



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2014-0129990
(43) 공개일자 2014년11월07일

(51) 국제특허분류(Int. Cl.)

G06F 17/40 (2006.01)

(21) 출원번호 10-2013-0082123

(22) 출원일자 2013년07월12일

심사청구일자 2013년07월12일

(30) 우선권주장

1020130048702 2013년04월30일 대한민국(KR)

(71) 출원인

경희대학교 산학협력단

경기도 용인시 기흥구 덕영대로 1732, 국제캠퍼스 내 (서천동, 경희대학교)

(72) 발명자

이영구

경기 용인시 기흥구 덕영대로 1732, 전자정보관 314호 (서천동, 경희대학교)

홍지혜

충남 천안시 동남구 충절로 42, 103동 105호 (신부동, 신부경남아너스빌아파트)

한용구

경기 용인시 기흥구 덕영대로 1732, 전자정보관 349호 (서천동, 경희대학교)

(74) 대리인

서재승

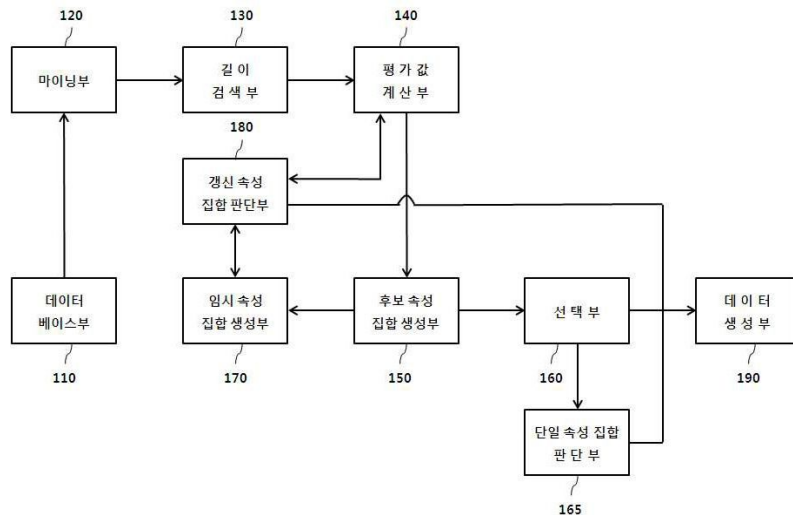
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 R D F 데이터의 데이터테이블 생성 방법

(57) 요약

본 발명은 데이터테이블의 생성 방법에 관한 것으로, 보다 구체적으로 데이터베이스부에 저장되어 있는 RDF 데이터를 서로 연관도가 높으며 동시에 서로 동일한 속성이 중복되지 않도록 데이터테이블을 생성하는 방법에 관한 것이다.

대표도



이 발명을 지원한 국가연구개발사업

과제고유번호 2012-049075

부처명 교육과학기술부

연구관리전문기관 한국연구재단

연구사업명 도약연구지원사업

연구과제명 u-Health 응용을 위한 유비쿼터스 DB 엔진 기술 연구

기 여 율 1/1

주관기관 경희대학교 산학협력단

연구기간 2012.07.01 ~ 2013.06.30

특허청구의 범위

청구항 1

데이터베이스에서 기설정된 최소 빈발 지지도 이상의 빈발 지지도를 가지는 빈발 속성 집합을 검색하는 단계;

상기 빈발 속성 집합의 길이와 빈발 지지도에 기초하여 상기 빈발 속성 집합의 연관 평가값을 계산하는 단계;

상기 빈발 속성 집합 중 임계 연관 평가값 이상을 가지는 후보 속성 집합을 생성하는 단계;

상기 후보 속성 집합의 널 비율 또는 상기 연관 평가값에 기초하여 상기 후보 속성 집합 중에서 선택 속성 집합을 선택하는 단계;

상기 데이터베이스에 저장되어 있는 속성 중 상기 빈발 속성 집합에 포함되지 않은 속성을 판단하여 상기 빈발 속성 집합에 포함되지 않은 속성에 대한 단일 속성 집합을 생성하는 단계; 및

상기 선택 속성 집합 및 상기 단일 속성 집합을 구분하여 상기 선택 속성 집합과 상기 단일 속성 집합의 데이터 테이블을 생성하는 단계를 포함하는 것을 특징으로 하는 데이터테이블 생성 방법.

청구항 2

제 1 항에 있어서, 상기 데이터베이스에는 객체, 속성 및 속성값으로 구분되어 데이터가 저장되어 있는 것을 특징으로 하는 데이터테이블 생성 방법.

청구항 3

제 2 항에 있어서, 상기 데이터베이스에 저장되어 있는 데이터는 리소스 기술 프레임워크(Resource Description Framework)에 따라 기술되어 있는 것을 특징으로 하는 데이터테이블 생성 방법.

청구항 4

제 3 항에 있어서, 상기 선택 속성 집합에 대한 데이터테이블은 객체별로 상기 선택 속성 집합을 구성하는 속성에 대한 속성값으로 생성되며, 상기 단일 속성 집합에 대한 데이터테이블은 객체별로 상기 단일 속성 집합을 구성하는 속성에 대한 속성값으로 생성되는 것을 특징으로 하는 데이터테이블 생성 방법.

청구항 5

제 4 항에 있어서, 상기 데이터테이블 생성 방법은

데이터베이스에서 상기 빈발 속성 집합 중 임계 길이 이상의 빈발 속성 집합을 2차 검색하는 단계를 더 포함하며,

2차 검색한 빈발 속성 집합 중에서 상기 후보 속성 집합을 생성하는 것을 특징으로 하는 데이터테이블 생성 방법.

청구항 6

제 4 항에 있어서, 상기 데이터테이블 생성 방법은

상기 후보 속성 집합 중 널 비율이 임계 널 비율보다 작은 후보 속성 집합, 상기 후보 속성 집합에서 가장 높은 연관 평가값을 가지는 최고 후보 속성 집합 또는 상기 최고 속성 집합과 공통된 속성이 존재하지 않는 후보 속성 집합을 선택 속성 집합으로 선택하는 것을 특징으로 하는 데이터테이블 생성 방법.

청구항 7

제 6 항에 있어서, 상기 데이터테이블 생성 방법은

상기 후보 속성 집합 중에서 상기 널 비율이 임계 널 비율보다 큰 후보 속성 집합 또는 상기 최고 후보 속성 집합과 공통된 속성이 존재하는 후보 속성 집합을 임시 속성 집합으로 분리하는 단계를 더 포함하는 것을 특징으로 하는 데이터테이블 생성 방법.

청구항 8

제 7 항에 있어서,

상기 후보 속성 집합 중에서 상기 최고 후보 속성 집합과 공통된 속성이 존재하는 후보 속성 집합은 상기 공통된 속성을 삭제하여 상기 임시 속성 집합으로 분리되는 것을 특징으로 하는 데이터테이블 생성 방법.

청구항 9

제 7 항에 있어서,

상기 후보 속성 집합 중에서 상기 널 비율이 임계 널 비율보다 큰 후보 속성 집합은 상기 후보 속성 집합을 구성하는 속성별 널 비율에 기초하여 높은 널 비율을 가지는 순서의 속성을 삭제하여 상기 후보 속성 집합의 널 비율이 상기 임계 널 비율보다 작도록 임시 속성 집합으로 분리되는 것을 특징으로 하는 데이터테이블 생성 방법.

청구항 10

제 8 항 또는 제 9 항에 있어서, 상기 데이터테이블 생성 방법은

상기 임시 속성 집합의 길이와 빈발 지지도에 기초하여 상기 임시 속성 집합의 연관 평가값을 계산하는 단계; 및

상기 임시 속성 집합 중 가장 높은 연관 평가값을 가지는 갱신 속성 집합을 선택하고, 상기 갱신 속성 집합의 데이터테이블을 생성하는 단계를 더 포함하는 것을 특징으로 하는 데이터테이블 생성 방법.

청구항 11

데이터베이스에서 기설정된 최소 빈발 지지도 이상의 빈발 지지도를 가지는 빈발 속성 집합을 검색하는 마이닝부;

상기 빈발 속성 집합의 길이와 빈발 지지도에 기초하여 상기 빈발 속성 집합의 연관 평가값을 계산하는 평가값 계산부;

상기 빈발 속성 집합 중 임계 연관 평가값 이상을 가지는 후보 속성 집합을 생성하는 후보 속성 집합 생성부;

상기 후보 속성 집합의 널 비율 또는 연관 평가값에 기초하여 상기 후보 속성 집합 중에서 선택 속성 집합을 선택하는 선택부;

상기 데이터베이스에 저장되어 있는 속성 중 상기 빈발 속성 집합에 포함되지 않은 속성을 판단하여 상기 빈발 속성 집합에 포함되지 않은 속성에 대한 단일 속성 집합을 생성하는 단일 속성 집합 생성부; 및

상기 선택 속성 집합 및 상기 단일 속성 집합을 구분하여 상기 선택 속성 집합과 상기 단일 속성 집합의 데이터테이블을 생성하는 데이터테이블 생성부를 포함하는 것을 특징으로 하는 데이터테이블 생성 장치.

청구항 12

제 11 항에 있어서, 상기 데이터베이스에는 객체, 속성 및 속성값으로 구분되어 데이터가 저장되어 있는 것을 특징으로 하는 데이터테이블 생성 장치.

청구항 13

제 12 항에 있어서, 상기 데이터베이스에 저장되어 있는 데이터는 리소스 기술 프레임워크(Resource Description Framework)에 따라 기술되어 있는 것을 특징으로 하는 데이터테이블 생성 장치.

청구항 14

제 13 항에 있어서, 상기 선택 속성 집합에 대한 데이터테이블은 객체별로 상기 선택 속성 집합을 구성하는 속성에 대한 속성값으로 생성되며, 상기 단일 속성 집합에 대한 데이터테이블은 객체별로 상기 단일 속성 집합을 구성하는 속성에 대한 속성값으로 생성되는 것을 특징으로 하는 데이터테이블 생성 장치.

청구항 15

제 14 항에 있어서, 상기 선택부는

상기 후보 속성 집합 중 널 비율이 임계 널 비율보다 작은 후보 속성 집합, 상기 후보 속성 집합에서 가장 높은 연관 평가값을 가지는 최고 후보 속성 집합 또는 상기 최고 속성 집합과 공통된 속성이 존재하지 않는 후보 속성 집합을 선택 속성 집합으로 선택하는 것을 특징으로 하는 데이터테이블 생성 장치.

청구항 16

제 14 항에 있어서, 상기 데이터테이블 생성 장치는

상기 데이터베이스에서 상기 빈발 속성 집합 중 임계 길이 이상의 빈발 속성 집합을 2차 검색하는 길이 검색부를 더 포함하며,

상기 평가값 계산부는 2차 검색한 빈발 속성 집합의 연관 평가값을 계산하는 것을 특징으로 하는 데이터테이블 생성 방법.

청구항 17

제 14 항에 있어서, 상기 후보 속성 집합 생성부는

상기 후보 속성 집합 중 널 비율이 임계 널 비율보다 작거나, 상기 후보 속성 집합에서 가장 높은 연관 평가값을 가지는 최고 후보 속성 집합과 공통된 속성이 존재하지 않는 후보 속성 집합을 상기 후보 속성 집합에서 제외하는 것을 특징으로 하는 데이터테이블 생성 장치.

청구항 18

제 14 항에 있어서, 상기 데이터테이블 생성 장치는

상기 후보 속성 집합 중에서 널 비율이 임계 널 비율보다 큰 후보 속성 집합 또는 상기 최고 후보 속성 집합과 공통된 속성이 존재하는 후보 속성 집합으로부터 임시 속성 집합을 생성하는 임시 속성 집합 생성부; 및

상기 임시 속성 집합의 길이와 빈발 지지도에 기초하여 계산한 상기 임시 속성 집합의 연관 평가값에 기초하여 상기 임시 속성 집합 중 가장 높은 연관 평가값을 가지는 갱신 속성 집합을 판단하는 갱신 속성 집합 판단부를 더 포함하며,

상기 데이터 생성부는 상기 갱신 속성 집합의 데이터테이블을 생성하는 것을 특징으로 하는 데이터테이블 생성 장치.

청구항 19

제 18 항에 있어서, 상기 임시 속성 집합 생성부는

상기 후보 속성 집합 중에서 상기 최고 후보 속성 집합과 공통된 속성이 존재하는 후보 속성 집합에서 상기 공통된 속성을 삭제하여 임시 속성 집합을 생성하는 것을 특징으로 하는 데이터테이블 생성 장치.

청구항 20

제 18 항에 있어서, 상기 임시 속성 집합 생성부는

상기 후보 속성 집합 중에서 상기 널 비율이 임계 널 비율보다 큰 후보 속성 집합에서 상기 후보 속성 집합을 구성하는 속성별 널 비율에 기초하여 높은 널 비율을 가지는 순서의 속성을 삭제하여 상기 후보 속성 집합의 널 비율이 상기 임계 널 비율보다 작도록 임시 속성 집합을 생성하는 것을 특징으로 하는 데이터테이블 생성 장치.

명 세 서

기술 분야

[0001]

본 발명은 데이터테이블의 생성 방법에 관한 것으로, 보다 구체적으로 데이터베이스부에 저장되어 있는 RDF 데이터를 서로 연관도가 높으며 동시에 서로 동일한 속성이 중복되지 않도록 데이터테이블을 생성하는 방법에 관

한 것이다.

배경 기술

- [0002] 확장 마크업언어(extensible markup language, XML)가 가진 비정규적 논리적 구조와 사용자들의 다양한 표현을 이용한 데이터를 기계가 이해하기는 쉽지 않다. 따라서 웹에 존재하는 데이터로부터 풍부하며 정확한 검색이 매우 어렵다. 이에 대한 해결책으로 웹에 수록된 데이터를 기술하기 위한 메타데이터(metadata)를 사용하는 것이다. 메타데이터는 데이터에 관한 구조화된 데이터로 속성정보라고도 하는데, 메타데이터에 대한 표준화의 많은 연구가 진행되고 있다. 특히 월드와이드 웹 컨소시엄(World Wide Web Consortium, W3C)에서는 메타데이터의 표준에 대한 연구로 RDF(Resource Description Framework)를 표준안으로 제시하였다.
- [0003] 기존의 HTML(Hypertext Markup Language)이나 XML이 가지는 장점과 더불어 메타데이터를 기술할 수 있는 장점을 더한 것이 RDF라 할 수 있다. 이처럼 RDF는 메타데이터를 처리하기 위한 표준으로 정보 리소스들의 의미와 이들 사이의 연관성을 기술하기 위한 수단을 제공해 주기 위해 개발됐다.
- [0004] RDF 형태의 데이터는 객체/속성/속성값이라는 반복된 구조로 구성되어 있다. 객체는 RDF에 의하여 표현되는 모든 것, 표현의 객체를 나타낸다. 객체는 모든 웹 페이지, 혹은 웹페이지의 일정 부분, 도서 정보, 사원 정보 등 거의 모든 것이 그 대상이다. 속성은 객체를 잘 기술하기 위한 특정한 성질, 특성, 관계 등을 의미한다. 속성값은 속성이 지니는 실제의 값을 의미한다.
- [0005] 이러한 RDF 형태의 데이터는 객체, 속성, 속성값에 의해 도 1에 도시되어 있는 바와 같이 트리플 형태의 그래프로 표현되는데, 도 2는 RDF 형태의 데이터를 그래프로 표현한 일 예를 도시하고 있다.
- [0006] 이러한 RDF 데이터를 효율적으로 저장하기 위해 RDF 데이터를 저장하는 저장 공간을 절약하며, 동시에 조인 연산을 줄여야 한다. 여기서 조인 연산이란 객체의 서로 다른 속성 집합이 각각 개별적으로 데이터테이블로 생성되어 저장되는 경우, 사용자의 검색 명령에 포함된 다수의 속성 질의에 따라 다수의 데이터테이블에 분산되어 있는 속성들을 통해 객체를 검색하는데 소요되는 연산을 의미한다.
- [0007] 조인 연산을 줄이기 위하여 도 3의 (a) 또는 (b)와 같이 연관도가 높은 속성 집합별로 데이터테이블을 생성하는 경우, 일부 속성은 중복하여 데이터테이블에 저장되어 저장 공간을 낭비하게 된다. 한편, 도 4의 (a) 내지 (e)와 같이 저장 공간의 낭비를 줄이기 위하여 모든 속성별로 개별적으로 데이터테이블을 생성하는 경우 조인 연산이 많아져 빠른 검색이 불가능하며 프로세스량이 많아진다는 문제점을 가진다.

발명의 내용

해결하려는 과제

- [0008] 따라서 조인 연산량을 줄이며 동시에 저장 공간을 절약하면서 RDF 데이터의 데이터테이블을 생성할 수 있는 새로운 데이터 관리 방법이 요구된다.
- [0009] 본 발명은 위에서 언급한 종래 RDF 데이터의 데이터 관리 기법이 가지는 문제점들을 해결하기 위한 것으로, 본 발명이 이루고자 하는 목적은 조인 연산을 줄일 수 있는 데이터테이블의 생성 방법을 제공하는 것이다.
- [0010] 본 발명이 이루고자 하는 다른 목적은 저장 공간을 절약할 수 있으며 동시에 조인 연산을 줄일 수 있는 데이터테이블의 생성 방법을 제공하는 것이다.

과제의 해결 수단

- [0011] 본 발명에 따른 데이터테이블 생성 방법은 데이터베이스에서 기설정된 최소 빈발 지지도 이상의 빈발 지지도를 가지는 빈발 속성 집합을 검색하는 단계와, 빈발 속성 집합의 길이와 빈발 지지도에 기초하여 상기 빈발 속성 집합의 연관 평가값을 계산하는 단계와, 빈발 속성 집합 중 임계 연관 평가값 이상을 가지는 후보 속성 집합을 생성하는 단계와, 후보 속성 집합의 널 비율 또는 연관 평가값에 기초하여 후보 속성 집합 중에서 선택 속성 집합을 선택하는 단계와, 데이터베이스에 저장되어 있는 속성 중 빈발 속성 집합에 포함되지 않은 속성을 판단하여 빈발 속성 집합에 포함되지 않은 속성에 대한 단일 속성 집합을 생성하는 단계와, 선택 속성 집합 및 단일 속성 집합을 구분하여 선택 속성 집합과 단일 속성 집합의 데이터테이블을 생성하는 단계를 포함하는 것을 특징으로 한다.
- [0012] 여기서 데이터베이스에는 객체, 속성 및 속성값으로 구분되어 데이터가 저장되어 있는 것을 특징으로 한다. 구

체적으로 데이터베이스에 저장되어 있는 데이터는 리소스 기술 프레임워크(Resource Description Framework, RDF)에 따라 기술되어 있는 것을 특징으로 한다.

- [0013] 선택 속성 집합에 대한 데이터테이블은 객체별로 선택 속성 집합을 구성하는 속성에 대한 속성값으로 생성되며, 단일 속성 집합에 대한 데이터테이블은 객체별로 단일 속성 집합을 구성하는 속성에 대한 속성값으로 생성되는 것을 특징으로 한다.
- [0014] 바람직하게, 본 발명에 따른 데이터테이블 생성 방법은 데이터베이스에서 빈발 속성 집합 중 임계 길이 이상의 빈발 속성 집합을 2차 검색하는 단계를 더 포함하며, 2차 검색한 빈발 속성 집합 중에서 후보 속성 집합을 생성하는 것을 특징으로 한다.
- [0015] 본 발명에 따른 데이터테이블 생성 방법은 후보 속성 집합 중 널 비율이 임계 널 비율보다 작은 후보 속성 집합, 후보 속성 집합에서 가장 높은 연관 평가값을 가지는 최고 후보 속성 집합 또는 최고 속성 집합과 공통된 속성이 존재하지 않는 후보 속성 집합을 선택 속성 집합으로 선택한다.
- [0016] 바람직하게, 본 발명에 따른 데이터테이블 생성 방법은 후보 속성 집합 중에서 널 비율이 임계 널 비율보다 큰 후보 속성 집합 또는 최고 후보 속성 집합과 공통된 속성이 존재하는 후보 속성 집합을 임시 속성 집합으로 분리하는 단계를 더 포함하는데, 후보 속성 집합 중에서 최고 후보 속성 집합과 공통된 속성이 존재하는 후보 속성 집합은 공통된 속성을 삭제하여 상기 임시 속성 집합으로 분리되거나, 후보 속성 집합 중에서 널 비율이 임계 널 비율보다 큰 후보 속성 집합은 후보 속성 집합을 구성하는 속성별 널 비율에 기초하여 높은 널 비율을 가지는 순서의 속성을 삭제하여 후보 속성 집합의 널 비율이 임계 널 비율보다 작도록 임시 속성 집합으로 분리되는 것을 특징으로 한다.
- [0017] 바람직하게, 본 발명에 따른 데이터테이블 생성 방법은 임시 속성 집합의 길이와 빈발 지지도에 기초하여 임시 속성 집합의 연관 평가값을 계산하는 단계와, 임시 속성 집합 중 가장 높은 연관 평가값을 가지는 갱신 속성 집합을 선택하고 갱신 속성 집합의 데이터테이블을 생성하는 단계를 더 포함하는 것을 특징으로 한다.
- [0018] 한편, 본 발명에 따른 데이터테이블 생성 장치는 데이터베이스에서 기설정된 최소 빈발 지지도 이상의 빈발 지지도 가지는 빈발 속성 집합을 검색하는 마이닝부와, 빈발 속성 집합의 길이와 지지도에 기초하여 빈발 속성 집합의 연관 평가값을 계산하는 평가값 계산부와, 빈발 속성 집합 중 임계 연관 평가값 이상을 가지는 후보 속성 집합을 생성하는 후보 속성 집합 생성부와, 후보 속성 집합의 널 비율 또는 연관 평가값에 기초하여 후보 속성 집합 중에서 선택 속성 집합을 선택하는 선택부와, 데이터베이스에 저장되어 있는 속성 중 빈발 속성 집합에 포함되지 않은 속성을 판단하여 빈발 속성 집합에 포함되지 않은 속성에 대한 단일 속성 집합을 생성하는 단일 속성 집합 생성부와, 선택 속성 집합 및 단일 속성 집합을 구분하여 선택 속성 집합과 단일 속성 집합의 데이터테이블을 생성하는 데이터테이블 생성부를 포함하는 것을 특징으로 한다.
- [0019] 바람직하게, 본 발명에 따른 데이터테이블 생성 장치는 데이터베이스에서 상기 빈발 속성 집합 중 임계 길이 이상의 빈발 속성 집합을 2차 검색하는 길이 검색부를 더 포함하는데, 평가값 계산부는 2차 검색한 빈발 속성 집합의 연관 평가값을 계산하는 것을 특징으로 한다.
- [0020] 바람직하게, 본 발명에 따른 데이터테이블 생성 장치는 후보 속성 집합 중에서 널 비율이 임계 널 비율보다 큰 후보 속성 집합 또는 최고 후보 속성 집합과 공통된 속성이 존재하는 후보 속성 집합으로부터 임시 속성 집합을 생성하는 임시 속성 집합 생성부와, 임시 속성 집합의 길이와 빈발 지지도에 기초하여 계산한 임시 속성 집합의 연관 평가값에 기초하여 임시 속성 집합 중 가장 높은 연관 평가값을 가지는 갱신 속성 집합을 판단하는 갱신 속성 집합 판단부를 더 포함하는데, 여기서 데이터 생성부는 갱신 속성 집합의 데이터테이블을 생성하는 것을 특징으로 한다.

발명의 효과

- [0021] 본 발명에 따른 데이터테이블 생성 방법 및 그 장치는 종래 데이터테이블 생성 방법과 비교하여 다음과 같은 효과들을 가진다.
- [0022] 첫째, 본 발명에 따른 데이터테이블 생성 방법은 최소 빈발 지지도 이상을 가지는 연관된 속성들을 1개의 데이터테이블로 생성함으로써, 조인 연산을 줄일 수 있다.
- [0023] 둘째, 본 발명에 따른 데이터테이블 생성 방법은 선택 속성 집합은 서로 공통된 속성을 가지지 않도록 선택하며 선택 속성 집합에 기초하여 데이터테이블을 생성함으로써, 속성의 중복 저장으로 인한 저장 공간의 낭비를 방지

할 수 있다.

[0024] 셋째, 본 발명에 따른 데이터테이블 생성 방법은 빈발 속성 집합의 빈발 지지도와 함께 빈발 속성 집합의 길이를 함께 고려함으로써, 연관이 높은 속성 집합이 서로 분리되어 데이터테이블로 생성되는 것을 방지할 수 있다.

도면의 간단한 설명

[0025] 도 1은 RDF 형태의 데이터를 객체, 속성, 속성값의 트리플 형태의 그래프로 표현한 일 예이다.
 도 2는 RDF 형태의 데이터를 그래프로 표현한 일 예를 도시하고 있다.
 도 3은 서로 연관도가 높은 속성들을 포함하는 데이터테이블의 일 예를 도시하고 있다.
 도 4는 모든 속성별로 개별적으로 데이터테이블을 생성한 일 예를 도시하고 있다.
 도 5는 본 발명에 따른 데이터테이블 생성 장치를 설명하기 위한 기능 블록도이다.
 도 6은 본 발명에 따른 데이터테이블 생성 방법을 설명하기 위한 흐름도이다.
 도 7은 본 발명에서 선택 속성 집합을 선택하는 방법의 일 예를 설명하기 위한 흐름도이다.
 도 8은 속성 집합의 데이터테이블의 일 예를 도시하고 있다.
 도 9는 본 발명에서 갱신 속성 집합을 선택하는 단계의 일 예를 설명하기 위한 흐름도이다.
 도 10은 후보 속성 집합의 연관 평가값의 일 예를 설명하기 위한 도면이다.
 도 11은 갱신 속성 집합을 생성하는 방법의 일 예를 설명하기 위한 도면이다.
 도 12는 선택 속성 집합, 단일 속성 집합 및 갱신 속성 집합으로 생성되는 데이터테이블의 일 예를 도시하고 있다.
 도 13은 최소 빈발 지지도 이상의 빈발 지지도를 가지는 빈발 속성 집합의 생성 과정을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0026] 이하 첨부한 도면을 참고로 본 발명에 따른 데이터테이블 생성 방법 및 그 장치에 대해 보다 구체적으로 설명한다.

[0027] 도 5는 본 발명에 따른 데이터테이블 생성 장치를 설명하기 위한 기능 블록도이다.

[0028] 도 5를 참고로 보다 구체적으로 살펴보면, 마이닝부(120)는 데이터베이스부(110)에 저장되어 있는 데이터를 마이닝하여 최소 빈발 지지도 이상으로 발생하는 빈발 속성 집합을 생성한다. 여기서 데이터베이스부(110)에 저장되어 있는 데이터는 객체, 속성, 속성값으로 구성되는 리소스 기술 프레임워크(RDF) 양식이며 객체, 속성, 속성값이 서로 연결되는 그래프 형태로 저장된다.

[0029] 마이닝부(120)는 데이터베이스부(110)에 저장되어 있는 그래프에서 각 객체에 대해 질의되는 속성에 기초하여 최소 빈발 지지도 이상으로 질의되는 속성 집합을 추출한다. 빈발 부분그래프 마이닝은 원래 그래프가 저장되어 있는 데이터베이스에서 최소 빈발 지지도 이상으로 출현하는 빈발 부분그래프를 생성하는 것인데, 그래프의 고유한 특성을 나타내어 그래프의 분류, 군집화, 인덱싱 등에서 사용된다. 빈발 부분그래프 마이닝은 그래프 데이터베이스에서 모든 빈발 부분그래프를 찾는 기법으로, 복잡한 구조를 가지는 그래프들이 대용량으로 저장되어 있는 그래프 데이터베이스로부터 최소 지지도 이상으로 출현하는 모든 부분그래프를 검색하는 과정을 의미한다. 빈발 부분그래프를 효율적으로 마이닝하기 위하여 AGM, FSG, gSpan 등의 다양한 종래 기법들이 연구되고 있다. 이 중에서 성능이 탁월한 것으로 알려진 gSpan이 널리 사용되고 있다. gSpan은 패턴 확장 방법으로 그래프 데이터베이스로부터 빈발 부분그래프를 마이닝하는 알고리즘으로, 깊이우선탐색(depth first search, 이하 DFS) 순서로 방문한 그래프의 에지들을 각 에지의 DFS 코드로 표현한다. 그래프는 탐색 시작 버텍스(vertex)에 따라 다양한 DFS 코드가 생성되는데, gSpan에서는 최소값을 가지는 DFS 코드를 정규 그래프로 정의하여 빈발 부분그래프가 최소 DFS코드와 다르게 확장된 경우 제거하여 많은 수의 중복 부분그래프들을 제거한다.

[0030] 마이닝부(120)는 이러한 마이닝 기법을 적용하여 객체가 가지는 다수의 속성으로 이루어진 그래프에서 사용자에게

의해 질의되는 객체별 속성에 기초하여 사용자에게 의해 동시에 질의되는 속성 중 최소 빈발 지지도(TH_{fr}) 이상으로 이루어지는 빈발 속성 집합을 생성한다. 도 13은 최소 빈발 지지도 이상의 빈발 지지도를 가지는 빈발 속성 집합의 생성 과정을 설명하기 위한 도면이다.

[0031] 평가값 계산부(140)는 빈발 속성 집합의 길이와 빈발 지지도에 기초하여 빈발 속성 집합의 연관 평가값을 계산한다. 여기서 빈발 속성 집합의 길이(d)란 빈발 속성 집합을 구성하는 속성의 수를 의미한다. 바람직하게, 길이 검색부(130)는 마이닝부(120)에서 생성된 빈발 속성 집합의 길이를 판단하여 빈발 속성 집합 중 임계 길이 이상을 가지는 빈발 속성 집합을 2차 검색하며, 평가값 계산부(140)는 빈발 속성 집합 중 2차 검색한 빈발 속성 집합에 대해서만 연관 평가값을 계산한다.

[0032] 후보 속성 집합 생성부(150)는 빈발 속성 집합 중 임계 연관 평가값 이상을 가지는 빈발 속성 집합을 판단하여 후보 속성 집합을 생성하며, 선택부(160)는 후보 속성 집합의 널($null$) 비율 또는 연관 평가값에 기초하여 후보 속성 집합 중에서 선택 속성 집합을 선택한다. 여기서 선택부(160)는 후보 속성 집합에서 가장 높은 연관 평가값을 가지는 최고 후보 속성 집합 또는 최고 속성 집합과 공통된 속성이 존재하지 않는 후보 속성 집합을 선택 속성 집합으로 선택하는데, 보다 바람직하게 최고 후보 속성 집합 또는 최고 속성 집합과 공통된 속성이 존재하지 않는 후보 속성 집합의 널 비율이 임계 널 비율보다 작은 경우 선택 속성 집합으로 선택한다. 여기서 널 비율이란 후보 속성 집합을 구성하는 속성들로 객체의 데이터테이블을 생성하는 경우 전체 데이터테이블 중 널이 차지하는 비율을 의미한다.

[0033] 단일 속성 집합 생성부(165)는 데이터베이스부(110)에 저장되어 있는 객체의 속성 중 빈발 속성 집합에 포함되지 않은 속성을 판단하여 빈발 속성 집합에 포함되지 않은 속성에 대한 단일 속성 집합을 생성한다.

[0034] 한편, 임시 속성 집합 생성부(170)는 후보 속성 집합 중에서 널 비율이 임계 널 비율보다 큰 후보 속성 집합 또는 최고 후보 속성 집합과 공통된 속성이 존재하는 후보 속성 집합을 임시 속성 집합으로 생성하며, 갱신 속성 집합 판단부(180)는 임시 속성 집합의 길이와 빈발 지지도에 기초하여 평가값 계산부(140)에서 계산한 임시 속성 집합의 연관 평가값에 기초하여 임시 속성 집합 중 가장 높은 연관 평가값을 가지는 갱신 속성 집합을 판단한다. 데이터 생성부(190)는 선택부(160)에서 선택한 선택 속성 집합, 단일 속성 집합 생성부(165)에서 생성한 단일 속성 집합 및 갱신 속성 집합 판단부(180)에서 판단한 갱신 속성 집합에 대한 데이터테이블을 개별적으로 구분하여 각각 생성한다.

[0035] 도 6은 본 발명에 따른 데이터테이블 생성 방법을 설명하기 위한 흐름도이다.

[0036] 도 6과 도 13을 참고로 보다 구체적으로 살펴보면, 데이터베이스부에 그래프 형식으로 저장되어 있는 RDF 데이터에서 최소 빈발 지지도(TH_{fr}) 이상의 빈발 지지도를 가지는 빈발 속성 집합을 생성하며(S110), 생성한 빈발 속성 집합 중에서 임계 길이(TH_d) 이상을 가지는 빈발 속성 집합을 검색하여 임계 길이 이상을 가지는 빈발 속성 집합을 생성한다(S120). 도 13을 참고로 살펴보면, 다수의 속성이 공통의 검색 질의로 입력되는 경우, 검색 질의를 구성하는 각 속성 중 최소 빈발 지지도 이상을 가지는 속성이 1차 빈발 속성 집합으로 생성된다. 1차 빈발 속성 집합은 1개의 속성으로 구성되는데 제1 길이($d=1$)를 가진다. 1차 빈발 속성 집합으로부터 확장되어 제2 길이($d=2$)를 가지는 2개의 속성으로 이루어진 최소 빈발 지지도 이상의 2차 빈발 속성 집합이 생성된다. 이와 같이 마이닝 기법에 따라 빈발 속성 집합은 계속 확장되어 빈발 속성 집합의 속성들을 모두 포함하는 새로운 자식 빈발 속성 집합을 생성한다.

[0037] 생성한 빈발 속성 집합의 길이와 빈발 지지도에 기초하여 빈발 속성 집합의 연관 평가값을 계산한다(S130). 본 발명에서 빈발 속성 집합의 연관 평가값(CE)은 아래의 수학적식(1)과 같이 계산된다.

[0038] [수학적식 1]

$$CE = P^{\alpha} \times d^{\beta}$$

[0040] 여기서 P 와 d 는 각각 빈발 속성 집합의 빈발 지지도와 길이를 의미하며, α , β 는 각각 빈발 지지도와 길이에 대한 기설정 가중치이다.

[0041] 빈발 속성 집합의 연관 평가값에 기초하여 빈발 속성 집합 중 임계 연관 평가값 이상을 가지는 빈발 속성 집합을 후보 속성 집합으로 생성하며(S140), 후보 속성 집합의 널 비율 또는 후보 속성 집합의 연관 평가값에 기초

하여 후보 속성 집합 중에서 선택 속성 집합을 선택한다(S150).

[0042] 데이터베이스에 저장되어 있는 속성 중 빈발 속성 집합에 포함되지 않은 속성을 판단하여 빈발 속성 집합에 포함되지 않은 속성을 단일 속성 집합으로 생성하며(S160), 후보 속성 집합에서 분리된 임시 속성 집합의 길이와 빈발 지지도에 기초하여 임시 속성 집합의 연관 평가값을 계산하고 임시 속성 집합 중 가장 높은 연관 평가값을 가지는 갱신 속성 집합을 선택한다(S170). 여기서 임시 속성 집합은 후보 속성 집합 중에서 널 비율이 임계 널 비율보다 큰 후보 속성 집합 또는 후보 속성 집합 중에서 연관 평가값이 가장 높은 최고 후보 속성 집합과 공통된 속성이 존재하는 후보 속성 집합을 임시 속성 집합으로 분리한다.

[0043] 선택 속성 집합, 단일 속성 집합 및 갱신 속성 집합 각각을 서로 구분하여 데이터테이블을 생성한다(S180).

[0044] 빈발 속성 집합이 확장될수록 빈발 속성 집합의 빈발 지지도는 감소하는 반면 빈발 속성 집합의 길이는 증가한다. 빈발 속성 집합에 대한 데이터테이블은 조인 연산을 줄이고 저장 공간의 낭비를 방지하기 위하여 높은 빈발 지지도를 가지며 동시에 많은 수의 속성으로 이루어지도록 생성하는 것이 가장 바람직한데, 본 발명에서는 임계 길이 이상을 가지는 빈발 속성 집합에 대해서만 데이터테이블을 생성하도록 제어한다. 또한, 빈발 속성 집합의 지지도뿐만 빈발 속성 집합의 길이도 함께 고려한 연관 평가값에 기초하여 데이터테이블을 생성함으로써, 연관도가 높은 속성들이 서로 분리되어 데이터테이블로 생성되는 것을 방지한다.

[0045] 도 7은 본 발명에서 선택 속성 집합을 선택하는 방법의 일 예를 설명하기 위한 흐름도이다.

[0046] 도 7을 참고로 보다 구체적으로 살펴보면, 후보 속성 집합의 연관 평가값에 기초하여 가장 높은 연관 평가값을 가지는 후보 속성 집합을 최고 후보 속성 집합으로 결정한다(S151). 후보 속성 집합 중 최고 후보 속성 집합을 제외한 다른 후보 속성 집합에 대해서는 먼저, 최고 후보 속성 집합을 구성하는 속성들과 서로 공통되는 속성이 포함되어 있는지를 판단하여 최고 후보 속성 집합과 서로 공통되는 속성이 포함되어 있는 경우 A단계(임시 속성 집합으로 분리 단계)로 진행한다(S153). 후보 속성 집합 중 최고 후보 속성 집합을 제외한 다른 후보 속성 집합이 최고 후보 속성 집합과 공통되는 속성이 포함되어 있지 않은 경우 다른 후보 속성 집합의 널 비율이 임계 널 비율보다 작은지 판단한다(S155). 널 비율은 속성 집합으로 데이터테이블을 생성하는 경우 데이터테이블에서 공백 비율, 즉 널의 비율을 의미하는데, 속성 집합을 구성하는 객체별 속성 중 속성값이 존재하지 않는 비율로 계산된다. 도 8은 속성 집합의 데이터테이블의 일 예를 도시하고 있는데, 도 8에 도시되어 있는 바와 같이 객체(사람1 내지 사람6)에 대해 3개의 속성, 즉 이름(p1), 주소(p2) 및 성별(p3)에 대해 속성값이 존재하는 경우와 존재하지 않는 경우 데이터테이블을 생성하면 이름(p1)에 대해서는 사람1, 사람3, 사람4, 사람6은 속성값을 가지는 반면, 사람2, 사람5는 속성값을 가지지 않는다. 따라서 이름(p1)에 대한 널 비율은 $33\%(2/6 \times 100)$ 로 계산된다. 이와 같이 주소(p2)에 대한 널 비율은 16.7%, 성별(p3)에 대한 널 비율은 50%로 도 8에 도시되어 있는 데이터테이블의 전체 널 비율은 33%로 계산된다.

[0047] 다시 도 7을 참고로 살펴보면, 다른 후보 속성 집합의 널 비율이 임계 널 비율보다 큰 경우에는 A 단계(임시 속성 집합으로 분리 단계)로 진행하며, 다른 후보 속성 집합의 널 비율이 임계 널 비율보다 작은 경우 다른 후보 속성 집합을 최고 후보 속성 집합과 함께 선택 속성 집합으로 선택한다(S157).

[0048] 도 9는 본 발명에서 갱신 속성 집합을 선택하는 단계의 일 예를 설명하기 위한 흐름도이다.

[0049] 도 9를 참고로 보다 구체적으로 살펴보면, 후보 속성 집합 중 최고 후보 속성 집합과 공통된 속성을 가져 임시 속성 집합으로 분리되는 경우, 후보 속성 집합에서 최고 후보 속성 집합과 공통된 속성을 삭제한다(S171). 한편, 후보 속성 집합 중 널 비율이 임계 널 비율보다 커 임시 속성 집합으로 분리되는 경우 널 비율이 높은 속성의 순서대로 널 비율이 가장 높은 속성을 삭제하고(S173), 널 비율이 가장 높은 속성을 삭제한 후보 속성 집합의 널 비율이 임시 널 비율보다 작은지 다시 판단한다(S174). 널 비율이 가장 높은 속성을 삭제한 후보 속성 집합의 널 비율이 임시 널 비율보다 큰 경우 다음 순서로 높은 널 비율의 속성을 순차적으로 삭제하여 널 비율이 임계 널 비율보다 작은 최종 임시 속성 집합을 생성한다.

[0050] 최종 임시 속성 집합의 연관 평가값을 계산하여(S175), 연관 평가값이 가장 높은 최종 임시 속성 집합을 갱신 속성 집합으로 선택한다(S177). 최종 임시 속성 집합을 구성하는 속성 중 갱신 속성 집합과 공통되지 않는 속성이 존재하는지 판단하여(S179) 최종 임시 속성 집합을 구성하는 속성 중 갱신 속성 집합과 공통되지 않는 속성이 존재하는 경우, 앞서 설명한 S171, S175, S177, S179를 반복한다.

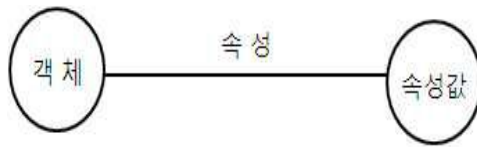
- [0051] 도 10 내지 도 12는 본 발명에 따라 모두 7개의 속성(p1, p2, p3, p4, p5, p6, p7)들 중 적어도 1개 이상의 속성을 가지는 객체들로 이루어진 데이터베이스로부터 데이터테이블을 생성하는 방법의 일 예를 설명하기 위한 도면이다.
- [0052] 먼저 도 10을 참고로 살펴보면, 그래프 형식으로 표현되는 RDF 데이터들이 저장되어 있는 데이터베이스를 마이닝하여 최소 빈발 지지도 이상의 빈발 지지도를 가지는 빈발 속성 집합을 생성하며, 빈발 속성 집합의 연관 평가값이 임계 연관 평가값 이상을 가지는 후보 속성 집합을 생성한다. 각 후보 속성 집합의 길이와 빈발 지지도에 기초하여 각 후보 속성 집합의 연관 평가값을 계산하여 가장 높은 연관 평가값(0.9)을 가지는 최고 후보 속성 집합({p2, p4, p5})을 선택한다. 최고 후보 속성 집합은 선택 속성 집합으로 선택된다. 여기서 p3는 빈발 속성 집합에 포함되지 않는 속성으로 빈발 속성 집합에 포함되지 않는 속성에 대한 단일 속성 집합을 생성한다.
- [0053] 다음으로 도 11을 참고로 살펴보면, 최고 후보 속성 집합을 제외한 다른 후보 속성 집합에서 최고 후보 속성 집합과 공통된 속성을 제외하여 최종 임시 속성 집합({p7}, {p1, p6}, {p1, p6, p7})을 생성한다. 여기서 설명의 간소화를 위하여 최고 후보 속성 집합을 제외한 다른 후보 속성 집합의 널 비율은 임계 널 비율보다 작은 것으로 가정한다.
- [0054] 최종 임시 속성 집합의 지지도와 길이에 기초하여 연관 평가값을 계산하고 가장 높은 연관 평가값을 가지는 최종 임시 속성 집합({p1, p6, p7})을 갱신 속성 집합으로 선택한다.
- [0055] 최종 임시 속성 집합을 구성하는 속성들 중 갱신 속성 집합의 속성과 공통되지 않은 속성이 존재하는지 판단하여, {p7}, {p1, p6}을 구성하는 속성들은 모두 갱신 속성 집합의 속성과 공통되므로 더 이상 갱신 속성 집합을 생성하지 않는다.
- [0056] 도 12의 (a), (b) 및 (c)는 선택 속성 집합, 단일 속성 집합 및 갱신 속성 집합으로 생성되는 데이터테이블의 일 예를 도시하고 있다. 7개의 속성으로 생성한 빈발 속성 집합 중 빈발 속성 집합의 연관 평가값에 기초하여 서로 연관도가 높은 속성들을 1개의 데이터테이블로 작성하며 동시에 서로 생성되는 데이터테이블에 속성값이 서로 중복되지 않도록 데이터테이블을 생성한다.
- [0057] 한편, 상술한 본 발명의 실시 예들은 컴퓨터에서 실행될 수 있는 프로그램으로 작성 가능하고, 컴퓨터로 읽을 수 있는 기록 매체를 이용하여 상기 프로그램을 동작시키는 범용 디지털 컴퓨터에서 구현될 수 있다.
- [0058] 상기 컴퓨터로 읽을 수 있는 기록 매체는 마그네틱 저장 매체(예를 들어, 롬, 플로피 디스크, 하드디스크 등), 광학적 판독 매체(예를 들면, 시디롬, 디브이디 등) 및 캐리어 웨이브(예를 들면, 인터넷을 통한 전송)와 같은 저장 매체를 포함한다.
- [0059] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

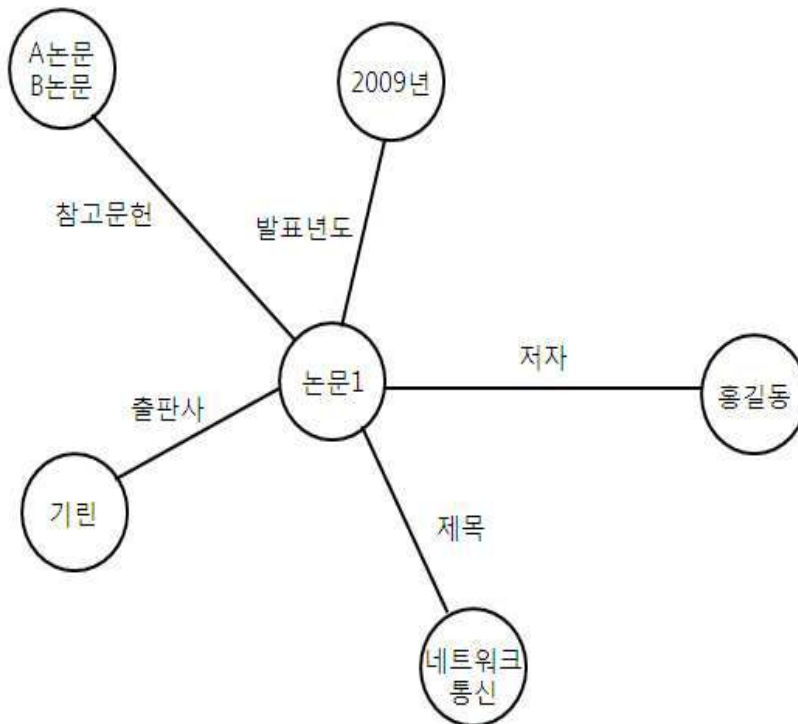
- | | |
|---------------------|-------------------|
| [0060] 110: 데이터베이스부 | 120: 마이닝부 |
| 130: 길이 검색부 | 140: 평가값 계산부 |
| 150: 후보 속성 집합 생성부 | 160: 선택부 |
| 165: 단일 속성 집합 생성부 | 170: 임시 속성 집합 생성부 |
| 180: 갱신 속성 집합 생성부 | 190: 데이터 생성부 |

도면

도면1



도면2



도면3

논문 식별자	저자	제목	발행년도
논문1	A-1	B-1	-
논문2	A-2	-	-
논문3	A-3	B-3	-
논문4	-	B-4	C-4
⋮	⋮	⋮	⋮
논문n	A-n	B-n	C-n

(a)

논문 식별자	저자	출판사	인용문헌
논문1	A-1	D-1	-
논문2	A-2	D-2	E-2
논문3	A-3	-	-
논문4	-	D-4	E-4
⋮	⋮	⋮	⋮
논문n	A-n	D-n	E-n

(b)

도면4

논문 식별자	저자	논문 식별자	제목	논문 식별자	발행년도
논문1	A-1	논문1	B-1	논문1	-
논문2	A-2	논문2	-	논문2	-
논문3	A-3	논문3	B-3	논문3	-
논문4	-	논문4	B-4	논문4	C-4
⋮	⋮	⋮	⋮	⋮	⋮
논문n	A-n	논문n	B-n	논문n	C-n

(a)

(b)

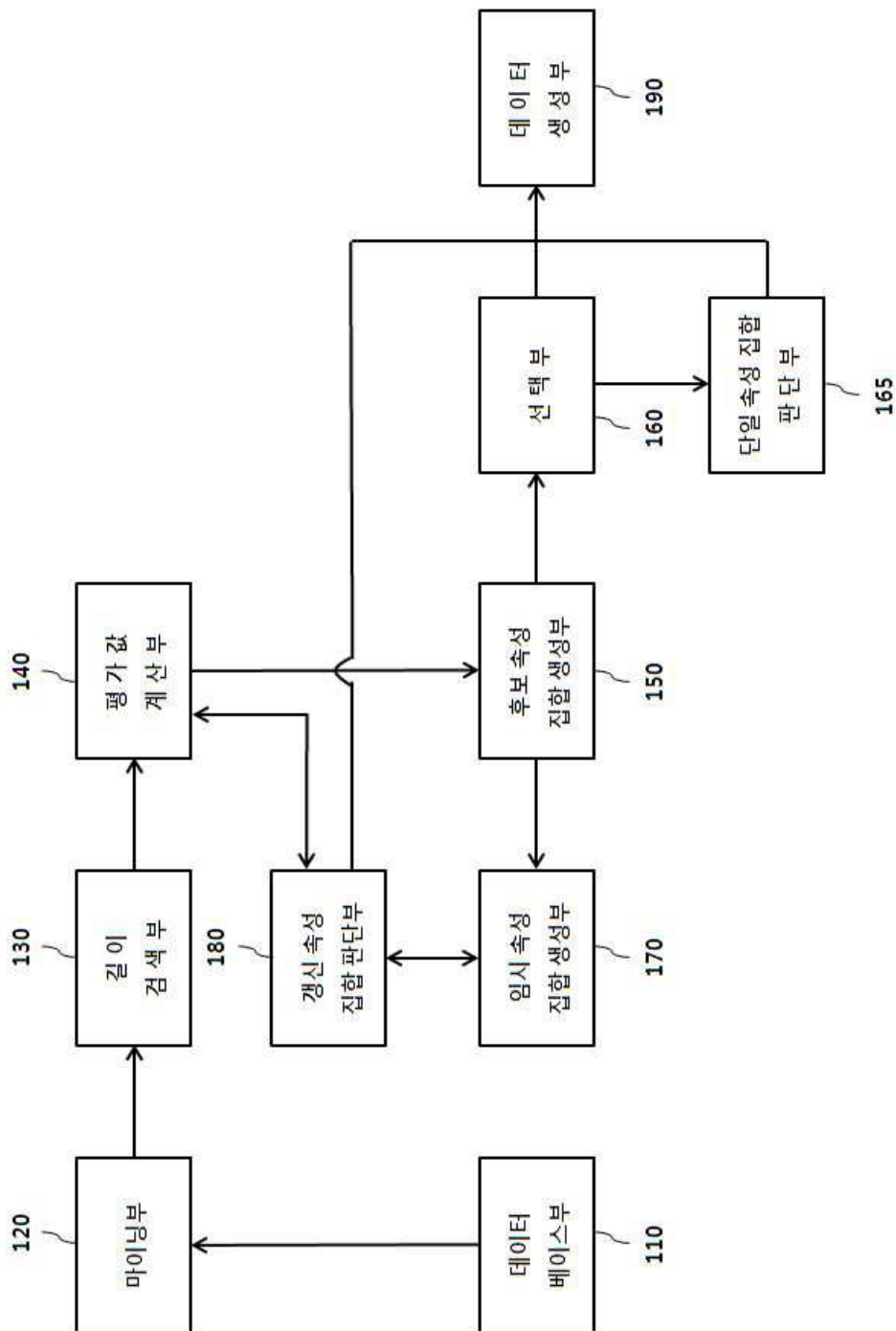
(c)

논문 식별자	출판사	논문 식별자	인용문헌
논문1	D-1	논문1	-
논문2	D-2	논문2	E-2
논문3	-	논문3	-
논문4	D-4	논문4	E-4
⋮	⋮	⋮	⋮
논문n	D-n	논문n	E-n

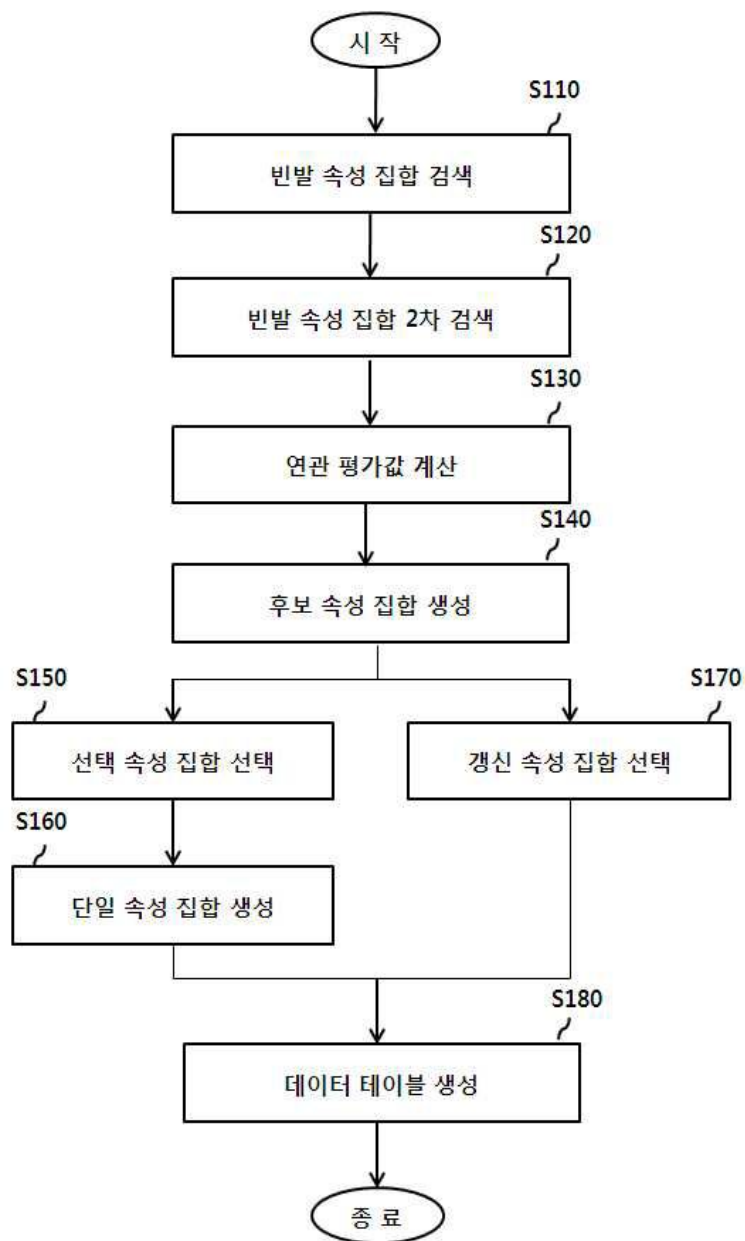
(d)

(e)

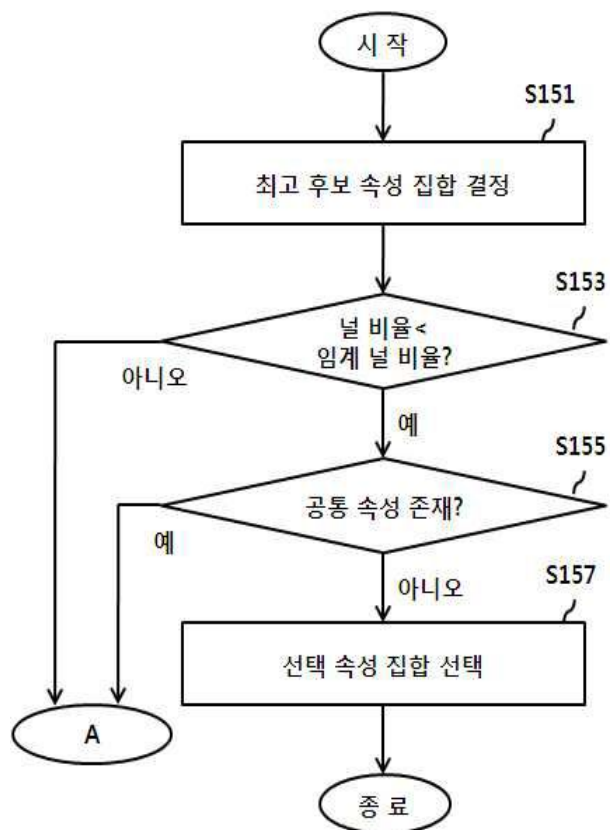
도면5



도면6



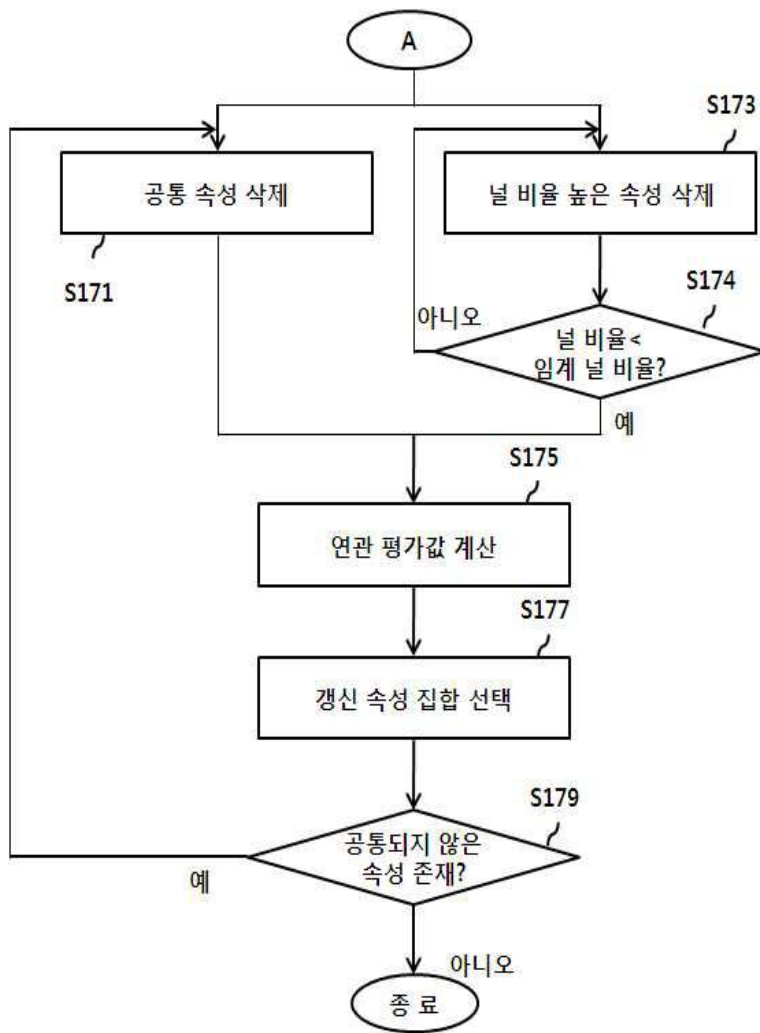
도면7



도면8

객체	이름(p1)	주소(p2)	성별(p3)
사람1	A-1	-	-
사람2	-	-	남
사람3	A-3	-	-
사람4	A-4	B-4	남
사람5	-	-	-
사람6	A-6	-	여

도면9



도면10

후보 속성 집합	지지도	길이	연관 평가값
{p2, p4, p5}	30%	3	0.9
{p2, p4, p5, p7}	10%	4	0.4
{p1, p2, p6}	25%	3	0.75
{p1, p2, p6, p7}	20%	4	0.8

도면11

임시 속성 집합	지지도	길이	연관 평가값
{p7}	40%	1	0.4
{p1, p6}	30%	2	0.6
{p1, p6, p7}	25%	3	0.75

도면12

객체	p2	p4	p5
1	A-1	B-1	-
2	-	B-2	C-2
3	A-3	B-3	C-3
4	A-4	B-4	C-4
⋮	⋮	⋮	⋮
n	A-n	B-n	C-n

(a)

객체	p3
1	D-1
2	D-2
3	D-3
4	D-4
m	D-m

(b)

객체	p1	p6	p7
1	E-1	F-1	-
2	E-2	-	G-2
3	-	F-3	G-3
4	E-4	F-4	-
	E-5	F-5	G-5
l	E-l	F-l	G-l

(c)

도면13

