(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2008-0005286
(43) 공개일자 2008년01월10일

(51) Int. Cl.
G06F 17/40 (2006.01)
(21) 출원번호 10-2007-7027185
(22) 출원일자 2007년11월22일
심사청구일자 2007년11월22일
번역문제출일자 2007년11월22일
(86) 국제출원번호 PCT/US2006/015413
국제출원일자 2006년04월24일
(87) 국제공개번호 WO 2006/116273
국제공개일자 2006년11월02일
(30) 우선권주장
11/112,716 2005년04월22일 미국(US)

(71) 출원인
구글 잉크.
미국 캘리포니아 마운틴 뷰 앰피씨어터 파크웨이 1600 (우편번호 94043)
(72) 발명자
게르킹, 다비드
미국 캘리포니아 91316-4319, 엔시노, 아론조 플레이스 17742
라우, 칭
미국 캘리포니아 90025, 로스앤젤레스 아파트 #12, 스토너 애비뉴1737
맥스웰, 앤드류
미국 캘리포니아 90026, 로스앤젤레스, 알리슨 애비뉴 1440
(74) 대리인
이범래

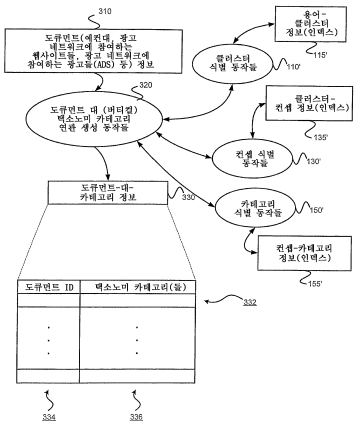
전체 청구항 수 : 총 26 항

(54) 카테고리화로부터 유도된 택소노미 및 데이터 구조들에관련된, 도큐먼트 및/또는 클러스터들과 같은 오브젝트들의카테고리화

(57) 요 약

웹사이트는 (a) 웹사이트 정보를 얻고, (b) 웹사이트 정보를 사용하여 웹사이트에 대한 스코어된 클러스터(예컨대, 의미, 용어 동시 발생 등)의 세트를 결정하고, (c) 클러스터들의 세트의 적어도 일부를 사용하여 미리정의된 택소노미(taxonomy)의 적어도 하나의 카테고리(예컨대, 버티컬(vertical) 카테고리)를 결정함으로써 자동으로 카테고리화될 수 있다. 의미 클러스터(예컨대, 용어 동시 발생 클러스터)는 (a) 의미 클러스터를 얻고, (b) 얻어진 클러스터를 사용하여 하나 이상의 스코어된 컨셉들의 세트를 식별하고, (c) 하나 이상의 스코어된 컨셉들(concepts)의 적어도 일부를 사용하여 하나 이상의 카테고리들의 세트를 식별하고, (d) 의미 클러스터와 하나 이상의 카테고리들의 적어도 일부를 연관시킴으로써 미리정의된 택소노미의 하나 이상의 카테고리(예컨대, 버티컬 카테고리들)와 자동으로 연관될 수 있다. 프로퍼티(예컨대, 웹사이트)는 (a) 프로퍼티에 관한 정보를 얻고, (b) 얻어진 프로퍼티 정보를 사용하여 하나 이상의 스코어된 의미 클러스터들(예컨대, 용어 동시 발생 클러스터들)의 세트를 식별하고, (c) 하나 이상의 스코어된 의미 클러스터들의 적어도 일부를 사용하여 하나 이상의 카테고리들(예컨대, 버티컬 카테고리들)의 세트를 식별하고, (d) 프로퍼티와 하나 이상의 카테고리들의 적어도 일부를 연관시킴으로써 미리정의된 택소노미의 하나 이상의 카테고리들(예컨대, 버티컬 카테고리들)과 연관될 수 있다.

대표도 - 도3



특허청구의 범위

청구항 1

웹사이트를 자동으로 카테고리화하기 위한 컴퓨터 구현 방법(computer-implemented method)에 있어서,

a) 웹사이트 정보를 얻는 단계;

b) 상기 웹사이트 정보를 사용하여 상기 웹사이트에 대한 스코어된 클러스터들(scored clusters)의 세트를 결정하는 단계; 및

c) 상기 클러스터들의 세트의 적어도 일부를 사용하여 미리정의된 택소노미(taxonomy)의 적어도 하나의 카테고리를 결정하는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 2

제 1 항에 있어서, 상기 웹사이트 정보를 사용하여 상기 웹사이트에 대한 스코어된 클러스터들의 세트를 결정하는 단계는 상기 웹사이트의 개별 웹페이지들 상의 페이지뷰들(pageviews) 및 활성 스코어들(activation scores)을 사용하는, 컴퓨터 구현 방법.

청구항 3

제 1 항에 있어서, 상기 클러스터들의 적어도 일부를 사용하여 미리정의된 택소노미의 적어도 하나의 카테고리를 결정하는 단계는:

i) 상기 스코어된 클러스터들의 세트를 사용하여 하나 이상의 컨셉들(concepts)의 세트를 결정하는 단계, 및

ii) 상기 하나 이상의 컨셉들의 세트의 적어도 일부를 사용하여 상기 적어도 하나의 카테고리를 결정하는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 4

제 1 항에 있어서, 상기 클러스터들의 적어도 일부를 사용하여 미리정의된 택소노미의 적어도 하나의 카테고리를 결정하는 단계는 하나 이상의 카테고리들을 룩업(look up)하기 위해 상기 클러스터들의 적어도 일부의 정보를 사용하는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 5

제 4 항에 있어서, 상기 미리정의된 택소노미는 계층적이며,

상기 클러스터들의 적어도 일부를 사용하여 미리정의된 택소노미의 적어도 하나의 카테고리를 결정하는 단계는,

- 상기 하나 이상의 카테고리들의 적어도 일부에 대해, (1) 상기 카테고리의 인트라-카테고리 클러스터 스코어, 및 (2) 상기 계층적인 택소노미에서의 상기 카테고리의 후손들(descendants)인 카테고리들의 인트라-카테고리 클러스터 스코어들을 포함하는 스코어를 결정하는 단계를 더 포함하는, 컴퓨터 구현 방법.

청구항 6

제 5 항에 있어서, 상기 클러스터들의 적어도 일부를 사용하여 미리정의된 택소노미의 적어도 하나의 카테고리를 결정하는 단계는,

- 미리결정된 임계치보다 큰 결정된 스코어를 갖는 가장 깊은 계층 레벨 카테고리(deepest hierarchical level category)를 결정하는 단계를 더 포함하는, 컴퓨터 구현 방법.

청구항 7

제 1 항에 있어서, 상기 미리정의된 택소노미의 카테고리들은 (A) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 프로덕트들(products), (B) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 서비스들, (C) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 산업들, (D) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 토픽들, 및 (E) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 콘텐트 포맷들(예컨대, 포럼들(forums), 블로그들 등) 중 적어도 하나에 대응하는, 컴퓨터 구현 방법.

청구항 8

미리정의된 택소노미의 하나 이상의 카테고리와 의미 클러스터(semantic cluster)를 연관시키는 컴퓨터 구현 방법에 있어서,

a) 의미 클러스터를 얻는 단계;

b) 상기 얻어진 클러스터를 사용하여 하나 이상의 스코어된 컨셉들의 세트를 식별하는 단계;

c) 상기 하나 이상의 스코어된 컨셉들의 적어도 일부를 사용하여 하나 이상의 카테고리들의 세트를 식별하는 단계; 및

d) 상기 의미 클러스터와 상기 하나 이상의 카테고리들의 적어도 일부를 연관시키는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 9

제 8 항에 있어서, 상기 의미 클러스터는 용어 동시 발생 클러스터(term co-occurrence cluster)인, 컴퓨터 구현 방법.

청구항 10

제 8 항에 있어서, 상기 의미 클러스터는 검색 엔진 상의 검색 세션(search session)에서 동시 발생하는 경향이 있는 용어들을 포함하는, 컴퓨터 구현 방법.

청구항 11

제 8 항에 있어서, 상기 의미 클러스터는 월드 와이드 웹(World Wide Web) 상에서 이용가능한 도큐먼트들(documents)에서 동시 발생하는 경향이 있는 용어들을 포함하는, 컴퓨터 구현 방법.

청구항 12

제 8 항에 있어서, 상기 의미 클러스터와 상기 하나 이상의 카테고리들의 적어도 일부를 연관시키는 단계는 상기 하나 이상의 카테고리들의 적어도 일부의 각각에 상기 의미 클러스터를 매핑하는 인덱스 엔트리(index entry)를 생성 및 저장하는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 13

제 8 항에 있어서, 상기 미리정의된 택소노미의 카테고리들은 (A) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 프로덕트들, (B) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 서비스들, (C) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 산업들, 및 (D) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 토픽들 중 적어도 하나에 대응하는, 컴퓨터 구현 방법.

청구항 14

미리정의된 택소노미의 하나 이상의 카테고리들과 프로퍼티(property)를 연관시키는 컴퓨터 구현 방법에 있어서,

a) 상기 프로퍼티에 관한 정보를 얻는 단계;

b) 상기 얻어진 프로퍼티 정보를 사용하여 하나 이상의 스코어된 의미 클러스터들의 세트를 식별하는 단계;

c) 상기 하나 이상의 스코어된 의미 클러스터들의 적어도 일부를 사용하여 하나 이상의 카테고리들의 세트를 식별하는 단계; 및

d) 상기 프로퍼티와 상기 하나 이상의 카테고리들의 적어도 일부를 연관시키는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 15

제 14 항에 있어서, 상기 프로퍼티는 웹페이지인, 컴퓨터 구현 방법.

청구항 16

제 14 항에 있어서, 상기 프로퍼티는 다수의 웹페이지를 포함하는 웹사이트인, 컴퓨터 구현 방법.

청구항 17

제 14 항에 있어서, 상기 프로퍼티와 상기 하나 이상의 카테고리들의 적어도 일부를 연관시키는 단계는 상기 하나 이상의 카테고리들의 적어도 일부의 각각에 프로퍼티 정보를 매핑하는 인덱스 엔트리를 생성 및 저장하는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 18

제 14 항에 있어서, 상기 미리정의된 택소노미의 카테고리들은 (A) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 프로덕트들, (B) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 서비스들, (C) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 산업들, (D) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 토픽들, 및 (E) 웹사이트 콘텐트에서 발견될 가능성이 있는 관련 콘텐트 포맷들 중 적어도 하나에 대응하는, 컴퓨터 구현 방법.

청구항 19

다수의 엔트리들을 포함하는 기계 판독가능 정보(machine-readable information)를 저장하는 기계 판독가능 매체로서, 각각의 엔트리는 의미 클러스터 정보 및 미리정의된 택소노미의 하나 이상의 연관된 카테고리들을 식별하는 정보를 포함하는, 기계 판독가능 매체.

청구항 20

제 19 항에 있어서, 상기 미리정의된 택소노미의 카테고리들은 (A) 웹사이트에서 발견될 가능성이 있는 관련 프로덕트들, (B) 웹사이트에서 발견될 가능성이 있는 관련 서비스들, (C) 웹사이트에서 발견될 가능성이 있는 관련 산업들, (D) 웹사이트에서 발견될 가능성이 있는 관련 토픽들, 및 (E) 웹사이트에서 발견될 가능성이 있는 관련 콘텐트 포맷들 중 적어도 하나에 대응하는, 기계 판독가능 매체.

청구항 21

제 19 항에 있어서, 상기 의미 클러스터는 용어 동시 발생 클러스터인, 기계 판독가능 매체.

청구항 22

제 19 항에 있어서, 상기 의미 클러스터는 검색 엔진 상의 검색 세션에서 동시 발생하는 경향이 있는 용어들을 포함하는, 기계 판독가능 매체.

청구항 23

제 19 항에 있어서, 상기 의미 클러스터는 월드 와이드 웹 상에서 이용가능한 도큐먼트들에서 동시 발생하는 경향이 있는 용어들을 포함하는, 기계 판독가능 매체.

청구항 24

웹사이트를 자동으로 카테고리화하기 위한 장치에 있어서,

a) 웹사이트 정보를 얻기 위한 수단;

b) 상기 웹사이트 정보를 사용하여 상기 웹사이트에 대한 스코어된 클러스터들의 세트를 결정하기 위한 수단; 및

c) 상기 클러스터들의 세트의 적어도 일부를 사용하여 미리정의된 택소노미의 적어도 하나의 카테고리를 결정하기 위한 수단을 포함하는, 장치.

청구항 25

미리정의된 택소노미의 하나 이상의 카테고리들과 의미 클러스터를 연관시키기 위한 장치에 있어서,

a) 의미 클러스터를 얻기 위한 수단;

b) 상기 얻어진 클러스터를 사용하여 하나 이상의 스코어된 컨셉들의 세트를 식별하기 위한 수단;

c) 상기 하나 이상의 스코어된 컨셉들의 적어도 일부를 사용하여 하나 이상의 카테고리들의 세트를 식별하기 위한 수단; 및

d) 상기 의미 클러스터와 하나 이상의 카테고리들의 적어도 일부를 연관시키기 위한 수단을 포함하는, 장치.

청구항 26

미리정의된 택소노미의 하나 이상의 카테고리들과 프로퍼티를 연관시키기 위한 장치에 있어서,

a) 상기 프로퍼티에 관한 정보를 얻기 위한 수단;

b) 상기 얻어진 프로퍼티 정보를 사용하여 하나 이상의 스코어된 의미 클러스터들의 세트를 식별하기 위한 수단;

c) 상기 하나 이상의 스코어된 의미 클러스터들의 적어도 일부를 사용하여 하나 이상의 카테고리들의 세트를 식별하기 위한 수단; 및

d) 상기 프로퍼티와 상기 하나 이상의 카테고리들의 적어도 일부를 연관시키기 위한 수단을 포함하는, 장치.

명 세 서

기 술 분 야

<1> 본 발명은 정보를 조직화하는 것에 관한 것이다. 특히, 본 발명은 택소노미에 관련하여 용어들(terms), 어구들(phrases), 도큐먼트들(documents) 및/또는 용어 동시 발생 클러스터들(term co-occurrence clusters)을 카테고리화하고, 그러한 카테고리화된 도큐먼트들 및/또는 클러스터들을 사용하는 것에 관한 것이다.

배 경 기 술

<2> "택소노미(taxonomy)"는 카테고리들 또는 클래스들(카테고리화(categorization) 또는 분류법(classification)에 기초한 원리들)의 구조화된, 일반적으로 계층적인 세트이다. 택소노미들은 그것들이 다양한 것들(간단히 "오브젝트들"라 함) 사이의 관계들을 표현하는데 사용될 수 있으므로 유용하다. 예를 들어, 택소노미들은 서로 다른 오브젝트들이 함께 "속하는"지 여부를 결정하거나, 서로 다른 오브젝트들이 얼마나 밀접하게 관련되는지 여부를 결정하는데 사용될 수 있다.

<3> 불행히도, 오브젝트들을 택소노미의 적절한 카테고리 또는 카테고리들에 할당하는 것은 어려울 수 있다. 이것은 서로 다른 타입의 오브젝트들이 택소노미에 할당되는 경우에 특히 참(true)이다. 또한, 이것은 카테고리화에 사용되는 오브젝트들의 속성들이 시간이 지남에 따라 변할 수 있는 경우 또는 많은 오브젝트들이 카테고리화될 다수의 오브젝트들로부터 부가 및/또는 제거되는 경우에 특히 참이다. 예를 들어, 웹사이트들은 월드 와이드 웹(World Wide Web)으로부터 지속적으로 부가되고 제거된다. 또한, 웹사이트의 콘텐트는 자주 변한다. 그러므로, 웹사이트들을 카테고리화하는 것은 매력적일 수 있다.

<4> 앞의 관점에서, 오브젝트들(예컨대, 웹사이트들), 가능하게는 서로 다른 타입의 오브젝트들을 적절한 택소노미의 카테고리들에 할당하기 위한 자동화된 수단을 제공하는 것은 유용하다.

발명의 상세한 설명

<5> 본 발명에 따른 적어도 몇몇 실시예들은 웹사이트를 자동으로 카테고리화할 수 있다. 이러한 실시예들은 (a) 웹사이트 정보를 받고, (b) 웹사이트 정보를 사용하여 웹사이트에 대해 스코어된 클러스터(scored clusters)(예컨대, 의미(semantic), 용어 동시 발생 등)의 세트를 결정하고, (c) 클러스터들의 세트의 적어도 일부를 사용하여 미리정의된 택소노미의 적어도 하나의 카테고리(예컨대, 버티컬 카테고리(vertical category))를 결정함으로써 그렇게 할 수 있다.

<6> 본 발명에 따른 적어도 몇몇 실시예들은 미리정의된 택소노미의 하나 이상의 카테고리들(예컨대, 버티컬 카테고리)과 의미 클러스터(예컨대, 용어 동시 발생 클러스터)를 연관시킬 수 있다. 이러한 실시예들은 (a) 의미 클러스터를 받고, (b) 받은 클러스터를 사용하여 하나 이상의 스코어된 컨셉들(concepts)의 세트를 식별하고, (c) 하나 이상의 스코어된 컨셉들의 적어도 일부를 사용하여 하나 이상의 카테고리들의 세트를 식별하고, (d) 의미

클러스터와 하나 이상의 카테고리들의 적어도 일부를 연관시킴으로써 그렇게 할 수 있다.

<7> 본 발명에 따른 적어도 몇몇 실시예들은 미리정의된 택소노미의 하나 이상의 카테고리들(예컨대, 버티컬 카테고리들)에 프로퍼티(property)(예컨대, 웹사이트)를 연관시킬 수 있다. 이러한 실시예들은 (a) 상기 프로퍼티에 대한 정보를 얻고, (b) 얻어진 프로퍼티 정보를 사용하여 하나 이상의 스코어된 의미 클러스터들(예컨대, 용어 동시 발생 클러스터들)의 세트를 식별하고, (c) 하나 이상의 스코어된 의미 클러스터들의 적어도 일부를 사용하여 하나 이상의 카테고리들(예컨대, 버티컬 카테고리들)을 식별하고, (d) 상기 프로퍼티와 하나 이상의 카테고리들의 적어도 일부를 연관시킴으로써 그렇게 할 수 있다.

실 시 예

<19> 본 발명은 예컨대 카테고리화로부터 유도된 택소노미 및 데이터 구조들에 관련하여, 예를 들어 워드들, 어구들, 도큐먼트들 및/또는 클러스터들과 같은, 오브젝트들을 카테고리화기 위한 신규한 방법들, 장치, 메시지 포맷들과 관련될 수 있다. 아래의 설명은 기술분야의 당업자가 본 발명을 행하고 사용할 수 있도록 제공되며, 특정한 어플리케이션들 및 그것들의 요구사항들의 문맥에서 제공된다. 그러므로, 본 발명에 따른 실시예들의 아래 설명은 예시 및 설명을 제고하지만, 본 발명을 개시되는 정확한 형태로 제한하고자 하는 것이 아니다. 개시된 실시예들의 다양한 변형들은 당업자에게 자명할 것이며, 아래에서 설명되는 일반적인 원리들은 다른 실시예들 및 응용들에 적용될 수 있다. 예를 들어, 비록 일련의 동작들은 흐름도를 참조하여 설명될 수 있지만, 동작들의 순서는 한 동작의 수행이 또 다른 동작의 완성에 의존하지 않을 때 다른 구현들에서 상이할 수 있다. 또한, 비의존성 동작들은 병행하여 수행될 수 있다. 설명에 사용되는 요소(element), 동작 또는 지시는 그와 같이 정확하게 설명되지 않는 경우에 본 발명에 대해 중요하거나 필수적인 것으로 고려되지 않는다. 또한, 여기에서 사용되는 바와 같이, 관사 "a"는 하나 이상의 아이템들을 포함하도록 의도된다. 하나의 아이템만이 의도되는 경우에는, 용어 "하나(one)" 또는 유사한 언어가 사용된다. 그러므로, 본 발명은 도시된 실시예에 제한되지 않으며, 발명자들은 그들의 발명을 설명된 임의의 잠재적인 주제로서 고려한다.

<20> 아래에서, 명세서에서 사용될 수 있는 정의들이 §4.1에 제공된다. 이어서, 본 발명에 따른 예시적인 실시예들이 §4.2에 설명된다. 본 발명에 따른 예시적인 실시예에서의 동작을 설명하는 예가 §4.3에 제공된다. 마지막으로, 본 발명에 관한 몇몇 결론들이 §4.4에서 설명된다.

<21> **§4.1 정의들**

<22> "프로퍼티(property)"는 광고가 제공될 수 있는 어떤 것이다. 프로퍼티는 온라인 콘텐트(예컨대, 웹사이트, MP3 오디오 프로그램, 온라인 게임들 등), 오프라인 콘텐트(예컨대, 신문, 잡지, 영화 제작, 콘서트, 스포츠 이벤트 등) 및/또는 오프라인 오브젝트들(예컨대, 게시판, 경기장 점수판, 및 광고판, 화물 트레일러의 측면 등)을 포함할 수 있다. 콘텐트를 갖는 프로퍼티들(예컨대, 잡지, 신문, 웹사이트, 이메일, 메시지들 등)은 "매체 프로퍼티들"이라 할 수 있다. 비록 프로퍼티들이 오프라인인 그것 자체일 수 있지만, 프로퍼티에 관한 적절한 정보(예컨대, 속성(들), 토픽(들), 컨셉(들), 카테고리(들), 키워드(들), 관련 정보, 지원되는 광고의 타입(들) 등)는 이용가능한 온라인일 수 있다. 예를 들어, 실외 재즈 음악 페스티벌은 토픽들 "음악" 및 "재즈", 콘서트들의 위치, 콘서트들의 시간, 페스티벌에 출현하기로 스케줄된 아티스트들, 이용가능한 광고 장소의 타입들(예컨대, 인쇄된 프린트 내의 장소, 무대 상의 장소, 의자 뒷면 상의 장소, 스폰서들의 오디오 알림들(audio announcements) 등)을 입력했다.

<23> "도큐먼트(document)"는 임의의 기계-판독가능 및 기계-저장가능 작업 프로덕트(work product)를 포함하도록 광범위하게 해석되어야 한다. 도큐먼트는 파일, 파일들의 조합, 다른 파일에의 내장된 링크들을 갖는 하나 이상의 파일들 등일 수 있다. 파일들은 텍스트 HTML, XML, 오디오, 이미지, 비디오 등과 같은 임의의 타입일 수 있다. 최종 사용자에게 렌더(render)되는 도큐먼트들의 일부들은 도큐먼트의 "콘텐트"로서 고려될 수 있다. 도큐먼트는 콘텐트(워드들, 픽처들 등) 및 그 콘텐트를 의미의 일부 표시(예컨대, 이메일 필드 및 관련 데이터, HTML 태그들 및 관련 데이터 등) 둘 모두를 포함하는 "구조화된 데이터(structured data)"를 포함할 수 있다. 도큐먼트에서 광고 장소들은 내장된 정보 또는 지시들에 의해 정의될 수 있다. 인터넷의 문맥에서, 일반적인 도큐먼트는 웹 페이지이다. 웹 페이지들은 종종 콘텐트를 포함하고, 내장된 정보(메타 정보, 하이퍼링크들 등) 및/또는 내장된 지시들(자바스크립트 등)을 포함할 수 있다. 많은 경우에, 도큐먼트는 고유하고 어드레스가능한 저장 위치를 가지며, 그에 따라, 이 어드레스가능한 위치에 의해 고유하게 식별될 수 있다. URL(universal resource locator)은 인터넷 상의 정보에 액세스하는데 사용되는 고유한 어드레스이다. 도큐먼트의 또 다른 예는 다수의 관련된 (예컨대, 링크된) 웹 페이지들을 포함하는 웹사이트이다. 도큐먼트의 또 다른 예는 광고이다.

<24> "웹 도큐먼트"는 웹 상에 공개된 임의의 도큐먼트를 포함한다. 웹 도큐먼트들의 예들은 예컨대, 웹사이트 또는 웹 페이지를 포함한다.

<25> "도큐먼트 정보"는 도큐먼트에 포함된 임의의 정보, 도큐먼트에 포함된 정보로부터 유도가능한 정보("도큐먼트 유도 정보"라 함), 및/또는 도큐먼트에 관련된 정보("도큐먼트 관련 정보"라 함)뿐만 아니라, 이러한 정보의 확장들(예컨대, 관련 정보로부터 유도된 정보)을 포함할 수 있다. 도큐먼트 유도 정보의 예는 도큐먼트의 텍스트 콘텐트에 기초한 분류이다. 도큐먼트 관련 정보의 예들은 인스턴트 도큐먼트(instant document)로의 링크를 갖는 다른 도큐먼트들로부터의 도큐먼트 정보뿐만 아니라, 인스턴트 도큐먼트가 링크하는 다른 도큐먼트들로부터의 도큐먼트 정보를 포함한다.

<26> "버티컬들(verticals)"은 웹사이트 콘텐트에서 또는 웹사이트 콘텐트에 대해 발견될 수 있는 관련된 프로덕트들, 서비스들, 산업들, 콘텐트 포맷들, 시청자 인구통계(audience demographics) 및/또는 토픽들의 그룹들이다.

<27> "클러스터(cluster)"는 서로 밀접하게 발생하는 경향이 있는 요소들의 그룹이다. 예를 들어, 클러스터는 (예컨대, 웹 페이지 상에서, 검색 쿼리들(search queries)에서, 프로덕트 카탈로그들에서, 스피치(speech)에서의 아티클(article)(온라인 또는 오프라인)에서, 토론에서, 또는 이메일 스레드들(e-mail threads)에서 등등) 종종 동시 발생하는 경향이 있는 용어들의 세트일 수 있다.

<28> "컨셉(concept)"은 (특정 언어에서 특정 워드와 같은, 의미의 대행자(agent)에 반대되는 것으로서) 의미의 전달자이다. 그러므로, 예컨대, 단일 컨셉은 임의 수의 언어들로써 표현되거나, 주어진 언어에서 대안의 방식으로 표현될 수 있다. 예를 들어, 워드들 STOP, HALT, ANSCHLAG, ARRESTO, 및 PARADA 모두가 동일한 컨셉에 속한다. 컨셉들은 그것들이 그것들의 확장에서 차이점을 생략하고 마치 그것들이 동일한 것처럼 그것들을 취급한다는 점에서 추상적이다. 컨셉들은 그것들의 확장에서 모든 것에 동일하게 적용한다는 점에서 일반적이다.

<29> "택소노미(taxonomy)"는 카테고리들 또는 분류들(또는 카테고리화 또는 분류법에 기초한 원리들)의 구조적인, 일반적으로 계층적인(그러나 평평(flat)할 수 있는) 세트이다. "카테고리"는 택소노미의 "노드(node)"에 대응할 수 있다.

<30> "스코어(score)"는 오브젝트에 할당된 임의의 숫자 값일 수 있다. 그러므로, 스코어는 공식에 의해 결정된 수를 포함할 수 있고, "공식적인 스코어(formulaic score)"라고 할 수 있다. 스코어는 오브젝트들의 순서화된 세트에서 오브젝트의 랭킹(ranking)을 포함할 수 있고,"차례를 나타내는 스코어(ordinal score)"라고 할 수 있다.

<31> **§4.2 본 발명에 따른 예시적인 실시예**

<32> 도 1은 본 발명에 따른 예시적인 실시예들에 제공될 수 있는 동작들뿐만 아니라, 이러한 동작들에 의해 사용 및/또는 생성될 수 있는 정보를 도시한다. 용어 동시 발생 기반의 클러스터 생성/식별 동작들(term co-occurrence based cluster generation/identification operations))(110)은 문맥 내 용어(terms in context)(105)를 얻고, 용어-클러스터 정보(예컨대, 인덱스)(115)를 생성할 수 있다. 이러한 정보(115)가 일단 생성되면, 용어 동시 발생 생성/식별 동작들(110)은 입력 용어(들)(105)에 응답하여 하나 이상의 클러스터들(예컨대, 용어들)(120)을 식별하는데 사용될 수 있다. 필터링/데이터 감소 동작들(122)은 "양호한" 클러스터들(122)의 서브세트를 생성하는데 사용될 수 있다.

<33> 컨셉 생성/식별 동작들(130)은 클러스터들(120 또는 124)을 얻고, 클러스터-컨셉 정보(예컨대, 인덱스)(135)를 생성할 수 있다. 이러한 정보(135)가 일단 생성되면, 컨셉 생성/식별 동작들(130)은 입력 클러스터들(120,124)에 응답하여 하나 이상의 컨셉들(140)을 식별하는데 사용될 수 있다. 필터링/데이터 감소 동작들(142)은 "양호한" 컨셉들(144)의 서브세트를 생성하는데 사용될 수 있다.

<34> 카테고리 생성/식별 동작들(150)은 컨셉들(140 또는 144)을 얻고, 컨셉-카테고리 정보(예컨대, 인덱스)(155)를 생성할 수 있다. 이러한 정보(155)가 일단 생성되면, 카테고리 생성/식별 동작(150)은 입력 컨셉들(140 또는 144)에 응답하여 하나 이상의 카테고리들(160)을 식별하는데 사용될 수 있다. 이들 카테고리들은 택소노미의 노드들일 수 있다. 카테고리 필터링/감소 동작(162)은 "양호한" 카테고리들(164)의 서브세트를 생성하는데 사용될 수 있다.

<35> 문맥 내 용어들(105)의 많은 예들이 존재한다. 예를 들어, 문맥 내 용어들은 검색 쿼리에 포함되는 워드들 및/또는 어구들일 수 있고, 및/또는 하나 이상의 검색 쿼리들을 포함하는 검색 세션(search session)으로 이루어질 수 있다. 또 다른 예로서, 문맥 내 용어들은 도큐먼트(예컨대, 웹 페이지) 또는 도큐먼트들의 콜렉션

(collection) 또는 그룹(예컨대, 웹사이트)에서 발견되는 워드들 및/또는 어구들일 수 있다. 또 다른 예로서, 문맥 내 용어들은 독창적인 광고에 있어서의 워드들 및/또는 어구들일 수 있다.

<36> 용어 동시 발생 기반의 클러스터 생성/식별 동작(110)을 다시 참조하면, 일부 문맥 또는 문맥들 내 용어들(예컨대, 검색 쿼리들, 검색 세션들, 웹 페이지들, 웹사이트들, 아티클들, 블로그들, 토론 스레드들 등)의 동시 발생은 워드들의 그룹들 또는 클러스터들을 생성하는데 사용될 수 있다. 이들 클러스터들이 일단 정의되면, 워드-대-클러스터 인덱스가 저장될 수 있다. 이러한 인덱스를 사용하면, 주어진 워드 또는 워드들, 워드들을 포함하는 하나 이상의 클러스터들이 식별될 수 있다. 이러한 클러스터들을 생성 및/또는 식별하는데 사용되는 동작들의 예는, 발명자가 조지스 해리크(Georges Harik)와 노암 샤지어(Noam Shazeer)인, 2002년 10월 3일에 출원된 발명의 명칭이 "Methods and Apparatus for Probabilistic Hierarchical Inferential Learner"인 미국 가출원번호 제60/416,144호("'144 가출원"으로서 기재되고, 참조문헌으로써 본 명세서에 포함됨) 및 2003년 9월 30일에 출원된 발명의 명칭이 "Methods and Apparatus for Characterizing Documents Based on Cluster Related Words"인 미국특허출원 제10/676,571호("'571출원"으로서 기재되고, 참조문헌으로써 본 명세서에 포함됨)에 개시되어 있는 바와 같은, 개연적 계층 추론 러너(probabilistic hierarchical inferential learner)("PHIL"라 함)이다.

<37> PHIL의 한가지 예시적인 실시예는 www.google.com 검색 세션들에서 함께 발생하는 경향이 있는 용어들의 상호관련된 클러스터들의 시스템이다. 이러한 클러스터 내의 용어는 클러스터에 대해 그것이 통계적으로 얼마나 중요한지에 의해 가중될 수 있다. 이러한 클러스터들은 소수의 용어들에서부터 수천 개의 용어들까지 가질 수 있다. PHIL 모델의 일 실시예는 수 천만개의 클러스터들을 포함하고, 그것들의 검색 빈도에 비례하여 모든 언어들을 포괄한다. 클러스터들은 STOP(예컨대, 작다는 의미를 전달하는 "the," "a," "an," 등등과 같은 워드들을 대부분 포함함), PORN, NEGATIVE("폭탄," "자살," 등과 같은 부정적이고, 우울하고, 또는 민감한 아티클들에서 종종 나타나는 워드들을 포함함), LOCATION 등과 같은, 응용(예컨대, 온라인 광고 서빙 시스템)에 의해 사용되는 할당된 속성들일 수 있다. PHIL의 또 다른 실시예에서, 유지 및 업데이팅을 단순화하는 모델이 각각의 언어에 대해 유지되다.

<38> PHIL 서버는 도큐먼드(예컨대, 웹페이지)를 입력으로서 취하고, 콘텐트에 "매칭"하는 클러스터들을 리턴할 수 있다. 또한, 그것은 독창적인 광고 및/또는 타겟팅 키워드들(targeting keywords)을 입력으로서 취하고, 매칭하는 클러스터들을 리턴할 수 있다. 그러므로, 그것은 웹페이지들의 콘텐트에 광고를 매칭시키는데 사용될 수 있다.

<39> 컨셉 생성/식별 동작들(130) 및 카테고리 생성/식별 동작들(150)을 다시 참조하면, 이들 동작들은 하나 이상의 클러스터들을 얻고, 택소노미의 하나 이상의 카테고리들(예컨대, 노드들)을 식별할 수 있다. 용어 동시 발생 클러스터 식별 동작들(110)과 함께 사용될 때, 이들 동작들(130,150)은 하나 이상의 용어들을 얻고, 택소노미의 하나 이상의 카테고리들을 식별할 수 있다.

<40> 카테고리들을 생성 및/또는 식별하는데 사용되는 동작들(130,150)의 예는 발명자들이 아담 웨이스맨(Adam Weissman)과 길래드 이스라엘 엘배즈(Gilad Isreal Elbaz)이고, 발명의 명칭이 "Meaning-Based Information Organization and Retrieval"인, 미국특허 제6,453,315호(참조문헌으로써 본 명세서에 포함됨) 및 발명자들이 아담 웨이스맨 및 길래드 이스라엘 엘배즈이고, 발명의 명칭이 "Meaning-Based Advertising and Document Relevant Determination"인, 미국특허 제6,816,857호(참조문헌으로써 본 명세서에 포함됨)에 개시되어 있는 바와 같이, 의미 인식 엔진이다.

<41> 예시적인 의미 인식 엔진(이하 "시르카디아(Circadia)"라 함)은 도큐먼트를 시험하고 그것을 임의의 택소노미로 카테고리화할 수 있다. 시르카디아는 수천 만개의 상호관련된 컨셉들 및 대응하는 용어들의 소유권 온톨로지(proprietary ontology)를 포함한다. 시르카디아 온톨로지의 컨셉들은 언어 독립적(language-independent)이다. 언어 특정인(language-specific) 용어들은 이들 컨셉들에 관련된다. 시르카디아 서버는 두 개의 주요 동작들("감지" 및 "탐색")을 지원한다. 이 감지 동작은 입력으로서 도큐먼트 또는 텍스트의 스트링(string)을 얻고, 출력으로서 입력에 대한 컨셉들의 가중된 세트("기스트(gist)"라 함)를 리턴할 수 있다. 그러므로, 시르카디아에서의 감지 동작은 컨셉 식별 동작들(130)의 예이다. 이 기스트는 탐색 요청 입력으로서 사용될 수 있다. 응답에서, 최상의 카테고리들 및 특정된 택소노미에서의 그것들 각각의 의미 스코어들은 리턴된다. 그러므로, 시르카디아에서의 탐색 동작은 카테고리 식별 동작들(150)(및 아마도 카테고리 필터링/감소 동작(162))의 예이다. 자연히, "ODP"(Open Directory Project) 택소노미와 같은 다른 택소노미들, "SIC"(Standard Industrial Classification) 택소노미 등이 사용될 수 있다.

<42> 도 2는 본 발명에 따른 예시적인 실시예들에 제공될 수 있는 동작들뿐만 아니라, 택소노미의 카테고리들과 클러스터들(예컨대 워드들 및/또는 용어들의 세트들)을 연관(예컨대, 매핑 또는 인덱싱)시키기 위해, 이러한 동작들에 의해 사용 및/또는 생성될 수 있는 정보를 도시한다. 클러스터 대 택소노미 카테고리 연관 생성 동작들(220)은 클러스터 정보(210)를 얻고, 클러스터-카테고리 정보(230)를 생성한다. 예를 들어, 동작들(220)은 클러스터 정보(예컨대, 클러스터 식별자들)를 컨셉 식별 동작들(130')로 보낼 수 있고, 이것은 하나 이상의 컨셉들을 얻기 위해 클러스터-컨셉 정보(예컨대, 인덱스)(135')를 사용할 수 있다. 이러한 동작들(130')은 컨셉(들)을 클러스터 대 택소노미 카테고리 연관 생성 동작들(220)에 리턴할 수 있다. 이들 동작들(220)은 컨셉 정보(예컨대, 컨셉 식별자들)를 카테고리 식별 동작들(150')로 보낼 수 있고, 그것은 하나 이상의 카테고리들을 얻기 위해 컨셉-카테고리 정보(예컨대, 인덱스)(155')를 사용할 수 있다. 이러한 동작들(150')은 클러스터 대 택소노미 카테고리 연관 생성 동작들(220)에 카테고리(들)를 리턴할 수 있다. 얻어진 클러스터 정보(210) 및 리턴된 카테고리 정보를 사용하여, 동작들(220)은 클러스터-대-카테고리 연관 정보(예컨대 매칭 또는 인덱스)(230)를 생성할 수 있다.

<43> 도시된 바와 같이, 본 발명에 관련된 적어도 한 실시예에서, 정보(230)는 복수의 엔트리들(232)을 포함하는 테이블일 수 있다. 엔트리들(232) 각각은 택소노미(236)의 하나 이상의 카테고리들(각각에 대한 식별자) 및 클러스터 식별자(234)를 포함할 수 있다. 도시되어 있지는 않지만, 각각의 카테고리를 하나 이상의 클러스터들에 매핑하는 반전된 인덱스(inverted index)가 또한 생성 및 저장될 수 있다.

<44> 도 3은 본 발명에 따른 예시적인 실시예들에 제공될 수 있는 동작들뿐만 아니라, 택소노미의 카테고리들과 도큐먼트(예컨대, 웹페이지들, 웹사이트들, 독창적인 광고) 정보를 연관시키기 위해, 이러한 동작들에 의해 사용 및/또는 생성될 수 있는 정보를 도시한다. 도큐먼트 대 택소노미 카테고리 연관 생성 동작들(320)은 도큐먼트 정보(320)를 얻고, 도큐먼트-대-카테고리 정보(330)를 생성한다. 예를 들어, 동작들(320)은 도큐먼트 정보를 클러스터 식별 동작들(110')로 보낼 수 있고, 그것은 하나 이상의 클러스터들을 식별하기 위해 용어 대 클러스터 정보(예컨대, 인덱스)(115')를 사용할 수 있다. 이러한 동작들(110')은 클러스터(들)를 도큐먼트 대 택소노미 카테고리 연관 생성 동작들(320)에 리턴할 수 있다. 이 동작들(320)은 클러스터 정보(예컨대, 클러스터 식별자들)를 컨셉 식별 동작들(130')로 보낼 수 있고, 그것은 하나 이상의 컨셉들을 얻기 위해 클러스터-컨셉 정보(예컨대, 인덱스(135'))를 사용할 수 있다. 이러한 동작들(130')은 컨셉(들)을 도큐먼트 대 택소노미 카테고리 연관 생성 동작들(320)로 리턴할 수 있다. 이들 동작들(320)은 컨셉 정보(예컨대, 컨셉 식별자들)를 카테고리 식별 동작들(150')로 보낼 수 있고, 그것은 하나 이상의 카테고리들을 얻기 위해 컨셉-카테고리 정보(예컨대, 인덱스)(155')를 사용할 수 있다. 이러한 동작들(150')은 카테고리(들)를 도큐먼트 대 택소노미 카테고리 연관 생성 동작들(320)에 리턴할 수 있다. 얻어진 도큐먼트 정보(310) 및 리턴된 카테고리 정보를 사용하여, 동작들(320)은 도큐먼트 대 카테고리 연관 정보(예컨대, 매핑 또는 인덱스)(330)를 생성할 수 있다.

<45> 도시된 바와 같이, 본 발명에 따른 적어도 하나의 실시예에서, 정보(330)는 다수의 엔트리들(332)을 포함하는 테이블일 수 있다. 엔트리들(332) 각각은 택소노미(336)의 하나 이상의 카테고리들(각각에 대한 식별자) 및 도큐먼트 식별자(334)를 포함할 수 있다. 도시되어 있지는 않지만, 각각의 카테고리를 하나 이상의 도큐먼트들에 매핑하는 반전된 인덱스가 또한 생성 및 저장될 수 있다.

<46> 도 4는 본 발명에 따른 예시적인 실시예들에 제공될 수 있는 대안의 동작들뿐만 아니라, 택소노미의 카테고리들과 도큐먼트들(예컨대, 웹페이지, 웹사이트들, 독창적인 광고)을 연관(예컨대, 매핑 또는 인덱싱)시키기 위해, 이러한 동작들에 의해 사용 및/또는 생성될 수 있는 정보를 도시한다. 도큐먼트 대 택소노미 카테고리 연관 생성 동작들(420)은 도큐먼트 정보(420)를 얻고, 도큐먼트-대-카테고리 정보(430)를 생성한다. 예를 들어, 동작들(420)은 도큐먼트 정보를 클러스터 식별 동작들(110')로 보낼 수 있고, 그것은 하나 이상의 클러스터들을 식별하기 위해 용어 대 클러스터 정보(예컨대, 인덱스)(115')를 사용할 수 있다. 이러한 동작들(110')은 도큐먼트 대 택소노미 카테고리 연관 생성 동작들(420)에 클러스터(들)를 리턴할 수 있다. 이들 동작들(420)은 클러스터-대-카테고리 정보(230')를 사용하여 하나 이상의 연관된 카테고리들을 찾기 위해 클러스터 정보(예컨대, 클러스터 식별자들)를 사용할 수 있다. 이 정보(230')는 예컨대 도 2에 도시된 매핑일 수 있다. 보다 특별히는, 각각의 클러스터 식별자는 하나 이상의 연관된 카테고리들(예컨대 도 2의 234 및 236을 상기하자)을 룩업(lookup)하기 위해 사용될 수 있다. 얻어진 도큐먼트 정보(410) 및 카테고리 정보를 사용하여, 동작들(420)은 도큐먼트-대-카테고리 연관 정보(예컨대, 매핑 또는 인덱스)(430)를 생성할 수 있다.

<47> 도시된 바와 같이, 도 3의 예시적인 실시예의 경우와 같이, 본 발명에 따른 적어도 하나의 실시예에서, 정보(430)는 다수의 엔트리들(432)을 포함하는 테이블일 수 있다. 엔트리들(432) 각각은 도큐먼트 식별자(434) 및 택소노미(436)의 하나 이상의 카테고리들(각각에 대한 식별자)을 포함할 수 있다. 도시되어 있지는 않지만, 각

각의 카테고리를 하나 이상의 도큐먼트들에 매핑하는 반전된 인덱스가 또한 생성 및 저장될 수 있다.

<48> **§4.2.1 예시적인 방법들**

<49> 도 5는 본 발명에 따른 방식으로, 하나 이상의 카테고리들과 하나 이상의 클러스터들을 연관시키는데 사용될 수 있는 예시적인 방법(500)의 흐름도이다. 도 2를 다시 참조하면, 방법(500)은 동작들(220)을 수행하기 위해 사용될 수 있다. 방법(500)의 주요 작용들은 다수의 클러스터들 각각에 대해 수행될 수 있다. 대안으로, 클러스터들은 그룹핑될 수 있고, 그룹으로 처리 및 취급된다. 하지만, 방법(500)의 설명을 단순화하기 위해, 단일 클러스터의 처리가 설명된다. 클러스터가 얻어지고(블록 510), 하나 이상의 컨셉들의 세트가 클러스터를 사용하여 식별된다(블록 520). 식별된 컨셉(들)은 감소 및/또는 필터링될 수 있다(블록 530). 이어서, 하나 이상의 카테고리들의 세트는 식별된 컨셉들을 사용하여 식별될 수 있다(블록 540). 식별된 카테고리(들)는 감소 및/또는 필터링될 수 있다(블록 550). 마지막으로, 얻어진 클러스터는, 방법(500)을 떠나기 전에(노드 570), 식별된(아마도 필터링된) 카테고리(들)와 연관될 수 있다(블록 560).

<50> 블록 510을 참조하면, 클러스터는 PHIL 클러스터, 또는 예컨대 검색 쿼리들 또는 검색 세션들에서 동시 발생하는 경향이 있는 용어들의 세트일 수 있다. 클러스터는 도큐먼트들에서 동시 발생하는 경향이 있는 용어들의 세트일 수 있다.

<51> 블록 530을 참조하면, 컨셉들은 예컨대, 그것들을 스코어링하고, 컨셉 스코어들을 하나 이상의 임계치들(절대적 및/또는 상대적)에 적용하고, 상부(top) N개의 스코어링 컨셉들만을 취하거나, 또는 앞의 것들을 조합함으로써 필터링 및/또는 감소될 수 있다. 유사하게, 블록 550을 참조하면, 카테고리들은 예컨대, 그것들을 스코어링하고, 카테고리 스코어들을 하나 이상의 임계치들(절대적 및/또는 상대적)에 적용하고, 상부 M개의 스코어링 컨셉들만을 취하거나, 또는 앞의 것들을 조합함으로써 필터링 및/또는 감소될 수 있다.

<52> 괄호로 표시된 바와 같이, 작용들(520-550)은 얻어진 클러스터를 사용하여 하나 이상의 카테고리들을 식별하는 단일 작용으로 조합될 수 있다. 하지만, 시르카디아는 "탐색" 동작에 선행하는 "감지" 동작을 사용하여 카테고리화하도록 설계된다. 클러스터들에서 카테고리들로 바로 가기보다는, 클러스터들로부터 컨셉들을 가장먼저 식별하고 이어서 컨셉들로부터 카테고리들을 식별하는 한가지 이점은, 중간 컨셉들("기스트")이 저장되면, 그것들이 감시 동작을 반복할 필요 없이 임의의 다수의 이용가능한 택소노미들을 분류하기 위해 직접 사용될 수 있다는 것이다. 즉, 컨셉이 일단 결정되면, 용어들, 카테고리들, 다른 컨셉들 등을 얻는 것이 용이하다.

<53> 블록 560을 참조하면, 클러스터는 클러스터(식별자)를 하나 이상의 카테고리들(식별자들)에 매핑하는 인덱스를 생성 및 저장함으로써 하나 이상의 카테고리들과 연관될 수 있다. 대안으로, 또는 부가적으로, 카테고리(식별자)를 하나 이상의 클러스터들(식별자들)에 매핑하는 반전된 인덱스가 생성 및 저장될 수 있다.

<54> 블록 510을 참조하면, 클러스터는 상부 T(예컨대, 50)개의 용어들(예컨대, 인터 클러스터 스코어링 및/또는 인트라 클러스터 스코어링에 기초함)만을 포함하도록 세분화될 수 있다. 여기에서, 인트라 클러스터 스코어링은 용어가 클러스터에서 나타나는 횟수가 증가함에 따라 증가할 수 있고, 도큐먼트(예컨대, 웹페이지들, 검색 쿼리들, 검색 세션들) 콜렉션에서 용어가 나타나는 횟수가 증가함에 따라 감소할 수 있다. 그러므로, 인트라 클러스터 스코어는 예컨대, count_in_cluster/count_in_search_query_collection으로서 정의될 수 있다. 또한, 각각의 클러스터에 대한 상부 용어들의 수(T)는 각 클러스터에 대해 고정된 동일한 수의 용어들 이외에, 인트라 클러스터 파이어링(intra cluster firing)에 기초하여 결정될 수 있다. '571 출원에 사용된 클러스터 스코어링들이 또한 사용될 수 있다.

<55> 도 5의 블록들 520-550을 다시 참조하면, 본 발명에 따른 적어도 하나의 예시적인 실시예에서, 컨셉들은 아래와 같이 시르카디아 서버를 사용하여 클러스터들로부터 결정될 수 있다.

<56> 도 5의 블록 520을 다시 참조하면, 시르카디아를 사용하는 카테고리화에 있어 제 1 단계는 "기스트"를 리턴하는, "감지" 동작을 행하는 것이다. 기스트는 시르카디아 온톨로지로부터 컨셉 매칭들의 내부 가중된 세트(internal weighted set)이다. 그러므로, 각 클러스터에 대한 기스트(예컨대, 50개의 용어들에 기초함)가 얻어진다.

<57> 도 5의 블록들 540 및 550을 다시 참조하면, 제 2 단계는 기스트가 주어지면, 열거된 택소노미로부터 상부 N(예컨대, N=2)개 및 대응하는 의미 스코어들을 요청하기 위해"탐색" 동작을 행하는 것을 포함한다.

<58> 본 발명에 따른 적어도 하나의 예시적인 실시예에서, 상부의 두 개의 카테고리들 및 그것들의 대응하는 의미 스코어들은 탐색 동작으로부터 요청된다. 이러한 예시적인 실시예(들)에서, 이들 상부 두 개의 카테고리들은 "기

본적인" 카테고리(상부 스코어링 하나에 대해) 및 각각의 클러스터에 대한 "2차적인" 카테고리로서 언급된다. 시르카디아가 클러스터에 대한 임의의 카테고리를 결정하지 않으면, 클러스터는 "NONE"의 기본적인 및 2차적인 카테고리들을 수신한다. 시르카디아가 기본적인 카테고리뿐만 아니라, 2차적인 카테고리를 결정하면, 2차적인 카테고리는 "NONE"로 설정된다.

<59> 블록 550을 다시 참조하면, 본 발명에 따른 적어도 하나의 실시예는 임계치보다 작은 스코어들을 갖는 카테고리들을 필터링한다. 임계치는 미리정해진 임계치일 수 있다. 또한, 임계치는, 사실상 시르카디아 콜(Circadia call)에 대해 통계적으로 중요한 측정치의 종류로서 각 클러스터에서 용어들의 수를 고려하는 원 클러스터(original cluster)에서 보다 많은 용어들이 존재하면 보다 낮게 설명될 수 있다. 예를 들어, 클러스터가 M(예컨대, 50)개 이상의 용어들을 가지면, 그것들의 상부 50만을 사용하여 양호한 대표적인 샘플들을 제공하고, 그것은 임계치를 느슨해지게 한다는 것을 확신할 수 있다. 하지만, 클러스터가 M개보다 작은 용어들을 가지면, 용어들의 샘플들이 보다 작으므로 임계치를 올리는 것을 권할만하고, 클러스터의 의미있는 용어들을 거의 포함하지 않을 수 있다.

<60> 도 6은 본 발명에 따른 방식에서, 하나 이상의 카테고리들과 하나 이상의 도큐먼트들을 연관시키는데 사용될 수 있는 예시적인 방법(600)의 흐름도이다. 도 3을 다시 참조하면, 방법(600)은 동작들(320)을 수행하는데 사용될 수 있다. 방법(600)의 주요 작용들은 다수의 도큐먼트들 각각에 대해 수행될 수 있다. 대안으로, 도큐먼트들은 그룹핑되고, 그룹으로서 처리 및 취급된다. 하지만, 방법(600)에 대한 설명을 단순화하기 위해, 단일 도큐먼트의 처리가 설명된다. 도큐먼트가 얻어지고(블록 610), 하나 이상의 클러스터들의 세트가 얻어진 도큐먼트(예컨대, 도큐먼트의 용어들)를 사용하여 식별된다(블록 620). 이어서, 클러스터(들)는 필터링 및/또는 감소될 수 있다(블록 630). 이어서, 하나 이상의 컨셉들의 세트는 클러스터들을 사용하여 식별된다(블록 640). 식별된 컨셉(들)은 감소 및/또는 필터링될 수 있다(블록 650). 이어서, 하나 이상의 카테고리들의 세트는 식별된 컨셉들을 사용하여 식별될 수 있다(블록 660). 식별된 카테고리(들)는 감소 및/또는 필터링될 수 있다(블록 670). 마지막으로, 얻어진 도큐먼트는, 방법(600)을 떠나기(노드 690) 전에, 식별된 카테고리(들)와 연관될 수 있다(블록 680).

<61> 블록 610을 참조하면, 도큐먼트는 웹페이지, 웹페이지로부터 추출된 콘텐트, 웹페이지의 일부(예컨대, 참조 또는 링크의 앵커 텍스트(anchor text)), 웹사이트, 웹사이트의 일부, 광고의 독창적인 텍스트 등일 수 있다.

<62> 블록 630을 참조하면, 클러스터들은 예를 들어, 그것들을 스코어링하고, 클러스터 스코어들을 하나 이상의 임계치들(절대적 및/또는 상대적)에 적용하고, 상부 N 개의 스코어링 클러스터만을 취하고, 또는 앞의 것들의 임의 조합을 취함으로써 필터링 및/또는 감소될 수 있다. 유사하게, 블록 650을 참조하면, 컨셉들은 예컨대, 그것들을 스코어링하고, 컨셉 스코어들을 하나 이상의 임계치들(절대적 및/또는 상대적)에 적용하고, 상부 N 개의 스코어링 컨셉들만을 취하고, 또는 앞의 것들의 임의 조합을 취함으로써 필터링 및/또는 감소될 수 있다. 유사하게, 블록 670을 참조하면, 카테고리들은 예컨대, 그것들을 스코어링하고, 카테고리 스코어들을 하나 이상의 임계치들(절대적 및/또는 상대적)에 적용하고, 상부 M 개의 스코어링 컨셉들만을 취하고, 또는 앞의 것들의 임의 조합을 취함으로써 필터링 및/또는 감소될 수 있다.

<63> 괄호로 나타내지는 바와 같이, 비록, 위에서 소개된 이유들에 대해 중간 컨셉들(예컨대, "기스트")을 결정하는 것이 유용할 수 있지만, 작용 640-670은 식별된 클러스터(들)를 사용하여 하나 이상의 카테고리들을 식별하는 단일 작용으로 조합될 수 있다.

<64> 블록 680을 참조하면, 도큐먼트는 도큐먼트(식별자)를 하나 상의 카테고리들(식별자들)에 매핑하는 인덱스를 생성 및 저장함으로써 하나 이상의 카테고리들과 연관될 수 있다. 대안으로, 또는 부가적으로, 카테고리(식별자)를 하나 이상의 도큐먼트들(식별자들)에 매핑하는 반전된 인덱스가 생성 및 저장될 수 있다.

<65> 도 7은 본 발명에 따른 방식에서, 하나 이상의 카테고리들과 하나 이상의 도큐먼트들을 연관시키는데 사용될 수 있는 예시적인 방법(700)의 흐름도이다. 도 4를 다시 참조하면, 방법(700)은 동작들(420)을 수행하는데 사용될 수 있다. 방법(700)의 주요 작용들은 다수의 도큐먼트들 각각에 대해 수행될 수 있다. 대안으로, 도큐먼트들은 그룹핑되고, 그룹으로서 처리 및 취급된다. 하지만, 방법(700)에 대한 설명을 단순화하기 위해, 단일 도큐먼트의 처리가 설명된다. 도큐먼트가 얻어지고(블록 710), 하나 이상의 클러스터들의 세트가 얻어진 도큐먼트(예컨대, 도큐먼트의 용어들)를 사용하여 식별된다(블록 720). 이어서, 클러스터(들)는 필터링 및/또는 감소될 수 있다(블록 730). 이어서, 하나 이상의 카테고리들의 세트는 식별된 클러스터들 및 클러스터 대 카테고리 연관 정보(cluster-to-category association information)를 사용하여 식별될 수 있다(블록 740). 식별된 카테고리(들)는 감소 및/또는 필터링될 수 있다(블록 750). 마지막으로, 얻어진 도큐먼트는, 방법(700)을 떠나

기(노드 770) 전에, 식별된 카테고리(들)와 연관될 수 있다(블록 760).

<66> 블록 710을 참조하면, 도큐먼트는 웹페이지, 웹페이지로부터 추출된 콘텐트, 웹페이지의 일부(예컨대, 참조 또는 링크의 앵커 텍스트), 웹사이트, 웹사이트의 일부, 광고의 독창적인 텍스트 등일 수 있다.

<67> 블록 730을 참조하면, 클러스터들은 예를 들어, 그것들을 스코어링하고, 클러스터 스코어들을 하나 이상의 임계치들(절대적 및/또는 상대적)에 적용하고, 상부 N 개의 스코어링 클러스터만을 취하고, 또는 앞의 것들의 임의 조합을 취함으로써 필터링 및/또는 감소될 수 있다. 유사하게, 블록 750을 참조하면, 카테고리들은 예컨대, 그것들을 스코어링하고, 카테고리 스코어들을 하나 이상의 임계치들(절대적 및/또는 상대적)에 적용하고, 상부 M 개의 스코어링 컨셉들만을 취하고, 또는 앞의 것들의 임의 조합을 취함으로써 필터링 및/또는 감소될 수 있다.

<68> 블록 740을 참조하면, 클러스터-대-카테고리 연관 정보는 다수의 클러스터들 각각을 하나 이상의 카테고리들에 매핑하는 인덱스일 수 있다. (예컨대, 도 2의 230 및 도 5의 560을 상기하자)

<69> 블록 760을 참조하면, 도큐먼트는 도큐먼트(식별자)를 하나이 상의 카테고리들(식별자들)에 매핑하는 인덱스를 생성 및 저장함으로써 하나 이상의 카테고리들과 연관될 수 있다. 대안으로, 또는 부가적으로, 카테고리(식별자)를 하나 이상의 도큐먼트들(식별자들)을 매핑하는 반전된 인덱스가 생성 및 저장될 수 있다.

<70> **§4.4.4 예시적인 장치**

<71> 도 25는 위에서 논의된 하나 이상의 동작들을 수행할 수 있는 기계(2500)의 블록도이다. 기계(2500)는 하나 이상의 프로세서들(2510), 하나 이상의 입력/출력 인터페이스 유닛들(2530), 하나 이상의 저장 디바이스들(2520), 및 결합된 요소들 사이에서 정보의 통신을 용이하게 하기 위한 하나 이상의 시스템 버스들 및/또는 네트워크들(2540)을 포함한다. 하나 이상의 입력 디바이스들(2532) 및 하나 이상의 출력 디바이스들(2534)은 하나 이상의 입력/출력 인터페이스들(2530)과 결합될 수 있다.

<72> 하나 이상의 프로세서들(2510)은, 본 발명의 하나 이상의 특징들에 영향을 미치도록, 기계 실행가능한 지시들(machine-executable instructions)(예컨대, 캘리포니아, 팔로 알토의 선 마이크로시스템즈로부터 이용가능한 솔라리스 오퍼레이팅 시스템(Solaris operating system) 상에서 운용되는 C 또는 C++, 노스 캘리포이아, 더럼(Durham)의 레드 해트(Red Hat). 인크(Inc)와 같은 다수의 판매자들로부터 광범위하게 이용가능한 리눅스 오퍼레이팅 시스템, 자바(Java), 어셈블리, 펄(Perl) 등)을 실행할 수 있다. 기계 실행가능 지시들의 적어도 일부는 하나 이상의 저장 디바이스들(2520) 상에 (일시적으로 또는 보다 영구적으로)저장될 수 있고, 또는 하나 이상의 입력 인터페이스 유닛들(2530)을 통해 외부 소스로부터 수신될 수 있다.

<73> 일 실시예에서, 기계(2500)는 하나 이상의 종래 개인용 컴퓨터, 모바일 전화기들, PAD들 등일 수 있다. 종래 개인용 컴퓨터의 경우에, 처리 유닛(2510)은 하나 이상의 마이크로프로세서들일 수 있다. 버스(2540)는 시스템 버스를 포함할 수 있다. 저장 디바이스들(2520)은 ROM(read only memory) 및/또는 RAM(random access memory)과 같은, 시스템 메모리를 포함할 수 있다. 저장 디바이스들(2530)은 또한, 하드디스크로부터 판독하고 하드디스크에 기록하기 위한 하드디스크 드라이브, (제거가능한) 자기 디스크로부터 판독하거나 그것에 기록하기 위한 자기 디스크 드라이브, 및 콤팩트디스크 또는 다른(자기 광 매체와 같은 제거가능한(자기) 광디스크 등을 포함할 수 있다.

<74> 사용자는 예컨대 키보드 및 포인팅 디바이스(예컨대, 마우스)와 같은, 입력 디바이스들(2532)을 통해 개인용 컴퓨터에 명령들 및 정보를 입력할 수 있다. 마이크로폰, 조이스틱, 게임 패드, 위성 접시, 스캐너 등과 같은 다른 입력 디바이스들이 또한 (또는 대안으로) 포함될 수 있다. 여러 가지 입력 디바이스들은 종종, 시스템 버스(2540)에 결합된 적절한 인터페이스(2530)를 통해 처리 유닛(들)(2510)에 접속된다. 출력 디바이스들(2534)은 모니터 또는 다른 타입의 디스플레이 디바이스를 포함할 수 있고, 또한, 적절한 인터페이스를 통해 시스템 버스(2540)에 접속될 수 있다. 모니터에 부가하여(또는 대신에), 개인용 컴퓨터는 예컨대 스피커들 및 프린터들과 같은 다른 (주변) 출력 디바이스들을 포함할 수 있다.

<75> 자연히, 위에서 설명된 입력 및 출력 수단들의 대다수는 본 발명에 따른 실시예의 적어도 몇몇 특징들의 문맥에서 필요하지 않을 수 있다.

<76> 위에서 설명된 다양한 동작들은 하나 이상의 기계들(2500)에 의해 수행될 수 있고, 위에서 설명된 다양한 정보는 하나 이상의 기계들(2500) 상에 저장될 수 있다. 이러한 기계들(2500)은 예컨대, 인터넷과 같은, 하나 이상의 네트워크들과 접속될 수 있다.

<77> **§4.2.3 개량 및 대안**

<78> 많은 실시예들이, 도큐먼트들, 특히 웹사이트 및 웹페이지와 같은 온라인 프로퍼티들의 문맥에서 설명되었지만, 본 발명에 따른 적어도 몇몇 실시예들은 심지어 비-매체 프로퍼티들(non-media properties)을 포함한, 오프라인 프로퍼티들을 지원할 수 있다.

<79> **§4.2.3.1 예시적인 인덱스 데이터 구조들**

<80> 도 8 내지 도 17은 다양한 예시적인 매핑들을 도시하며, 하나 이상의 매핑들이 본 발명에 따른 다양한 실시예들에서 인덱스들로서 저장될 수 있다. 도 8은 로드(예컨대, 알파벳 스트링, 음소 스트링(phonemic string), 용어, 어구 등)에서 하나 이상의 클러스터들(예컨대, PHIL 클러스터(들))의 세트로의 매핑을 도시한다. 도 9는 클러스터로부터 하나 이상의 워드들로의 매핑을 도시한다. 도 10은 도큐먼트(예컨대, 웹페이지(또는 그것의 부분)), 웹사이트(또는 그것의 부분), 앵커 텍스트, 독창적인 광고 텍스트 등)에서 택소노미의 하나 이상의 카테고리들의 세트로의 매핑을 도시한다. (예컨대, 도 3의 330 및 332, 도 4의 430 및 432를 상기하자). 도 11은 택소노미의 카테고리에서 하나 이상의 도큐먼트들의 세트로의 매핑을 도시한다. 도 12는 클러스터에서 택소노미의 하나 이상의 카테고리들로의 매핑을 도시한다. (예컨대, 도 2의 230 및 232, 도 4의 230'을 상기하자). 도 13은 택소노미의 카테고리에서 하나 이상의 클러스터들로의 매핑을 도시한다. 도 14는 도큐먼트에서 하나 이상의 클러스터들의 세트로의 매핑을 도시한다. 도 15는 클러스터에서 하나 이상의 도큐먼트들의 세트로의 매핑을 도시한다. 도 16은 워드(예컨대, 알파벳 스트링, 음소 스트링, 용어, 어구 등)에서 택소노미의 하나 이상의 카테고리들의 세트로의 매핑을 도시한다. 도 17은 택소노미의 카테고리에서 하나 이상의 워드들의 세트로의 매핑을 도시한다.

<81> **§4.2.3.2 카테고리들을 임의의 클러스터들에 할당하기 위해 클러스터 속성들을 사용**

<82> 본 발명에 따른 적어도 하나의 실시예에서, 하나 이상의 클러스터들은 이러한 클러스터(들)에 대한 자동적인 카테고리 결정을 무효로 하는(또는 보충하는), 택소노미의 하나 이상의 카테고리들에 수동으로 매핑될 수 있다. 예를 들어, 이러한 실시예에서, PORN 속성을 갖는 클러스터들은, 자동으로 결정된 카테고리가 상이한 경우에도, "/Adult/Porn" 카테고리에 할당될 수 있다. 유사하게, NEGATIVE 속성을 갖는 클러스터들은, 자동으로 결정된 카테고리가 상이한 경우에도, "News & Current Events/News Subjects(Sensitive)" 카테고리에 할당될 수 있다. 유사하게, LOCATION 속성을 갖는 클러스터들은, 자동으로 결정된 카테고리가 상이한 경우에도, "/Local Services/City & Regional Guides/LOC(Locations)" 카테고리에 할당될 수 있다. 이러한 클러스터들은 수동으로 생성되고, 수동으로 수정되고, 및/또는 수동으로 리뷰(review)될 수 있다.

<83> **§4.2.3.3 콘텐트 관련 광고 서빙 로그들(content-relevant ad serving logs)로부터 웹사이트 클러스터 매핑 및 스코어들을 추출**

<84> 도 1의 용어 클러스터 정보(인덱스)(115)를 다시 참조하면, 클러스터들의 가중된 세트는 아래와 같이, 웹사이트들(예컨대, 캘리포니아, 마운티 뷰의 구글로부터 AsSense와 같은, 콘텐트 관련 광고 서빙 네트워크에 참여하는 웹사이트들)을 위해 생성될 수 있다.

<85> 로그 기록(log record)은 웹페이지 디스플레이(예컨대, AsSense) 광고들에 대한 각각의 페이지뷰(pageview)를 위해 생성될 수 있다. 웹페이지를 위해 스코어된(PHIL) 클러스터들의 세트는 그 로그 기록으로 기록될 수 있다. 주어진 웹페이지에 대해, 다수(예컨대, 1과 12개 사이)의 클러스터들이 존재할 수 있고, 각각의 클러스터는 연관된 활성 스코어(activation score)를 갖는다. (예컨대, "활성"을 설명하는 '571 출원을 참조). 활성 스코어는 분석되는 클러스터에 대해 주어진 클러스터가 개념적으로 얼마나 중요한지에 대한 측정치이다. 낮은 값의 활성 스코어들은 낮은 개념적인 중요도를 나타내고, 높은 값의 활성 스코어들은 높은 개념적인 중요도를 나타낸다.

<86> **§4.2.3.4 각각의 웹사이트에 대해 스코어된 클러스터들의 세트를 결정**

<87> (위에서 논의된 바와 같은)웹페이지에 대해 적어도 미리정해진 값(예컨대 1.0)의 활성 스코어를 갖지 않는 클러스터들은 무시될 수 있다. (도 1의 동작들(122)을 상기하자). 미리정해진 값은 서빙 광고들에서 광고 서빙 시스템에 의해 사용되는 최소 임계치로 설정될 수 있다. 임의의 특별한 경우 클러스터들(예컨대, STOP으로서 마킹된 것들)은 또한 무시될 수 있다.

<88> 남아 있는 클러스터들("자격부여 클러스터들(qualifying clusters)"이라 함)에 대해, 이들 클러스터들의 활성 스코어들의 합은 결정될 수 있다. 웹페이지에 대한 각각의 자격부여 클러스터는 "스코어"를 얻는다. 클러스터 스코어는 (a) 웹페이지 상의 자격부여 클러스터의 활성 스코어와 (b) 웹사이트가 수신한 페이지뷰들의 수의 곱(product)으로서 정의될 수 있다.

<89> 아래의 예는 자격부여된 클러스터들이 위에서 설명된 것과 같이 어떻게 스코어될 수 있는지를 나타낸다. 주어진 클러스터(c_1)가 웹사이트 내에서 두 개의 웹페이지 상에 활성된다고 가정하자. 클러스터가 웹페이지(p_1) 상에서 10.0의 활성 스코어를 갖고, 웹페이지(p_2) 상에서 20.0의 활성 스코어를 갖는다고 가정하자. 주일(week) 동안, 웹페이지(p_1)는 1000개의 페이지뷰들을 쉰하고, 웹페이지(p_2)는 1500개의 페이지뷰들을 수신한다. 주일동안 웹사이트에 대한 클러스터 스코어와 페이지뷰 곱들의 합은 100,000이다. 이어서, 클러스터는 웹사이트에 대해 아래의 전체 스코어를 수신한다.

<90> 스코어 = ((10.0 활성/페이지뷰*1000페이지뷰들) + (20.0 활성/페이지뷰*1500페이지뷰들))/100,000활성

<91> = (10,000+30,000)/100,000

<92> = 0.4

<93> 이것은 웹사이트의 개별 웹페이지 상에서 페이지뷰들 및 활성 스코어들에 의해 웹사이트에 대한 총 클러스터 스코어들을 효과적으로 가중한다. 웹사이트에 대한 클러스터 스코어들의 세트는 1로 합산한다. 이 접근법의 한가지 단점은, 주어진 웹페이지에 대한 높은 트래픽(traffic)이, 그 웹페이지가 카테고리화 관점에서 낮은 트래픽 웹페이지보다 대표적이라는 것을 필연적으로 의미하는 것이 아니라는 점이다. 그러므로, 페이지뷰 파마미터를 조절하고 또는 클러스터 웹페이지 활성 스코어에 보다 높은 가중을 부여하는 것이 바람직할 수 있다. 자연히, 활성 스코어들은 본 발명에 따른 실시예가 사용되는 문맥에서 적당한 하나 이상의 팩터들(factors)의 함수로서 가중될 수 있다.

<94> 웹사이트에 대한 스코어된 클러스터들의 세트가 얻어진 후에, 클러스터들의 수는 상부 S(예컨대, 25)개의 가장 높은 스코어링 클러스터들(S개의 클러스터들보다 작은 웹사이트들의 모든 클러스터들)만을 선택함으로써 감소될 수 있다. 이 세트는 스코어의 용어들에서 식별된 세트의 상부 Y%(예컨대, 70%)를 보충하는 가장 큰 스코어링 클러스터들만을 유지함으로써 추가로 감소될 수 있다.

<95> 남아있는 클러스터들의 스코어들은 그것들이 1로 합산하도록 정규화될 수 있다.

<96> **§4.2.3.5 각각의 웹사이트에 대한"최상의" 카테고리들을 결정**

<97> 도 1의 동작들(162)을 다시 참조하면, 카테고리들(예컨대, 기본적인 및 2차적인 카테고리들)의 감소된 세트는 각각의 웹사이트에 대해 결정될 수 있다. 이 동작에 입력으로서 기능하는 카테고리들(160)은 웹사이트(이미 위에서 설명됨)에 대해 스코어된 카테고리들(PHIL 클러스터들과 연관되고, 아래에서 "클러스터 카테고리들"로서 언급됨)의 삭감된 세트일 수 있다. 통상적으로, 본 발명에 따른 하나의 예시적인 실시예에서, 웹사이트당 약 10개의 클러스터 카테고리들의 최종 세트가 존재한다. 일반적으로, 클러스터 카테고리들의 중첩이 존재하지만, 각각의 클러스터들이 완전히 상이한 카테고리들을 갖는 것이 가능하다.

<98> 본 발명에 따른 일 실시예에서, 카테고리들은 브랜치(branch)당 Z(예컨대 5) 레벨들까지 포함하는 계층적인 택소노미의 파트(part)이다. 이러한 실시예에서, 택소노미의 서로 다른 "브랜치들" 중에서 결정하는 것을 제외하고는, 브랜치 중에서 최상의 레벨이 또한 결정된다. 예를 들어, 카테고리가 "/Automotive" 브랜치 내의 어딘가에 있다는 것이 명백하지만, 문제는 "Automtive", "/Automotive/Auto Parts", "/Automotive/Auto Parts/Vehicle Tires", 또는 "/Automotive/Vehicle Maintenance" 중 어느 것이 최상인가 하는 것이다. 각각의 입력 클러스터에 대한 스코어는 웹사이트의 전체 카테고리화에 대해, 그것의 대응하는 기본적인 및 2차적인 클러스터 카테고리들의 중요도에 기여한다.

<99> 얼마나 많은 클러스터 카테고리들이 웹사이트 카테고리화를 위해 서로 경쟁하는지에 상관없이, 그것들 중 어느 것도 메리트(merit)가 선택되는데에 충분한 개념적인 중요성(예컨대, 그 카테고리에 대한 스코어들의 합에 의해 측정되는 것과 같이)을 갖지 않는 것이 가능하다. 달리 말해서, 가능한 카테고리들은 "승리(win)"하기 위해 임의의 단일의 하나에 대한 웹사이트 중에서 매우 희석된다(dilute). 본 발명에 따른 적어도 몇몇 실시예들에서, 이 최소 개념 중요성(minimum conceptual significance)은 임계값(예컨대, 부동 소수점 십진법(floating point decimal)으로서 저장됨)을 설정함으로써 요구사항으로서 강제될 수 있다. 주어진 웹사이트에 대한 클러스터 스코어들이 1로 합산하도록 정규화되면, 적어도 몇몇 실시예들에서, 0.24 또는 약 0.24의 최소 개념 중요성 임계값은 양호한 결과들을 생성할 수 있다. 이것은, 기본적인 또는 2차적인 카테고리에 대한 최상의 후보가 0.24보다 작은 합산된 스코어를 가지면, "NONE"의 카테고리가 할당된다는 것을 의미한다. 이 임계값이 웹사이트 상의 클러스터들을 스코어하는데 사용되는 방법에 기초하여 조정될 수 있다는 것에 유의하자.

<100> 적어도 몇몇 실시예들에서, 웹사이트들을 카테고리화하기 위해 기본적인 또는 2차적인 클러스터 카테고리들 둘 모두를 사용하는 대신에, 웹사이트들을 카테고리화하기 위해 2차적인 클러스터 카테고리들을 생략하는 것이 바람직할 수 있다.

<101> 이하의 용어가 아래의 예시적인 실시예에 대한 설명에서 사용된다. 폼(form) /level-1/level-2/.../level-m의 계층적인 카테고리 경로(여기서, m은 경로에서 가장 깊은 레벨의 수)가 주어지면, "subsume-level-n"은, 만약 n<m이면, 레벨-n까지의 경로의 서브썸션(subsumption)이고, 만약 n >= m이면, 경로의 서브썸션이 아니다. 예를 들어, n<m인 경우에 대해, 카테고리 경로 "/Automotive/Auto Parts/Vehicle Tires"의 서브썸션-레벨-2는 ""/Automotive/Auto Parts/"이다. 또 다른 예로서, n >= m 인 경우에 대해, "/Automotive/Auto Parts/Vehicle Tires"의 서브썸션-레벨-4는 변경되지 않은, 바로"/Automotive/Auto Parts/Vehicle Tires" 자체이다.

<102> 레벨-n 카테고리는 그 자체의 인트라 카테고리 클러스터 스코어(들)뿐만 아니라 임의의 서브썸되고 보다 깊은 층인 카테고리들의 스코어들을 포함한다. 이들 클러스터 스코어들의 합은 레벨-n 카테고리에 대해 "self&subsumed category cluster score"(또는 "S&S category cluster score")로서 언급된다.

<103> 얼마나 많은 카테고리들이 도큐먼트(예컨대, 웹사이트) 카테고리화를 위해 서로 경재하는지에 상관없이, 그것들 중 어느 것도, 메리트가 선택되는데에 충분한, S&S 카테고리 클러스터 스코어에 의해 측정된 개념적인 중요성을 갖지 않는 것이 가능하다. 달리 말해서, 웹사이트에 대한 클러스터들은 명백히 웹사이트에 대한 최상의 카테고리로서 고려될 임의의 하나의 카테고리에 대하여 가능한 카테고리들 중에서 매우 희석된다.

<104> 본 발명에 따른 적어도 몇몇 실시예들에서, 최소 개념 중요성 요구사항(minimum conceptual significance requirement)은 임계값의 세팅을 통해 부과될 수 있다. 자연히, 보다 높은 서브썸-레벨들에서 임계치를 통과하는(pass) 카테고리들을 얻는 것이 보다 쉬운데, 왜냐하면, 그것들이 보다 일반적인 카테고리들에 대응하기 때문이다. 본 발명에 따른 몇몇 실시예들에서, 임계값은 다양한 서브썸-레벨들에 걸치는 전체 품질(quality)을 최소화하도록 선택되지만, 그러한 카테고리들이 비록 가장 적절할 수 있지만, 카테고리화 서브썸-레벨 스코어들이 높은 레벨들보다 낮은 레벨들에서 자연적으로 보다 낮아지게 되므로, 보다 낮은 서브썸-레벨쪽으로 약간 바이어스된다.

<105> 본 발명의 하나의 예시적인 실시예에서, 주어진 웹사이트에 대한 클러스터 스코어들이 1(위에서 언급된 바와 같음)로 합산하고, 500 노드들에 속하여 5 층 카테고리 택소노미가 사용된다고 가정하면, 약 0.24의 최소 개념 중요성 임계값은 잘 작용된다. 0.20 내지 0.30의 최소 개념 중요성 임계값이 작 작용한다고 믿어진다. 이것은, 주어진 서브썸-레벨에서 기본적인 또는 2차적인 카테고리에 대한 최상의 후보가 임계치보다 작은 합산된 스코어를 가지면, "NONE"의 카테고리가 할당된다는 것을 의미한다. 적절한 임계치의 결정은 카테고리화되는 도큐먼트 상의 클러스터들을 스코어하는데 사용된 방법에 의존할 수 있다.

<106> 몇몇 용어들을 소개하면, 본 발명에 따른 방식에서, 도큐먼트들에 대한 "최상의" 카테고리들을 결정하는 예시적인 방법이 이제 설명된다. t를 최소 개념 중요성 임계값이라 하자. d를 택소노미에서 가장 깊은 레벨이라고 하자. "최상의" 기본적인 카테고리는 아래와 같이 결정될 수 있다. 최상의 서브썸-레벨-1 및 그것에 대응하는 S&S 카테고리 클러스터 스코어가 결정된다. 이것은 레벨 d까지 모든 라벨들에 대해 반복된다. 최상의 서브썸-레벨-p 카테고리가 S&S 카테고리 클러스터 스코어≥t를 갖는 p의 가장 큰(가장 깊은) 값이 선택된다. 대안으로, S&S 카테고리 클러스터 스코어들은 가장 깊은 카테고리 레벨에서 상부(가장 일반적인) 카테고리 레벨로 분석된다. 이런 식으로, 상기 방법은, S&S 카테고리 스코어 ≥ t인 레벨을 처리한 후에 정지한다. ν(최상의 기본적인 카테고리)를 최상의 서브썸-레벨-p 카테고리라고 하고, 또한, 카테고리가 임계치를 만족하지 못하는 경우 "NONE"이다.

<107> "최상의" 2차적인 카테고리는 아래와 같이 정의될 수 있다. 만약, ν(최상의 기본적인 카테고리)가 "NONE"이면, 최상의 2차적인 카테고리는 "NONE"일 것이다. 만약 ν이 "NONE"이면, 최상의 서브썸-레벨-1 및 그것에 대응하는 서브썸-레벨-1-스코어가 결정되고, 여기서 서브썸-레벨-1은 ν와 같지 않다. 이것은 레벨 d까지의 모든 레벨들에 대해, ν와 같지 않은 서브썸-레벨-n의 제한(restriction)에 의해 반복되고 강제된다. 최상의 서브썸-레벨-q 카테고리가 S&S 카테고리 클러스터 스코어 >= t를 갖는 q의 가장 큰(가장 깊은) 값이 선택된다. w(최상의 2차적인 카테고리)를 최상의 서브썸-레벨-q 카테고리라고 하고, 또는 카테고리가 임계치를 만족하지 못하는 경

우 "NONE"이다.

<108> **§4.3 본 발명에 따른 예시적인 실시예에서 동작들의 예**

<109> 도 8 내지 도 23은 본 발명에 따른 예시적인 사용자 인터페이스의 다양한 디스플레이 스크린들을 도시한다. 도 18은 사용자가 블록(1810)에서 택소노미(이 경우에, "기본적인 버티컬 노드 명칭(primary vertical node name)"의 카테고리를 입력하는 스크린(1800)을 도시한다. 응답에서, 다양한 PHIL 클러스터들(1820)이 출력된다. (이 예에서, 클러스터 명칭은 단순히 클러스터에서 6개의 가장 중요하거나 가장 높은 스코어링 용어들이다). 이 출력은 예컨대, 도 13에 도시된 바와 같은 매핑들을 포함하는 인덱스를 사용하여 생성될 수 있다. 버티컬 노드(즉, 택소노미의 카테고리)와 클러스터의 연관은 체크 박스들(1830)에 의해 나타내지는 바와 같이, 수동적인 승인을 받을 수 있다.

<110> 도 19는 사용자가 블록(1910)에서 웹사이트(홈페이지) 어드레스를 입력할 수 있는 스크린(1900)을 도시한다. 응답에서, 다양한 PHIL 클러스터들(1920)이 출력된다. 이 출력은 예컨대, 도 14에 도시된 바와 같은 매핑들을 포함하는 인덱스를 사용하여 생성될 수 있다. 도큐먼트(예컨대, 웹사이트)와 클러스터의 연관은 체크 박스들(1930)에 의해 나타내진 바와 같은 수동적인 승인을 받을 수 있다.

<111> 도 20은 사용자가 관련된 버티컬 카테고리들과 웹사이트들을 얻기 위해 블록(2010)에서 하나 이상의 워드들을 입력할 수 있는 스크린(2000)을 도시한다. 도 21은 출력 버티컬 카테고리들(2110)과 웹사이트들(2120)을 포함하는 스크린(2100)을 도시한다. 예컨대, 도 8 및 도 12에 도시된 바와 같은 매핑들을 포함하는 인덱스들이 입력 워드로부터 카테고리들의 세트를 출력하는데 사용된다. 대안으로, 웹사이트들의 워드들의 인덱스들이 (예컨대, 검색 엔진에서)공통이므로, 박스(2010)에서의 워드들은 하나 이상의 웹사이트들의 세트로 매핑되고, 그것의 일부는 택소노미의 카테고리들을 얻기 위해 도 10에 도시된 바와 같은 매핑들을 포함하는 인덱스와 함께 사용될 수 있다. 도시된 바와 같이, 웹사이트 정보(2120)는 웹사이트 명칭들(2122)과 스코어들(2124)을 포함할 수 있다.

<112> 도 22는 사용자가 관련된 버티컬 카테고리들과 웹사이트들을 얻기 위해 블록(2210)에서 하나 이상의 웹사이트들을 입력할 수 있는 스크린(2200)(도 18의 스크린(1800)과 같음)을 도시한다. 도 23은 출력 버티컬 카테고리들(2310)과 웹사이트들(2320)을 포함하는 스크린(2300)을 도시한다. 예를 들어, 도 10에 도시된 바와 같은 매핑들을 포함하는 인덱스는 입력 웹사이트로부터 카테고리들의 세트를 출력하는데 사용된다. 또한, 도 11에 도시된 바와 같은 매핑들을 포함하는 인덱스는 결정된 카테고리(들)로부터 추가적인 웹사이트(들)를 생성하는데 사용된다. 도시된 바와 같이, 웹사이트 정보(2320)는 웹사이트 명칭들(2322)과 스코어들(2324)을 포함할 수 있다.

<113> 앞의 예들이 설명하는 바와 같이, 다양한 인덱스들이 사용되고 또는 제 1 타입의 입력 오브젝트들로부터 제 2 타입의 관련된 오브젝트들을 얻기 위해 조합하여 사용될 수 있다. 다양한 타입들의 오브젝트들은 택소노미의 카테고리들(예컨대, 노드들)과 연관될 수 있다.

<114> 웹사이트의 기본적인 및 2차적인 카테고리를 선택하기 위한, 위의 §4.2.3.5에서 설명된 바와 같은, 예시적인 기술을 나타내는 예는 도 24를 참조하여 설명된다. 0.24의 임계치가 사용된다고 가정하자. 또한, 웹사이트에 대한 클러스터들 및 대응하는 기본적인 카테고리들과 클러스터-카테고리 스코어가 아래와 같다고 가정하자.

<115> 클러스터 ID	기본적인 카테고리	클러스터 스코어
<116> 6937542	/컴퓨터 및 테크놀러지(2410)	0.13
<117> 6922978	/컴퓨터 및 테크놀러지/가전제품/	
<118>	오디오 장비/MP3 플레이어(2448)	0.14
<119> 6976937	/컴퓨터 및 테크놀러지/가전제품/	
<120>	카메라 및 캠코더/캠코더(2442)	0.07
<121> 6922928	/컴퓨터 및 테크놀러지/가전제품/	
<122>	카메라 및 캠코더/캠코더(2444)	0.06
<123> 6922526	/컴퓨터 및 테크놀러지/가전제품/	
<124>	카메라 및 캠코더/캠코더(2442)	0.09

<125> 6946862 /컴퓨터 및 테크놀러지/가전제품/

<126> 퍼스널 일렉트로닉스(2432) 0.16

<127> 6926006 /컴퓨터 및 테크놀러지/가전제품/

<128> 퍼스널 일렉트로릭스/휴대용 및 PAD(2434) 0.06

<129> 6922985 /컴퓨터 및 테크놀러지/하드웨어/데스크탑(2434) 0.08

<130> 6922448 /컴퓨터 및 테크놀러지/하드웨어/랩탑(2435) 0.05 0.05

<131> 6936814 /뉴스 및 현재 이벤트/뉴스 소스(도시되지 않음) 0.16

<132> 기본적인 카테고리의 유도시에 포함되는 중간 결과들은 아래와 같다.

<133> 서브썸-레벨 1 카테고리: /컴퓨터 및 테크놀러지

<134> S&S 카테고리 클러스터 스코어: 0.84

<135> 서브썸-레벨 2 카테고리: /컴퓨터 및 테크놀러지/가전제품

<136> S&S 카테고리 클러스터 스코어: 0.58

<137> 서브썸-레벨 3 카테고리: /컴퓨터 및 테크놀러지/가전제품/카메라 및 캠코더

<138> S&S 카테고리 클러스터 스코어: 0.22

<139> 서브썸-레벨 4 카테고리: /뉴스 및 현재 이벤트/뉴스 소스

<140> S&S 카테고리 클러스터 스코어: 0.16

<141> 서브썸-레벨 5 카테고리: /뉴스 및 현재 이벤트/뉴스 소스

<142> S&S 카테고리 클러스터 스코어: 0.16

<143> 층 4 및 5 카테고리들에서, n>m임의 유의하자. 앞의 예에서, 승리하는 기본적인 카테고리(winning Primary Category)는 그것이 0.24의 임계치를 초과하는 S&S 카테고리 클러스터 스코어를 갖는 가장 깊은(가장 특정한) 레벨이었으므로, "/컴퓨터 및 테크놀러지/가전제품"이다.

<144> **§4.4 결론**

<145> 앞에서 이해되는 바와 같이, 본 발명에 따른 몇몇 실시예들은 택소노미의 카테고리들(노드들)을 서로 다른 타입들의 오브젝트들을 연관시키는데 사용될 수 있다. 이들 연관들이 일단 행해지면, 본 발명에 따른 몇몇 실시예들은 택소노미의 카테고리들과 오브젝트들 사이의 연관을 사용하여, 아마도 서로 다른 타입의 "관련된" 오브젝트들을 찾는데 사용될 수 있다. 예를 들어, 본 발명에 따른 실시예들은 표준화된 산업 버티컬 카테고리들(standardized industry vertical categories)의 계층적인 택소노미로 웹사이트들이 카테고리화되게 하는데 사용될 수 있다. 이러한 계층적인 택소노미는 잠재적으로 많이 사용된다. 또한, 서로 다른 타입들의 오브젝트들(예컨대, 광고들, 쿼리들, 웹페이지들, 웹사이트들 등)이 카테고리화될 수 있으면, 이들 서로 다른 타입들의 오브젝트들 사이의 관계들(예컨대, 유사성들)이 결정되고, (예컨대, 웹페이지 또는 웹사이트에 적절한 광고들을 결정하는데 또는 그 반대의 경우에)사용될 수 있다.

<146> 이 택소노미로 웹사이트들 및 클러스터들을 카테고리화한 후에, 다른 차원들(예컨대, 언어, 국가 등)이(예컨대, 온라인 분석 처리(online analytical processing:OLAP) 데이터베이스 및 데이터 웨어하우징 스타 스키머들(data warehousing star schemas)의 방식으로)부가될 수 있다. 카테고리 차원은 계층적인 레벨들에 의해 정의될 수 있지만, 언어와 같은 다른 차원들의 일부는 평평하다. 이들 다양한 차원들을 유도한 후에, 메트릭들(metrics)(예컨대, 페이지뷰들, 광고 느낌, 광고 클릭들, 비용 등)이 그것들로 집결(aggregate)될 수 있다.

도면의 간단한 설명

<8> 도 1은 본 발명에 따른 예시적인 실시예들에 제공될 수 있는 동작들뿐만 아니라, 이러한 동작들에 의해 사용 및/또는 생성될 수 있는 정보를 도시하는 도면.

<9> 도 2는 본 발명에 따른 예시적인 실시예들에 제공될 수 있는 동작들뿐만 아니라, 택소노미의 카테고리들과 클러

스터들(예컨대, 워드들(words) 및/또는 용어들의 세트들)을 연관시키기 위해, 이러한 동작들에 의해 사용 및/또는 생성될 수 있는 정보를 도시하는 도면.

<10> 도 3은 본 발명에 따른 예시적인 실시예들에 제공될 수 있는 동작들뿐만 아니라, 택소노미의 카테고리들과 도큐먼트들을 연관시키기 위해, 이러한 동작들에 의해 사용 및/또는 생성될 수 있는 정보를 도시하는 도면.

<11> 도 4는 본 발명에 따른 예시적인 실시예들에 제공될 수 있는 동작들뿐만 아니라, 택소노미의 카테고리들과 도큐먼트들을 연관시키기 위해, 이러한 동작들에 의해 사용 및/또는 생성될 수 있는 정보를 도시하는 도면.

<12> 도 5는 본 발명에 따른 방식으로, 하나 이상의 택소노미 카테고리들과 하나 이상의 클러스터들을 연관시키기 위해 사용될 수 있는 예시적인 방법(500)의 흐름도.

<13> 도 6은 본 발명에 따른 방식으로, 하나 이상의 택소노미 카테고리들과 하나 이상의 도큐먼트들을 연관시키기 위해 사용될 수 있는 예시적인 방법(600)의 흐름도.

<14> 도 7은 본 발명에 따른 방식으로, 하나 이상의 택소노미 카테고리들과 하나 이상의 클러스터들을 연관시키기 위해 사용될 수 있는 예시적인 방법(700)의 흐름도.

<15> 도 8 내지 도 17은 본 발명에 따른 인덱스들(indexes)로서 저장될 수 있는 다양한 예시적인 매핑들(mappings)을 도시하는 도면.

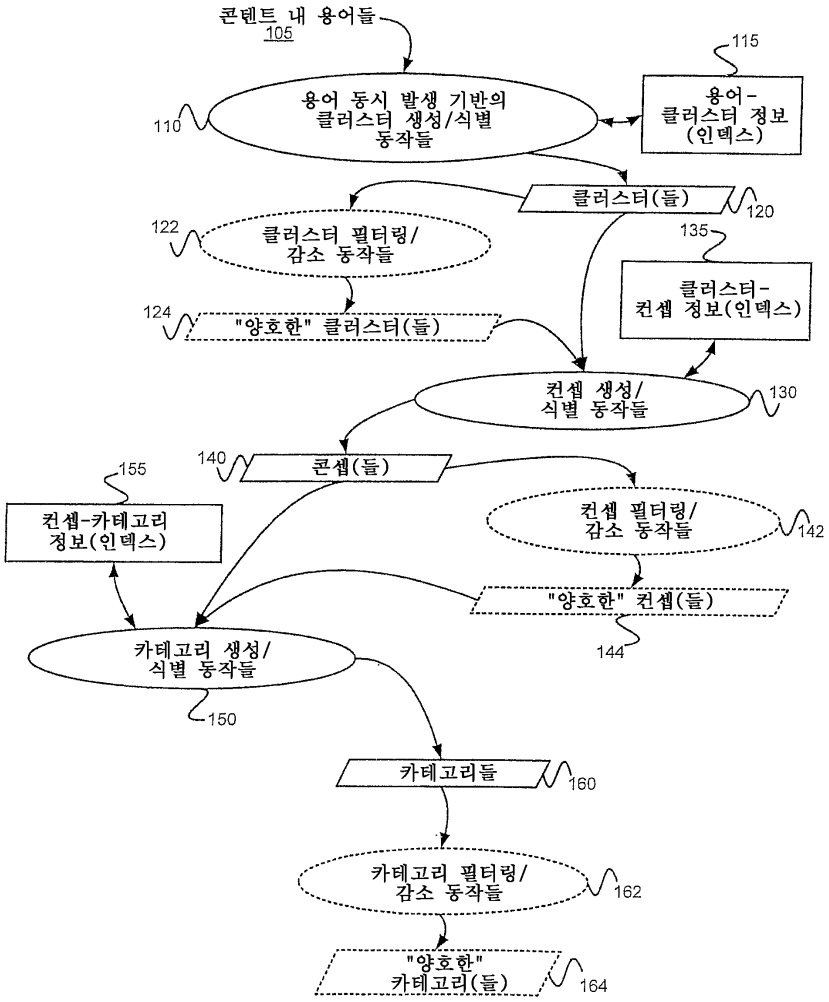
<16> 도 18 내지 도 23은 본 발명에 따른 예시적인 사용자 인터페이스의 다양한 디스플레이 스크린들을 도시하는 도면.

<17> 도 24는 본 발명에 따른 예시적인 실시예를 사용하여 "최상"의 카테고리가 어떻게 결정될 수 있는지를 설명하기 위해 사용되 택소노미의 부분을 도시하는 도면.

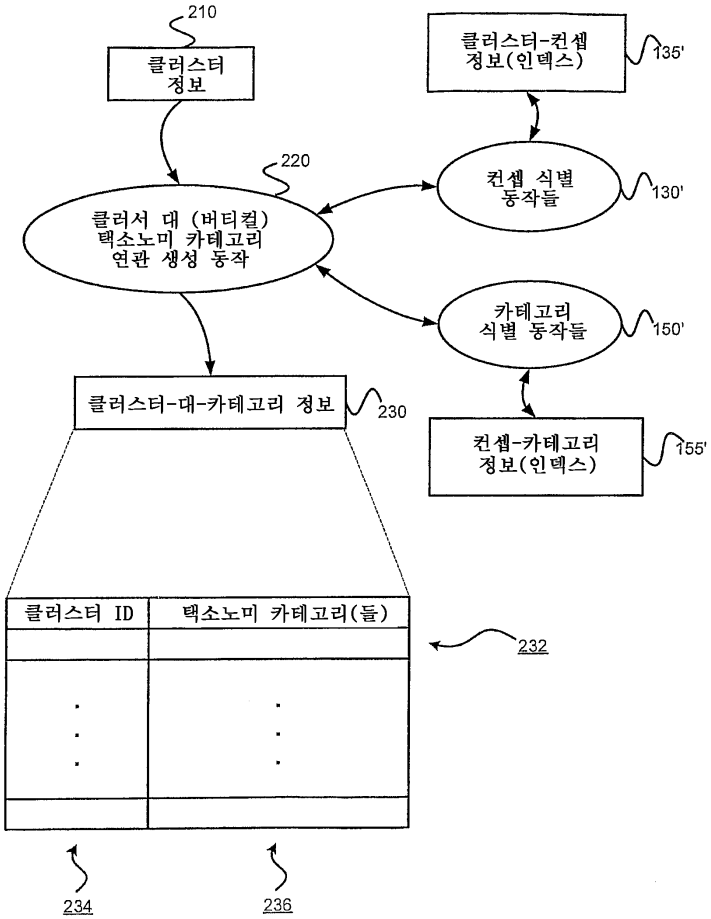
<18> 도 25는 본 발명에 따른 예시적인 실시예들에서 동작들을 수행 및/또는 정보를 저장하기 위해 사용될 수 있는 예시적인 장치의 블록도.

도면

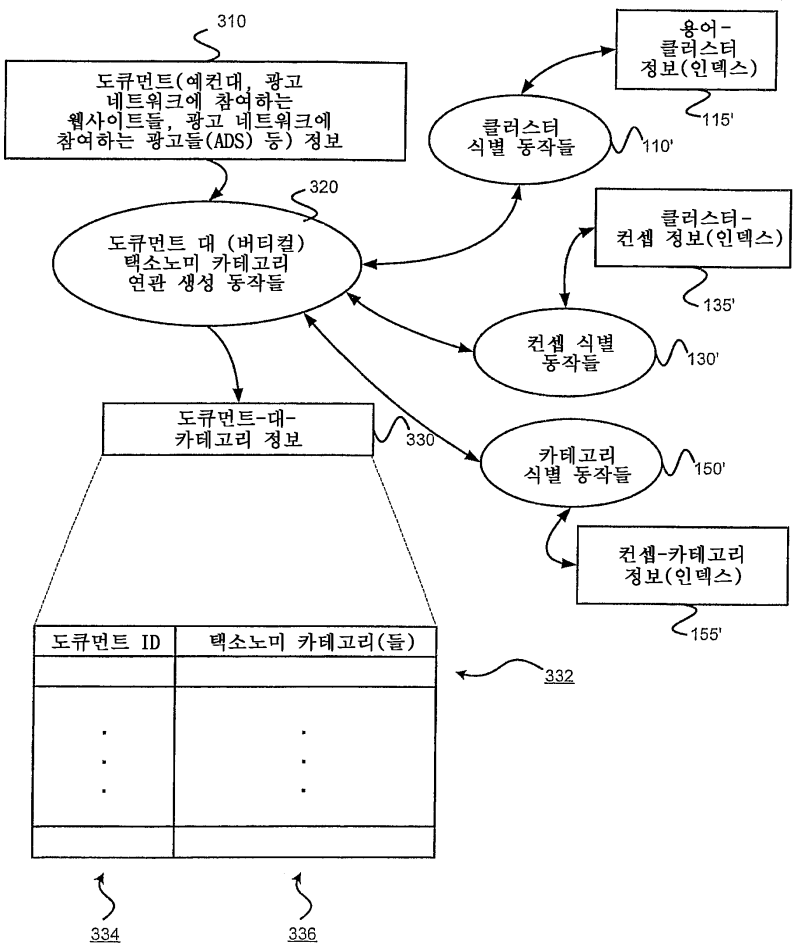
도면1


콘텐트 내 용어들
105
115
용어 동시 발생 기반의
클러스터 생성/식별
동작들
110
용어-
클러스터 정보
(인덱스)
클러스터(들)
120
122
클러스터 필터링/
감소 동작들
135
클러스터-
컨셉 정보(인덱스)
124
"양호한" 클러스터(들)
컨셉 생성/
식별 동작들
130
140
콘셉(들)
155
컨셉-카테고리
정보(인덱스)
컨셉 필터링/
감소 동작들
142
"양호한" 컨셉(들)
144
카테고리 생성/
식별 동작들
150
카테고리들
160
카테고리 필터링/
감소 동작들
162
"양호한"
카테고리(들)
164

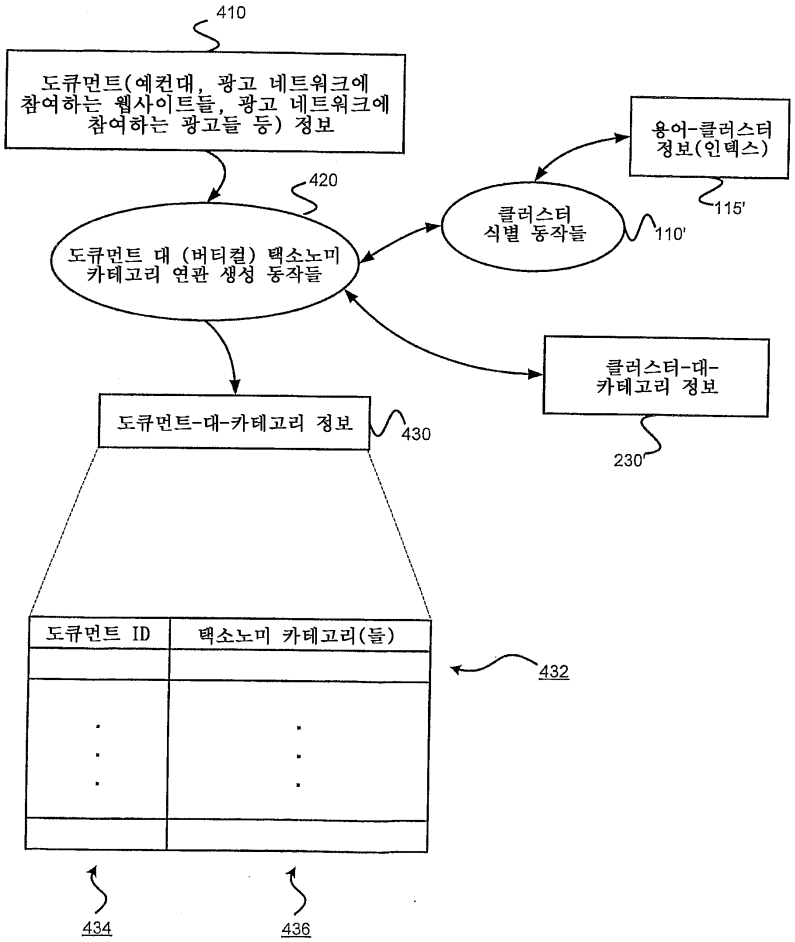
도면2


210
클러스터 정보
클러스터-컨셉 정보(인덱스)
135'
220
컨셉 식별 동작들
130'
클러서 대 (버티컬) 택소노미 카테고리 연관 생성 동작
카테고리 식별 동작들
150'
클러스터-대-카테고리 정보
230
컨셉-카테고리 정보(인덱스)
155'
클러스터 ID
택소노미 카테고리(들)
232
234
236

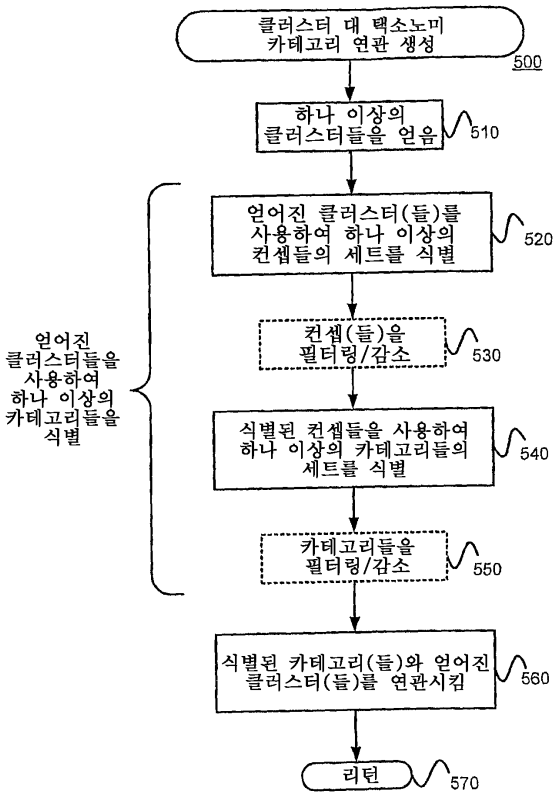
도면3


310
도큐먼트(예컨대, 광고 네트워크에 참여하는 웹사이트들, 광고 네트워크에 참여하는 광고들(ADS) 등) 정보
320
도큐먼트 대 (버티컬) 택소노미 카테고리 연관 생성 동작들
용어-클러스터 정보(인덱스)
115'
클러스터 식별 동작들
110'
클러스터-컨셉 정보(인덱스)
135'
컨셉 식별 동작들
130'
도큐먼트-대-카테고리 정보
330
카테고리 식별 동작들
150'
컨셉-카테고리 정보(인덱스)
155'
도큐먼트 ID
택소노미 카테고리(들)
332
334
336

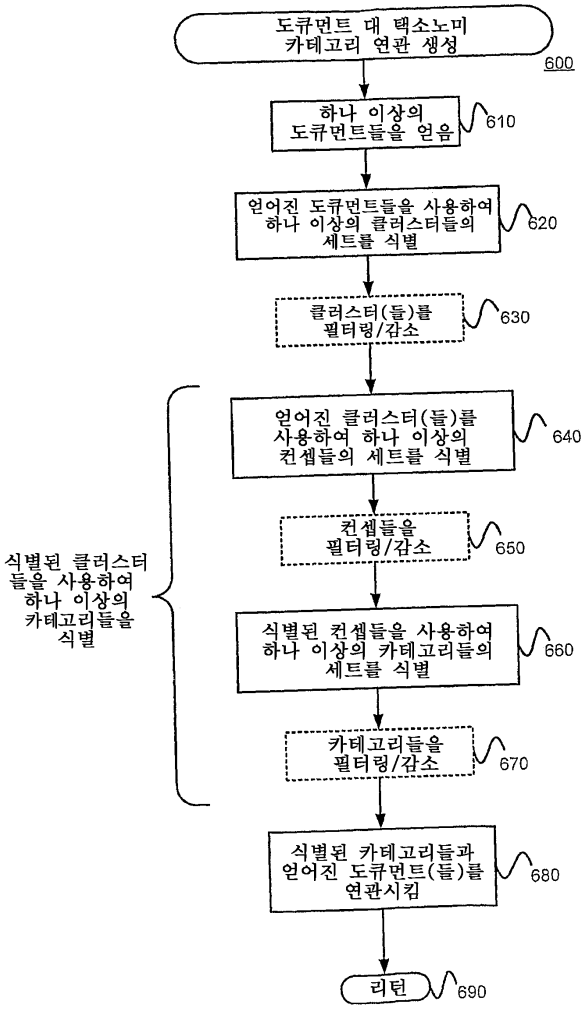
도면4


410
도큐먼트(예컨대, 광고 네트워크에 참여하는 웹사이트들, 광고 네트워크에 참여하는 광고들 등) 정보
420
도큐먼트 대 (버티컬) 택소노미 카테고리 연관 생성 동작들
용어-클러스터 정보(인덱스)
115'
클러스터 식별 동작들
110'
클러스터-대-카테고리 정보
230'
도큐먼트-대-카테고리 정보
430
도큐먼트 ID
택소노미 카테고리(들)
432
434
436

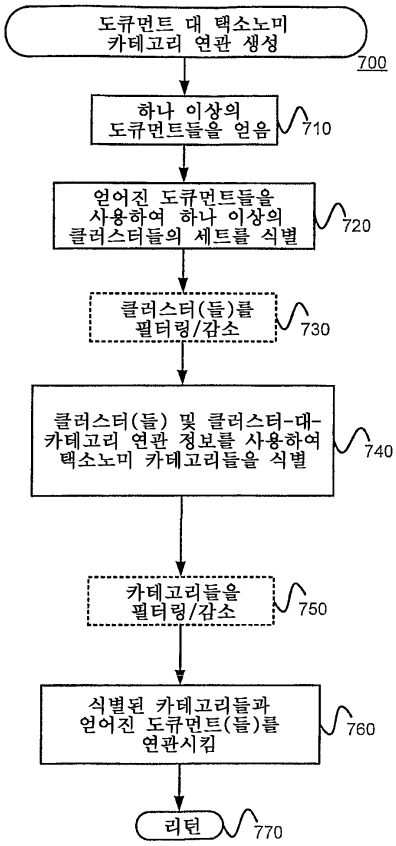
도면5


클러스터 대 택소노미 카테고리 연관 생성
500
하나 이상의 클러스터들을 얻음
510
얻어진 클러스터(들)를 사용하여 하나 이상의 컨셉들의 세트를 식별
520
컨셉(들)을 필터링/감소
530
식별된 컨셉들을 사용하여 하나 이상의 카테고리들의 세트를 식별
540
카테고리들을 필터링/감소
550
얻어진 클러스터들을 사용하여 하나 이상의 카테고리들을 식별
식별된 카테고리(들)와 얻어진 클러스터(들)를 연관시킴
560
리턴
570

도면6

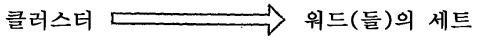

도큐먼트 대 택소노미 카테고리 연관 생성
600
하나 이상의 도큐먼트들을 얻음
610
얻어진 도큐먼트들을 사용하여 하나 이상의 클러스터들의 세트를 식별
620
클러스터(들)를 필터링/감소
630
얻어진 클러스터(들)를 사용하여 하나 이상의 컨셉들의 세트를 식별
640
컨셉들을 필터링/감소
650
식별된 컨셉들을 사용하여 하나 이상의 카테고리들의 세트를 식별
660
카테고리들을 필터링/감소
670
식별된 클러스터들을 사용하여 하나 이상의 카테고리들을 식별
식별된 카테고리들과 얻어진 도큐먼트(들)를 연관시킴
680
리턴
690

도면7

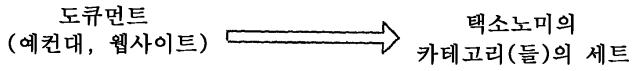


도면8

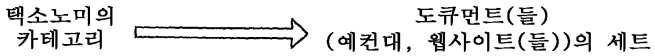
도면9



도면10

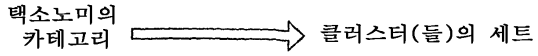


도면11

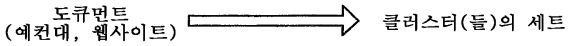


도면12

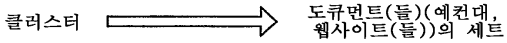
도면13



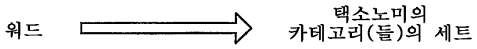
도면14



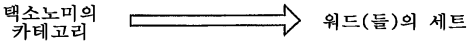
도면15



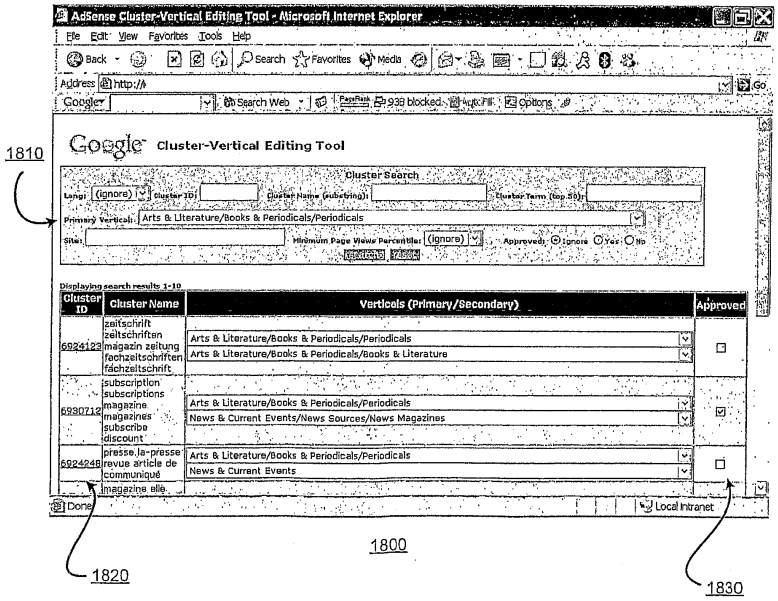
도면16



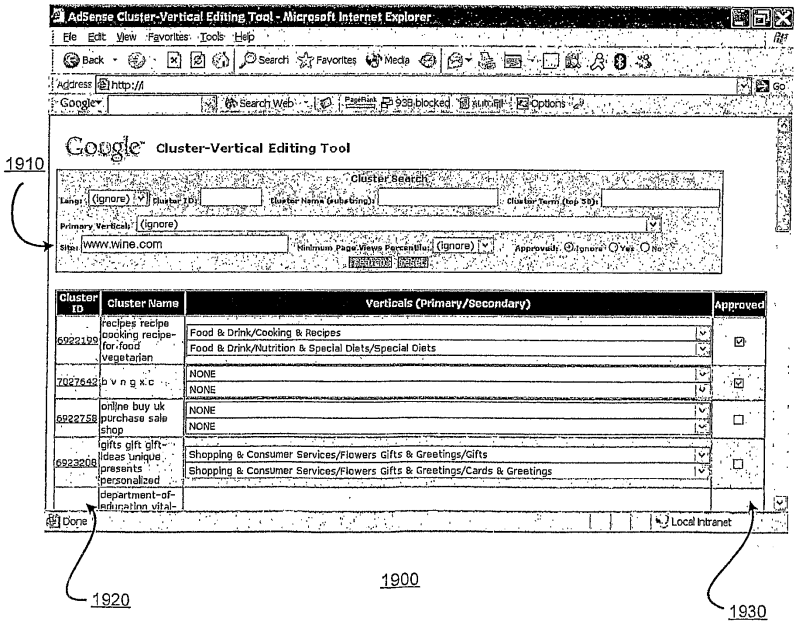
도면17



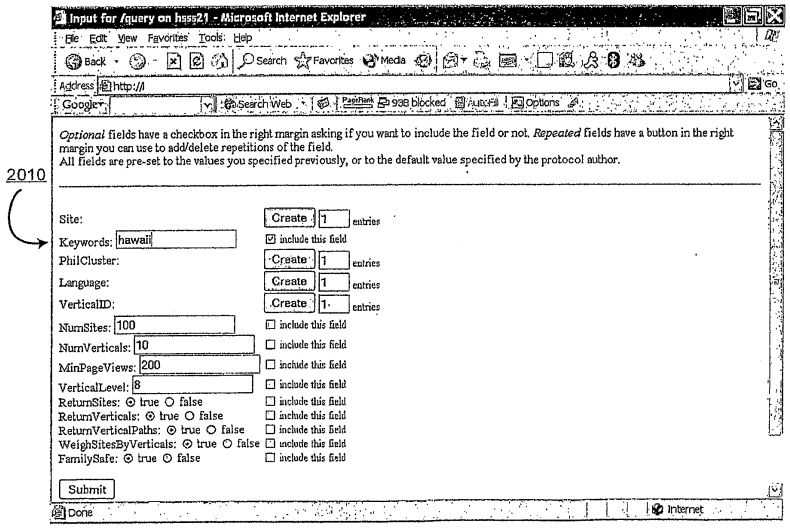
도면18


AdSense Cluster-Vertical Editing Tool - Microsoft Internet Explorer
Google Cluster-Vertical Editing Tool
Cluster Search
Displaying search results 1-10
Cluster ID
Cluster Name
Verticals (Primary/Secondary)
Approved
1810
1800
1820
1830

도면19

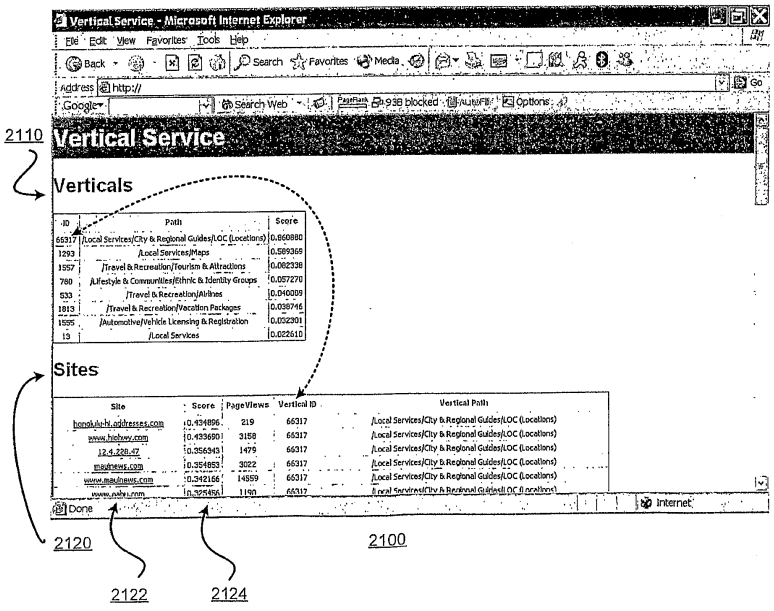

AdSense Cluster-Vertical Editing Tool - Microsoft Internet Explorer
Google Cluster-Vertical Editing Tool
Cluster Search
www.wine.com
Cluster ID
Cluster Name
Verticals (Primary/Secondary)
Approved
1910
1900
1920
1930

도면20



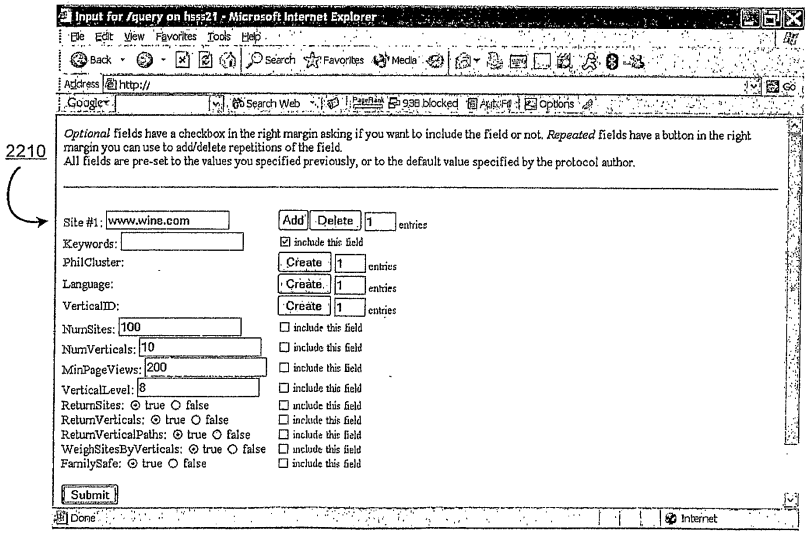
2000

도면21



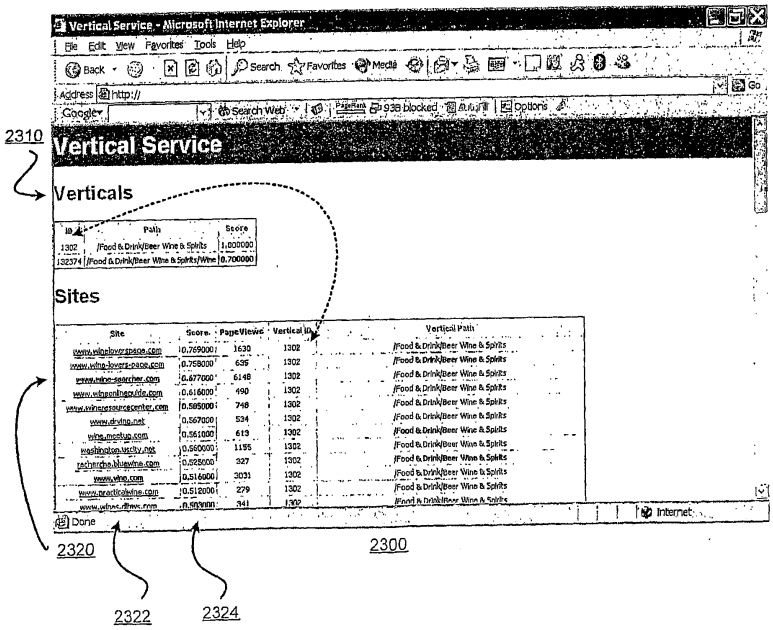
2100

도면22

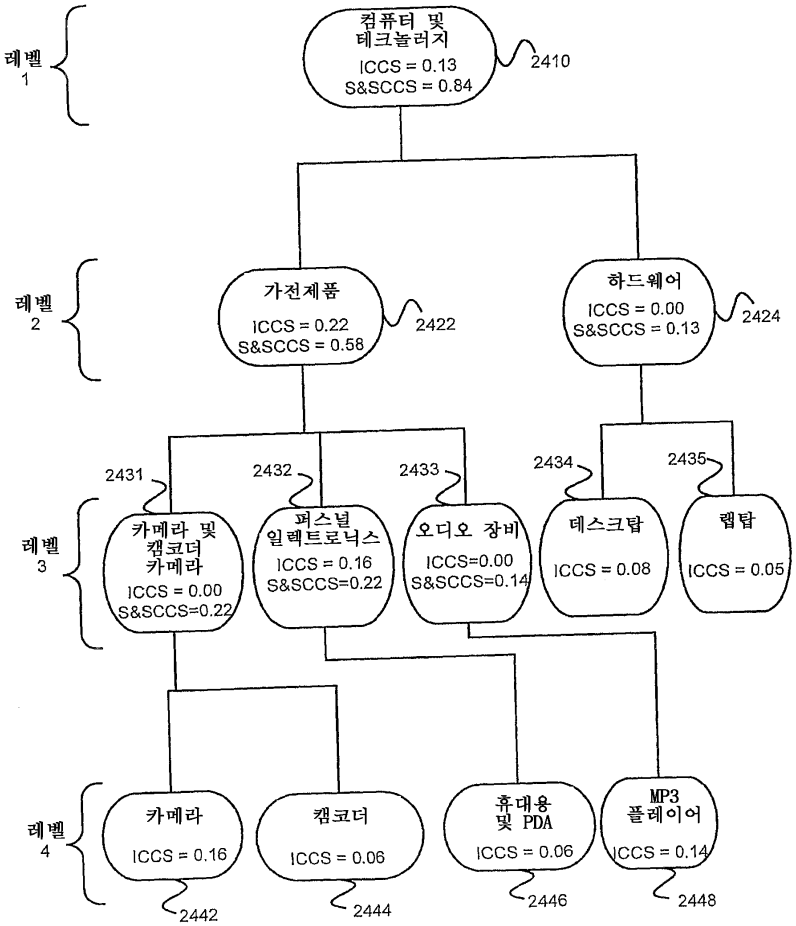


2200

도면23



도면24



ICCS: 인트라-카테고리 클러스터 스코어
S&SCCS : 자체의&서브썸된 카테고리 클러스터 스코어

도면25