



(19) 대한민국특허청(KR)  
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0100449  
(43) 공개일자 2023년07월05일

(51) 국제특허분류(Int. Cl.)  
G06F 16/901 (2019.01) G06F 16/26 (2019.01)  
H03M 7/30 (2006.01)  
(52) CPC특허분류  
G06F 16/9024 (2019.01)  
G06F 16/26 (2019.01)  
(21) 출원번호 10-2021-0190331  
(22) 출원일자 2021년12월28일  
심사청구일자 2021년12월28일

(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)  
권기진  
경기도 용인시 기흥구 서그내로 23-5, 501호(서천동)  
(74) 대리인  
두호특허법인

전체 청구항 수 : 총 18 항

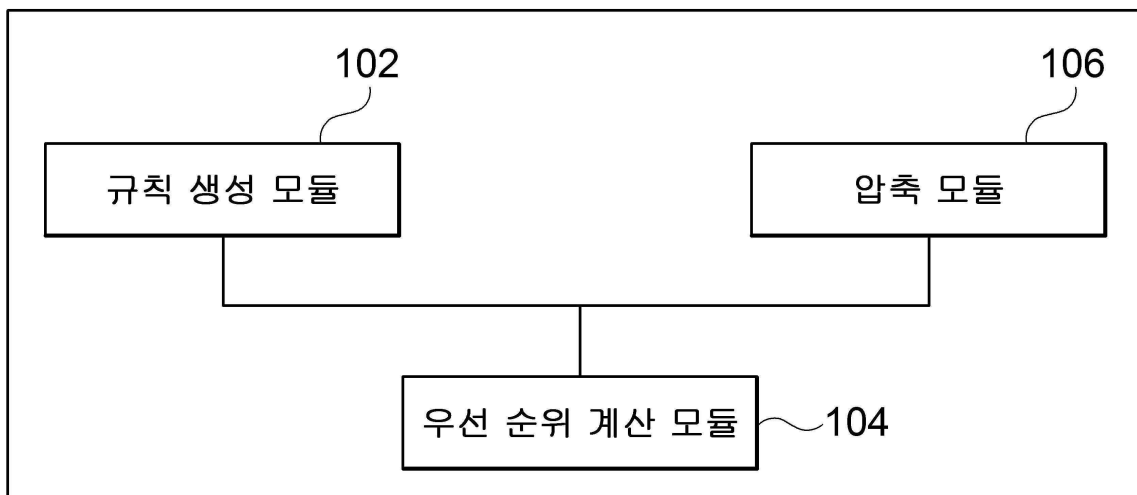
(54) 발명의 명칭 자원 기술 프레임워크 데이터의 압축 장치 및 방법

(57) 요약

자원 기술 프레임워크 데이터의 압축 장치 및 방법이 제공된다. 일 실시예에 따른 RDF 데이터의 압축 장치는 자원 기술 프레임워크(RDF, Resource Description Framework) 그래프에 포함된 복수 개의 트리플에 대한 마이닝을 수행하여 복수의 규칙을 생성하는 규칙 생성 모듈; 상기 복수의 규칙간의 우선 순위를 계산하는 우선 순위 계산 모듈; 및 상기 우선 순위에 따라 상기 복수의 규칙을 상기 RDF 그래프에 적용하여 상기 RDF 그래프를 압축하는 압축 모듈을 포함한다.

대표도 - 도1

100



(52) CPC특허분류  
*H03M 7/30* (2019.01)

이 발명을 지원한 국가연구개발사업

|             |                                       |
|-------------|---------------------------------------|
| 과제고유번호      | 1711134745                            |
| 과제번호        | 2021-0-00859-001                      |
| 부처명         | 과학기술정보통신부                             |
| 과제관리(전문)기관명 | 정보통신기획평가원                             |
| 연구사업명       | SW컴퓨팅산업원천기술개발(R&D, 정보화)               |
| 연구과제명       | 초거대 그래프의 지능적 고속 처리를 위한 그래프 DBMS 기술 개발 |
| 기 여 율       | 1/1                                   |
| 과제수행기관명     | 포항공과대학교 산학협력단                         |
| 연구기간        | 2021.04.01 ~ 2021.12.31               |

---

## 명세서

### 청구범위

#### 청구항 1

자원 기술 프레임워크(RDF, Resource Description Framework) 그래프에 포함된 복수 개의 트리플에 대한 마이닝을 수행하여 복수의 규칙을 생성하는 규칙 생성 모듈;

상기 복수의 규칙간의 우선 순위를 계산하는 우선 순위 계산 모듈; 및

상기 우선 순위에 따라 상기 복수의 규칙을 상기 RDF 그래프에 적용하여 상기 RDF 그래프를 압축하는 압축 모듈을 포함하는, RDF 데이터의 압축 장치.

#### 청구항 2

청구항 1에 있어서,

상기 압축 모듈은, 상기 복수 개의 트리플 중 상기 복수의 규칙의 헤드와 일치하는 트리플을 상기 RDF 그래프에서 삭제하는, RDF 데이터의 압축 장치.

#### 청구항 3

청구항 1에 있어서,

상기 압축 모듈은, 상기 복수 개의 트리플 중 특정 트리플이 상기 복수의 규칙 중 어느 하나에 대한 유효한 객체이면서 해당 규칙의 헤드와 일치하는 트리플이 상기 RDF 그래프에 존재하는 경우, 상기 특정 트리플을 활성 그래프로 이동시키는, RDF 데이터의 압축 장치.

#### 청구항 4

청구항 1에 있어서,

상기 압축 모듈은, 상기 복수의 규칙 중 어느 것과도 일치하지 않는 트리플을 휴면 그래프로 이동시키는, RDF 데이터의 압축 장치.

#### 청구항 5

청구항 1에 있어서,

상기 우선 순위 계산 모듈은, 상기 복수의 규칙 각각의 규칙 적용 가능성 점수에 기반하여 상기 우선 순위를 계산하는, RDF 데이터의 압축 장치.

#### 청구항 6

청구항 5에 있어서,

상기 압축 모듈은, 상기 우선 순위에 따라 순차적으로 상기 복수의 규칙을 적용하여 상기 RDF 데이터를 압축하는, RDF 데이터의 압축 장치.

#### 청구항 7

청구항 5에 있어서,

상기 규칙 적용 가능성 점수는, 각 규칙을 적용할 경우 삭제될 트리플의 최대 개수, 각 규칙의 선행 항목에 존재하는 조건의 개수 및 각 규칙을 적용할 경우 삭제를 방지할 수 있는 최대 규칙 수에 기반하여 계산되는, RDF 데이터의 압축 장치.

#### 청구항 8

청구항 1에 있어서,

상기 규칙 생성 모듈은,

기 설정된 지지도 임계값에 기반하여 상기 복수의 규칙의 개수를  $k$ 개( $k$ 는 1 이상의 자연수)로 제한하는, RDF 데이터의 압축 장치.

#### 청구항 9

청구항 8에 있어서,

상기 지지도 임계값은, 마이닝된 각 규칙의 헤드 커버리지에 기반하여 설정되는, RDF 데이터의 압축 장치.

#### 청구항 10

컴퓨터상에서 수행되는 방법으로서,

자원 기술 프레임워크(RDF, Resource Description Framework) 그래프에 포함된 복수 개의 트리플에 대한 마이닝을 수행하여 복수의 규칙을 생성하는 단계;

상기 복수의 규칙간의 우선 순위를 계산하는 단계; 및

상기 우선 순위에 따라 상기 복수의 규칙을 상기 RDF 그래프에 적용하여 상기 RDF 그래프를 압축하는 단계를 포함하는, RDF 데이터의 압축 방법.

#### 청구항 11

청구항 10에 있어서,

상기 압축 단계는, 상기 복수 개의 트리플 중 상기 복수의 규칙의 헤드와 일치하는 트리플을 상기 RDF 그래프에서 삭제하도록 구성되는, RDF 데이터의 압축 방법.

#### 청구항 12

청구항 10에 있어서,

상기 압축 단계는, 상기 복수 개의 트리플 중 특정 트리플이 상기 복수의 규칙 중 어느 하나에 대한 유효한 객체이면서 해당 규칙의 헤드와 일치하는 트리플이 상기 RDF 그래프에 존재하는 경우, 상기 특정 트리플을 활성 그래프로 이동시키도록 구성되는, RDF 데이터의 압축 방법.

#### 청구항 13

청구항 10에 있어서,

상기 압축 단계는, 상기 복수의 규칙 중 어느 것과도 일치하지 않는 트리플을 휴면 그래프로 이동시키도록 구성되는, RDF 데이터의 압축 방법.

#### 청구항 14

청구항 10에 있어서,

상기 우선 순위 계산 단계는, 상기 복수의 규칙 각각의 규칙 적용 가능성 점수에 기반하여 상기 우선 순위를 계산하도록 구성되는, RDF 데이터의 압축 방법.

#### 청구항 15

청구항 14에 있어서,

상기 압축 단계는, 상기 우선 순위에 따라 순차적으로 상기 복수의 규칙을 적용하여 상기 RDF 데이터를 압축하도록 구성되는, RDF 데이터의 압축 방법.

#### 청구항 16

청구항 14에 있어서,

상기 규칙 적용 가능성 점수는, 각 규칙을 적용할 경우 삭제될 트리플의 최대 개수, 각 규칙의 선행 항목에 존재하는 조건의 개수 및 각 규칙을 적용할 경우 삭제를 방지할 수 있는 최대 규칙 수에 기반하여 계산되는, RDF 데이터의 압축 방법.

#### 청구항 17

청구항 10에 있어서,

상기 규칙 생성 단계는,

기 설정된 지지도 임계값에 기반하여 상기 복수의 규칙의 개수를  $k$ 개( $k$ 는 1 이상의 자연수)로 제한하도록 구성되는, RDF 데이터의 압축 방법.

#### 청구항 18

청구항 17에 있어서,

상기 지지도 임계값은, 마이닝된 각 규칙의 헤드 커버리지에 기반하여 설정되는, RDF 데이터의 압축 방법.

### 발명의 설명

#### 기술 분야

[0001] 개시되는 실시예들은 자원 기술 프레임워크(RDF, Resource Description Framework) 데이터를 압축하기 위한 기술과 관련된다.

#### 배경 기술

[0002] 최근 몇 년간 컴퓨터로 처리 가능한(machine operable) 비독점(non-proprietary) RDF 데이터로 인하여 다양한 데이터 자원 간의 연계 능력이 향상되었다. 정부를 포함한 여러 기업들이 데이터를 RDF 형식으로 공유하여 재사용성을 높이고 여러 애플리케이션과 통합하고 있다. 이는 웹에서 사용 가능한 공개 RDF 데이터의 규모와 다양성 측면에서의 성장을 가속화하고 있다. 이러한 RDF 데이터의 증가는 시맨틱 웹에서는 긍정적이지만, 필연적으로 RDF 데이터 관리 측면에서의 성능 저하를 불러일으킨다. 그 결과 구조화된 데이터를 압축해야 할 필요성이 증가하고 있다.

[0003] 기존의 RDF 데이터의 압축 기법들은 주로 데이터의 구조적 중복성을 고려하는 방향으로 연구되었다. 또한 RDF 데이터의 간결한 표현에 초점을 맞춘 연구들도 존재한다. 그러나 압축 과정에서 RDF 데이터의 의미적 특성을 고려하는 기술은 찾기 어렵다.

### 발명의 내용

#### 해결하려는 과제

[0004] 개시되는 실시예들은 RDF 데이터의 다양한 의미적 연관 규칙을 이용하여 RDF 데이터 세트에 대한 압축을 수행하기 위한 기술적인 수단을 제공하기 위한 것이다.

#### 과제의 해결 수단

[0005] 예시적인 실시예에 따르면, 자원 기술 프레임워크(RDF, Resource Description Framework) 그래프에 포함된 복수 개의 트리플에 대한 마이닝을 수행하여 복수의 규칙을 생성하는 규칙 생성 모듈; 상기 복수의 규칙간의 우선순위를 계산하는 우선 순위 계산 모듈; 및 상기 우선 순위에 따라 상기 복수의 규칙을 상기 RDF 그래프에 적용하여 상기 RDF 그래프를 압축하는 압축 모듈을 포함하는, RDF 데이터의 압축 장치가 제공된다.

[0006] 상기 압축 모듈은, 상기 복수 개의 트리플 중 상기 복수의 규칙의 헤드와 일치하는 트리플을 상기 RDF 그래프에서 삭제할 수 있다.

[0007] 상기 압축 모듈은, 상기 복수 개의 트리플 중 특정 트리플이 상기 복수의 규칙 중 어느 하나에 대한 유효한 객체이면서 해당 규칙의 헤드와 일치하는 트리플이 상기 RDF 그래프에 존재하는 경우, 상기 특정 트리플을 활성 그래프로 이동시킬 수 있다.

- [0008] 상기 압축 모듈은, 상기 복수의 규칙 중 어느 것보다도 일치하지 않는 트리플을 휴먼 그래프로 이동시킬 수 있다.
- [0009] 상기 우선 순위 계산 모듈은, 상기 복수의 규칙 각각의 규칙 적용 가능성 점수에 기반하여 상기 우선 순위를 계산할 수 있다.
- [0010] 상기 압축 모듈은, 상기 우선 순위에 따라 순차적으로 상기 복수의 규칙을 적용하여 상기 RDF 데이터를 압축할 수 있다.
- [0011] 상기 규칙 적용 가능성 점수는, 각 규칙을 적용할 경우 삭제될 트리플의 최대 개수, 각 규칙의 선행 항목에 존재하는 조건의 개수 및 각 규칙을 적용할 경우 삭제를 방지할 수 있는 최대 규칙 수에 기반하여 계산될 수 있다.
- [0012] 상기 규칙 생성 모듈은, 기 설정된 지지도 임계값에 기반하여 상기 복수의 규칙의 개수를 k개(k는 1 이상의 자연수)로 제한할 수 있다.
- [0013] 상기 지지도 임계값은, 마이닝된 각 규칙의 헤드 커버리지에 기반하여 설정될 수 있다.
- [0014] 다른 예시적인 실시예에 따르면, 컴퓨터상에서 수행되는 방법으로서, 자원 기술 프레임워크(RDF, Resource Description Framework) 그래프에 포함된 복수 개의 트리플에 대한 마이닝을 수행하여 복수의 규칙을 생성하는 단계; 상기 복수의 규칙간의 우선 순위를 계산하는 단계; 및 상기 우선 순위에 따라 상기 복수의 규칙을 상기 RDF 그래프에 적용하여 상기 RDF 그래프를 압축하는 단계를 포함하는, RDF 데이터의 압축 방법이 제공된다.
- [0015] 상기 압축 단계는, 상기 복수 개의 트리플 중 상기 복수의 규칙의 헤드와 일치하는 트리플을 상기 RDF 그래프에서 삭제하도록 구성될 수 있다.
- [0016] 상기 압축 단계는, 상기 복수 개의 트리플 중 특정 트리플이 상기 복수의 규칙 중 어느 하나에 대한 유효한 객체이면서 해당 규칙의 헤드와 일치하는 트리플이 상기 RDF 그래프에 존재하는 경우, 상기 특정 트리플을 활성 그래프로 이동시키도록 구성될 수 있다.
- [0017] 상기 압축 단계는, 상기 복수의 규칙 중 어느 것보다도 일치하지 않는 트리플을 휴먼 그래프로 이동시키도록 구성될 수 있다.
- [0018] 상기 우선 순위 계산 단계는, 상기 복수의 규칙 각각의 규칙 적용 가능성 점수에 기반하여 상기 우선 순위를 계산하도록 구성될 수 있다.
- [0019] 상기 압축 단계는, 상기 우선 순위에 따라 순차적으로 상기 복수의 규칙을 적용하여 상기 RDF 데이터를 압축하도록 구성될 수 있다.
- [0020] 상기 규칙 적용 가능성 점수는, 각 규칙을 적용할 경우 삭제될 트리플의 최대 개수, 각 규칙의 선행 항목에 존재하는 조건의 개수 및 각 규칙을 적용할 경우 삭제를 방지할 수 있는 최대 규칙 수에 기반하여 계산될 수 있다.
- [0021] 상기 규칙 생성 단계는, 기 설정된 지지도 임계값에 기반하여 상기 복수의 규칙의 개수를 k개(k는 1 이상의 자연수)로 제한하도록 구성될 수 있다.
- [0022] 상기 지지도 임계값은, 마이닝된 각 규칙의 헤드 커버리지에 기반하여 설정될 수 있다.

### 발명의 효과

- [0023] 개시되는 실시예들에 따르면 RDF 데이터의 다양한 의미적 연관 규칙을 이용하여 효과적으로 RDF 데이터 세트에 대한 압축을 수행할 수 있다.

### 도면의 간단한 설명

- [0024] 도 1은 일 실시예에 따른 자원 기술 프레임워크(RDF, Resource Description Framework) 데이터의 압축 장치(100)를 설명하기 위한 블록도
- 도 2는 개시되는 실시예에 따른 RDF 데이터 압축 장치(100)의 RDF 데이터 압축 과정을 설명하기 위한 예시도
- 도 3은 일 실시예에 따른 RDF 데이터의 압축 장치에서 RDF 그래프의 압축에 사용되는 각 규칙간의 종속성을 나타낸 그래프

도 4는 top-k 접근 방식을 사용하여 최저 지지도 임계값을 설정하는 과정을 설명하기 위한 예시도

### 발명을 실시하기 위한 구체적인 내용

- [0025] 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.
- [0026] 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.
- [0027] 도 1은 일 실시예에 따른 자원 기술 프레임워크(RDF, Resource Description Framework) 데이터의 압축 장치(100)를 설명하기 위한 블록도이다. 일 실시예에 따른 RDF 데이터의 압축 장치(100)는 복수 개의 트리플로 구성되는 RDF 그래프를 압축하기 위한 장치이다. 도시된 바와 같이, 일 실시예에 따른 RDF 데이터의 압축 장치(100)는 규칙 생성 모듈(102), 우선 순위 계산 모듈(104) 및 압축 모듈(106)을 포함한다.
- [0028] 규칙 생성 모듈(102)은 RDF 그래프에 포함된 복수 개의 트리플에 대한 마이닝을 수행하여 복수의 규칙을 생성한다.
- [0029] 우선 순위 계산 모듈(104)은 상기 복수의 규칙간의 우선 순위를 계산한다.
- [0030] 압축 모듈(106)은 상기 우선 순위에 따라 상기 복수의 규칙을 상기 RDF 그래프에 적용하여 상기 RDF 그래프를 압축한다.
- [0031] 이하에서는 상기와 같이 구성되는 일 실시예에 따른 RDF 데이터의 압축 장치를 좀 더 상세히 설명한다.
- [0032] 자원 기술 프레임워크(RDF, Resource Description Framework)는 다양한 종류의 시맨틱 데이터(semantic data)를 저장, 검색 및 배포하는데 널리 사용되는 데이터 모델이다. 그러나 그 막대한 용량 때문에, 스토리지에서 데이터를 효율적으로 관리하거나 자료를 가져오는 것은 여전히 극복해야 할 과제이다. RDF 기반의 지식 베이스(KB, Knowledge Base)를 효율적으로 압축 및 관리하기 위하여, 개시되는 실시예에서는 혼 규칙(Horn rules), 문법 기반 시스템(GBS), 규칙 간의 순위 및 top-k 접근 방식을 사용하게 된다.
- [0033] **혼 규칙**
- [0034] 원자(atom)는  $R(A, B)$  형태를 가지는 사실(fact)이다. 여기서 R은 관계(relation), A와 B는 상수 또는 변수를 나타낸다. 원자를 인스턴스화할 때 적어도 하나의 인자(argument)는 상수여야 한다(예를 들어,  $\text{livesIn}(a, \text{Korea})$ ). 반면 기본 원자(ground atom)는 두 개의 상수를 인자로 가진다(예를 들어,  $\text{livesIn}(a, b)$ ). 어떤 원자 X가 변수 집합을 반환할 경우,  $\text{var}(X)$ 로 정의된다. 결합 쿼리(conjunctive query)  $\wedge$ 는 원자의 결합을 정의한다( $X_1 \wedge X_2 \wedge \dots \wedge X_n$ ). 대체  $\mu$ 는 지식 베이스(KB)  $\kappa$ 에서 변수에서 상수로의 부분 매핑을 나타낸다(예를 들어,  $\mu(X_i) \in \kappa, \exists_i \in \{1, \dots, n\}$ ).
- [0035] 혼 규칙은 헤드(단일 원자)와 바디(원자 집합)로 구성된다. 이 규칙의 공식은  $X \Rightarrow H$  와 같이 나타낼 수 있으며, X와 H는 각각 바디 원자와 헤드 원자의 쿼리이다. 개시되는 실시예들에서는 X와 H에 대한 규칙을 ( $?x$  property ?y)의 형태로 나타낸다. 여기서 property(속성)는 RDF 데이터의 주체(subject)와 객체(object) 간의 관계를 정의하는 술어(predicates)를 나타낸다.
- [0036] RDF 데이터셋(또는 RDF 그래프)으로부터 마이닝된 규칙은 두 개의 변수로 구성된 일반적인 패턴, 예를 들어 ( $?x$  property ?y)의 형태를 가지거나 또는 변수가 아닌 특정한 값(specific value)을 가지도록 구성될 수 있다. 특정한 값을 가지는 규칙의 예는 다음과 같이 기술될 수 있다.
- [0037] ( $?x \text{ ref} : \text{type DistinctiveEvents}$ )  $\Rightarrow$  ( $?x \text{ ref} : \text{type Events}$ )

[0038] 이와 같은 규칙은 클래스 계층 구조(Class Hierachy), 예를 들어 DistinctiveEvents  $\subseteq$  Events를 기술하는 데 도움을 준다. 개시되는 실시예들은 의미적 연관을 고려하여 RDF 데이터를 압축하기 위한 것이다. 따라서, 클래스 계층 구조를 설명하기 위하여 특정한 값을 가지는 규칙만을 사용한다. 규칙을 최소한의 수준으로 유지하기 위하여 다른 규칙들은 무시된다.

[0039] RDF 데이터는 트리플(triple)로도 불리는, 주어-술어-목적어 형태로 이루어진 데이터 포맷을 의미한다. 트리플은 종종 선례(precedent)의 객체화이지만, 트리플이 규칙의 헤드와 일치하더라도 해당 트리플이 그래프에 존재하지 않는 경우가 있을 수 있다. 규칙의 헤드와 일치하는 트리플이 그래프 내에 존재하지 않을 경우를 잘못된 객체화(invalid instantiation), 규칙의 헤드와 일치하는 트리플이 그래프 내에 존재하는 경우를 유효한 객체화(valid instantiation)라고 한다. 따라서, 어떤 규칙에 대한 지지도(support)는 유효한 객체의 총 개수이다. 추가적으로, 신뢰도(confidence)는 선례에 대한 모든 객체화에 대한 유효한 객체의 비율이다. 지지도와 신뢰도는 다음의 수학식으로 나타낼 수 있다.

[0040] [수학식 1]

$$\text{Support } (X \Rightarrow H) = \text{number of}(a,b): \forall r_1, \dots, r_n: X \wedge H$$

[0041]

[0042] [수학식 2]

$$\text{Confidence } (X \Rightarrow H) = \frac{\text{Support } (X \Rightarrow H)}{\text{number of}(a,b): \forall r_1, \dots, r_n: X}$$

[0043]

[0044] 상기 수학식에서, a, b,는 헤드 변수이고,  $r_1$  내지  $r_n$ 은 헤드 변수가 아닌 변수이다.

#### [0045] 규칙 패턴의 표현성

[0046] 개시되는 실시예에서는 특정 작업에 필요한 규칙을 찾기 위한 규칙 패턴의 표현성을 향상시키기 위하여, 복잡한

$$GBS = (M, \Delta, \Pi, \Phi, R, S_t)$$

규칙 기반 패턴에 대한 문법 기반 패턴 시스템(GBS)을 사용한다. 예를 들어, 를 원자의 유효한 패턴을 생성하기 위한 문법이라고 가정하자. 예를 들어 원자의 주어(S), 술어(P), 목적어(O) 항목에 대

$$M, \Delta, \Pi, \Phi, R$$

한 문법은 (S P O)이다. 여기서, 및 은 각각 전부(all); 비단말 기호(non-terminal symbols); 입력 지식 베이스(KGs)에 나타나는 리터럴(literals)과 리소스(resources)에 대한 단말 기호(terminal symbols); 그

$$S_t$$

를 위한 특수 단말 기호; 이 문법에 적합한 규칙 집합을 나타낸다. 는 전체 항목 패턴을 나타내는 시작 기호이다. 이를 정리하면 다음과 같다.

[0047] [수학식 3]

$$M = \{E, F, G\}$$

[0048]

$$\Delta = \text{constants from the input KGs}$$

[0049]

$$\Pi = \{?x, ?y, \dots, ?z\}$$

[0050]

$$\Phi = \{?, ?_v, ?_c, [, ], \neg\}$$

[0051]

$$R = \{E \rightarrow ?; E \rightarrow ?_v; E \rightarrow ?_c; E \rightarrow x \in \Pi; A \rightarrow x \in \Delta; F \rightarrow E; F \rightarrow F, F; G \rightarrow E; G \rightarrow [F]; G \rightarrow \neg[F]\}$$

[0052]

$$S_t = \{G\}$$

[0053]

[0054]

여기서, 심볼  $\tau$ 는 모든 항목의 패턴을 나타내고,  $\tau_v$ 는 변수의 패턴,  $\tau_c$ 는 모든 상수의 패턴을 나타낸다. 구체적 인 변수를 나타내기 위해서는 기호  $\tau$ 가 앞에 붙은 단일 알파벳 문자를 사용한다. 비단말 기호 E는 모든 단말 패턴(terminal patterns)을 나타낸다. 패턴 집합에는 대괄호 안에 있는 항목(예를 들어, [ $\langle \text{Seoul} \rangle, \langle \text{SouthKorea} \rangle$ ])과 일치하는 하나의 내부 패턴이 있어야 한다.  $\neg$  기호는 항목과 일치하는 내부 패턴이 없는 패턴 집합의 여집합(complement)을 표현하기 위하여 사용된다(예를 들어,  $\neg [\langle \text{Seoul} \rangle, \langle \text{SouthKorea} \rangle]$ ). 규칙 패턴(RP)의 우변에는 헤드 원자 패턴( $A_{ph}$ )이, 좌변에는 바디 원자 패턴( $A_{p1}$  내지  $A_{pn}$ )의 결합(conjunction)이 위치한다. 이를 수식으로 나타내면 다음과 같다.

[0055]

[수학식 4]

$$RP = A_{p1} \wedge \dots \wedge A_{pn} \Rightarrow A_{ph}$$

[0056]

[0057]

### 압축

[0058]

RDF 데이터 압축 장치(100)의 압축 모듈(106)은 혼 규칙과 문법 기반 시스템의 규칙 패턴을 이용하여 RDF 데이터 세트(RDF 그래프)에 대한 압축을 수행한다. 이 과정에서 높은 압축률을 달성하기 위해 규칙 간의 다양한 의미적 연관성을 사용한다. RDF 데이터 압축 장치(100)는 검색 공간과 반복 계산을 줄이기 위하여 문법 기반 시스템(GBS, Grammar Based System)을 사용한다. 아울러 RDF 데이터 압축 장치(100)는 규칙 간의 순위(the ranking of rules)와 규칙 간의 상호의존성에 기반한 top-k 접근 방식을 사용함으로써 모든 규칙들을 효과적으로 사용하고 사용자가 설정한 최소 지원 임계값(minimum support threshold)로 인해 발생하는 조합 폭발의 문제를 해결하도록 구성된다.

[0059]

일 실시예에서, 압축 모듈(106)은 RDF 그래프에 포함된 복수 개의 트리플 중 규칙 생성 모듈(102)에서 생성한 복수의 규칙의 헤드와 일치하는 트리플을 상기 RDF 그래프에서 삭제할 수 있다.

[0060]

만약 복수 개의 트리플 중 특정 트리플이 상기 복수의 규칙 중 어느 하나에 대한 유효한 객체인데서 해당 규칙의 헤드와 일치하는 트리플이 상기 RDF 그래프에 존재하는 경우, 압축 모듈(106)은 해당 특정 트리플을 활성 그래프( $G_{ac}$ )로 이동시킨다. 이와 달리 상기 복수의 규칙 중 어느 것과도 일치하지 않는 트리플의 경우 휴면 그래프( $G_{d0}$ )로 이동한다.

[0061]

압축 해제시에는 위 과정을 반대로 적용, 즉 혼 규칙을 이용하여 삭제된 트리플을 복원하게 된다.

[0062]

도 2는 개시되는 실시예에 따른 RDF 데이터 압축 장치(100)의 RDF 데이터 압축 과정을 설명하기 위한 예시도이다.

[0063]

RDF 그래프 G는 복수의 중복되지 않은 RDF 데이터로 구성된다. RDF 데이터 압축 장치(100)의 압축 모듈(106)은 RDF 그래프의 마이닝을 통해 도출된 규칙을 이용하여 RDF 그래프를 압축한다. 구체적으로 RDF 데이터 압축 장치(100)는 각 규칙의 헤드와 일치하는 트리플들을 RDF 그래프에서 삭제한다. 압축 모듈(106)은 RBC(Rule Based Compression) 시스템 중 하나인 AMIE+를 적용하여 RDF 그래프 G를 활성 그래프( $G_{ac}$ )와 휴면 그래프( $G_{d0}$ )로 분할한다. 압축 모듈(106)은 규칙에 대한 유효한 객체인데서 일부 규칙의 헤드와 일치하는 트리플이 RDF 그래프 내에 존재하는 트리플을 활성 그래프로 이동한다.

[0064]

R을 압축 해제 규칙의 집합이라 할 때,  $G = G_{d0} \cup R(G_{ac})$ 를 만족할 경우 무손실 압축을 달성할 수 있다. 압축 모듈(106)은 어떤 규칙과도 일치하지 않는 트리플들을 휴면 그래프로 이동시킨다. 전술한 바와 같이 RDF 데이터 압축 장치(100)는 AMIE+를 사용하기는 하지만, 활성 그래프와 휴면 그래프를 조정하여 성능을 더욱 향상시킬 수 있다. 예를 들어, 압축 모듈(106)은 AMIE+에서 허용하지 않는 신뢰도가 1 이하인 트리플을 활성 그래프에 저장하는 것을 허용한다. 압축 해제 과정에서, RDF 데이터 압축 장치(100)는 전체 활성 그래프와 매칭되는 규칙만을 허용한다.

[0065]

전술한 바와 같이 개시되는 실시예의 경우 압축 과정에서 신뢰도(confidence)가 1 이하인 규칙을 허용한다. 따라서 압축 해제 과정에서는 원래 RDF 그래프에 존재하지 않던 트리플이 생성될 수 있다. 이런 경우는 규칙, 예를 들어  $R_1$ 에 대한 잘못된 객체화(invalid instantiation)에 해당하는 트리플이 활성 그래프에 저장될 때 발생

할 수 있다. 이러한 트리플이 활성 그래프에 저장되는 이유는 해당 트리플이 몇몇 추가 규칙, 예를 들어  $R_2$ 에 대한 유효한 객체화의 일부이기 때문이다. 이러한 트리플을 충돌 트리플(conflicting triple)이라 한다. 예를 들어, 다음의 트리플들을 포함하는 그래프가 존재한다고 가정하자.

$t_1$ : <Kim hasFather Joe>

$t_2$ : <Kim hasParent Joe>

$t_3$ : <Quak hasFather Carol>

$t_4$ : <Quak hasParent Carol>

$t_5$ : <Lee hasFather Carol>

$t_6$ : <Carol hasChild Lee>

트리플  $t_5$ 는 규칙  $R_1$ ,  $(?x \text{ hasFather } y \Rightarrow ?x \text{ hasParent } ?y)$ 의 일부이지만 해당 규칙의 헤드와 일치하는 트리플은 그래프 내에 존재하지 않는다. 그러나, 만약 규칙  $R_2$ ,  $(?x \text{ hasFather } ?y \Rightarrow ?y \text{ hasChild } ?x)$ 을 적용할 경우  $t_5$ 는  $t_6$ 과 함께 유효한 객체화를 구성하므로 시스템을  $t_5$ 를 활성 그래프로 잘못 전달한다. 그 결과  $R_1$ 는 압축 해제 과정에서 <Lee hasChild Carol>이라는, 원래는 없던 트리플을 생성하게 된다. 즉 여기서  $t_5$ 는 충돌 트리플이 된다. 개시되는 실시예들에서 RDF 데이터 압축 장치(100)는 이러한 충돌 트리플을 감지하고 해당 트리플이 활성 그래프로 잘못 이동하는 것을 방지하도록 구성된다.

#### 규칙간 순위

특정 규칙의 선행 조건이 다른 규칙의 헤드에 존재하는 경우, 특정 규칙은 다른 규칙에 종속된다. 예를 들어 다음과 같이  $R_1$ 부터  $R_6$ 까지의 6개의 규칙이 있다고 가정하자.

$R_1$ :  $(?x \text{ isPopularFor } ?y) \Rightarrow (?x \text{ generated } ?y)$  [1000]

$R_2$ :  $(?x \text{ isBornIn } ?y) \Rightarrow (?x \text{ livesIn } ?y)$  [6000]

$R_3$ :  $(?a \text{ isstudentAt } ?y) \wedge (?a \text{ livesIn } ?y) \Rightarrow (?x \text{ islocated } ?y)$  [20000]

$R_4$ :  $(?x \text{ livesIn } ?y) \Rightarrow (?x \text{ isCitizenOf } ?y)$  [8000]

$R_5$ :  $(?x \text{ isPopularFor } ?y) \Rightarrow (?x \text{ generated } ?y)$  [12000]

$R_6$ :  $(?f \text{ isCitizenOf } ?x) \wedge (?f \text{ isBornIn } ?x) \Rightarrow (?x \text{ hasCapitalOf } ?y)$  [3000]

대괄호[]에 정의된 숫자는 규칙 카디널리티로서 규칙의 헤드와 일치하는 트리플 수를 나타낸다.

도 3은 전술한 각 규칙간의 종속성을 나타낸 그래프이다. 그래프에서는 화살표를 이용하여 규칙 간의 종속성을 나타내었다. 도시된 바와 같이 술어 livesIn은  $R_2$ 의 헤드,  $R_3$ 와  $R_4$ 의 선행 조건에 각각 사용된 것을 알 수 있다. 즉,  $R_3$ 와  $R_4$ 는 각각  $R_2$ 에 종속된다. 따라서 만약  $R_3$ 와  $R_4$ 를 적용하기 전에  $R_2$ 를 먼저 적용하면 압축률을 높일 수 있다.  $R_2$ 를 통해 해당 규칙의 헤드에 일치하는 모든 트리플을 식별할 수 있기 때문이다. 즉 특정 규칙의 선행 조건(선행 항목)은 종속성으로 인해 다른 많은 규칙에도 나타나므로 규칙 간의 순위를 지정할 경우 압축률을 높일 수 있다. 어떤 트리플이 활성 그래프로 이동하면 해당 트리플을 삭제하는 것은 불가능하다. 또한 어떤 트리플이 삭제되면, 해당 트리플은 다른 트리플을 삭제하는데 사용될 수 없다. 따라서 개시되는 실시예들에서 RDF 데이터 압축 장치(100)는 최대한 많은 수의 트리플을 삭제할 수 있도록 활성 그래프에 가능한 한 적은 수의 트리플을 추가하도록 구성된다. 이를 위해 규칙 적용 가능성 점수(rule applicability score)가 사용된다. 즉 우선 순위 계산 모듈(104)은, 상기 복수의 규칙 각각의 규칙 적용 가능성 점수에 기반하여 상기 우선 순위를 계산하도록 구성되며, 압축 모듈(106)은, 상기 우선 순위에 따라 순차적으로 상기 복수의 규칙을 적용하여 상기 RDF 데이터를 압축하게 된다.

상기 규칙 적용 가능성 점수는, 각 규칙을 적용할 경우 삭제될 트리플의 최대 개수, 각 규칙의 선행 항목에 존

제하는 조건의 개수 및 각 규칙을 적용할 경우 삭제될 트리플의 최대 개수에 기반하여 계산될 수 있다. 구체적으로 rule<sub>j</sub>의 규칙 적용 가능성 점수는 다음과 같이 계산될 수 있다.

[수학식 5]

$$\text{Rule applicability score}(j) = \frac{1}{|\text{Antecedent}(j)|} * \frac{M(j)}{N(j)}$$

여기서, M(j)는 rule<sub>j</sub>를 적용할 경우 삭제될 트리플의 최대 개수이고, Antecedent(j)는 rule<sub>j</sub>의 선행 항목에 존재하는 조건의 개수이다. M(j)는 다음의 수학식을 이용하여 계산될 수 있다.

[수학식 6]

$$M(j) = \text{Cardinality}(j) + \sum_{G_i \in M1(j)} \text{MaxCardinality}(G_i)$$

또한 N(j)는 rule<sub>j</sub>를 적용하여 삭제를 방지할 수 있는 최대 규칙 수를 의미한다. N(j)는 다음의 수학식을 이용하여 계산될 수 있다.

[수학식 7]

$$N(j) = \sum_{G_i \in P1(j)} \text{MaxCardinality}(G_i) + \sum_{G_i \in P2(j)} \text{MaxCardinality}(G_i)$$

P1(j)와 P2(j)는 rule<sub>j</sub>에 종속된 규칙의 집합이다.

#### top-k 접근 방식

일 실시예에서 규칙 생성 모듈(102)은 기 설정된 지지도 임계값에 기반하여 상기 복수의 규칙의 개수를 k개(k는 1 이상의 자연수)로 제한할 수 있다. 이때 상기 지지도 임계값은, 마이닝된 각 규칙의 헤드 커버리지에 기반하여 설정될 수 있다. 이를 좀 더 상세히 설명하면 다음과 같다.

규칙을 이용하여 RDF 데이터를 압축하기 위해서는 RDF 데이터에 대한 마이닝 과정이 필요하다. 개시되는 실시예에서는 이른바 top-k 접근 방식을 이용하여 사용자가 올바른 마이닝 임계값(유효한 객체를 구별하는 기준)을 찾는 데 도움을 줄 뿐 아니라 마이닝 시간을 크게 단축하도록 구성된다. 마이닝 과정은 특정한 헤드 커버리지(헤드와 일치하는 트리플 범위) 임계값으로부터 도출되는 지지도(support) 임계값에 도달한 규칙을 찾는 과정을 포함한다.

만약 헤드 커버리지 임계값 또는 지지도 임계값이 설정되지 않은 경우, RDF 데이터 압축 장치(100)의 규칙 생성 모듈(102)은 반환될 최대 규칙 수를 식별하기 위해 k 값을 설정할 수 있다. 이 경우, 규칙 마이닝 알고리즘은 탐색된 모든 적절한 규칙을 고정 길이 k를 갖는 우선 순위 큐(priority queue)에 저장한다. 여기서 헤드 규칙은 헤드 커버리지가 가장 작은 규칙이 차지한다. 우선 순위 큐의 용량이 다 차면, 최저 지지도 임계값은 헤드 규칙의 헤드 커버리지 임계값으로 설정된다. 만약 최저 지지도 임계값을 초과하는 새로운 규칙이 입력되면, 헤드 규칙은 새로운 규칙으로 대체되고 최저 지지도 임계값은 우선 순위 큐의 다음 헤드 규칙의 헤드 커버리지로 변경된다. 결과적으로, 지지도 임계값은 마이닝이 계속됨에 따라 지속적으로 상승하고 최종적으로 헤드 커버리지가 가장 높은 k개의 규칙이 남게 된다. 규칙 생성 모듈(102)은 이와 동일한 방식으로 규칙들의 신뢰도 임계값을 설정할 수 있다. 즉, 신뢰도가 가장 높은 k개의 규칙들의 신뢰값 중 가장 낮은 신뢰값이 신뢰도 임계값으로 설정될 수 있다. 만약 최저 신뢰도 임계값을 높이면 신뢰도 계산의 속도도 빨라지게 된다. 최저 신뢰도 임계값(minimumConfidence)이 설정되면 도 2의 수학식은 다음과 같이 재작성될 수 있다.

[수학식 8]

$$\text{number of}(a, b): \forall_{r_1 \dots r_n}: X \leq \text{Support}(X \Rightarrow H) / \text{minimumConfidence}(X \Rightarrow H)$$

- [0099] 도 4는 top-k 접근 방식을 사용하여 최저 지지도 임계값을 설정하는 과정을 설명하기 위한 예시도이다. 도시된 예시도에서는 k가 5로 설정된 예를 나타내었다. 도 4의 (a)에 도시된 우선 순위 큐의 경우 규칙  $R_1$ 이 가장 낮은 지지도(3)를 가지므로, 이 경우 최저 지지도 임계값은 3으로 설정된다. 만약 지지도가 8인 규칙  $R_6$ 이 입력되면, 도 4의 (b)와 같이  $R_1$ 은  $R_6$ 로 대체되고, 최저 지지도 임계값은  $R_2$ 의 지지도인 7로 재설정된다.
- [0100] 도 5는 일 실시예에 따른 RDF 데이터의 압축 방법(500)을 설명하기 위한 흐름도이다. 도시된 방법은 하나 이상의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치, 예컨대 전술한 RDF 데이터의 압축 장치(100)에서 수행될 수 있다. 도시된 흐름도에서는 상기 방법 또는 과정을 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 추가되어 수행될 수 있다.
- [0101] 단계 502에서, RDF 데이터의 압축 장치(100)의 규칙 생성 모듈(102)은 RDF 그래프에 포함된 복수 개의 트리플에 대한 마이닝을 수행하여 복수의 규칙을 생성한다.
- [0102] 단계 504에서, 우선 순위 계산 모듈(104)은 상기 복수의 규칙간의 우선 순위를 계산한다.
- [0103] 단계 506에서, 압축 모듈(106)은 상기 우선 순위에 따라 상기 복수의 규칙을 상기 RDF 그래프에 적용하여 상기 RDF 그래프를 압축한다.
- [0104] 이상에서 대표적인 실시예를 통하여 본 발명에 대하여 상세하게 설명하였으나, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 전술한 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.

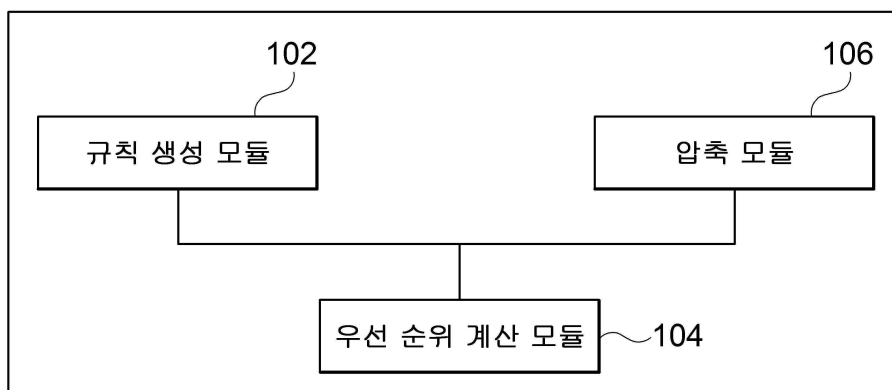
## 부호의 설명

- [0105] 100: RDF 데이터의 압축 장치  
102: 규칙 생성 모듈  
104: 우선 순위 계산 모듈  
106: 압축 모듈

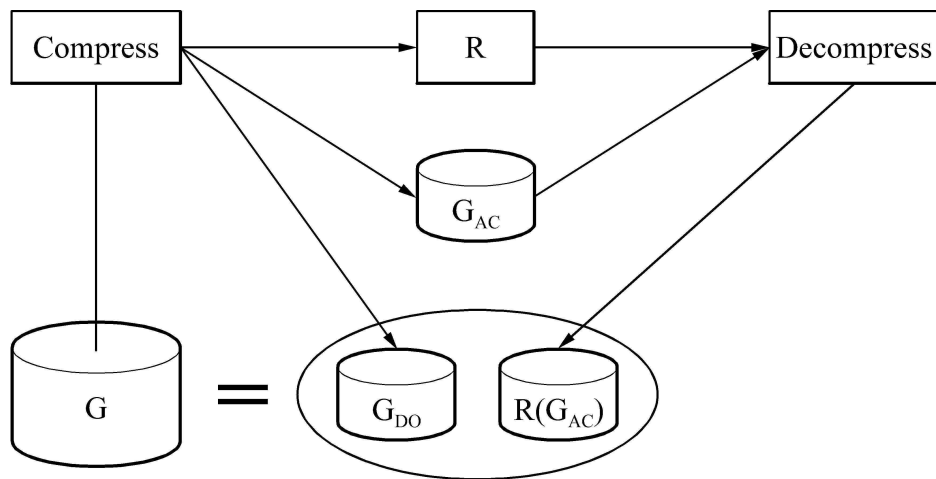
## 도면

### 도면1

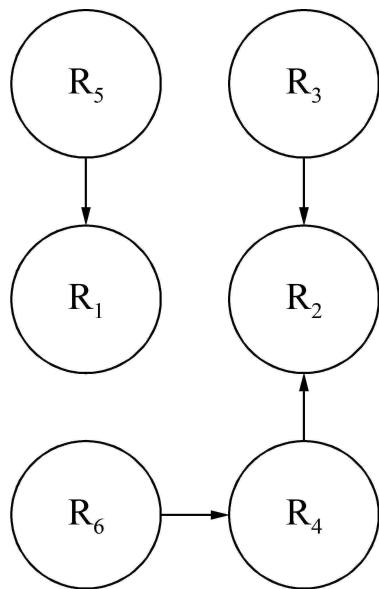
#### 100



도면2



도면3



도면4

|                                    |                                     |                                     |                                     |                                      |                                      |
|------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|--------------------------------------|--------------------------------------|
| <b>Head</b><br>Minimum Support : 3 | <b>R<sub>1</sub></b><br>Support : 3 | <b>R<sub>2</sub></b><br>Support : 7 | <b>R<sub>3</sub></b><br>Support : 9 | <b>R<sub>4</sub></b><br>Support : 14 | <b>R<sub>5</sub></b><br>Support : 17 |
| <b>Head</b><br>Minimum Support : 7 | <b>R<sub>2</sub></b><br>Support : 7 | <b>R<sub>6</sub></b><br>Support : 8 | <b>R<sub>3</sub></b><br>Support : 9 | <b>R<sub>4</sub></b><br>Support : 14 | <b>R<sub>5</sub></b><br>Support : 17 |

도면5

500

