듈 및 객체 매핑 모듈을 통해 RDF 데이터 셋에서 중복된 부분을 감소시켜 메모리 공간 사용을 최적화 하면서 압축 효율을 높이고 데이터 손실을 줄일 수 있게 된다.

도면의 간단한 설명

[0037] 도 1은 본 발명의 일 실시예에 따른 RDF 데이터 압축 장치의 구성을 나타낸 도면

도 2은 본 발명의 일 실시예에 따른 지명 그래프를 갖는 RDF 데이터 셋을 나타낸 도면

도 3은 본 발명의 일 실시예에서 HDT 확장 모듈에 의한 HDT 사전 및 그래프 정보를 나타낸 도면

도 4는 본 발명의 일 실시예에서 HDT 확장 모듈이 트리플 인터프리터(TI) 및 그래프 인터프리터(GI)의 조합을 통해 그래프 정보를 나타낸 것을 표현한 도면

도 5는 본 발명의 일 실시예에 따른 RDF 트리플의 직렬화 된 값을 나타낸 도면

도 6은 본 발명의 일 실시예에서 직렬화 된 트리플을 동일한 주제를 기준으로 트리 형태로 배열한 상태를 나타낸 도면

도 7은 본 발명의 일 실시예에서 그룹화 된 술어 세트에 ID가 부여된 상태를 나타낸 도면

도 8은 본 발명의 일 실시예에서 동일한 술어 세트를 기준으로 주제를 재정렬한 상태를 나타낸 도면

도 9는 본 발명의 일 실시예에서 동일한 술어 세트를 기준으로 재정렬된 각 주제에 대해 새로운 ID가 부여된 상태를 나타낸 도면

도 10은 본 발명의 일 실시예에 따른 압축 모듈에서 재배열된 RDF 트리플을 압축시키는 과정을 나타낸 도면

도 11은 본 발명의 실시예에서 객체 매핑 모듈을 통해 직렬화된 RDF 트리플에서 객체(Object)의 반복을 탐지하는 과정을 나타낸 도면

도 12는 본 발명의 일 실시예에 따른 RDF 데이터 압축 방법을 설명하기 위한 흐름도

도 13은 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도

발명을 실시하기 위한 구체적인 내용

[0038] 이하, 도면을 참조하여 본 발명의 구체적인 실시형태를 설명하기로 한다. 이하의 상세한 설명은 본 명세서에서 기술된 방법, 장치 및/또는 시스템에 대한 포괄적인 이해를 돕기 위해 제공된다. 그러나 이는 예시에 불과하며 본 발명은 이에 제한되지 않는다.

[0039] 본 발명의 실시예들을 설명함에 있어서, 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 그리고, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다. 상세한 설명에서 사용되는 용어는 단지 본 발명의 실시예들을 기술하기 위한 것이며, 결코 제한적이어서는 안 된다. 명확하게 달리 사용되지 않는 한, 단수 형태의 표현은 복수 형태의 의미를 포함한다. 본 설명에서, "포함" 또는 "구비"와 같은 표현은 어떤 특성들, 숫자들, 단계들, 동작들, 요소들, 이들의 일부 또는 조합을 가리키기 위한 것이며, 기술된 것 이외에 하나 또는 그 이상의 다른 특성, 숫자, 단계, 동작, 요소, 이들의 일부 또는 조합의 존재 또는 가능성을 배제하도록 해석되어서는 안 된다.

[0040] 또한, 제1, 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용

어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로 사용될 수 있다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성 요소는 제2 구성 요소로 명명될 수 있고, 유사하게 제2 구성 요소도 제1 구성 요소로 명명될 수 있다.

[0042] 도 1은 본 발명의 일 실시예에 따른 RDF 데이터 압축 장치의 구성을 나타낸 도면이다.

[0043] 도 1을 참조하면, RDF 데이터 압축 장치(100)는 HDT(Header-Dictionary-Triples) 확장 모듈(102), 중복 감소 모듈(104), 압축 모듈(106), 및 객체 매핑 모듈(108)을 포함할 수 있다.

[0044] HDT 확장 모듈(102)은 RDF 데이터 세트인 지명 그래프(Named Graphs)들을 인코딩 하도록 마련될 수 있다. HDT 확장 모듈(102)은 기존의 HDT에서 지명 그래프를 인코딩(바이너리 직렬화) 하기 위해 확장된 기능을 포함할 수 있다. 지명 그래프는 URI를 사용하여 일련의 RDF(Resource Description Framework)를 식별하는 시맨틱 웹 아키텍처의 핵심 개념으로, 컨텍스트, 출처 정보 또는 기타 메타 데이터와 같은 정보들을 포함한다.

[0045] 도 2은 본 발명의 일 실시예에 따른 지명 그래프를 갖는 RDF 데이터 셋을 나타낸 도면이다. 도 2의 (a)를 참조하면, RDF 데이터 셋은 그래프 X(Graph X) 및 그래프 Y(Graph Y)를 가지며, 도 2의 (b)는 RDF 데이터 셋에 대한 트리플(Triples)을 나타내었다.

[0046] HDT 확장 모듈(102)은 HDT 사전(HDT Dictionary)에 새로운 구성 요소를 추가하여 각 지명 그래프의 그래프 이름을 저장할 수 있다. 구체적으로, HDT 확장 모듈(102)은 각 지명 그래프의 그래프 이름을 나타내기 HDT 사전에 별도의 섹션(G_i)을 추가할 수 있다. 또한, HDT 확장 모듈(102)은 특정 지명 그래프에 특정 트리플이 있는지 여부에 대한 정보를 나타내기 위한 그래프 정보(Graph Information)를 추가할 수 있다.

[0047] 도 3은 본 발명의 일 실시예에서 HDT 확장 모듈(102)에 의한 HDT 사전(도 3의 (a)) 및 그래프 정보(도 3의 (b))를 나타낸 도면이다. 도 3의 (a)를 참조하면, HDT 확장 모듈(102)은 그래프 X(Graph X) 및 그래프 Y(Graph Y)에 대한 각 트리플의 용어들을 분석하여 기 설정된 섹션 즉, SO(Subject, Object), S(Subject), O(Object), P(Predicate), 및 G(Graphname)의 섹션으로 분류하고, 각 섹션 별로 용어들에 대해 번호를 할당할 수 있다. 여기서, G(Graphname) 섹션은 각 지명 그래프의 그래프 이름을 나타내기 HDT 사전에 추가된 섹션이다.

[0048] 도 3의 (b)를 참조하면, HDT 확장 모듈(102)은 그래프 X 및 그래프 Y에 대해 그래프 정보를 추가할 수 있다. 여기서, 그래프 정보(GI)는 각 그래프에 특정 트리플이 있는지 없는지의 정보를 나타내기 위한 부울 행렬(Boolean Matrix)로 표현될 수 있다. 즉, 도 2에서 RDF 데이터 셋은 총 11개의 트리플을 가지며(각 트리플은 고유의 ID가 부여될 수 있음), 그래프 X 및 그래프 Y가 각각 11개의 특정 트리플을 가지고 있는지 여부를 그래프 정보로 나타낼 수 있다. 여기서, 각 그래프에 특정 트리플이 있으면 1로 나타내고, 특정 트리플이 없으면 0으로 나타낼 수 있다.

[0049] 여기서, HDT 확장 모듈(102)은 그래프 정보의 행렬을 보다 쉽고 효율적으로 나타내기 위해 트리플 인터프리터(Triple Interpreter : TI) 및 그래프 인터프리터(Graph Interpreter : GI)를 포함할 수 있다. 도 4는 본 발명의 일 실시예에서 HDT 확장 모듈(102)이 트리플 인터프리터(TI) 및 그래프 인터프리터(GI)의 조합을 통해 그래프 정보를 나타낸 것을 표현한 도면이다.

[0050] 도 4를 참조하면, 트리플 인터프리터(TI)는 각 트리플(t_i)을 기준으로 해당 트리플이 존재하는 그래프를 비트맵으로 나타내기 위한 것이다. 여기서, 비트맵의 크기는 그래프의 개수와 동일하다. 트리플 인터프리터(TI)는 비트맵 B_m^{TI}로 나타낼 수 있다. 여기서, m은 데이터 셋의 트리플 개수일 수 있다. 트리플 인터프리터(TI)는 각 트리플 당 비트맵을 생성하므로 11개의 비트맵을 생성할 수 있다. $B_i^{TI}[j]$ =1은 i번째 트리플이 j번째 그래프에 있다는 것을 나타내고, $B_i^{TI}[j]$ =0은 i번째 트리플이 j번째 그래프에 없다는 것을 나타낸다.

[0051] 예를 들어, 1번째 트리플(t_1)에 대한 트리플 인터프리터(IT)의 비트맵(B_1^{TI})은 {1,1}이 되고, 2번째 트리플(t_2)에 대한 트리플 인터프리터(IT)의 비트맵(B_2^{TI})은 {1,1}이 되며, 3번째 트리플(t_3)에 대한 트리플 인터프리터(IT)의 비트맵(B_3^{TI})은 {1,0}이 된다.

[0052] 그래프 인터프리터(GI)는 트리플 인터프리터(TI)와 직교하는 구성을 가질 수 있다. 그래프 인터프리터(GI)는 각 그래프를 기준으로 각 그래프가 어떤 트리플을 갖는지의 여부를 비트맵으로 나타내기 위한 것이다. 여기서, 비트맵의 크기는 트리플의 개수와 동일하다. 그래프 인터프리터(GI)는 비트맵 B_g^{GI}로 나타낼 수 있다. 여기서, g는 데이터 셋의 그래프 개수일 수 있다. 그래프 인터프리터(GI)는 각 그래프당 비트맵을 생성하므로 2개의 비트맵을 생성할 수 있다.

[0053] $B_i^{GI}[j]$ =1은 j번째 트리플이 i번째 그래프에 있다는 것을 나타내고, $B_i^{GI}[j]$ =0은 j번째 트리플이 i번째 그래프에 없다는 것을 나타낸다. 도 4에서, 그래프 X에 대한 그래프 인터프리터(GI)의 비트맵(B_1^{GI})은 {1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1}이고, 그래프 Y에 대한 그래프 인터프리터(GI)의 비트맵(B_2^{GI})은 {1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1}이 된다.

[0054] 이와 같이, 트리플 인터프리터(TI) 및 그래프 인터프리터(GI)를 조합하여 그래프 정보를 표현함으로써, 데이터 셋에서 각 그래프 별로 어떤 트리플을 포함하고 있는지에 대한 정보를 쉽게 찾을 수 있게 된다.

[0055] 중복 감소 모듈(104)은 RDF 데이터에서 중복적인 부분을 제거할 수 있다. 즉, RDF 데이터에는 그 특성 상 의미론적(Semantic), 기호적(Symbolic), 구문론적(Syntactic) 중복이 발생하게 된다. 의미론적 중복은 더 적은 수의 트리플을 사용하여 정보를 표시할 때 발생하고, 기호적 중복은 데이터 셋에 불필요한 기호 반복이 있을 때 발생하며, 구문론적 중복은 RDF 그래프 직렬화 및 기본 그래프 구조에 따라 발생하게 된다.

[0056] 중복 감소 모듈(104)은 RDF 트리플의 직렬화 된 값을 이용하여 RDF 트리플을 재배열 및 재구성하는 것을 통해 RDF 트리플에 포함된 중복을 제거 할 수 있다. 예시적인 실시예에서, RDF 트리플은 터틀(Turtle)을 통해 직렬화 할 수 있으나, 이에 한정되는 것은 아니며 기존의 다양한 직렬화 방식을 사용할 수 있음은 물론이다.

[0057] 예를 들어, 터틀(Turtle) 직렬화 방식의 경우, <Subject, Predicate, Object>는 공백으로 구분되고 마침표(.)로 끝나게 된다. 그리고, Subject, Predicate, Object를 나타내기 위해 십진수 값이 각각 ID(식별자)로 사용될 수 있다. 도 5는 본 발명의 일 실시예에 따른 RDF 트리플의 직렬화 된 값을 나타낸 도면이다.

[0058] 중복 감소 모듈(104)은 직렬화된 트리플들을 동일한 주제(Subject)를 기준으로 그룹화 할 수 있다. 이때, 중복 감소 모듈(104)은 동일한 주제(Subject)를 기준으로 트리플들을 트리 형태로 배열할 수 있다. 즉, 특정 주제(Subject)를 루트로 하고, 그에 대응하는 술어(Predicate) 및 객체(Object)를 트리로 하는 트리 형태로 배열할 수 있다.

[0059] 도 6은 본 발명의 일 실시예에서 직렬화 된 트리플을 동일한 주제를 기준으로 트리 형태로 배열한 상태를 나타낸 도면이다. 도 6을 참조하면, 직렬화 된 트리플이 주제(Subject) 1부터 주제 9까지 트리 형태로 배열된 것을 볼 수 있다. 그리고, 주제(Subject) 1의 술어(Predicate) 목록은 1, 3, 6, 7, 8이며, 이와 동일한 5개의 술어가 주제 5에서도 사용되는 것을 볼 수 있다. 즉, 주제(Subject) 1과 주제(Subject) 5는 동일한 술어 구조를 사용한다. 여기서, 동일한 술어 구조에서 술어 ID가 제일 높은 술어(도 6에서는 술어 9)를 RDF:Type 술어로 나타낼 수 있다.

[0060] 중복 감소 모듈(104)은 주제를 기준으로 그룹화 한 트리 배열에서 동일한 술어 세트를 그룹화 하고, 그룹화 된 술어 세트에 ID를 부여할 수 있다. 도 7은 본 발명의 일 실시예에서 그룹화 된 술어 세트에 ID가 부여된 상태를 나타낸 도면이다. 먼저, 도 6을 보면, 주제를 그룹으로 그룹화 된 트리 배열에서 동일한 술어 세트가 4개({1,3,6,7}, {2,4,5}, {1,5,8,9}, {3,7})가 있음을 알 수 있다.

[0061] 여기서, 중복 감소 모듈(104)은 도 7에 도시된 바와 같이 술어 세트 {1,3,6,7}(Group1)에는 ID 1을 부여하고, 술어 세트 {2,4,5}(Group2)에는 ID 2를 부여하며, 술어 세트 {1,5,8,9}(Group3)에는 ID 3을 부여하고, 술어 세트 {3,7}(Group4)에는 ID 4를 부여할 수 있다.

[0062] 중복 감소 모듈(104)은 술어 세트에 부여된 ID를 기반으로 술어 매핑 시퀀스(Predicate Mapping Sequence)를 생성할 수 있다. 중복 감소 모듈(104)은 술어 세트에 부여된 ID를 순차적으로 배열하여 술어 매핑 시퀀스(1-2-3-2-1-3-2-4-3)를 생성할 수 있다.

[0063] 중복 감소 모듈(104)은 동일한 술어 세트를 기준으로 주제를 재정렬할 수 있다. 도 8은 본 발명의 일 실시예에서 동일한 술어 세트를 기준으로 주제를 재정렬한 상태를 나타낸 도면이다. 도 8을 참조하면, 중복 감소 모듈

(104)은 술어 세트의 ID가 동일한 주제끼리 재정렬하여 그룹화 할 수 있다. 즉, 술어 세트 ID가 1인 주제 1과 주제 5를 그룹 1로 묶고, 술어 세트 ID가 2인 주제 2, 주제 4, 주제 7을 그룹 2로 묶고, 술어 세트 ID가 3인 주제 3, 주제 6, 주제 9를 그룹 3으로 묶고, 술어 세트 ID가 4인 주제 8을 그룹 4로 묶을 수 있다.

[0064] 중복 감소 모듈(104)은 재정렬된 주제의 ID를 기반으로 주제 매핑 시퀀스(Subject Mapping Sequence)를 생성할 수 있다. 중복 감소 모듈(104)은 재정렬된 주제의 ID를 순차적으로 배열하여 주제 매핑 시퀀스(1-5-2-4-7-3-6-9-8)를 생성할 수 있다.

[0065] 중복 감소 모듈(104)은 동일한 술어 세트를 기준으로 재정렬된 각 주제에 대해 새로운 ID를 부여할 수 있다. 중복 감소 모듈(104)은 동일한 술어 세트를 기준으로 그룹화된 주제들을 대상으로 그룹 별로 각 주제에 대해 새로운 식별자(즉, ID)를 부여할 수 있다. 도 9는 본 발명의 일 실시예에서 동일한 술어 세트를 기준으로 재정렬된 각 주제에 대해 새로운 ID가 부여된 상태를 나타낸 도면이다.

[0066] 도 9를 참조하면, 중복 감소 모듈(104)은 술어 세트 ID가 1인 주제 1과 주제 5를 포함하는 그룹 1에서 각 주제에 대해 주제 1^*, 주제 2^*라는 새로운 ID를 부여할 수 있다. 또한, 중복 감소 모듈(104)은 술어 세트 ID가 2인 주제 2, 주제 4, 주제 7을 포함하는 그룹 2에서 각 주제에 대해 주제 1^*, 주제 2^*, 주제 3^*이라는 새로운 ID를 부여할 수 있다. 또한, 중복 감소 모듈(104)은 술어 세트 ID가 3인 주제 3, 주제 6, 주제 9를 포함하는 그룹 3에서 각 주제에 대해 주제 1^*, 주제 2^*, 주제 3^*이라는 새로운 ID를 부여할 수 있다.

[0067] 이때, 중복 감소 모듈(104)은 각 그룹 별로 각 주제에 대해 원본 ID와 새로 부여된 ID를 매핑하여 저장할 수 있다. 예를 들어, 중복 감소 모듈(104)은 그룹 1에서 주제 1과 주제 1^*을 매핑하고, 주제 5와 주제 2^*를 매핑하여 저장할 수 있다.

[0068] 압축 모듈(106)은 중복 감소 모듈(104)에 의해 재배열된 RDF 트리플을 압축시킬 수 있다. 예시적인 실시예에서, 압축 모듈(106)은 차등 압축기(Differential Compressor)를 이용하여 재배열된 RDF 트리플을 압축시킬 수 있다. 압축 모듈(106)은 음의 차등 연산(Negative Differential Operation)을 통해 재배열된 RDF 트리플을 압축하고, 양의 차등 연산(Positive Differential Operation)을 통해 압축된 RDF 트리플을 압축 해제할 수 있다.

[0069] 압축 모듈(106)은 다음의 수학식 1 및 수학식 2에 의한 음의 차등 연산을 통해 재배열된 RDF 트리플을 압축할 수 있다.

[0070] (수학식 1)

[0071] $$\gamma = \alpha - \beta = (\gamma_1, .., \gamma_i, .., \gamma_n)$$

[0072] (수학식 2)

[0073] $$\Gamma_i \in [1,n], \gamma_i = \begin{cases} " \ " & if(\alpha_i = \beta_i) \\ \alpha_i & else. \end{cases}$$

[0074] 여기서, α = (α_1, α_2, 쩐 , α_n)으로 주제(Subject)에 대응하는 입력 벡터이고, β = (β_1, β_2, ... , β_n)으로 객체(Object)에 대응하는 입력 벡터일 수 있다. 수학식 1 및 2에 의하면, 벡터 ɣ는 입력 벡터 α와 입력 벡터 β의 값이 같으면 빈 문자열("")을 출력하고, 입력 벡터 α와 입력 벡터 β의 값이 같지 않으면 입력 벡터 α의 값을 출력할 수 있다. 예를 들어, α = (1, 5, 8)이고, β = (3, 5, 9)인 경우, ɣ = α - β = (1, "", 8)이 된다.

[0075] 또한, 압축 모듈(106)은 다음의 수학식 3 및 수학식 4에 의한 양의 차등 연산을 통해 압축된 RDF 트리플을 압축 해제할 수 있다.

[0076] (수학식 3)

[0077] $$\gamma = \alpha + \beta = (\gamma_1, .., \gamma_i, .., \gamma_n)$$

[0078] (수학식 4)

[0079] $$\Gamma_i \in [1,n], \gamma_i = \begin{cases} \alpha_i & if(\alpha_i \neq " \ ") \\ \beta_i & else. \end{cases}$$

[0080] 수학식 3 및 4에 의하면, 벡터 γ는 입력 벡터 α의 값이 빈 문자열("")이 아닌 경우 입력 벡터 α의 값을 출력하고, 입력 벡터 α의 값이 빈 문자열("")인 경우 입력 벡터 β의 값을 출력할 수 있다. 예를 들어, α = (5, "", "")이고, β = (4, 2, 3)인 경우, γ = α + β = (5, 2, 3)이 된다.

[0081] 한편, 도 10은 본 발명의 일 실시예에 따른 압축 모듈(106)에서 재배열된 RDF 트리플을 압축시키는 과정을 나타낸 도면이다. 도 10을 참조하면, 먼저 압축 모듈(106)은 중복 감소 모듈(104)에 의해 재배열된 RDF 트리플에서 각 주제 별 트리에 따른 트리플 값을 추출할 수 있다(도 10에서 (a)→(b)의 과정).

[0082] 예를 들어, 압축 모듈(106)은 도 9의 그룹 1에 속한 주제 1^*에 대해 다음과 같은 트리플 값 (1^*, 1, 1), (1^*, 1, 6), (1^*, 1, 8), (1^*, 1, 7), (1^*, 1, 5)을 추출할 수 있다. 또한, 압축 모듈(106)은 도 9의 그룹 1에 속한 주제 2^*에 대해 다음과 같은 트리플 값 (2^*, 1, 4), (2^*, 1, 8), (2^*, 1, 11), (2^*, 1, 5)을 추출할 수 있다.

[0083] 다음으로, 압축 모듈(106)은 각 주제 별 트리에 따른 트리플 값들을 순차적으로 배열할 수 있다(도 10의 (b)→(c)의 과정). 예를 들어, 압축 모듈(106)은 그룹 1에 속한 주제 1^*의 트리플 값 (1^*, 1, 1), (1^*, 1, 6), (1^*, 1, 8), (1^*, 1, 7), (1^*, 1, 5)들을 (1^*, 1, 1, 1^*, 1, 6, 1^*, 1, 8, 1^*, 1, 7, 1^*, 1, 5)으로 배열할 수 있다. 또한, 압축 모듈(106)은 그룹 1에 속한 주제 2^*의 트리플 값 (2^*, 1, 4), (2^*, 1, 8), (2^*, 1, 11), (2^*, 1, 5)들을 (2^*, 1, 4, 2^*, 1, 8, 2^*, 1, 11, 2^*, 1, 5)으로 배열할 수 있다.

[0084] 다음으로, 압축 모듈(106)은 각 주제 별 트리의 트리플들에 대한 합계 값(Summed Value)을 산출할 수 있다(도 10에서 (d)의 과정). 여기서, 합계 값은 해당 주제의 새로 부여된 ID 값과 해당 주제의 트리플들 중 객체의 ID 값을 합산하여 산출할 수 있다.

[0085] 예를 들어, 그룹 1에 속한 주제 1^*의 트리플들에 대한 합계 값은 해당 주제의 새로 부여된 ID 값인 1과 각 트리플에서 객체의 ID 값인 1, 6, 8, 7, 5을 합산한 28(1+1+6+8+7+5=28)이 된다. 또한, 그룹 1에 속한 주제 2^*의 트리플들에 대한 합계 값은 해당 주제의 새로 부여된 ID 값인 2와 각 트리플에서 객체의 ID 값인 4, 8, 11, 5를 합산한 30(2+4+8+11+5=30)이 된다.

[0086] 다음으로, 압축 모듈(106)은 산출한 합계 값에 기반하여 각 주제 별 트리의 트리플들을 기 설정된 형태의 데이터로 재구성할 수 있다(도 10의 (e) 과정). 예시적인 실시예에서, 압축 모듈(106)은 산출한 합계 값으로 해당 주제의 트리플들의 술어(Predicate)의 ID 값을 대체할 수 있다.

[0087] 압축 모듈(106)은 각 주제 별 트리의 트리플들을 <합계 값, {주제(Subject), 객체1($Object_1$), 객체2($Object_2$),..., 객체n($Object_n$)}>의 집합으로 재구성할 수 있다. 예를 들어, 압축 모듈(106)은 그룹 1에 속한 주제 1^*의 트리플 값 (1^*, 1, 1, 1^*, 1, 6, 1^*, 1, 8, 1^*, 1, 7, 1^*, 1, 5)을 (28, 1^*, 1, 6, 8, 7, 5)로 재구성할 수 있다. 또한, 압축 모듈(106)은 그룹 1에 속한 주제 2^*의 트리플 값 (2^*, 1, 4, 2^*, 1, 8, 2^*, 1, 11, 2^*, 1, 5)을 (30, 2^*, 4, 8, 11, 5)로 재구성할 수 있다.

[0088] 다음으로, 압축 모듈(106)은 각 주제 별 트리의 트리플들에 대한 재구성 데이터에서 동일한 합계 값이 존재하는지 여부를 확인할 수 있다. 동일한 합계 값이 존재하는 경우, 압축 모듈(106)은 음의 차등 연산을 이용하여 중복된 객체의 값을 빈 문자열로 대체할 수 있다(도 10의 (f) 과정).

[0089] 즉, 도 10의 (e)를 참조하면, 2번째 트리에 대한 트리플들의 재구성 데이터의 합계 값과 7번째 트리에 대한 트리플들의 재구성 데이터의 합계 값이 30으로 동일한 것을 볼 수 있다. 이 경우, 압축 모듈(106)은 뒤에 나오는 7번째 트리에 대한 트리플들의 재구성 데이터에서 2번째 트리에 대한 트리플들의 재구성 데이터와 중복되는 객체의 값인 8을 빈 문자열로 대체할 수 있다. 즉, 7번째 트리에 대한 트리플들의 재구성 데이터는 (30, 2^*, 7, 8, 9, 4)에서 (30, 2^*, 7, , 9, 4)로 변경될 수 있다.

[0090] 또한, 4번째 트리에 대한 트리플들의 재구성 데이터의 합계 값과 9번째 트리에 대한 트리플들의 재구성 데이터의 합계 값이 27로 동일한 것을 볼 수 있다. 이 경우, 압축 모듈(106)은 뒤에 나오는 9번째 트리에 대한 트리플들의 재구성 데이터에서 4번째 트리에 대한 트리플들의 재구성 데이터와 중복되는 객체의 값인 9와 4를 각각 빈

문자열로 대체할 수 있다. 즉, 9번째 트리에 대한 트리플들의 재구성 데이터는 (27, 1*, 9, 4, 13)에서 (27, 1*, , , 13)로 변경될 수 있다.

[0091] 다음으로, 압축 모듈(106)은 각 주제 별 트리의 트리플들에 대한 재구성 데이터에 기반하여 해쉬 테이블을 생성하여 중복 감소 모듈(104)에 의해 재배열된 RDF 트리플을 압축할 수 있다(도 10의 (g) 과정).

[0092] 여기서, 해쉬 테이블은 Total과 Values로 구성될 수 있다. Total에는 재구성 데이터 중 합계 값이 입력될 수 있다. Values에는 재구성 데이터 중 합계 값을 제외한 데이터 즉, {주제(Subject), 객체1($Object_1$), 객체2($Object_2$),..., 객체n($Object_n$)}가 입력될 수 있다.

[0093] 압축 모듈(106)은 각 주제 별 트리의 트리플들에 대한 재구성 데이터들에 대해 순차적으로 Total과 Values를 입력하여 해쉬 테이블을 생성할 수 있다. 이때, 재구성 데이터들 중 동일한 합계 값이 나오는 경우 이전의 Values를 새로운 재구성 데이터의 {주제(Subject), 객체1($Object_1$), 객체2($Object_2$),..., 객체n($Object_n$)}으로 대체할 수 있다.

[0094] 즉, 6번째 트리의 트리플들에 대한 재구성 데이터는 그 합계 값이 3번째 트리의 트리플들에 대한 재구성 데이터의 합계 값과 동일(32로 동일)하다. 이 경우, 압축 모듈(106)은 해쉬 테이블에서 3번째 트리의 Values를 6번째 트리의 트리플들에 대한 재구성 데이터 중 합계 값을 제외한 데이터(즉, 1*, 3, 2, 10, 11, 5))를 입력할 수 있다.

[0095] 객체 매핑 모듈(108)은 직렬화된 RDF 트리플에서 객체(Object)의 반복을 탐지할 수 있다. 도 11은 본 발명의 실시예에서 객체 매핑 모듈(108)을 통해 직렬화된 RDF 트리플에서 객체(Object)의 반복(또는 객체의 규칙성)을 탐지하는 과정을 나타낸 도면이다.

[0096] 도 11을 참조하면, 먼저 객체 매핑 모듈(108)은 술어 세트의 그룹(Predict Group) 별로 해당 그룹에 속하는 술어들에 대응하도록 RDF 트리플의 객체를 재배열할 수 있다(도 11의 (a)). 즉, 술어 세트의 제1 그룹은 {1,3,6,7}의 술어를 포함하고, 술어 세트의 제2 그룹은 {2,4,5}의 술어를 포함하며, 술어 세트의 제3 그룹은 {1,5,8,9}의 술어를 포함하고, 술어 세트의 제4 그룹은 {3,7}의 술어를 포함한다. 예를 들어, 술어 세트의 제1 그룹에서 술어 1에 대응하는 객체는 1, 4가 있으며, 술어 세트의 제1 그룹에서 술어 3에 대응하는 객체는 6, 8, 8이 있고, 술어 세트의 제1 그룹에서 술어 6에 대응하는 객체는 7, 11이 있고, 술어 세트의 제1 그룹에서 술어 7에 대응하는 객체는 5,5가 있다. 이와 같이, 각 술어 세트의 그룹 별로 해당 그룹에 속하는 술어들에 대응하도록 객체를 재배열할 수 있다.

[0097] 다음으로, 객체 매핑 모듈(108)은 술어 세트를 구성하는 각 술어(Predicate) 별로 순차적으로 그에 대응하는 객체를 재배열할 수 있다(도 11의 (b)). 즉, 술어 1은 술어 세트의 제1 그룹과 제3 그룹에 속하여 있고 그에 대응하는 객체는 각각 1, 4 및 3, 7, 5가 된다. 술어 2는 술어 세트의 제2 그룹에 속하여 있고 그에 대응하는 객체는 8, 9, 1이 된다. 술어 3은 술어 세트의 제1 그룹과 제4 그룹에 속하여 있고 그에 대응하는 객체는 각각 6, 8, 8 및 12가 된다. 이런 방식으로, 술어 세트를 구성하는 각 술어(Predicate) 별로 순차적으로 그에 대응하는 객체를 재배열할 수 있다.

[0098] 다음으로, 객체 매핑 모듈(108)은 술어 세트를 구성하는 각 술어(Predicate) 별로 순차적으로 재배열된 객체의 값(객체의 ID 값)들을 기 설정된 방식으로 정규화하여 매핑 시킬 수 있다(도 11의 (c)). 즉, 상기 재배열된 객체의 값들을 보다 작은 범위의 값으로 정규화하여 매핑 시킬 수 있다. 예를 들어, 도 11의 (b)에서 술어 5에서 술어 세트의 제2 그룹에 대응하는 객체 (2, 10, 8, 2)를 [1,2,3,4] 범위에 속하는 더 작은 값을 가지도록 정규화 하여 매핑하면 (1, 3, 2, 1)이 된다. 또한, 술어 4에 대응하는 객체(13, 4, 3)을 [1,2,3] 범위에 속하는 더 작은 값을 가지도록 정규화 하여 매핑하면 (3, 2, 1)이 된다. 객체 매핑 모듈(108)은 이러한 매핑을 추적하기 위해 객체 매핑 시퀀스(Object Mapping Sequence)를 생성할 수 있다. 객체 매핑 시퀀스는 매핑 전 객체의 원래 ID 값이 어느 술어에 속하여 있는지에 대한 정보를 포함하도록 생성될 수 있다.

[0099] 개시되는 실시예에 의하면, HDT 확장 모듈(102)을 통해 각 지명 그래프의 그래프 이름을 나타내기 HDT 사전에 별도의 섹션을 추가하고 그래프 정보를 트리플 인터프리터 및 그래프 인터프리터를 통해 나타냄으로써, RDF 데이터 셋에서 지명 그래프도 압축 및 관리를 할 수 있으며, 각 그래프 별 트리플의 포함 여부를 효율적으로 검색할 수 있게 된다.

[0100] 또한, 중복 감소 모듈(104) 및 객체 매핑 모듈(108)을 통해 RDF 데이터 셋에서 중복된 부분을 감소시켜 메모리

공간 사용을 최적화 하면서 압축 효율을 높이고 데이터 손실을 줄일 수 있게 된다.

[0101] 본 명세서에서 모듈이라 함은, 본 발명의 기술적 사상을 수행하기 위한 하드웨어 및 상기 하드웨어를 구동하기 위한 소프트웨어의 기능적, 구조적 결합을 의미할 수 있다. 예컨대, 상기 "모듈"은 소정의 코드와 상기 소정의 코드가 수행되기 위한 하드웨어 리소스의 논리적인 단위를 의미할 수 있으며, 반드시 물리적으로 연결된 코드를 의미하거나, 한 종류의 하드웨어를 의미하는 것은 아니다.

[0103] 도 12는 본 발명의 일 실시예에 따른 RDF 데이터 압축 방법을 설명하기 위한 흐름도이다. 도시된 흐름도에서는 상기 방법을 복수 개의 단계로 나누어 기재하였으나, 적어도 일부의 단계들은 순서를 바꾸어 수행되거나, 다른 단계와 결합되어 함께 수행되거나, 생략되거나, 세부 단계들로 나뉘어 수행되거나, 또는 도시되지 않은 하나 이상의 단계가 부가되어 수행될 수 있다.

[0104] 도 12를 참조하면, RDF 데이터 압축 장치(100)는 RDF 데이터를 직렬화 하고 HDT 사전(HDT Dictionary)에 각 지명 그래프의 그래프 이름을 저장하기 위한 섹션을 추가할 수 있다(S 101).

[0105] 다음으로, RDF 데이터 압축 장치(100)는 각 지명 그래프에 특정 트리플이 있는지 여부를 나타내는 그래프 정보를 생성할 수 있다(S 103). 여기서, 그래프 정보는 트리플 인터프리터(Triple Interpreter : TI) 및 그래프 인터프리터(Graph Interpreter : GI)를 통해 나타낼 수 있다.

[0106] 다음으로, RDF 데이터 압축 장치(100)는 직렬화 된 RDF 트리플을 동일한 주제를 기준으로 그룹화 한 후(S 105), 주제를 기준으로 그룹화 된 트리 배열에서 동일한 술어 세트를 그룹화 하며, 그룹화 된 술어 세트에 ID를 부여할 수 있다(S 107).

[0107] 다음으로, RDF 데이터 압축 장치(100)는 동일한 술어 세트를 기준으로 주제를 재정렬할 수 있다(S 109). 즉, RDF 데이터 압축 장치(100)는 술어 세트의 ID가 동일한 주제끼리 재정렬하여 그룹화 할 수 있다.

[0108] 다음으로, RDF 데이터 압축 장치(100)는 동일한 술어 세트를 기준으로 재정렬된 각 주제에 대해 새로운 ID를 부여할 수 있다(S 111). 이때, RDF 데이터 압축 장치(100)는 각 그룹 별로 각 주제에 대해 원본 ID와 새로 부여된 ID를 매핑하여 저장할 수 있다.

[0109] 다음으로, RDF 데이터 압축 장치(100)는 각 주제 별 트리에 따른 트리플 값을 추출하여 순차적으로 배열할 수 있다(S 113).

[0110] 다음으로, RDF 데이터 압축 장치(100)는 각 주제 별 트리의 트리플들에 대한 합계 값(Summed Value)을 산출할 수 있다(S 115). 여기서, 합계 값은 해당 주제의 새로 부여된 ID 값과 해당 주제의 트리플들 중 객체의 ID 값을 합산하여 산출할 수 있다.

[0111] 다음으로, RDF 데이터 압축 장치(100)는 산출한 합계 값에 기반하여 각 주제 별 트리의 트리플들을 기 설정된 형태의 데이터로 재구성할 수 있다(S 117).

[0112] 이때, RDF 데이터 압축 장치(100)는 각 주제 별 트리의 트리플들을 <합계 값, {주제(Subject), 객체1($Object_1$), 객체2($Object_2$),..., 객체n($Object_n$)}>의 집합으로 재구성할 수 있다.

[0113] 다음으로, RDF 데이터 압축 장치(100)는 각 주제 별 트리의 트리플들에 대한 재구성 데이터에서 동일한 합계 값이 존재하는 경우, 음의 차등 연산을 이용하여 중복된 객체의 값을 빈 문자열로 대체할 수 있다(S 119).

[0114] 다음으로, RDF 데이터 압축 장치(100)는 각 주제 별 트리의 트리플들에 대한 재구성 데이터에 기반하여 해쉬 테이블을 생성하여 RDF 트리플을 압축할 수 있다(S 121).

[0115] 한편, RDF 데이터 압축 방법에는 직렬화된 RDF 트리플에서 객체의 반복(또는 규칙성)을 탐지하는 위한 과정이 추가될 수 있다.

[0117] 도 13은 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경(10)을 예시하여 설명하기 위한 블록도이다. 도시된 실시예에서, 각 컴포넌트들은 이하에 기술된 것 이외에 상이한 기능 및 능력을 가질 수 있고, 이하에 기술된 것 이외에도 추가적인 컴포넌트를 포함할 수 있다.

[0118] 도시된 컴퓨팅 환경(10)은 컴퓨팅 장치(12)를 포함한다. 일 실시예에서, 컴퓨팅 장치(12)는 RDF 데이터 압축 장

치(100)일 수 있다.

[0119] 컴퓨팅 장치(12)는 적어도 하나의 프로세서(14), 컴퓨터 판독 가능 저장 매체(16) 및 통신 버스(18)를 포함한다. 프로세서(14)는 컴퓨팅 장치(12)로 하여금 앞서 언급된 예시적인 실시예에 따라 동작하도록 할 수 있다. 예컨대, 프로세서(14)는 컴퓨터 판독 가능 저장 매체(16)에 저장된 하나 이상의 프로그램들을 실행할 수 있다. 상기 하나 이상의 프로그램들은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 상기 컴퓨터 실행 가능 명령어는 프로세서(14)에 의해 실행되는 경우 컴퓨팅 장치(12)로 하여금 예시적인 실시예에 따른 동작들을 수행하도록 구성될 수 있다.

[0120] 컴퓨터 판독 가능 저장 매체(16)는 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능 저장 매체(16)에 저장된 프로그램(20)은 프로세서(14)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능 저장 매체(16)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 컴퓨팅 장치(12)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.

[0121] 통신 버스(18)는 프로세서(14), 컴퓨터 판독 가능 저장 매체(16)를 포함하여 컴퓨팅 장치(12)의 다른 다양한 컴포넌트들을 상호 연결한다.

[0122] 컴퓨팅 장치(12)는 또한 하나 이상의 입출력 장치(24)를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(22) 및 하나 이상의 네트워크 통신 인터페이스(26)를 포함할 수 있다. 입출력 인터페이스(22) 및 네트워크 통신 인터페이스(26)는 통신 버스(18)에 연결된다. 입출력 장치(24)는 입출력 인터페이스(22)를 통해 컴퓨팅 장치(12)의 다른 컴포넌트들에 연결될 수 있다. 예시적인 입출력 장치(24)는 포인팅 장치(마우스 또는 트랙패드 등), 키보드, 터치 입력 장치(터치패드 또는 터치스크린 등), 음성 또는 소리 입력 장치, 다양한 종류의 센서 장치 및/또는 촬영 장치와 같은 입력 장치, 및/또는 디스플레이 장치, 프린터, 스피커 및/또는 네트워크 카드와 같은 출력 장치를 포함할 수 있다. 예시적인 입출력 장치(24)는 컴퓨팅 장치(12)를 구성하는 일 컴포넌트로서 컴퓨팅 장치(12)의 내부에 포함될 수도 있고, 컴퓨팅 장치(12)와는 구별되는 별개의 장치로 컴퓨팅 장치(12)와 연결될 수도 있다.

[0124] 이상에서 본 발명의 대표적인 실시예들을 상세하게 설명하였으나, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 상술한 실시예에 대하여 본 발명의 범주에서 벗어나지 않는 한도 내에서 다양한 변형이 가능함을 이해할 것이다. 그러므로 본 발명의 권리범위는 설명된 실시예에 국한되어 정해져서는 안 되며, 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등한 것들에 의해 정해져야 한다.

부호의 설명

[0126] 100 : RDF 데이터 압축 장치

102 : HDT 확장 모듈

104 : 중복 감소 모듈

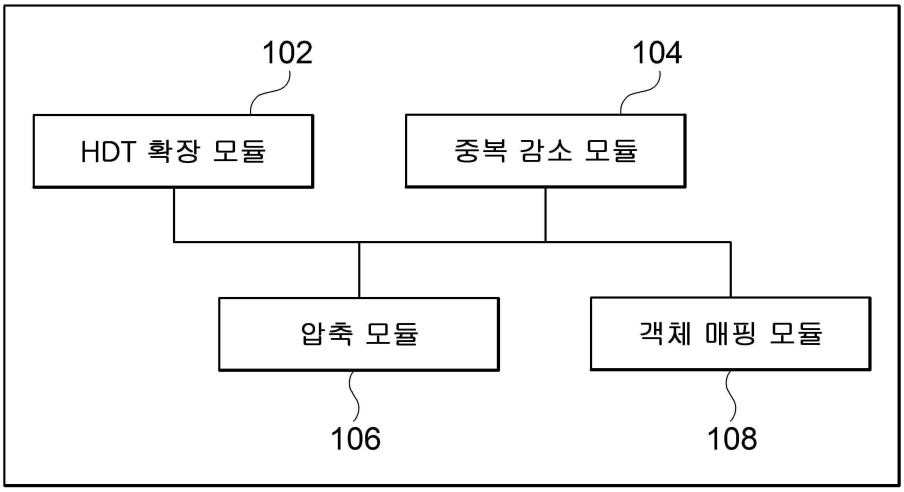
106 : 압축 모듈

108 : 객체 매핑 모듈

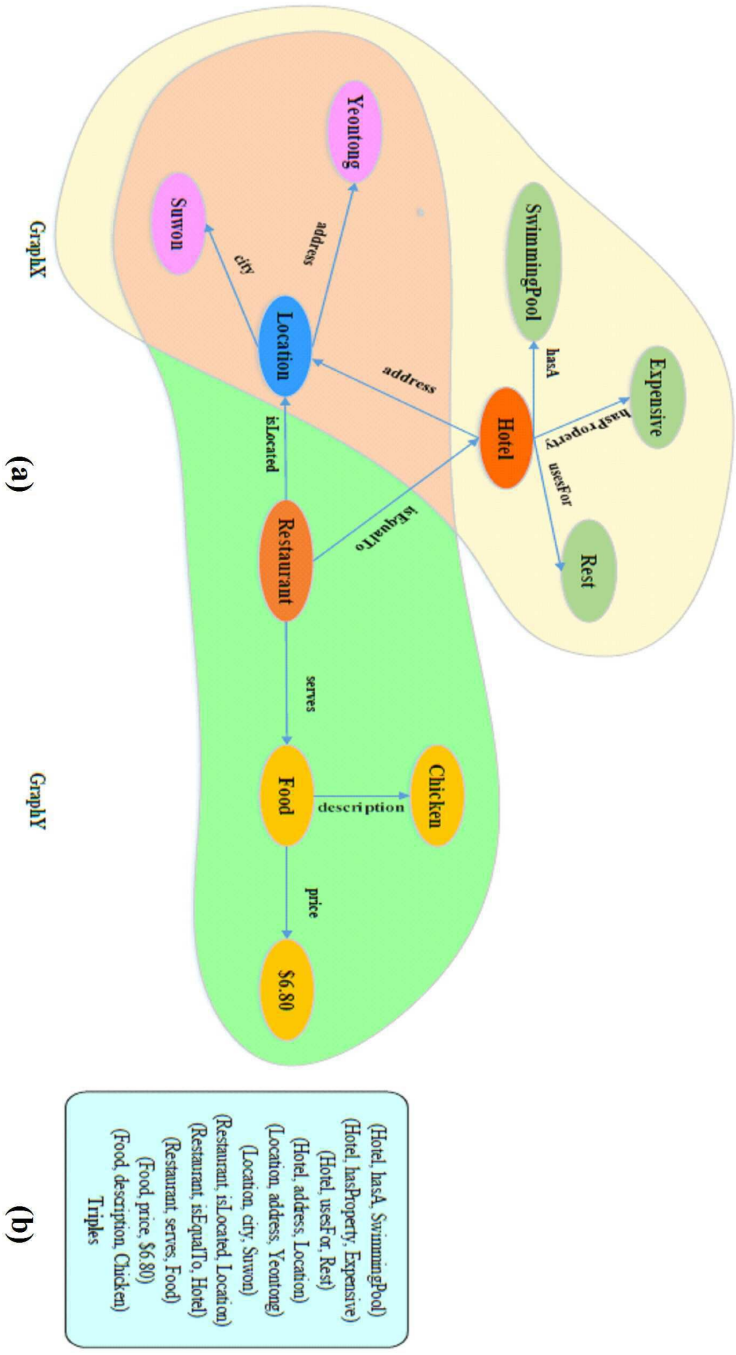
도면

도면1

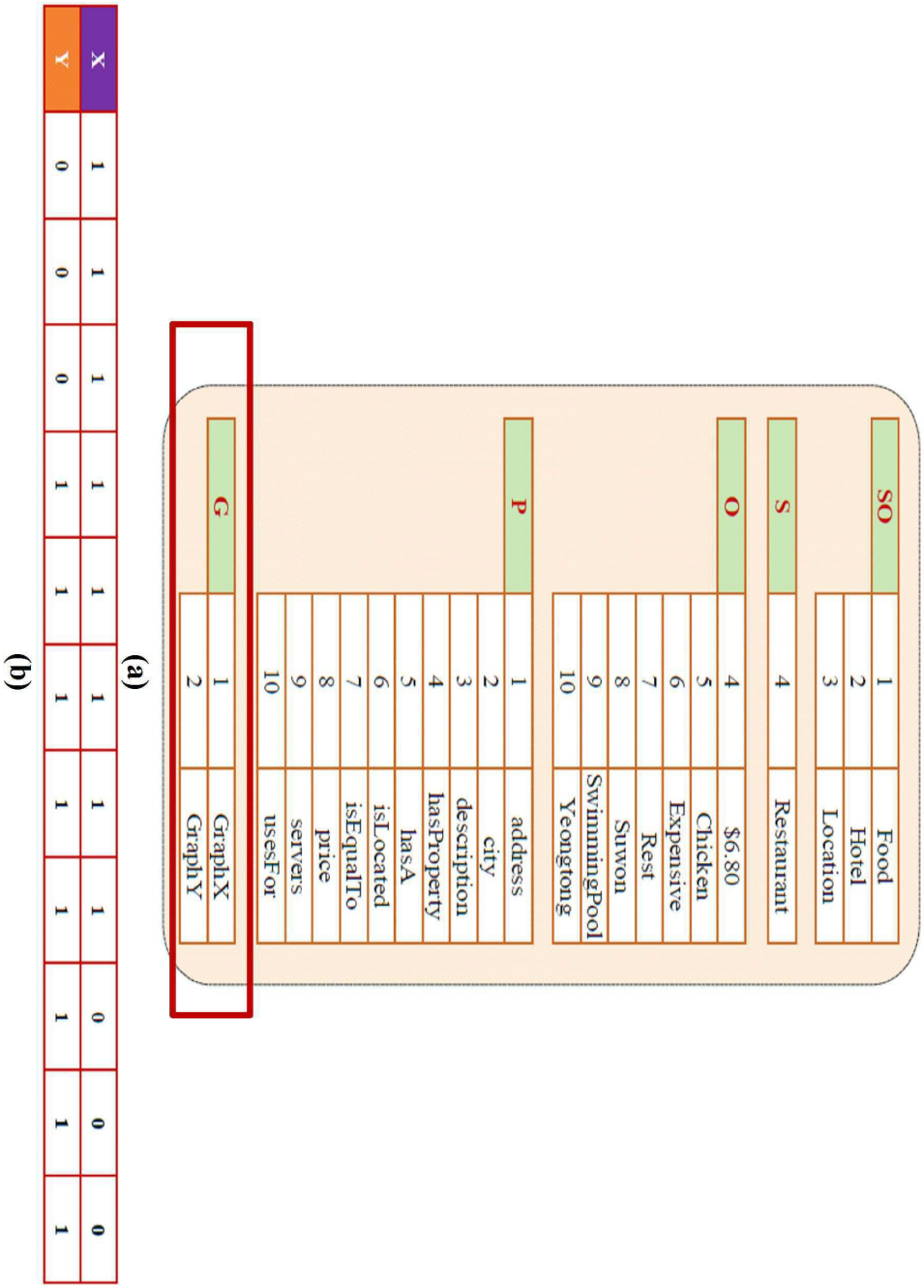
100



도면2


(a)
GraphX
GraphY
Yeontong
Suwon
address
city
Location
address
SwimmingPool
hasA
Expensive
hasProperty
Hotel
usesFor
Rest
isLocated
Restaurant
isEqualTo
serves
Food
description
Chicken
price
$6.80
(b)
(Hotel, hasA, SwimmingPool)
(Hotel, hasProperty, Expensive)
(Hotel, usesFor, Rest)
(Hotel, address, Location)
(Location, address, Yeontong)
(Location, city, Suwon)
(Restaurant, isLocated, Location)
(Restaurant, isEqualTo, Hotel)
(Restaurant, serves, Food)
(Food, price, $6.80)
(Food, description, Chicken)
Triples

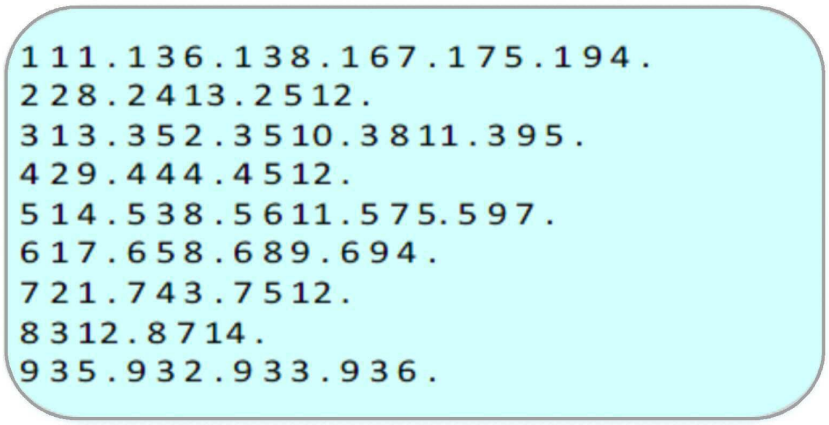
도면3


SO
1 Food
2 Hotel
3 Location
S
4 Restaurant
O
4 $6.80
5 Chicken
6 Expensive
7 Rest
8 Suwon
9 SwimmingPool
10 Yeongtong
P
1 address
2 city
3 description
4 hasProperty
5 hasA
6 isLocated
7 isEqualTo
8 price
9 servers
10 usesFor
G
1 GraphX
2 GraphY
(a)
(b)
X
Y
1 1 1 1 1 1 1 1 1 0
0 1 1 1 1 1 1 1 0 0
0 1 1 1 1 1 1 1 0 0
1 1 1 1 1 1 1 1 1 0
1 1 1 1 1 1 1 1 1 0
0 0 0 0 0 0 0 0 0 0

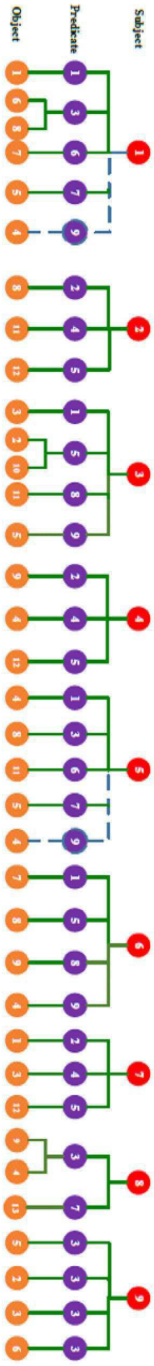
도면4

Triple	X	Y	TI
t_1	1	0	B^{TI}_1
t_2	1	0	B^{TI}_2
t_3	1	0	B^{TI}_3
t_4	1	1	B^{TI}_4
t_5	1	1	B^{TI}_5
t_6	1	1	B^{TI}_6
t_7	1	1	B^{TI}_7
t_8	1	1	B^{TI}_8
t_9	0	1	B^{TI}_9
t_{10}	0	1	B^{TI}_{10}
t_{11}	0	1	B^{TI}_{11}
GI	B^{GI}_1	B^{GI}_2	

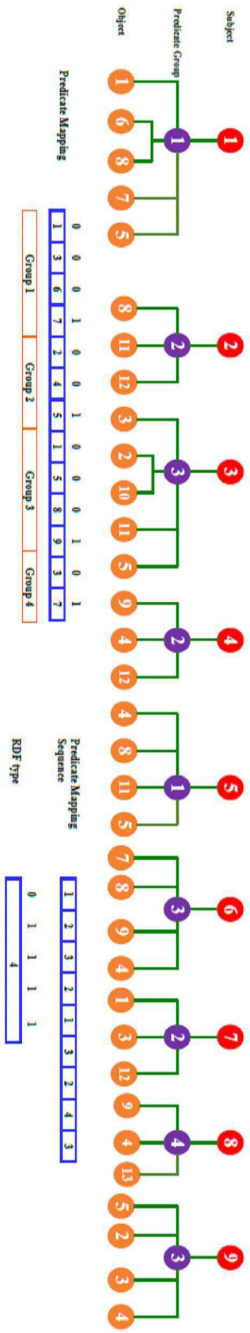
도면5


111.136.138.167.175.194.
228.2413.2512.
313.352.3510.3811.395.
429.444.4512.
514.538.5611.575.597.
617.658.689.694.
721.743.7512.
8312.8714.
935.932.933.936.

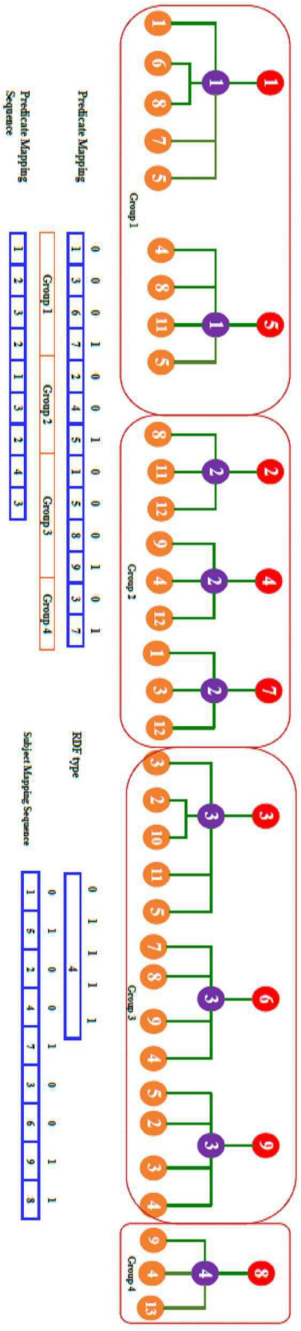
도면6



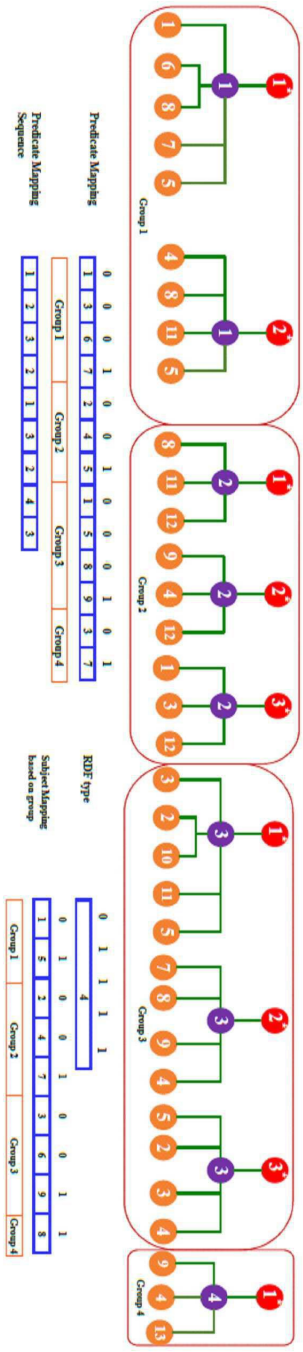
도면7


Subject
Predicate Group
Object
Predicate Mapping
Group 1
Group 2
Group 3
Group 4
Predicate Mapping Sequence
RDF type

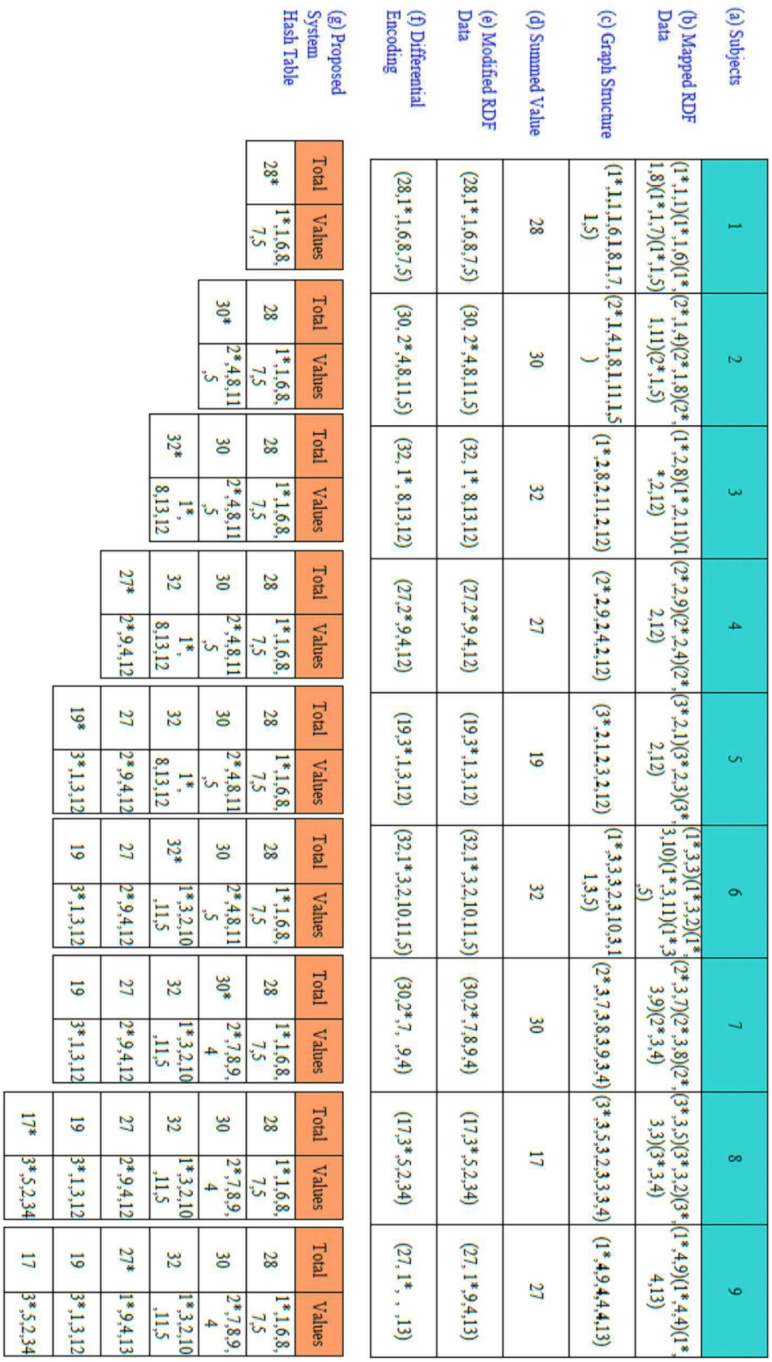
도면8


Predicate Mapping
Predicate Mapping Sequence
Group 1
Group 2
Group 3
Group 4
RDF type
Subject Mapping Sequence

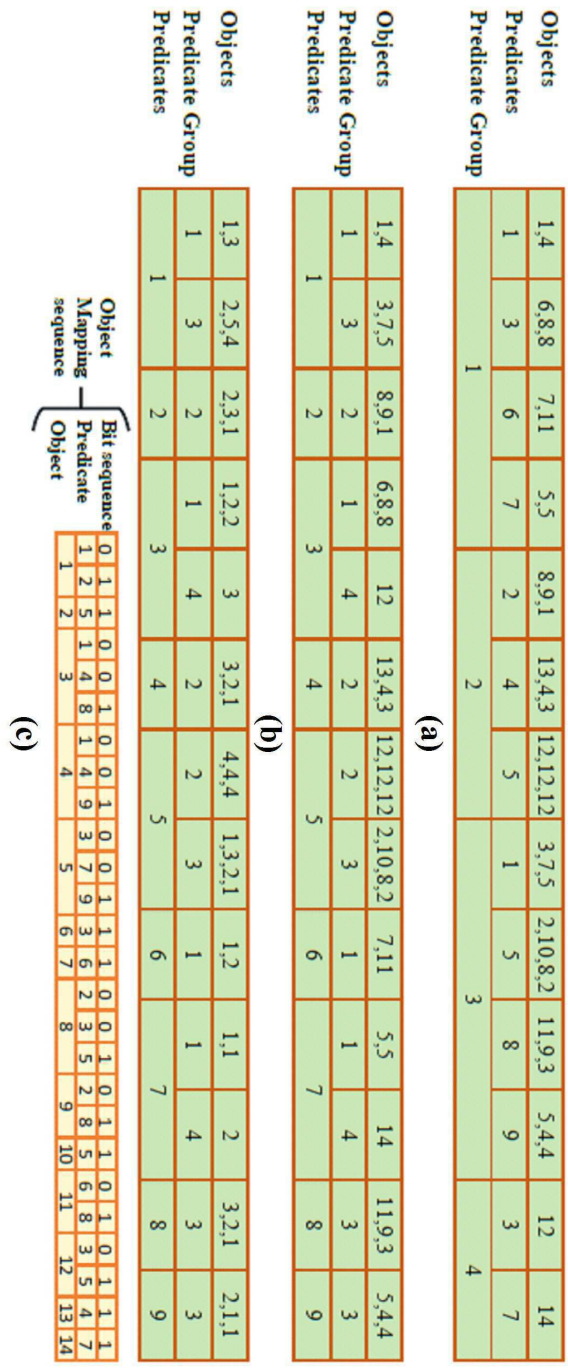
도면9


Predicate Mapping
Predicate Mapping Sequence
Group 1
Group 2
Group 3
Group 4
RDF type
Subject Mapping based on group

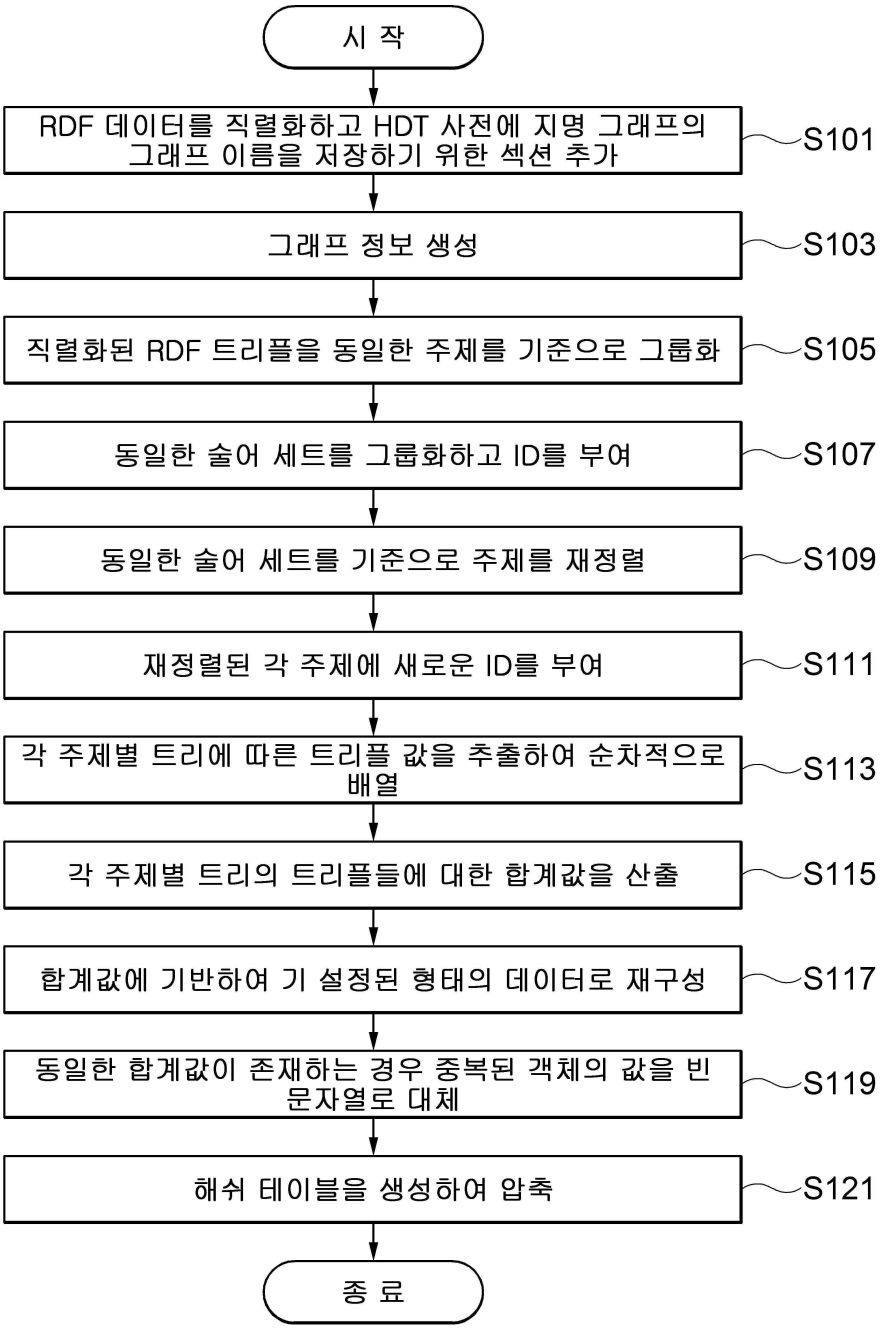
도면10


(a) Subjects
1
2
3
4
5
6
7
8
9
(b) Mapped RDF Data
(1*,1,1)(1*,1,6)(1*,1,8)(1*,1,7)(1*,1,5)
(2*,1,4)(2*,1,8)(2*,1,11)(2*,1,5)
(1*,2,8)(1*,2,11)(1*,2,12)
(2*,2,9)(2*,2,4)(2*,2,12)
(3*,2,1)(3*,2,3)(3*,2,12)
(1*,3,3)(1*,3,2)(1*,3,10)(1*,3,11)(1*,3,5)
(2*,3,7)(2*,3,8)(2*,3,9)(2*,3,4)
(3*,3,5)(3*,3,2)(3*,3,3)(3*,3,4)
(1*,4,9)(1*,4,4)(1*,4,13)
(c) Graph Structure
(1*,1,1,6,1,8,1,7,1,5)
(2*,1,4,1,8,1,11,1,5)
(1*,2,8,2,11,2,12)
(2*,2,9,2,4,2,12)
(3*,2,1,2,3,2,12)
(1*,3,3,2,3,10,3,11,3,5)
(2*,3,7,3,8,3,9,3,4)
(3*,3,5,3,2,3,3,3,4)
(1*,4,9,4,4,4,13)
(d) Summed Value
28
30
32
27
19
32
30
17
27
(e) Modified RDF Data
(28,1*,1,6,8,7,5)
(30, 2*,4,8,11,5)
(32, 1*, 8,13,12)
(27,2*,9,4,12)
(19,3*,1,3,12)
(32,1*,3,2,10,11,5)
(30,2*,7, 8,9,4)
(17,3*,5,2,34)
(27, 1*,9,4,13)
(f) Differential Encoding
(28,1*,1,6,8,7,5)
(30, 2*,4,8,11,5)
(32, 1*, 8,13,12)
(27,2*,9,4,12)
(19,3*,1,3,12)
(32,1*,3,2,10,11,5)
(30,2*,7, 9,4)
(17,3*,5,2,34)
(27, 1*, , ,13)
(g) Proposed System Hash Table
Total
Values
28*
1*,1,6,8,7,5
30*
28
1*,1,6,8,7,5
2*,4,8,11,5
32*
30
28
1*,1,6,8,7,5
2*,4,8,11,5
1*,8,13,12
27*
32
30
28
1*,1,6,8,7,5
2*,4,8,11,5
1*,8,13,12
2*,9,4,12
19*
32
30
28
1*,1,6,8,7,5
2*,4,8,11,5
1*,8,13,12
2*,9,4,12
3*,1,3,12
19
32*
30
28
1*,1,6,8,7,5
2*,4,8,11,5
1*,3,2,10,11,5
2*,9,4,12
3*,1,3,12
19
30*
28
1*,1,6,8,7,5
2*,7,8,9,4
1*,3,2,10,11,5
2*,9,4,12
3*,1,3,12
17*
32
30
28
1*,1,6,8,7,5
2*,7,8,9,4
1*,3,2,10,11,5
2*,9,4,12
3*,1,3,12
3*,5,2,34
27*
32
30
28
1*,1,6,8,7,5
2*,7,8,9,4
1*,3,2,10,11,5
2*,9,4,12
3*,1,3,12
17
3*,5,2,34

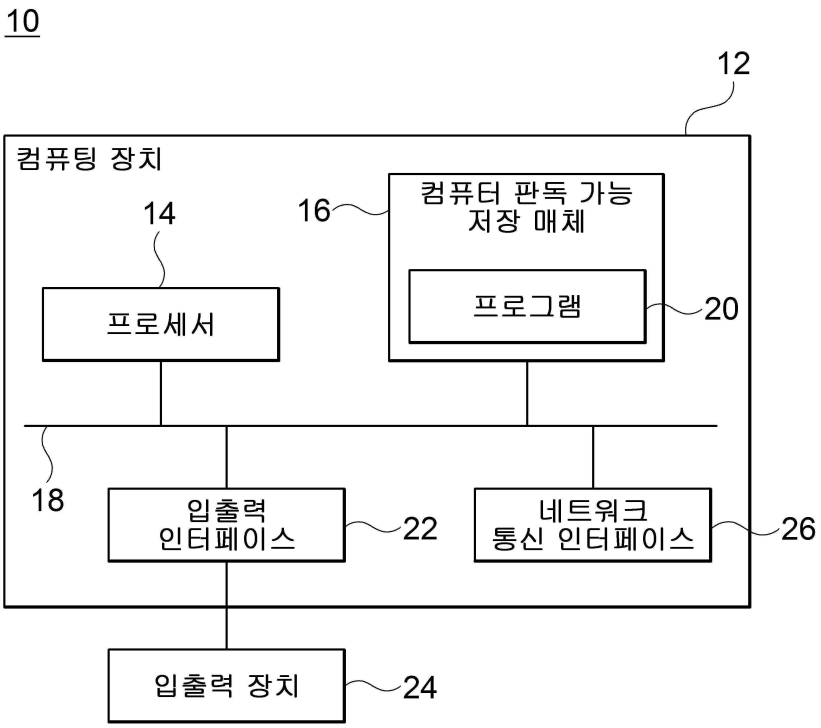
도면11


Objects
Predicates
Predicate Group
1,4 6,8,8 7,11 5,5 8,9,1 13,4,3 12,12,12 3,7,5 2,10,8,2 11,9,3 5,4,4 12 14
1 3 6 7 2 4 5 1 5 8 9 3 7
1 2 3 4
(a)
Objects
Predicate Group
Predicates
1,4 3,7,5 8,9,1 6,8,8 12 13,4,3 12,12,12 2,10,8,2 7,11 5,5 14 11,9,3 5,4,4
1 3 2 1 4 2 3 1 2 1 4 3 3
1 2 3 4 5 6 7 8 9
(b)
Objects
Predicate Group
Predicates
1,3 2,5,4 2,3,1 1,2,2 3 3,2,1 4,4,4 1,3,2,1 1,2 1,1 2 3,2,1 2,1,1
1 3 2 1 4 2 2 3 1 1 4 3 3
1 2 3 4 5 6 7 8 9
(c)
Object Mapping sequence
Bit sequence
Predicate
Object
0 1 1 0 0 1 0 0 1 0 0 1 1 1 0 0 1 1 0 1 1 0 1 1 1 1
1 2 5 1 4 8 1 4 9 3 7 9 3 6 2 3 5 2 8 5 6 8 3 5 4 7
1 2 3 4 5 6 7 8 9 10 11 12 13 14

도면12


시 작
RDF 데이터를 직렬화하고 HDT 사전에 지명 그래프의 그래프 이름을 저장하기 위한 섹션 추가
S101
그래프 정보 생성
S103
직렬화된 RDF 트리플을 동일한 주제를 기준으로 그룹화
S105
동일한 술어 세트를 그룹화하고 ID를 부여
S107
동일한 술어 세트를 기준으로 주제를 재정렬
S109
재정렬된 각 주제에 새로운 ID를 부여
S111
각 주제별 트리에 따른 트리플 값을 추출하여 순차적으로 배열
S113
각 주제별 트리의 트리플들에 대한 합계값을 산출
S115
합계값에 기반하여 기 설정된 형태의 데이터로 재구성
S117
동일한 합계값이 존재하는 경우 중복된 객체의 값을 빈 문자열로 대체
S119
해쉬 테이블을 생성하여 압축
S121
종 료

도면13


10
12
컴퓨팅 장치
14
프로세서
16
컴퓨터 판독 가능
저장 매체
프로그램
20
18
입출력
인터페이스
22
네트워크
통신 인터페이스
26
입출력 장치
24