(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2012년01월30일
(11) 등록번호 10-1109225
(24) 등록일자 2012년01월17일

(51) Int. Cl.
G06F 17/30 (2006.01)
(21) 출원번호 10-2005-0040134
(22) 출원일자 2005년05월13일
심사청구일자 2010년05월07일
(65) 공개번호 10-2006-0047885
(43) 공개일자 2006년05월18일
(30) 우선권주장
10/846,396 2004년05월14일 미국(US)
(56) 선행기술조사문헌
JP05054083 A
JP2002207655 A
전체 청구항 수 : 총 16 항

(73) 특허권자
마이크로소프트 코포레이션
미국 워싱턴주 (우편번호 : 98052) 레드몬드 원 마이크로소프트 웨이
(72) 발명자
웬, 지-롱
미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내
마, 웨이-잉
미국 98052 워싱턴주 레드몬드 원 마이크로소프트 웨이마이크로소프트 코포레이션 내
(74) 대리인
제일특허법인

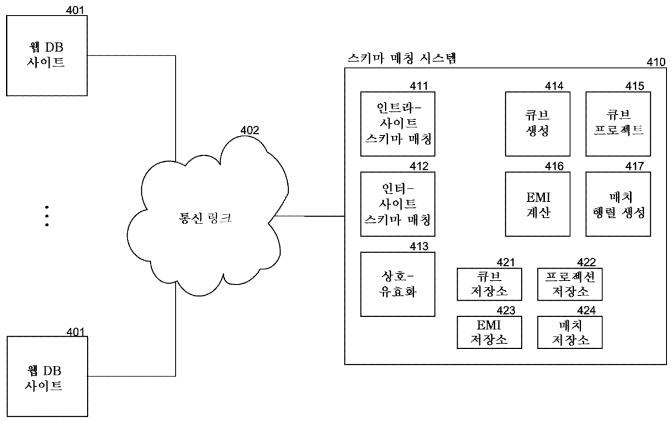
심사관 : 이명진

(54) 웹 데이타베이스의 스키마 매칭을 위한 방법 및 시스템

(57) 요 약

웹 데이타베이스의 스키마들을 식별하는 방법 및 시스템이 제공된다. 스키마 매칭 시스템은 기본적 데이타베이스 스키마를 나타내기 위해 사용되는 웹 데이타베이스의 인터페이스 스키마와 결과 스키마 간의 매핑을 생성한다. 스키마 매칭 시스템은 또한 웹 데이타베이스의 인터페이스 속성들과 결과 속성들을, 의미가 알려진 글로벌 스키마의 글로벌 속성들로 매핑을 생성한다. 이들 매핑들을 사용하여, 검색 엔진 서비스는 글로벌 속성들을 사용하여 쿼리들(queries)을 생성할 수 있고, 이들 쿼리들을 대응하는 인터페이스 속성들에 매핑할 수 있고, 쿼리를 제출할 수 있고, 및 원하는 글로벌 속성들에 대응하는 결과 속성들로부터의 값들을 검색할 수 있다.

대 표 도



특허청구의 범위

청구항 1

컴퓨터 시스템에서 데이타베이스들의 스키마 매칭을 위해 이용되는 어커런스 큐브(occurrence cube)를 생성하는 방법으로서,

데이타베이스의 도메인의 글로벌 속성(global attribute) 각각에 대하여,

상기 데이타베이스의 인터페이스 속성 각각에 대해, 상기 컴퓨터 시스템의 프로세싱 유닛에서, 통신 링크를 통해 상기 데이타베이스에 쿼리들(queries)을 제출하는 단계 - 각각의 쿼리는, 상기 데이타베이스의 인터페이스 속성의 값을 상기 데이타베이스의 도메인의 글로벌 속성의 글로벌 속성 값으로 설정함 -; 및

각각의 제출된 쿼리의 결과 각각에 대하여, 상기 컴퓨터 시스템의 프로세싱 유닛에서, 상기 글로벌 속성의 값이 상기 결과의 각각의 결과 속성(result attribute) 내에서 출현하는 횟수를 카운트하는 단계; 및

글로벌 속성, 인터페이스 속성 및 결과 속성 조합 각각에 대하여, 상기 컴퓨터 시스템의 프로세싱 유닛에서, 상기 인터페이스 속성을 상기 글로벌 속성의 글로벌 속성 값으로 설정하여 제출된 쿼리로부터의 결과에서 각각의 결과 속성 내에 상기 글로벌 속성의 값이 출현하는 횟수의 카운트들의 축적(accumulation)을 상기 어커런스 큐브의 요소로서 메모리 디바이스에 저장하는 단계

를 포함하며,

저장된 요소들은 상기 어커런스 큐브를 형성하는 방법.

청구항 2

제1항에 있어서, 상기 컴퓨터 시스템의 프로세싱 유닛에서, 상기 어커런스 큐브로부터 글로벌 속성들 및 인터페이스 속성들과 연관된 어커런스 행렬(occurrence matrix)을 생성하는 단계를 포함하는 방법.

청구항 3

제1항에 있어서, 상기 컴퓨터 시스템의 프로세싱 유닛에서, 상기 어커런스 큐브로부터 글로벌 속성들 및 결과 속성들과 연관된 어커런스 행렬을 생성하는 단계를 포함하는 방법.

청구항 4

제1항에 있어서, 상기 컴퓨터 시스템의 프로세싱 유닛에서, 상기 어커런스 큐브로부터 인터페이스 속성들 및 결과 속성들과 연관된 어커런스 행렬을 생성하는 단계를 포함하는 방법.

청구항 5

제1항에 있어서, 글로벌 속성 값과 인터페이스 속성의 각각의 조합에 대해 쿼리가 제출되는 방법.

청구항 6

제1항에 있어서, 상기 어커런스 큐브는, 글로벌 속성, 인터페이스 속성 및 결과 속성 조합 각각에 대한 카운트를 포함하는 방법.

청구항 7

컴퓨터 시스템에서 데이타베이스들의 스키마 매칭을 위해 도메인 내의 데이타베이스의 속성들을 식별하는 방법으로서,

상기 컴퓨터 시스템의 프로세싱 유닛에서, 상기 도메인의 글로벌 스키마의 글로벌 속성들 및 상기 데이타베이스의 인터페이스 스키마의 인터페이스 속성들과 결과 스키마의 결과 속성들과 연관된 어커런스들의 카운트들을 메모리 디바이스에 저장함으로써 제공하는 단계 - 각각의 카운트는, 글로벌 속성, 인터페이스 속성 및 결과 속성 조합 각각에 대하여, 상기 인터페이스 속성을 글로벌 속성 값으로 설정하여 상기 데이타베이스에 제출되는 쿼리의 결과에서 상기 글로벌 속성에 대한 상기 글로벌 속성 값이 상기 결과 속성의 값으로서 출현하는 어커런스들

의 수를 나타냄 -;

상기 컴퓨터 시스템의 프로세싱 유닛에서, 상기 제공된 카운트들에 기초하여 스키마들의 쌍들 간에 상호 정보를 추정하는 단계;

상기 컴퓨터 시스템의 프로세싱 유닛에서, 상기 추정된 상호 정보로부터 어느 속성들이 매칭하는지를 식별하는 단계; 및

상기 컴퓨터 시스템의 프로세싱 유닛에서, 상기 매칭하는 속성들의 표시를 상기 메모리 디바이스에 저장하는 단계

를 포함하는 방법.

청구항 8

제7항에 있어서, 상기 카운트들을 제공하는 단계는, 글로벌 속성들, 인터페이스 속성들 및 결과 속성들과 연관된 어커런스들의 카운트를 스키마들의 쌍들과 연관된 행렬로 제공하는 어커런스 큐브를 프로젝트하는 단계를 포함하는 방법.

청구항 9

제8항에 있어서, 상기 컴퓨터 시스템의 프로세싱 유닛에서, 상기 인터페이스의 속성들의 값들을 상기 글로벌 속성들의 글로벌 속성 값들로 설정하여 상기 데이타베이스에 쿼리들을 제출함으로써 상기 어커런스 큐브를 생성하는 단계를 포함하는 방법.

청구항 10

제9항에 있어서, 상기 어커런스 큐브 내의 상기 어커런스들의 카운트는 쿼리의 인터페이스 속성의 값으로서 사용되는 글로벌 속성의 글로벌 속성 값들이 상기 쿼리의 결과의 결과 속성에 나타나는 횟수를 나타내는 방법.

청구항 11

제7항에 있어서, 상기 인터페이스 속성들은 HTML 입력-관련 요소들에 기초하여 식별되는 방법.

청구항 12

제7항에 있어서, 상기 결과 속성들은 정규 표현 래퍼(regular expression wrapper)를 사용하여 식별되는 방법.

청구항 13

제7항에 있어서, 상기 어커런스들의 카운트들은 인터페이스 속성들의 값들을 상기 글로벌 속성들의 글로벌 속성 값들로 설정하여 상기 데이타베이스에 쿼리들을 제출함으로써 제공되는 방법.

청구항 14

제7항에 있어서, 상기 상호 정보는 다음 식:

$$EMI(S_{1i},S_{2j})=\frac{m_{ij}}{M}\log\frac{\frac{m_{ij}}{M}}{\frac{m_{i+}}{M}*\frac{m_{+j}}{M}}$$

에 의해 추정되어 산출되는 방법.

청구항 15

제7항에 있어서, 스키마들의 쌍의 속성들 간의 매칭은, 다른 스키마의 한 속성에 대해 가장 높은 추정된 상호 정보를 갖는 한 스키마의 속성이 상기 다른 스키마의 다른 속성에 대해 더 높은 추정된 상호 정보를 갖지 않을 때 식별되는 방법.

청구항 16

인트라-사이트(intra-site) 컴포넌트, 인터-사이트(inter-site) 컴포넌트 및 상호-유효화(cross-validate) 컴포넌트를 포함하며, 통신 링크를 통해 웹 데이타베이스 사이트에 접속되는 스키마 매칭 시스템으로서,

상기 인트라-사이트 컴포넌트는,

어커런스 큐브를 생성하는 수단;

상기 어커런스 큐브에 기초하여, 글로벌-대-인터페이스, 글로벌-대-결과 및 인터페이스-대-결과 어커런스 행렬들을 생성하는 수단;

상기 어커런스 행렬들에 기초하여 추정된 벡터 유사성을 계산하는 수단; 및

상기 추정된 벡터 유사성에 기초하여 속성들의 어느 쌍들이 매칭하는지를 식별하는 수단

을 포함하고,

상기 인터-사이트 컴포넌트는,

상기 어커런스 행렬들을 이용하여 추정된 벡터 유사성을 계산하는 수단; 및

상기 추정된 벡터 유사성에 기초하여 속성들의 어느 쌍들이 매칭하는지를 식별하는 수단

을 포함하며,

상기 상호-유효화 컴포넌트는, 부정확하게 매칭된 것으로 나타나는 속성들에 대한 매칭을 변경하는 수단을 포함하는 스키마 매칭 시스템.

청구항 17

삭제

청구항 18

삭제

청구항 19

삭제

청구항 20

삭제

청구항 21

삭제

청구항 22

삭제

청구항 23

삭제

청구항 24

삭제

청구항 25

삭제

청구항 26

삭제

청구항 27

삭제

청구항 28

삭제

청구항 29

삭제

청구항 30

삭제

청구항 31

삭제

청구항 32

삭제

청구항 33

삭제

청구항 34

삭제

청구항 35

삭제

청구항 36

삭제

청구항 37

삭제

청구항 38

삭제

명 세 서

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

[0029] 본 기재의 기술은 일반적으로 웹 데이타베이스의 스키마를 결정하는 것에 관한 것이다.

[0030] WWW(World Wide Web)("웹")은 웹 페이지들을 통해 액세스가능한 대량의 정보를 제공한다. 웹 페이지들은 정적 컨텐트 또는 동적 컨텐트를 포함할 수 있다. 정적 컨텐트는 일반적으로 웹 페이지들의 많은 액세스에서 동일하게 남아있는 정보를 일컫는다. 동적 컨텐트는 일반적으로 웹 데이타베이스에 저장되어 있어서 검색 요구에 응

답하여 웹 페이지에 추가되는 정보를 일컫는다. 동적 컨텐트는 딥 웹(deep web) 또는 히든 웹(hidden web)으로서 지칭되는 것들을 나타낸다.

[0031] 다수의 검색 엔진 서비스들은 사용자가 웹의 정적 컨텐트를 검색할 수 있게 한다. 사용자가 검색 용어들을 포함하는 검색 요구나 쿼리(query)를 제출한 후에, 검색 엔진 서비스는 그 검색 용어들에 관련될 수 있는 웹 페이지들을 식별한다. 이들 웹 페이지들은 검색 결과이다. 관련 웹 페이지들을 신속하게 식별하기 위해,검색 엔진 서비스들은 키워드들을 웹 페이지들에 매핑하는 것을 계속할 것이다. 이 매핑은 웹을 "크롤링(crawling)"하여 생성되어 각각의 웹 페이지의 키워드들을 식별할 수 있다. 웹을 크롤링하기 위해, 검색 엔진 서비스는 루트(root) 웹 페이지들의 리스트를 사용하여 이들 루트 웹 페이지들을 통해 액세스가능한 모든 웹 페이지들을 식별할 수 있다. 임의의 특정 웹 페이지의 키워드들은, 헤드라인의 워드들, 웹 페이지의 메타데이타에 제공된 워드들, 하이라이트된 워드들 등을 식별하는 것과 같은, 다양한 잘 공지된 정보 검색 기술을 사용하여 식별될 수 있다.

[0032] 그러나, 이들 검색 엔진 서비스들은 일반적으로 또한 크롤링할 수 없는 컨텐트으로 고려되는 동적 컨텐트의 검색을 제공하지 않는다. 동적 컨텐트의 검색에서의 한 가지 문제점은 웹 데이타베이스를 제공하는 웹 사이트의 협조없이 대응하는 웹 데이타베이스들의 스키마들을 직접적으로 얻는 것은 어렵거나 또는 불가능하다는 점이다. 스키마는 데이타베이스에 저장된 정보 또는 속성들을 정의한다. 예를 들어, 서적상을 위한 웹 데이타베이스는 각각의 책에 대한 표제 속성 및 저자 속성을 포함하는 책들의 그 자신의 카탈로그(즉, 웹 데이타베이스)를 위한 스키마를 가질 수 있다. 그 스키마를 알지 않고는, 검색 엔진 서비스가 웹 데이타베이스의 컨텐트를 크롤링하여 검색을 위해 무슨 정보가 이용가능한지를 결정하기가 매우 어려울 것이다. 웹 데이타베이스의 스키마가 알려졌다고 해도, 검색 엔진 서비스는 여전히 웹 데이타베이스의 컨텐트를 수취하기 위해 웹 데이타베이스를 크롤링하는 방법을 결정할 필요가 있다. 검색 엔진이 웹 데이타베이스들의 컨텐트를 검색할 수 있다고 가정해도, 검색 엔진 서비스는 여전히 언제 다른 스키마들의 속성들이 의미상으로 동격인 속성들을 나타내는지를 식별할 필요가 있을 것이다. 예를 들어, 서적상 웹 사이트들은 책이 종이 표지인지, 두꺼운 표지인지, 또는 컴팩트 디스크인지를 명시하는 카탈로그들을 가지고 있을 수 있다. 한 서적상의 웹 사이트는 이 속성을 "유형(type)"으로 명명할 수 있고, 다른 서적상의 웹 사이트는 동일한 속성을 "형식(format)"으로 명명할 수 있다. 복수 웹 사이트들에서 동적 컨텐트를 효과적으로 검색하도록 하기 위해, 검색 엔진 서비스는 웹 데이타베이스들의 속성들의 뜻 또는 의미를 아는 것이 필요하다.

[0033] 웹 데이타베이스들과 연관된 스키마들을 자동으로 식별하는 기술을 구비하고, 동일한 의미적 컨텐트를 나타내는 상이한 스키마들의 속성들을 식별하는 것은 바람직할 것이다.

발명이 이루고자 하는 기술적 과제

[0034] 웹 데이타베이스들의 스키마들을 식별하는 방법 및 시스템이 제공된다. 스키마 매칭 시스템은 기본적 데이타베이스 스키마를 나타내기 위해 사용되는 웹 데이타베이스의 인터페이스 스키마와 결과 스키마 간의 매핑을 생성한다. 스키마 매칭 시스템은 또한 웹 데이타베이스의 인터페이스 속성들과 결과 속성들을, 의미가 알려진 글로벌 스키마의 글로벌 속성들에 매핑을 생성한다. 이들 매핑들을 사용하여, 검색 엔진 서비스는 글로벌 속성들을 사용하여 쿼리들을 생성할 수 있고, 이들 쿼리들을 대응하는 인터페이스 속성들에 매핑할 수 있고, 쿼리를 제출할 수 있고, 원하는 글로벌 속성들에 대응하는 결과 속성들로부터의 값들을 검색할 수 있다.

발명의 구성 및 작용

[0035] 웹 데이타베이스들의 스키마들을 식별하는 방법 및 시스템이 제공된다. 일 실시예에서, 스키마 매칭 시스템은 기본적(underlying) 데이타베이스 스키마를 표현하기 위해 사용되는 웹 데이타베이스의 인터페이스 스키마와 결과 스키마 간의 매핑을 생성한다. 웹 데이타베이스의 인터페이스 스키마는 검색을 위해 사용될 수 있는 데이타베이스의 속성들을 나타낸다. 웹 데이타베이스의 결과 스키마는 검색 결과의 일부로서 디스플레이되는 데이타베이스의 속성들을 나타낸다. 매핑은 어느 인터페이스 속성이 어느 결과 속성과 동일한 의미를 갖는지를(또한 대응하거나 매치하는 것으로서 지칭되는지를) 나타낸다. 스키마 매칭 시스템은 또한 웹 데이타베이스의 인터페이스 속성들과 결과 속성들을 의미가 알려진 글로벌 스키마의 글로벌 속성들로의 매핑을 생성한다. 이들 매핑들을 사용하여, 검색 엔진 서비스는 글로벌 속성들을 사용하여 쿼리들을 공식화할 수 있고, 이들 쿼리들을 대응하는 인터페이스 속성들에 매핑할 수 있고, 쿼리를 제출할 수 있고, 및 원하는 글로벌 속성들에 대응하는 결과 속성들로부터 값들을 검색할 수 있다. 이 방식으로, 스키마 매칭 시스템은 웹 데이타베이스를 검색하기 위해 사용될 수 있는 웹 데이타베이스의 스키마들을 식별한다.

[0036] 도 1은 서적상을 위한 웹 데이타베이스의 다양한 스키마들을 도시하는 도면이다. 웹 데이타베이스는 데이타베이스 스키마(101), 인터페이스 스키마(102), 및 결과 스키마(103)를 포함한다. 데이타베이스 스키마는, 이 예에서 표제, 저자, 발행인, ISBN, 형식, 및 발행일 속성들을 포함하는, 웹 데이타베이스의 기본적 스키마를 나타낸다. 웹 사이트는 검색 웹 페이지를 제공하여, 사용자가 책들을 검색하기 위해 방문할 수 있다. 이 웹 데이타베이스를 위한 인터페이스 스키마는 표제, 저자, 형식, 및 ISBN 속성들을 포함한다. 사용자는 책 데이타베이스의 검색을 위한 인터페이스 속성들의 임의의 조합을 위한 검색 문자열들을 명시할 수 있다. 웹 페이지의 "당신의 검색(your search)"이라는 필드는 사용자가 웹 데이타베이스의 모든 속성들 내에서 검색할 수 있게 한다. 검색 결과는 결과 웹 페이지에 디스플레이된다. 이 웹 데이타베이스를 위한 결과 스키마는 표제, 저자, 발행인, 형식, 및 발행일을 포함한다. 검색 결과는 통상적으로 검색 요구와 매치하는 데이타베이스의 각 엔트리에 대한 복수 개의 엔트리들을 제공할 것이다. 결과의 각 엔트리는 통상적으로 결과 속성들의 각각에 대한 값을 포함한다. 이 예에서, 인터페이스 스키마는 결과 스키마에 포함되지 않은 속성(즉, ISBN)을 가지며, 결과 스키마는 인터페이스 스키마에 포함되지 않는 속성(즉, 발행일)을 가진다.

[0037] 웹 데이타베이스를 위한 인터페이스 스키마와 결과 스키마의 사용에 추가하여, 스키마 매칭 시스템은 또한 도메인 특정(domain-specific) 글로벌 스키마를 사용한다. 도메인을 위한 글로벌 스키마는 그 도메인 내의 웹 데이타베이스들에 의해 일반적으로 사용되는 일련의 속성들을 나타낸다. 예를 들어, 책 도메인 내의 웹 데이타베이스들은 통상적으로 표제, 저자, 및 발행인을 포함하는 속성들을 가지며, 자동차 도메인 내의 웹 데이타베이스들은 통상적으로 제조자, 모델, 및 제조년도를 포함하는 속성들을 가진다. 글로벌 스키마는 또한 그것과 연관된 샘플 글로벌 속성 값들을 가질 수 있다. 예를 들어, 책 도메인의 발행인 속성은 "랜덤 하우스"와 "MIT 출판사"를 포함하는 글로벌 속성 값들을 가질 수 있다.

[0038] 매핑들을 생성하기 위해, 스키마 매칭 시스템은 초기에 도메인에 대한 웹 데이타베이스의 글로벌 스키마 및 웹 데이타베이스의 인터페이스 스키마 및 결과 스키마를 식별한다. (이들 스키마들을 식별하는 기술들은 아래 기재된다.) 스키마 매칭 시스템은 글로벌 속성들의 글로벌 속성 값들로부터(예를 들어, 값들의 샘플 세트로부터) 쿼리들을 생성하고, 이들 쿼리들을 인터페이스 웹 페이지를 통해 웹 데이타베이스에 제출한다(예를 들어, 검색 웹 페이지를 통해 쿼리의 제출에 대응하는 HTTP 요구를 전송함). 스키마 매칭 시스템은 결과 웹 페이지에 의해 제공된 결과를 분석하여, 어느 인터페이스 속성들이 어느 결과 속성들에 대응하는지("인터페이스-대-결과 대응"), 어느 글로벌 속성들이 어느 인터페이스 속성들에 대응하는지("글로벌-대-인터페이스 대응"), 및 어느 글로벌 속성들이 어느 결과 속성들에 대응하는지("글로벌-대-결과 대응")를 결정한다. 인터페이스 및 결과 스키마들은 한 개의 웹 사이트의 스키마들에 대응하므로, 이들 대응들은 "인트라-사이트(intra-site)" 매칭으로서 불리운다. 스키마 매칭 시스템은 결과 속성의 값이 검색시에 사용되는 인터페이스 속성의 값과 매치할 때에 기초하여 인터페이스 속성이 결과 속성에 대응할 수 있음을 식별한다. 예를 들어, 표제 인터페이스 속성에 "해리 포터"의 값이 주어질 때, 결과의 많은 엔트리들은 표제 결과 속성에 "해리 포터"의 값을 가질 것이다. 반면, 저자 인터페이스 속성에 검색을 위해 "해리 포터"의 값이 주어지면, 결과의 단지 몇 개의 엔트리들만이 표제 인터페이스 속성에 "해리 포터"의 값을 가질 것이다. 그렇게 해서, 표제 인터페이스 속성은 표제 결과 속성에 대응할 수 있지만, 저자 인터페이스 속성은 표제 결과 속성에 대응할 수 없을 것이다.

[0039] 일 실시예에서, 스키마 매칭 시스템은 또한 상이한 웹 사이트들의 인터페이스 스키마들과 결과 스키마들 간의 매핑들을 생성할 것이다. 스키마 매칭 시스템은 상술된 바와 같이 제출된 쿼리들의 결과들을 분석하여, 한 웹 사이트 스키마의 어느 인터페이스 속성이 다른 웹 사이트 스키마의 어느 인터페이스 속성과 대응하는지("인터페이스-대-인터페이스 대응"), 및 한 웹 사이트 스키마의 어느 결과 속성이 다른 웹 사이트 스키마의 어느 결과 속성과 대응하는지("결과-대-결과 대응")를 식별한다. 예를 들어, 스키마 매칭 시스템은 한 웹 사이트의 유형 인터페이스 속성이 다른 웹 사이트의 형식 인터페이스 속성에 대응할 것임을 식별할 수 있다. 스키마들이 상이한 웹 사이트들 간에 매치되므로, 이들 대응들은 "인터-사이트(inter-site)" 매칭으로서 불리운다. 인터-사이트 매칭 정보는 도메인 내의 복수 개의 웹 데이타베이스들을 검색할 때 사용될 수 있다. 인터-사이트 매칭 정보는 또한 인트라-사이트 매칭이 정확한지를 유효화하는 것을 돕기 위해 사용될 수 있다.

[0040] 도 2는 일 실시예에서 인트라-사이트 및 인터-사이트 매칭을 도시한다. 타원웹 데이타베이스 도메인(202)은 책 도메인 내의 웹 데이타베이스들과 관련된 스키마들을 나타낸다. 각각의 사이트 1...N은 인터페이스 스키마("IS") 및 결과 스키마("RS")를 가지며, 도메인은 글로벌 스키마("GS")를 가진다. 스키마들의 표현들 사이의 라인들은 인트라-사이트 및 인터-사이트 매칭을 나타낸다. 예를 들어, 사이트 1의 IS와 GS 사이의 라인은 인트라-사이트 글로벌-대-인터페이스 대응을 나타내며, 사이트 1의 IS와 사이트 1의 RS 사이의 라인은 인트라-사이트 인터페이스-대-결과 대응을 나타내고, 사이트 1의 IS와 사이트 2의 IS 사이의 라인은 사이트 1과 사이트 2

간의 인터-사이트 인터페이스-대-인터페이스 대응을 나타낸다.

[0041] 일 실시예에서, 스키마 매칭 시스템은, 웹 데이타베이스의 글로벌 속성, 인터페이스 속성, 및 결과 속성의 각각 조합에 대해, 글로벌 속성 값이 검색시에 해당 인터페이스 속성 값으로서 사용될 때 해당 글로벌 속성의 글로벌 속성 값이 그 결과 속성에서 출현하는 횟수를 식별하는 어커런스 큐브(occurrence cube)를 생성한다. 각각의 인터페이스 속성에 대해, 스키마 매칭 시스템은 복수 개의 쿼리들을 제출한다. 각각의 쿼리는 인터페이스 속성 값이 다른 글로벌 속성 값으로 설정되도록 한다. 예를 들어, 글로벌 속성들이 종이 표지, 두꺼운 표지, 및 컴팩트 디스크의 값들을 갖는 형식 속성, 및 롤링의 값을 갖는 저자 속성을 포함한다면, 스키마 매칭 시스템은 표제 속성이 종이 표지로 설정된 쿼리, 표제 속성이 두꺼운 표지로 설정된 쿼리, 표제 속성이 컴팩트 디스크로 설정된 쿼리, 및 표제 속성이 롤링으로 세트된 쿼리를 제출한다. 각각의 다른 인터페이스 속성에 대해, 스키마 매칭 시스템은 종이 표지, 두꺼운 표지, 컴팩트 디스크, 및 롤링의 글로벌 속성 값들에 대한 쿼리들을 제출한다. 각각의 쿼리 결과에 대해, 스키마 매칭 시스템은 쿼리의 글로벌 속성 값이 각각의 결과 속성의 값으로서 출현하는 횟수를 카운트한다. 예를 들어, 표제 인터페이스 속성이 종이 표지로 설정된 쿼리가 제출될 때, 매칭은 극소수만이 발견되거나 전혀 발견되지 않을 수 있고, 이것은 표제 인터페이스 속성이 아마도 형식 글로벌 속성과 매치하지 않음을 나타낸다. 반면, 쿼리가 형식 인터페이스 속성이 종이 표지로 설정되어 제출될 때, 다수의 매치들이 발견되어, 검색 용어 "종이 표지"는 형식 결과 속성 내의 결과의 다수의 엔트리들에서 발견될 수 있고, 이것은 형식 글로벌 속성, 형식 인터페이스 속성, 및 형식 결과 속성이 서로 대응할 수 있음을 나타낸다. 특히 다른 조합들에 상대적으로 특정 글로벌 속성, 인터페이스 속성, 및 결과 속성 조합에 대한 높은 카운트는 이들 속성들이 대응할 수 있음, 즉, 그들은 동일한 의미의 컨텐트를 나타낸다는 것을 나타낼 수 있다.

[0042] 어커런스 큐브를 생성한 후에, 스키마 매칭 시스템은 글로벌-대-인터페이스 대응, 글로벌-대 결과 대응, 및 인터페이스-대 결과 대응을 위한 어커런스 행렬들(occurrence matrices)을 생성한다. 일 실시예에서, 스키마 매칭 시스템은 어커런스 큐브의 차원을 평면에 프로젝트하여 어커런스 행렬을 생성한다. 글로벌-대-인터페이스 대응에 대한 어커런스 행렬을 생성하기 위해, 스키마 매칭 시스템은 각각의 글로벌 속성과 인터페이스 속성의 조합에 대해 모든 결과 속성들에 대한 어커런스 카운트를 합한다. 스키마 매칭 시스템은 유사한 방식으로 글로벌-대-결과 대응과 인터페이스-대-결과 대응을 위한 어커런스 행렬들을 생성한다. 표 1은 글로벌-대-인터페이스 대응을 위한 어커런스 행렬의 예이다.

표 1

	$Title_{GS}$	$Author_{GS}$	$Publisher_{GS}$	$ISBN_{GS}$
$Author_{IS}$	93	498	534	0
$Title_{IS}$	451	345	501	0
$Publisher_{IS}$	62	184	468	2
$Keyword_{IS}$	120	248	143	275
$ISBN_{IS}$	0	0	0	258

[0043]

[0044] 카운트의 크기는 속성들의 쌍들 간의 대응을 나타내지만, 상대적 크기는 절대적 크기보다 매치를 더 잘 나타낸다. 더 구체적으로, 더 높은 어커런스 카운트는 대응하는 속성들을 나타내지 않을 수 있다. 예를 들어, $Author_{IS}$와 $Publisher_{GS}$(534)에 대한 행렬 요소는 그 행렬에서 최고 값이지만, $Author_{IS}$와 $Publisher_{GS}$는 서로 의미적으로 대응하지는 않는다. 일반적으로, 특정 행렬 요소 m_{ij}가 주어지면, 그것의 인터페이스 속성 i와 글로벌 속성 j에 대한 모든 요소들 중의 그것의 상대적 크기는 그것의 절대적 크기보다 더 중요하다. 예를 들어, $Keyword_{IS}$는 "당신의 검색"을 포함할 수 있고 책 도메인에 대한 실제 속성이 아니며, 모든 글로벌 속성들에 대해 유사한 성과를 가지며, 이것은 글로벌 속성들 중의 임의의 것에 대해서도 좋은 매치가 아닐 것임을 나타낸다. $Publisher_{IS}$와 $Publisher_{GS}$(468)의 요소는 $Publisher_{GS}$의 요소들 중에 가장 높은 것이 아니다. 그러나, $Publisher_{IS}$에 대한 그외의 요소들보다는 비교적 크다.

[0045] 속성들 중의 어느 쌍이 대응하는지를 식별하기 위해, 스키마 매칭 시스템은 속성들의 쌍의 상호 정보 컨텐트를 추정한다. 상호 정보는 또한 상호-엔트로피(cross-entropy) 및 정보 이득으로서 지칭된다. 스키마 매칭 시스템은 스키마 속성들에 의해 웹 데이타베이스의 분할을 나타내는 각각의 스키마를 고려한다. 스키마들의 분할들이 가장 많이 겹치는 상이한 스키마들로부터의 속성들의 쌍들은 대응할 수 있다. 일 실시예에서, 스키마 매칭 시스템은 다음식에 따라 속성들의 쌍 간의 상호 정보를 추정한다:

수학식 1

$$EMI(S_{1i},S_{2j})=\frac{m_{ij}}{M}\log\frac{\frac{m_{ij}}{M}}{\frac{m_{i+}}{M}*\frac{m_{+j}}{M}}$$

[0046]

[0047] 여기서, EMI 는 스키마 S_{1i} 의 i 번째 속성과 S_{2j} 의 j 번째 속성 간의 추정된 상호 정보이고, M 은 $\sum_{i,j} m_{ij}$ 이고, m_{i+} 는 $\sum_{i} m_{ij}$ 이고, 및 m_{+j} 는 $\sum_{j} m_{ij}$ 이다. 표 1의 어커런스 행렬의 EMI 행렬은 표 2에 도시된다.

표 2

	$Title_{GS}$	$Author_{GS}$	$Publisher_{GS}$	$ISBN_{GS}$
$Author_{IS}$	-0.04	0.11	0.06	0.00
$Title_{IS}$	0.19	-0.03	-0.01	0.00
$Publisher_{IS}$	-0.03	-0.02	0.14	-0.01
$Keyword_{IS}$	-0.01	0.01	-0.07	0.17
$ISBN_{IS}$	0.00	0.00	0.00	0.32

[0048]

[0049] 스키마 매칭 시스템은, 한 EMI 행렬 요소가 동일 인터페이스 속성에 대한 다른 요소들(즉, 동일 행에서)보다 더 크고, 또한 동일 글로벌 속성에 대한 다른 요소들(즉, 동일 열에서)보다 더 클 때, 속성들 간의 매치를 탐지한다. 대응하는 속성들은 사각형들로 나타낸 바와 같이 그들이 반대쪽 스키마의 다른 속성들과 겹치기보다는 서로 간에 정보 컨텐트에서 더 많이 겹친다. 예를 들어, $Author_{IS}$와 $Author_{GS}$에 대한 EMI 행렬 요소(즉, 0.11)는 저자 인터페이스 속성들과 저자 글로벌 속성들 모두에 대해 가장 큰 것이고, 이것은 정확한 매치이다. 속성들의 매치는 다음식으로 표현된다:

수학식 2

$$MAP(S_{1i},S_{2j}) = \text{match when } e_{ij} \geq e_{ik} \mid k \neq j \text{ and } e_{ij} \geq e_{ik} \mid k \neq i$$

[0050]

[0051] 여기서, MAP은, 스키마 S_1 의 i 번째 속성이 스키마 S_2 의 j 번째 속성과 매치하는지, e_{ij}가 스키마 S_1의 i 번째 속성에 대한 EMI 행렬 요소인지, 스키마 S_2 의 j 번째 속성에 대한 EMI 행렬 요소인지를 나타낸다.

[0052] 일 실시예에서, 스키마 매칭 시스템은 상이한 웹 데이타베이스들의 속성들 간의 매치들을 식별한다. 스키마 매칭 시스템은 웹 데이타베이스들에 대한 대응하는 어커런스 행렬들의 벡터들 간의 유사성에 기초한 매치들을 식별한다. 예를 들어, 표 3은 스키마 S_1 에 대한 글로벌-대-인터페이스 어커런스 행렬을 나타내고, 표 4는 스키마 S_2 에 대한 글로벌-대-인터페이스 어커런스 행렬을 나타낸다. 글로벌 스키마 GS는 {Title, Author, Publisher, ISBN}이고, 사이트 1에 대한 인터페이스 스키마 IS_1은 {$Author_1$, $Title_1$, $Publisher_1$, $Keyword_1$, $ISBN_1$} 이고, 사이트 2에 대한 인터페이스 스키마 IS_2는 {$Title_2$, $Author_2$, $ISBN_2$}이다.

표 3

	T_G	A_G	P_G	I_G
A_1	93	498	534	0
T_1	451	345	501	0
P_1	62	184	468	2
K_1	120	248	143	275
I_1	0	0	0	258

[0053]

표 4

	T_G	A_G	P_G	I_G
T_2	166	177	118	0
$A_2(P)$	39	331	406	0
I_2	0	0	0	18

[0054]

[0055] 속성 A1은 표 3의 제1 행의 벡터에 의해 표현되고, 속성 A2는 표 4의 제2 행의 벡터에 의해 표현된다. 스키마 매칭 시스템은 다음식을 사용하여 2개의 속성들 간의 유사성을 계산한다:

수학식 3

$$EVS(S_{1i}, S_{2j}) = \frac{\sum_k a_{ik} b_{jk}}{\sqrt{\sum_k a_{ik}^2} * \sqrt{\sum_k b_{jk}^2}}$$

[0056]

[0057] 여기서, EVS는 스키마 S_1 의 i 번째 속성과 스키마 S_2 의 j 번째 속성 간의 추정된 벡터 유사성이고, a_{ik}는 스키마 S_1에 대한 어커런스 행렬의 값들을 나타내고, b_{jk} 는 스키마 S_2 에 대한 어커런스 행렬의 값들을 나타낸다.

[0058] 표 5는 표 3과 표 4로부터 유도된 추정된 벡터 유사성들을 나타낸다.

표 5

	T_2	$A_2(P)$	I_2
A_1	0.84	0.99	0
T_1	0.96	0.84	0
P_1	0.71	0.95	0.01
K_1	0.72	0.67	0.66
I_1	0	0	1.00

[0059]

[0060] 스키마 매칭 시스템은 한 개의 EVS 행렬 요소가 한 웹 사이트의 동일 인터페이스 속성의 다른 요소들보다 크고, 또한 다른 웹 사이트의 동일 인터페이스 속성에 대한 다른 요소들보다도 클 때 속성들 간의 매치를 탐지한다. 표 5의 사각형들은 그것의 행과 열 모두에서 가장 큰 유사성 값들을 나타내고, 이것은 또한 정확한 매칭을 나타낸다. IS_2의 제2 속성, $Author_2$, 가 GS의 $Publisher_2$에 부정확하게 매치되지만, 스키마 매칭 시스템은 인터-사이트 매칭을 사용하여 이 매칭을 교정한다.

[0061] 일 실시예에서, 스키마 매칭 시스템은 글로벌-대-인터페이스 대응, 글로벌-대-결과 대응, 인터페이스-대-결과 대응, 인터페이스-대-인터페이스 대응, 및 결과-대-결과 대응을 상호-유효화하여(cross-validate) 부정확할 수 있는 매치들을 식별하여 교정한다. 스키마 매칭 시스템은 인터페이스 속성들(및 유사하게 결과 속성들)을 그들이 매치하는 글로벌 속성들에 기초하여 복수 개의 클러스터들로 클러스터링한다. 예를 들어, 특정 글로벌 속성에 매치되었던 다양한 웹 데이타베이스들의 속성들은 한 개의 클러스터를 나타낸다. 이 클러스터링은 인트라-사이트 매칭에 기초한다. 인터-사이트 매칭은 또한 클러스터들을 상호-유효화하기 위해 사용될 수 있다. 인트라-사이트 및 인터-사이트 매칭이 완전히 정확하면, 웹 데이타베이스의 각각의 속성은 동일 클러스터 내에 있는 그외의 웹 데이타베이스들의 속성들만으로 매핑된다. 부연하면, 웹 데이타베이스들의 속성들은 서로 간에 그리고 글로벌 속성들로 일관적으로 매핑할 것이다. 일 실시예에서, 스키마 매칭 시스템은 웹 데이타베이스 스키마들의 속성들을 꼭지점들로 나타내고, 인터-사이트 매칭을 꼭지점들 사이의 에지들로서 나타낸다. 스키마 매칭 시스템은 에지-컷(edge-cut)이 최소화하도록 꼭지점들을 분할한다. 에지-컷은 분할들 간의 모든 에지들의 가중치들의 합이다(즉, 각각의 에지는 동일 가중치를 가짐). 에지-컷을 최소화함으로써, 스키마 매칭 시스템은 상이한 클러스터들의 꼭지점들 사이의 에지들의 수를 최소화한다.

[0062] 일 실시예에서, 스키마 매칭 시스템은 초기 클러스터들을 초기 분할로서 사용하고, 컷(cut)들의 수가 감소하는 한 한 클러스터로부터 다른 클러스터로 꼭지점들을 이동하여 에지-컷의 최소화를 근사시킨다. 일반적으로, 꼭지점은 그것의 대부분의 이웃들이 존재하는 클러스터로 이동된다. 이웃 꼭지점들은 그들 사이에 에지를 가진다. 꼭지점은 그것의 다수의 이웃들이 이동되면 이동될 필요가 있기 때문에, 스키마 매칭 시스템은 복수

패스들(passes)을 사용하여 에지-컷이 로컬 최적점(local optimum)에 수렴하도록 한다. 에지-컷이 수렴할 때, 스키마 매칭 시스템은 클러스터 간 매치를 무시하고 C_1에 클러스터링된 사이트 S_2의 속성 B_k에 A_i을 재매칭함으로써, 또는 그 반대로 해서, 2개의 클러스터들 C_1와 C_2에 포함된 사이트 S_1의 속성 A_i과 사이트 S_2의 속성 B_j 간의 클러스터 간의 매칭을 해결한다.

[0063] 도 3은 일 실시예에서 스키마 매칭 시스템의 분할의 일 패스를 도시한다. 이 예에서, 글로벌 스키마는 2개의 속성들 {Author, Publisher}을 포함하고, 5개의 웹 데이타베이스들은 IS 속성들 $IS_1 = \{A_a\}$, $IS_2 = \{B_a, B_p\}$, $IS_3 = \{C_a, C_p\}$, $IS_4 = \{D_a, D_p\}$, 및 $IS_5 = \{E_a, E_p\}$를 포함한다. 클러스터들(301, 302)은 속성들이 어느 글로벌 속성에 매치하는지(인트라-사이트 매칭에 의해)에 기초하여 속성들의 초기 클러스터들(꼭지점들로서 표현됨)을 도시하고, 속성들의 쌍들 사이의 에지들은 속성들이 매치함(인터-사이트 매칭에 의해)을 나타낸다. 초기 상태에서, A_a, 는 Publisher 글로벌 속성에 잘못 매치되고, 또한 B_p에 잘못 매치되는 한편, 그것은 Author 카테고리의 다른 3개의 속성들에 정확하게 매칭된다. 그러므로, 스키마 매칭 시스템은 A_a로 이동하여 클러스터들에 걸쳐 에지들의 수를 3에서 1로 감소시킨다. 그 이동은 A_a의 매칭 속성을 Publisher로부터 Author 글로벌 속성으로 교정한다. 이동 후, 스키마 매칭 시스템은 A_a과 B_p 사이의 에지를 제거하고, A_a과 B_a 간의 새 에지를 추가한다(Author 글로벌 속성에 매치되는 사이트 2의 속성). 클러스터들(311, 312)은 교정된 대응들을 나타낸다.

[0064] 글로벌 스키마들, 인터페이스 스키마들, 및 결과 스키마들은 다양한 기술들을 사용하여 식별될 수 있다. 글로벌 스키마들을 식별하는 일부 기술들은 속성들의 이름들과 요소들의 구조에 의존한다. (본 명세서에서 참조되는, 이종 데이타 소스의 글로벌 일람(Global Viewing of Heterogeneous Data Sources), S. Castano, V. Antonellis, and S. Vimercati, IEEE Trans. Data and Knowledge Eng., Vol. 13, no. 2, 2001; 및 웹 쿼리 인터페이스에서 통계적 스키마 매칭(Statistical Schema Matching across Web Query Interfaces), B. He and C.C. Chang, Proc. ACM SIGMOD Conf., 2003를 참조.) 그외의 기술들은 형식적 온탈로지들(ontologies)에 의존한다. (본 명세서에서 참조되는, 웹 쿼리 인터페이스에서 통계적 스키마 매칭(Statistical Schema Matching across Web Query Interface), B. He and C.C. Chang, Proc. ACM SIGMOD Conf., 2003; 및 형식적 온탈로지들을 사용한 글로벌 스키마 생성(Global Schema Generation Using Formal Ontologies), F. Hakimpour, and A. Geppert, Proc. 21st Conf. on Conceptual Modeling, 2002) 샘플 글로벌 속성 값들은 다양한 샘플 웹 데이타베이스들로부터 수집될 수 있거나, 또는 수동으로 생성될 수 있다. 웹 데이타베이스의 인터페이스 스키마는 HTML 스펙에 의해 정의되는 바와 같이 쿼리 웹 페이지의 입력-관련 태그들로부터 식별될 수 있다. 결과 스키마를 식별하는 일부 기술들은 동적 탬플릿-생성된(template-generated) 웹 페이지들로부터 구현된 반-구조화된 데이타 컨텐트를 발췌하기 위해 래퍼들(wrappers)을 생성한다. (본 명세서에 참조되는, 웹 페이지들로부터의 구조화된 데이타의 발췌(Extracting Structured Data from Web Pages), A. Arasu, and H. Garcia-Molina, Proc. ACM SIGMOD Conf., 2003; 패턴 발견에 기초한 정보 발췌(Information Extraction based on Pattern Discovery), C.H. Chang, and S.C. Lui, IEPAD, Proc. 10th World Wide Web Conf., 681-688, 2001; 거대 웹 사이트들로부터의 자동 데이타 발췌에 대해(Towards Automatic Data Extraction from Large Web Sites), V. Crescenzi, G. Mecca, and P. Merialdo, ROADRUNNER, Proc. 27th VLDB. Conf., 109-118, 2001; 및 웹 데이타베이스들에 대한 데이타 발췌 및 레이블 할당(Data Extraction and Label Assignment for Web Databases), J. Wang and F. Lochovsky, Proc. 12th World Wide Web Conf., 187-196, 2003 참조.) 한 기술은 HTML 페이지들에서 네스트된(nested) 반복-패턴 발견에 기초하여 정규-표현 래퍼를 생성한다. (본 명세서에 참조되는, 웹 데이타베이스들에 대한 데이타 발췌 및 레이블 할당(Data Extraction and Label Assignment for Web Databases), J. Wang and F. Lochovsky, Proc. 12th World Wide Web Conf., 187-196, 2003 참조) 당업자라면 이들 스키마들의 각각은 또한 수동으로 또는 자동과 수동 수단의 조합으로 식별될 수 있음을 이해할 것이다.

[0065] 도 4는 일 실시예에서 스키마 매칭 시스템의 컴포넌트들을 도시하는 블럭도이다. 스키마 매칭 시스템(410)은 통신 연결(402)을 통해 다양한 웹 데이타베이스 사이트(401)들에 접속한다. 스키마 매칭 시스템은 인트라-사이트 매칭 컴포넌트(411), 인터-사이트 매칭 컴포넌트(412), 상호-유효화 컴포넌트(413), 큐브 생성 컴포넌트(414), 큐브 프로젝트 컴포넌트(415), EMI 계산 컴포넌트(416), 및 매치 행렬 생성 컴포넌트(417)를 포함한다. 스키마 매칭 시스템은 또한 큐브 저장소(421), 프로젝션 저장소(422), EMI 저장소(423), 및 매치 저장소(424)를 포함한다. 인트라-사이트 매칭 컴포넌트는 큐브 생성 컴포넌트를 호출하여 어커런스 큐브를 생성하고, 큐브 프

로젝트 컴포넌트를 호출하여 글로벌-대-인터페이스, 글로벌-대-결과, 및 인터페이스-대-결과 어커런스 행렬들을 생성한다. 인트라-사이트 매칭 컴포넌트는 또한 EMI 계산 컴포넌트를 호출하여 어커런스 행렬들에 기초한 추정 상호 정보를 계산하고, 매치 행렬 생성 컴포넌트를 호출하여 속성들의 어느 쌍들이 매치하는지를 식별한다. 인터-사이트 매칭 컴포넌트는 어커런스 행렬들을 사용하여 추정된 벡터 유사성을 계산하고, 매치 행렬 생성 컴포넌트를 호출하여 매치들을 식별한다. 상호-유효화 컴포넌트는 부정확하게 매치된 것으로 나타난 속성들에 대한 매칭을 변경한다. 큐브 저장소는 어커런스 큐브들을 포함하고, 프로젝션 저장소는 어커런스 행렬들을 포함하고, EMI 저장소는 EMI 행렬들을 포함하고, 매치 저장소는 매치 행렬들을 포함한다.

[0066] 스키마 매칭 시스템이 구현되는 컴퓨팅 디바이스는 중앙 프로세싱 유닛, 메모리, 입력 디바이스들(예를 들어, 키보드 및 포인팅 디바이스들), 출력 디바이스들(예를 들어, 디스플레이 디바이스들), 및 저장 디바이스들(예를 들어, 디스크 드라이브)을 포함할 수 있다. 메모리와 저장 디바이스들은 스키마 매칭 시스템을 구현하는 명령어들을 포함할 수 있는 컴퓨터-판독가능한 매체들이다. 또한, 데이타 구조들 및 메시지 구조들은, 통신 링크 상의 신호와 같은, 데이타 전송 매체를 통해 저장될 수 있거나 전송될 수 있다. 인터넷, 근거리 통신망, 원거리 통신망, 또는 점대점 다이얼-업(point-to-point dial-up) 접속과 같은 다양한 통신 링크들이 사용될 수 있다.

[0067] 스키마 매칭 시스템은 개인용 컴퓨터, 서버 컴퓨터, 핸드헬드 또는 랩톱 디바이스, 멀티프로세서 시스템, 마이크로프로세서-기본적 시스템, 프로그램-가능한 소비자 전자제품, 통신망 PC, 미니 컴퓨터, 메인프레임 컴퓨터, 상술된 시스템들과 디바이스들 중의 임의의 것을 포함하는 분산 컴퓨팅 환경 등을 포함하는 다양한 운영 환경들에서 구현될 수 있다.

[0068] 스키마 매칭 시스템은 한 개 이상의 컴퓨터들이나 그외의 디바이스들에 의해 실행되는 프로그램 모듈들과 같은 컴퓨터-실행가능한 명령어들의 일반적인 문맥으로 기재될 수 있다. 일반적으로, 프로그램 모듈들은 특정 작업들을 수행하거나 특정 추상 데이타 유형들을 구현하는 루틴, 프로그램, 객체, 컴포넌트, 데이타 구조 등을 포함한다. 통상적으로, 프로그램 모듈들의 기능성은 다양한 실시예들에서 원하는대로 조합되거나 분산될 수 있다.

[0069] 도 5는 일 실시예에서 인트라-사이트 매칭 컴포넌트의 프로세싱을 도시하는 흐름도이다. 컴포넌트는 웹 데이타베이스에 대한 글로벌-대-인터페이스, 글로벌-대-결과, 및 인터페이스-대-결과 대응들을 식별한다. 블럭(501)에서, 컴포넌트는 큐브 생성 컴포넌트를 호출하여 어커런스 큐브를 생성한다. 블럭(502) 내지 블럭(506)에서, 컴포넌트는 스키마들의 쌍들(즉, 글로벌과 인터페이스, 글로벌과 결과, 및 인터페이스와 결과)을 선택하여 순환(loop)시키고, 각각의 쌍의 대응을 나타내는 매치 행렬을 생성한다. 블럭(502)에서, 컴포넌트는 스키마들의 다음 쌍을 선택한다. 판정 블럭(503)에서, 스키마들의 모든 쌍들이 이미 선택되었다면, 컴포넌트는 완료하고, 그렇지 않으면 컴포넌트는 블럭(504)에서 계속한다. 블럭(504)에서, 컴포넌트는 큐브 프로젝트 컴포넌트를 호출하여, 스키마들의 선택된 쌍에 대한 어커런스 행렬을 생성한다. 블럭(505)에서, 컴포넌트는 EMI 계산 컴포넌트를 호출하여 스키마들의 선택된 쌍의 속성들의 쌍들 간의 상호 정보를 추정한다. 블럭(506)에서, 컴포넌트는 매치 행렬 생성 컴포넌트를 호출하여 스키마들의 선택된 쌍에 대한 속성 대응들을 나타내는 매치 행렬을 생성한다. 그 다음, 컴포넌트는 블럭(502)으로 순환하여 스키마들의 다음 쌍을 선택한다.

[0070] 도 6은 일 실시예에서 큐브 생성 컴포넌트의 프로세싱을 도시하는 흐름도이다. 컴포넌트는 글로벌 스키마, 인터페이스 스키마, 및 결과 스키마에 기초하여 웹 데이타베이스의 어커런스 큐브를 생성한다. 어커런스 큐브는 글로벌 속성, 인터페이스 속성, 및 결과 속성의 각각의 조합을 카운트에 매핑하는 3차원 행렬이다. 카운트는 해당 인터페이스 속성이 해당 글로벌 속성의 글로벌 속성 값으로 설정된 쿼리의 결과 엔트리가 그 결과 속성의 해당 글로벌 속성 값을 갖는 횟수이다. 블럭(601)에서, 컴포넌트는 다음 글로벌 속성을 선택한다. 판정 블럭(602)에서, 모든 글로벌 속성들이 이미 선택되었으면, 컴포넌트는 리턴하고, 그렇지 않으면 컴포넌트는 블럭(603)에서 계속한다. 블럭(603)에서, 컴포넌트는 선택된 글로벌 속성에 대한 다음 글로벌 속성 값을 선택한다. 판정 블럭(604)에서, 선택된 글로벌 속성에 대한 모든 글로벌 속성 값들이 이미 선택되었으면, 컴포넌트는 블럭(601)으로 순환하여 다음 글로벌 속성을 선택하고, 그렇지 않으면 컴포넌트는 블럭(605)에서 계속한다. 블럭(605) 내지 블럭(609)에서, 컴포넌트는 각각의 인터페이스 속성을 선택하는 단계와, 그 인터페이스 속성을 선택된 글로벌 속성 값들로 설정한 쿼리를 제출하는 단계를 순환한다. 당업자는 일부 인터페이스 속성들의 값들의 도메인이 제한적일 수 있음을 이해할 것이다. 예를 들어, 인터페이스 속성이 HTML SELECT 소자에 의해 표현되면, 그것의 값들의 도메인은 연관된 OPTION 소자의 값들로 제한될 것이다. 그런 경우, 컴포넌트는 옵션 값과 "유사한" 글로벌 속성 값들에 대한 쿼리들만을 제출할 수 있다. 글로벌 속성 값이 옵션 값을 포함하면 유사하다고 고려될 것이다. 당업자는 유사성의 다른 척도들이 사용될 수 있음을 이해할 것이다. CHECKBOX와 RADIOBOX 소자들에 대한 쿼리들은 유사한 방식으로 취급될 수 있다. TEXTBOX의 값들에 대한 도메인은 알려지지

않을 수 있으므로, 컴포넌트는 TEXTBOX로 표현되는 인터페이스 속성에 대한 모든 글로벌 속성 값들을 사용하여 총망라하여 쿼리들을 제출할 수 있다. 일 실시예에서, 컴포넌트는 각각의 쿼리에 대해 단지 한 개의 인터페이스 속성을 위한 값을 설정한다. 다른 인터페이스 속성들의 값들은 웹 사이트에 의해 정의되는 바와 같이 디폴트 값을 가질 수 있다. 블럭(605)에서, 컴포넌트는 다음 인터페이스 속성을 선택한다. 판정 블럭(606)에서, 모든 인터페이스 속성들이 이미 선택되었으면, 컴포넌트는 블럭(603)으로 순환하여 선택된 글로벌 속성에 대한 다음 글로벌 속성 값을 선택한다. 블럭(607)에서, 컴포넌트는 선택된 인터페이스 속성과 선택된 글로벌 속성 값을 사용하여 쿼리를 공식화한다. 블럭(608)에서, 컴포넌트는 생성된 쿼리를 웹 사이트에 제출한다. 블럭(609)에서, 컴포넌트는 쿼리의 결과에 기초하여 어커런스 큐브를 업데이트하고, 그 다음, 블럭(605)으로 순환하여 다음 인터페이스 속성을 선택한다.

[0071] 도 7은 일 실시예에서 큐브 업데이트 컴포넌트의 프로세싱을 도시하는 흐름도이다. 컴포넌트는 글로벌 속성, 글로벌 속성 값, 및 인터페이스 속성과 쿼리 결과의 표시가 전달된다. 블럭(701)에서, 컴포넌트는 결과의 다음 엔트리 또는 행을 선택한다. 판정 블럭(702)에서, 결과의 모든 엔트리들이 이미 선택되었으면, 컴포넌트는 리턴하고, 그렇지 않으면 컴포넌트는 블럭(703)에서 계속한다. 블럭(703)에서 컴포넌트는 다음 결과 속성이나 열을 선택한다. 판정 블럭(704)에서, 모든 결과 속성들이 이미 선택되었으면, 컴포넌트는 블럭(701)으로 순환하여 결과의 다음 엔트리를 선택하고, 그렇지 않으면 컴포넌트는 블럭(705)에서 계속한다. 블럭(705)에서, 글로벌 속성 값이 선택된 엔트리의 선택된 결과 속성의 값과 동일하면, 컴포넌트는 블럭(706)에서 계속하고, 그렇지 않으면 컴포넌트는 블럭(703)으로 순환하여 선택된 엔트리의 다음 결과 속성을 선택한다. 블럭(706)에서, 컴포넌트는 전달된 글로벌 속성, 전달된 인터페이스 속성, 및 선택된 결과 속성에 대한 어커런스 큐브 내의 카운트를 증분시킨다. 그 다음, 컴포넌트는 블럭(703)으로 순환하여 선택된 엔트리의 다음 결과 속성을 선택한다.

[0072] 도 8은 일 실시예에서 큐브 프로젝트 컴포넌트의 프로세싱을 도시하는 흐름도이다. 이 실시예에서, 컴포넌트는 글로벌-대-인터페이스 대응에 대한 어커런스 행렬을 생성한다. 스키마 매칭 시스템은 유사한 방식으로 글로벌-대-결과 대응 및 인터페이스-대-결과 대응에 대한 어커런스 행렬들을 생성할 수 있다. 이 실시예에서, 컴포넌트는 글로벌 속성과 인터페이스 속성 쌍에 대해 결과 속성들의 카운트를 합하여 어커런스 큐브의 3차원을 대응 행렬의 2차원으로 프로젝트한다. 당업자들은 쉬운 덧셈 외의 프로젝션 기술들이 사용될 수 있음을 이해할 것이다. 예를 들어, 가중치가 결과 스키마의 자동 식별 동안에 유도된 신뢰도에 기초하는 경우, 컴포넌트는 가중치 합을 사용할 수 있다. 블럭(801)에서, 컴포넌트는 다음 글로벌 속성을 선택한다. 판정 블럭(802)에서, 모든 글로벌 속성들이 이미 선택되었으면, 컴포넌트는 리턴하고, 그렇지 않으면 컴포넌트는 블럭(803)에서 계속한다. 블럭(803)에서, 컴포넌트는 다음 인터페이스 속성을 선택한다. 판정 블럭(804)에서, 모든 인터페이스 속성들이 선택되었으면, 컴포넌트는 블럭(801)으로 순환하여 다음 글로벌 속성을 선택하고, 그렇지 않으면 컴포넌트는 블럭(805)에서 계속한다. 블럭(805)에서, 컴포넌트는 다음 결과 속성을 선택한다. 판정 블럭(806)에서, 모든 결과 속성들이 이미 선택되었으면, 컴포넌트는 블럭(803)으로 순환하여 다음 인터페이스 속성을 선택하고, 그렇지 않으면 컴포넌트는 블럭(807)에서 계속한다. 블럭(807)에서, 컴포넌트는, 선택된 글로벌 속성, 인터페이스 속성, 및 결과 속성에 대한 어커런스 큐브로부터의 횟수에 의해 선택된 인터페이스 속성과 글로벌 속성에 대한 어커런스 행렬의 카운트를 증분시킨다. 그 다음, 컴포넌트는 블럭(805)으로 순환하여 다음 결과 속성을 선택한다.

[0073] 도 9는 일 실시예에서 EMI 계산 컴포넌트의 프로세싱을 도시하는 흐름도이다. 이 컴포넌트는 수학식 1을 사용하여 어커런스 행렬의 속성들의 쌍들에 대한 상호 정보를 추정한다. 당업자라면 다양한 기술들이 속성들의 쌍들이 매치하는 확률을 추정하는데 사용될 수 있음을 이해할 것이다. 컴포넌트는 어커런스 행렬을 전달하고, EMI 행렬을 리턴한다. 블럭(901)에서, 컴포넌트는 어커런스 행렬 내의 모든 카운트들의 합을 계산한다. 블럭(902)에서, 컴포넌트는 어커런스 행렬의 각각의 행 내의 카운트들의 합을 계산한다. 블럭(903)에서, 컴포넌트는 어커런스 행렬의 각각의 열 내의 카운트들의 합을 계산한다. 블럭(904) 내지 블럭(908)에서, 컴포넌트는 어커런스 행렬의 속성들의 각각의 쌍을 선택하는 단계를 순환하여, 속성들이 매치하는 확률을 결정한다. 블럭(904)에서, 컴포넌트는 어커런스 행렬의 다음 행을 선택한다. 판정 블럭(905)에서, 어커런스 행렬의 모든 행들이 이미 선택되었으면, 컴포넌트는 리턴하고, 그렇지 않으면 컴포넌트는 블럭(906)에서 계속한다. 블럭(906)에서, 컴포넌트는 어커런스 행렬의 다음 열을 선택한다. 판정 블럭(907)에서, 어커런스 행렬의 모든 열들이 이미 선택되었으면, 컴포넌트는 블럭(904)으로 순환하여 어커런스 행렬의 다음 행을 선택하고, 그렇지 않으면 컴포넌트는 블럭(908)에서 계속한다. 블럭(908)에서, 컴포넌트는 선택된 행과 열에 의해 표현된 속성들에 대한 추정된 상호 정보를 계산한다. 그 다음, 컴포넌트는 블럭(906)으로 순환하여 다음 열을 선택한다.

[0074] 도 10은 일 실시예에서 매치 행렬 생성 컴포넌트의 프로세싱을 도시하는 흐름도이다. 컴포넌트에는 속성들의

쌍들이 매치하는 확률을 나타내는, EMI 행렬과 같은 행렬이 전달된다. 속성들의 쌍에 대한 확률이 양측 속성들에 대해 가장 높은 확률이면(예를 들어, 한 개 속성을 나타내는 행에서 가장 높고, 다른 속성을 나타내는 열에서 가장 높음), 컴포넌트는 속성들이 매치함을 발견한다. 블럭(1001)에서, 컴포넌트는 전달된 행렬의 다음 행을 선택한다. 판정 블럭(1002)에서, 전달된 행렬의 전체 행들이 이미 선택되었으면, 컴포넌트는 리턴하고, 그렇지 않으면 컴포넌트는 블럭(1003)에서 계속한다. 블럭(1003)에서, 컴포넌트는 전달된 행렬의 다음 열을 선택한다. 판정 블럭(1004)에서, 전달된 행렬의 모든 열들이 이미 선택되었으면, 컴포넌트는 블럭(1001)으로 순환하여 전달된 행렬의 다음 행을 선택하고, 그렇지 않으면 컴포넌트는 블럭(1005)에서 계속한다. 판정 블럭(1005)에서, 선택된 행과 열의 값이 그 행에서 최고 높으면, 컴포넌트는 블럭(1006)에서 계속하고, 그렇지 않으면 컴포넌트는 블럭(1003)으로 순환하여 다음 열을 선택한다. 판정 블럭(1006)에서, 선택된 행과 열의 값이 그 열 내에서 최고 높으면, 컴포넌트는 블럭(1007)에서 계속하고, 그렇지 않으면 컴포넌트는 블럭(1003)으로 순환하여 다음 열을 선택한다. 블럭(1007)에서, 컴포넌트는 선택된 행과 열에 대한 매치 행렬 값을 설정하여 매치를 나타내고, 그 다음, 블럭(1003)으로 순환하여 선택된 행의 다음 열을 선택한다.

[0075] 도 11은 일 실시예에서 인터-사이트 매칭 컴포넌트의 프로세싱을 도시하는 흐름도이다. 컴포넌트는 한 웹 사이트의 어느 속성들이(인터페이스 및 결과) 다른 웹 사이트의 어느 속성들과 매치하는지를 식별한다. 컴포넌트는 웹 사이트들의 글로벌-대-인터페이스 대응의 어커런스 행렬을 사용하여 인터페이스 스키마들에 대한 매치들을 식별하고, 웹 사이트들의 글로벌-대-결과 대응의 어커런스 행렬을 사용하여 결과 스키마들의 매치들을 식별한다. 블럭(1101)에서, 컴포넌트는 큐브 생성 컴포넌트를 호출하여 사이트 A에 대한 어커런스 큐브를 생성한다. 블럭(1102)에서, 컴포넌트는 큐브 프로젝트 컴포넌트를 호출하여 사이트 A에 대한 어커런스 행렬들을 생성한다. 블럭(1103)에서, 컴포넌트는 큐브 생성 컴포넌트를 호출하여 사이트 B에 대한 어커런스 큐브를 생성한다. 블럭(1104)에서, 컴포넌트는 큐브 프로젝트 컴포넌트를 호출하여 사이트 B에 대한 어커런스 행렬들을 생성한다. 블럭(1105)에서, 컴포넌트는 인터페이스 속성들에 대한 추정된 벡터 유사성 계산 컴포넌트를 호출하여 사이트 A로부터 사이트 B로의 인터페이스 속성들의 쌍들이 매치하는 확률을 생성한다. 당업자는 다수의 다른 기술들이 이 확률을 추정하기 위해 사용될 수 있으며, 이 벡터 유사성은 단지 일 예일 뿐임을 이해할 것이다. 블럭(1106)에서, 컴포넌트는 인터페이스 속성들에 대한 추정된 벡터 유사성 행렬을 전달하는 매치 행렬 생성 컴포넌트를 호출하여 인터페이스 속성들의 쌍들이 매치함을 나타내는 행렬을 생성한다. 블럭(1107)에서, 컴포넌트는 추정된 벡터 유사성 계산 컴포넌트를 호출하여 결과 속성들에 대한 추정된 벡터 유사성 행렬을 생성한다. 블럭(1108)에서, 컴포넌트는 매치 행렬 생성 컴포넌트를 호출하여 결과 속성들의 쌍들이 매치함을 나타내는 행렬을 생성한다. 그 다음, 컴포넌트는 완료한다.

[0076] 도 12는 일 실시예에서 추정된 벡터 유사성 계산 컴포넌트의 프로세싱을 도시하는 흐름도이다. 컴포넌트에는 인터페이스-대-인터페이스 대응 또는 결과-대-결과 대응에 대한 어커런스 행렬이 전달되고, 속성들의 각각의 쌍이 매치하는 확률을 결정한다. 블럭(1201)에서, 컴포넌트는 사이트 A의 다음 속성을 선택한다. 판정 블럭(1202)에서, 사이트 A의 모든 속성들이 이미 선택되었으면, 컴포넌트는 리턴하고, 그렇지 않으면 컴포넌트는 블럭(1203)에서 계속한다. 블럭(1203)에서, 컴포넌트는 사이트 B의 다음 속성을 선택한다. 판정 블럭(1204)에서, 사이트 B의 모든 속성들이 이미 선택되었으면, 컴포넌트는 블럭(1201)으로 순환하여 사이트 A의 다음 속성을 선택하고, 그렇지 않으면 컴포넌트는 블럭(1205)에서 계속한다. 블럭(1205)에서, 컴포넌트는 수학식 3에 따라 선택된 속성들에 대한 추정된 벡터 유사성을 계산하고, 그 다음, 블럭(1203)으로 순환하여 사이트 B의 다음 속성을 선택한다.

[0077] 도 13은 일 실시예에서 상호-유효화 컴포넌트의 프로세싱을 도시하는 흐름도이다. 인트라-사이트 매치가 부정확하다고 인터-사이트 매치들이 나타낼 때, 컴포넌트는 속성들의 매치들을 변경한다. 블럭(1301)에서, 컴포넌트는 다음 글로벌 속성을 선택한다. 판정 블럭(1302)에서, 모든 글로벌 속성들이 이미 선택되었으면, 컴포넌트는 완료하고, 그렇지 않으면 컴포넌트는 블럭(1303)에서 계속한다. 블럭(1303)에서, 컴포넌트는 다음 웹 사이트를 선택한다. 판정 블럭(1304)에서, 모든 웹 사이트들이 이미 선택되었으면, 컴포넌트는 블럭(1301)으로 순환하여 다음 글로벌 속성을 선택하고, 그렇지 않으면 컴포넌트는 블럭(1305)에서 계속한다. 판정 블럭(1305)에서, 선택된 웹 사이트가 선택된 글로벌 속성과 매치하는 속성을 가지고 있으면, 컴포넌트는 블럭(1306)에서 계속하고, 그렇지 않으면 컴포넌트는 블럭(1303)으로 순환하여 다음 웹 사이트를 선택한다. 판정 블럭(1306)에서, 선택된 속성이 다른 글로벌 속성으로 이동되어야 하면, 컴포넌트는 블럭(1307)에서 계속하고, 그렇지 않으면 컴포넌트는 블럭(1303)으로 순환하여 다음 웹 사이트를 선택한다. 블럭(1307)에서, 컴포넌트는 선택된 속성을 변경하여 상이한 글로벌 속성에 매치시킨다. 블럭(1308)에서, 컴포넌트는 선택된 속성의 인트라-사이트 매치들을 변경한다. 그 다음, 컴포넌트는 블럭(1303)으로 순환하여 다음 웹 사이트를 선택한다.

[0078] 당업자라면, 스키마 매칭 시스템의 특정 실시예들이 설명을 목적으로 본 명세서에 기재되었지만, 본 발명의 사상 및 범주를 벗어나지 않고 다양한 변경을 행할 수 있음을 이해할 것이다. 따라서, 본 발명은 첨부된 청구범위에 의한 것이 아니고는 제한 받지 않는다.

발명의 효과

[0079] 본 발명에 따라 생성된 매핑들을 사용하여, 검색 엔진 서비스는 글로벌 속성들을 사용하여 쿼리들을 생성할 수 있고, 이들 쿼리들을 대응하는 인터페이스 속성들에 매핑할 수 있고, 쿼리를 제출할 수 있고, 및 원하는 글로벌 속성들에 대응하는 결과 속성들로부터의 값들을 검색할 수 있다.

도면의 간단한 설명

[0001] 도 1은 서적상(bookseller)을 위한 웹 데이타베이스의 다양한 스키마들을 도시하는 도면이다.

[0002] 도 2는 일 실시예에서 인트라-사이트(intra-site) 및 인터-사이트(inter-site) 매칭을 도시한다.

[0003] 도 3은 일 실시예에서 스키마 매칭 시스템의 분할(partition)의 일 패스(pass)를 도시한다.

[0004] 도 4는 일 실시예에서 스키마 매칭 시스템의 컴포넌트들을 도시하는 블럭도이다.

[0005] 도 5는 일 실시예에서 인트라-사이트 매칭(match intra-site) 컴포넌트의 프로세싱을 도시하는 흐름도이다.

[0006] 도 6은 일 실시예에서 큐브 생성(generate cube) 컴포넌트의 프로세싱을 도시하는 흐름도이다.

[0007] 도 7은 일 실시예에서 큐브 업데이트(update cube) 컴포넌트의 프로세싱을 도시하는 흐름도이다.

[0008] 도 8은 일 실시예에서 큐브 프로젝션(project cube) 컴포넌트의 프로세싱을 도시하는 흐름도이다.

[0009] 도 9는 일 실시예에서 EMI 계산(calculate EMI) 컴포넌트의 프로세싱을 도시하는 흐름도이다.

[0010] 도 10은 일 실시예에서 매치 행렬 생성(generate match matrix) 컴포넌트의 프로세싱을 도시하는 흐름도이다.

[0011] 도 11은 일 실시예에서 인터-사이트 매칭(match inter-site) 컴포넌트의 프로세싱을 도시하는 흐름도이다.

[0012] 도 12는 일 실시예에서 추정된 벡터 유사성 계산(calculate estimated vector similarity) 컴포넌트의 프로세싱을 도시하는 흐름도이다.

[0013] 도 13은 일 실시예에서 상호-유효화(cross-validate) 컴포넌트의 프로세싱을 도시하는 흐름도이다.

[0014] <도면의 주요 부분에 대한 부호의 간단한 설명>

[0015] 401: 웹 DB 사이트

[0016] 402: 통신 연결

[0017] 410: 스키마 매칭 시스템

[0018] 411: 인트라-사이트 스키마 매칭

[0019] 412: 인터-사이트 스키마 매칭

[0020] 413: 상호-유효화

[0021] 414: 큐브 생성

[0022] 415: 큐브 프로젝션

[0023] 416: EMI 계산

[0024] 417: 매치 행렬 생성

[0025] 421: 큐브 저장소

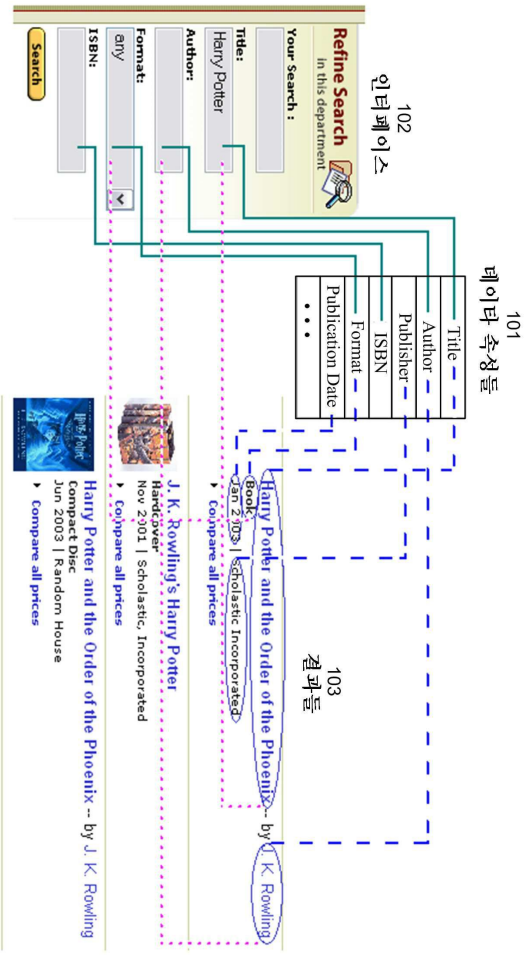
[0026] 422: 프로젝션 저장소

[0027] 423: EMI 저장소

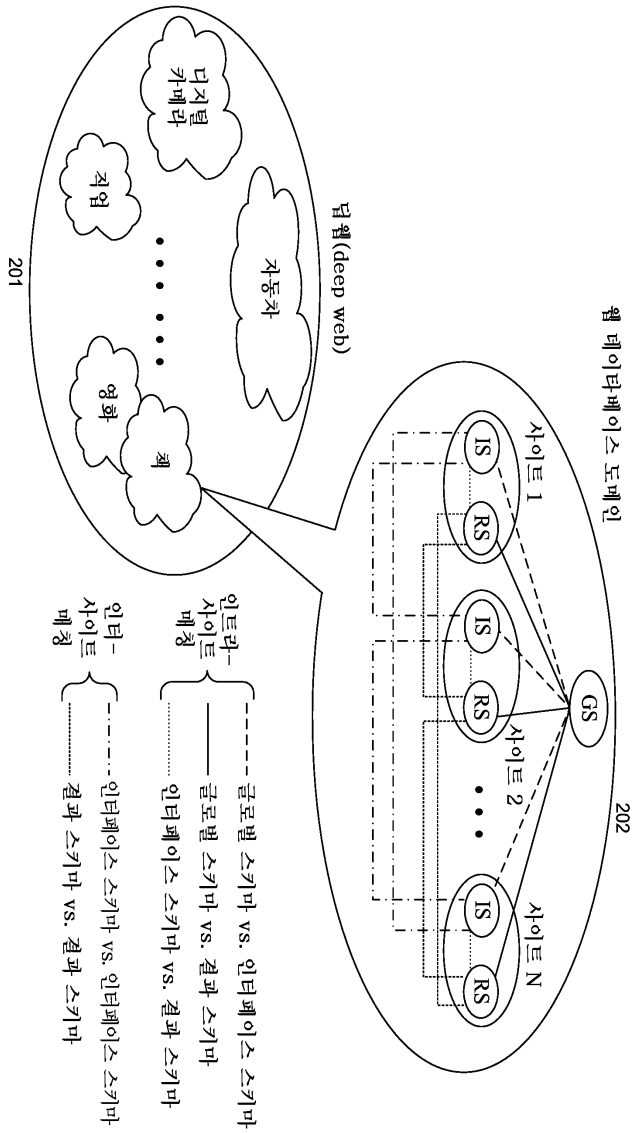
[0028] 424: 매치 저장소

도면

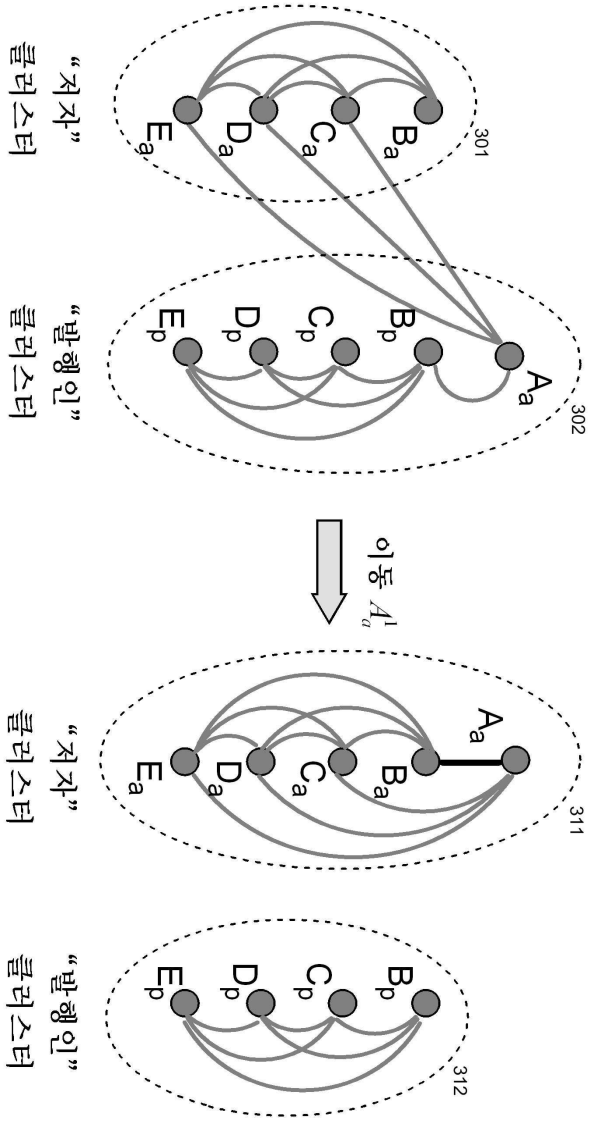
도면1



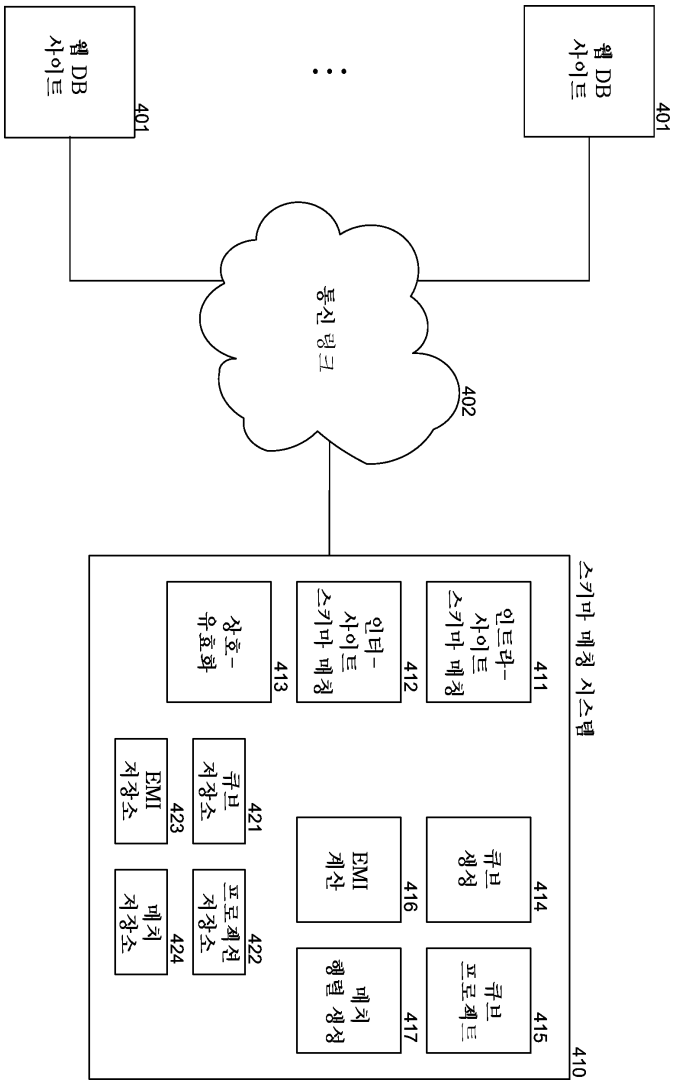
도면2


웹 데이터베이스 도메인
사이트 1
사이트 2
사이트 N
IS
RS
GS
202
딥웹(deep web)
디지털 카메라
직업
자동차
영화
책
201
인트라-사이트 매칭
인터-사이트 매칭
글로벌 스키마 vs. 인터페이스 스키마
글로벌 스키마 vs. 결과 스키마
인터페이스 스키마 vs. 결과 스키마
인터페이스 스키마 vs. 인터페이스 스키마
결과 스키마 vs. 결과 스키마

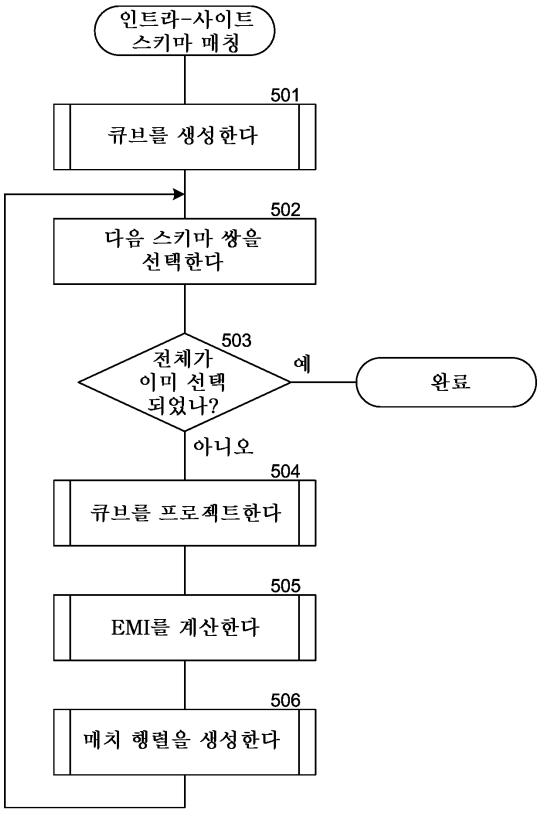
도면3



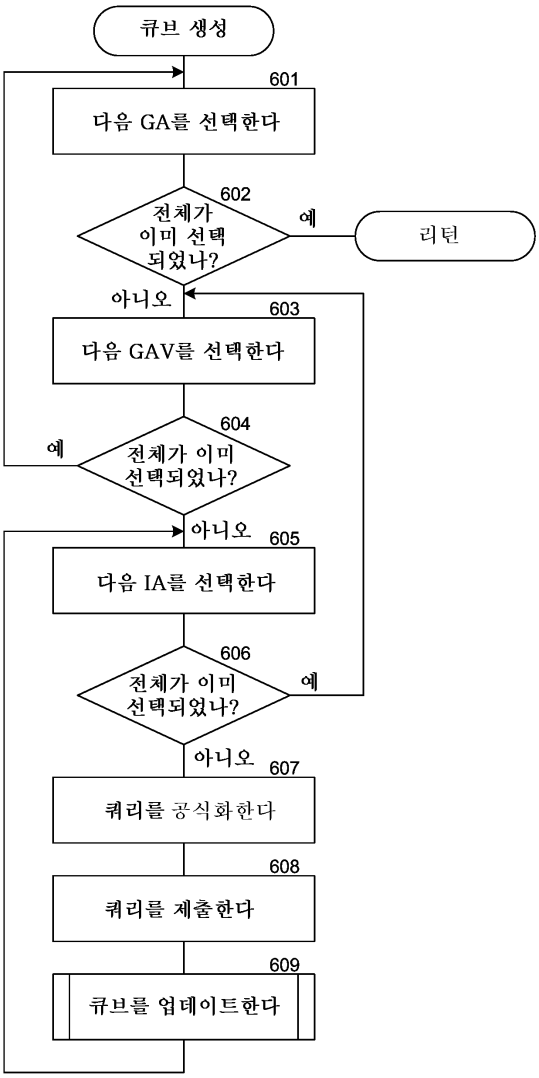
도면4


웹 DB 사이트
401
웹 DB 사이트
401
...
통신 링크
402
스키마 매칭 시스템
410
인트라-사이트 스키마 매칭
411
인터-사이트 스키마 매칭
412
상호-유효화
413
큐브 생성
414
큐브 프로젝트
415
EMI 계산
416
매치 행렬 생성
417
큐브 저장소
421
프로젝션 저장소
422
EMI 저장소
423
매치 저장소
424

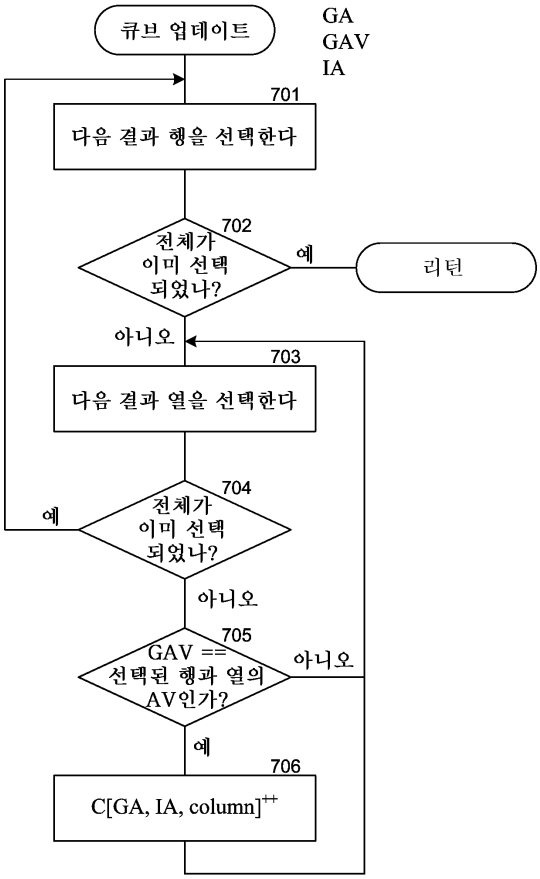
도면5


인트라-사이트 스키마 매칭
501
큐브를 생성한다
502
다음 스키마 쌍을 선택한다
503
전체가 이미 선택 되었나?
예
완료
아니오
504
큐브를 프로젝트한다
505
EMI를 계산한다
506
매치 행렬을 생성한다

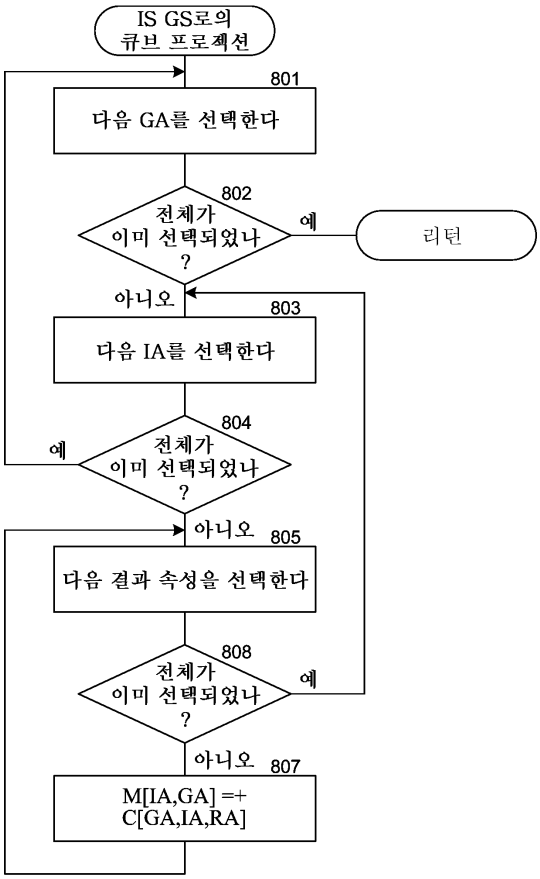
도면6


큐브 생성
601
다음 GA를 선택한다
602
전체가 이미 선택 되었나?
예
리턴
아니오
603
다음 GAV를 선택한다
604
전체가 이미 선택되었나?
예
아니오
605
다음 IA를 선택한다
606
전체가 이미 선택되었나?
예
아니오
607
쿼리를 공식화한다
608
쿼리를 제출한다
609
큐브를 업데이트한다

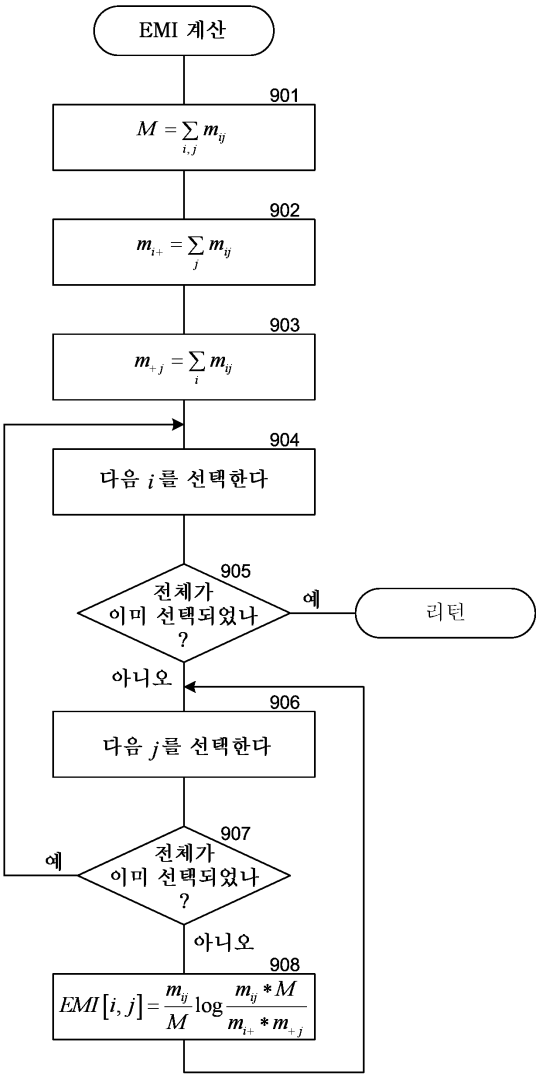
도면7


큐브 업데이트
GA
GAV
IA
701
다음 결과 행을 선택한다
702
전체가 이미 선택 되었나?
예
리턴
아니오
703
다음 결과 열을 선택한다
704
전체가 이미 선택 되었나?
예
아니오
705
GAV == 선택된 행과 열의 AV인가?
아니오
예
706
C[GA, IA, column]++

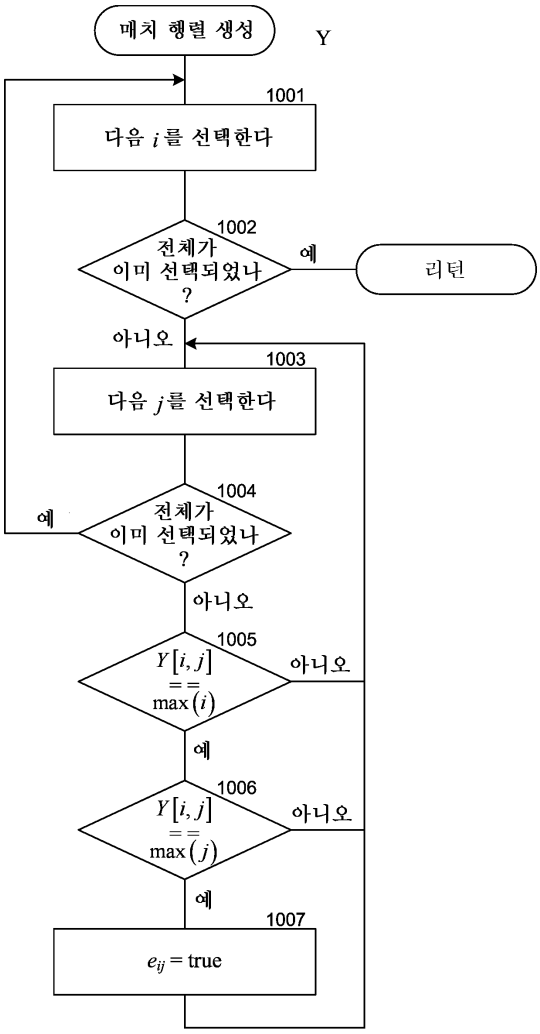
도면8


IS GS로의
큐브 프로젝션
801
다음 GA를 선택한다
802
전체가
이미 선택되었나
?
예
리턴
아니오
803
다음 IA를 선택한다
804
전체가
이미 선택되었나
?
예
아니오
805
다음 결과 속성을 선택한다
808
전체가
이미 선택되었나
?
예
아니오
807
M[IA,GA] =+
C[GA,IA,RA]

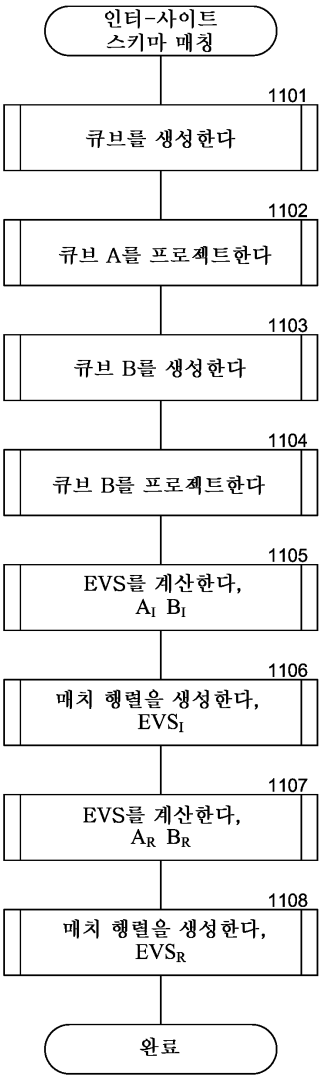
도면9


EMI 계산
901
$M = \sum_{i,j} m_{ij}$
902
$m_{i+} = \sum_{j} m_{ij}$
903
$m_{+j} = \sum_{i} m_{ij}$
904
다음 i를 선택한다
905
전체가 이미 선택되었나 ?
예
리턴
아니오
906
다음 j를 선택한다
907
전체가 이미 선택되었나 ?
예
아니오
908
$EMI[i,j] = \frac{m_{ij}}{M} \log \frac{m_{ij} * M}{m_{i+} * m_{+j}}$

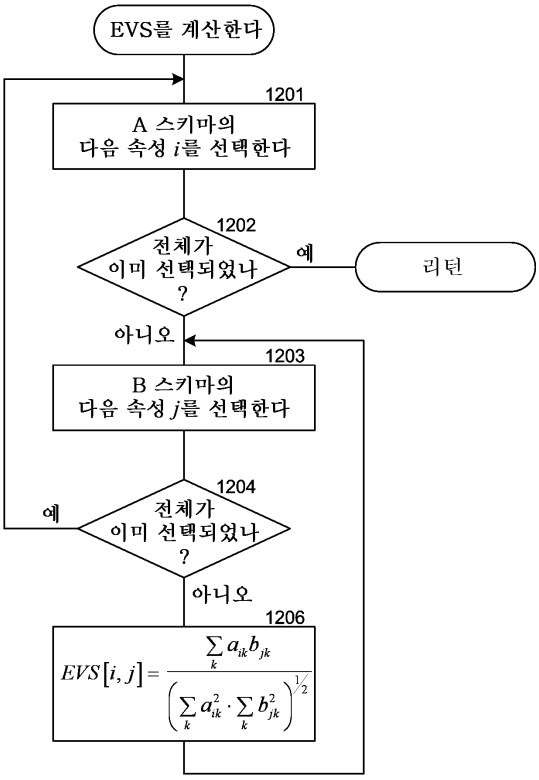
도면10


매치 행렬 생성
Y
1001
다음 i를 선택한다
1002
전체가 이미 선택되었나 ?
예
리턴
아니오
1003
다음 j를 선택한다
1004
전체가 이미 선택되었나 ?
예
아니오
1005
$Y[i,j] == \max(i)$
아니오
예
1006
$Y[i,j] == \max(j)$
아니오
예
1007
e_{ij} = true

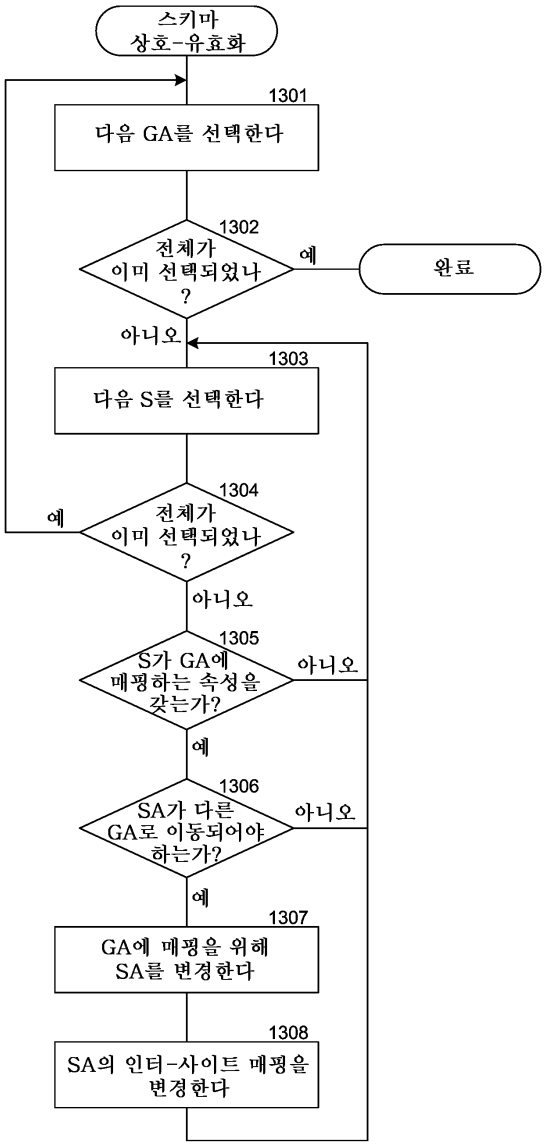
도면11


인터-사이트
스키마 매칭
1101
큐브를 생성한다
1102
큐브 A를 프로젝트한다
1103
큐브 B를 생성한다
1104
큐브 B를 프로젝트한다
1105
EVS를 계산한다,
A_I B_I
1106
매치 행렬을 생성한다,
EVS_I
1107
EVS를 계산한다,
A_R B_R
1108
매치 행렬을 생성한다,
EVS_R
완료

도면12


EVS를 계산한다
1201
A 스키마의
다음 속성 i를 선택한다
1202
전체가
이미 선택되었나
?
예
리턴
아니오
1203
B 스키마의
다음 속성 j를 선택한다
1204
전체가
이미 선택되었나
?
예
아니오
1206
$EVS[i,j] = \frac{\sum_k a_{ik} b_{jk}}{\left(\sum_k a_{ik}^2 \cdot \sum_k b_{jk}^2\right)^{1/2}}$

도면13


스키마
상호-유효화
1301
다음 GA를 선택한다
1302
전체가
이미 선택되었나
?
예
완료
아니오
1303
다음 S를 선택한다
1304
전체가
이미 선택되었나
?
예
아니오
1305
S가 GA에
매핑하는 속성을
갖는가?
아니오
예
1306
SA가 다른
GA로 이동되어야
하는가?
아니오
예
1307
GA에 매핑을 위해
SA를 변경한다
1308
SA의 인터-사이트 매핑을
변경한다