(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2009-0018487
(43) 공개일자 2009년02월20일

(51) Int. Cl.
G06F 17/30 (2006.01) *G06F 17/40* (2006.01)
(21) 출원번호 10-2007-0082961
(22) 출원일자 2007년08월17일
심사청구일자 2007년08월17일

(71) 출원인
울산대학교 산학협력단
울산 남구 무거2동 산29
(72) 발명자
옥철영
울산 남구 무거2동 835번지 쌍용스윗닷홈 101동 602호
최호섭
울산 울주군 범서읍 두산리 392-5번지
(뒷면에 계속)
(74) 대리인
장한특허법인

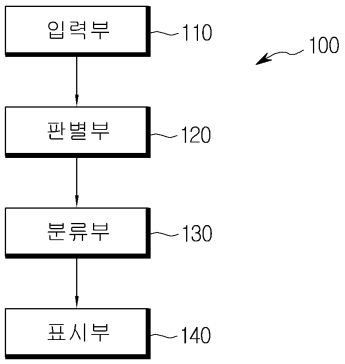
전체 청구항 수 : 총 5 항

(54) 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템

(57) 요 약

본 발명에 따르면, 개념 및 인스턴스 분류 대상에 해당되는 말뭉치 내의 각 용어에 관하여, 용어에 관한 한글표제어 정보, 용어에 대응되는 대역어인 영어표제어 정보 및 용어에 관한 한글뜻풀이 정보를 각 용어별로 입력받는 입력부; 입력부로부터 입력받은 한글표제어 정보, 영어표제어 정보, 한글뜻풀이 정보를 분석하여, 각 정보가 개념 속성 또는 인스턴스 속성을 갖는지의 여부를 각 용어별로 판별하여, 개념 속성인 경우 개념 설정코드를 부가하고 인스턴스 속성인 경우 인스턴스 설정코드를 부가하는 판별부; 및 각 용어별 각 정보에 관하여 개념 설정코드가 부가된 갯수와 인스턴스 설정코드가 부가된 갯수를 비교하는 것에 의해 해당 용어가 개념 또는 인스턴스인지를 판단하여, 각 용어를 개념 또는 인스턴스인 것으로 분류하는 분류부를 포함하는 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템이 제공된다. 개시된 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템에 따르면, 보다 정확한 온톨로지 구축을 위해 해당 용어가 개념인지 또는 인스턴스인지를 정확성 있게 자동 분류함에 따라, 구축된 온톨로지에 관한 신뢰성을 향상시킬 수 있다.

대 표 도 - 도1



(72) 발명자

임지희

울산 남구 무거1동 1484-17번지

배영준

부산 금정구 구서2동 일신아파트 3동 107호

이 발명을 지원한 국가연구개발사업

과제고유번호	C1090-0603-0039
부처명	정보통신부
연구사업명	대학 IT연구센터 육성.지원 사업
연구과제명	Augmented Reality 기반의 디지털 제조 정보기술 개발
주관기관	울산대학교 산학협력단
연구기간	2006년 09월 01일 ~ 2007년 08월 31일

특허청구의 범위

청구항 1

개념 및 인스턴스 분류 대상에 해당되는 말뭉치 내의 각 용어에 관하여, 상기 용어에 관한 한글표제어 정보, 상기 용어에 대응되는 대역어인 영어표제어 정보 및 상기 용어에 관한 한글뜻풀이 정보를 상기 각 용어별로 입력받는 입력부;

상기 입력부로부터 입력받은 상기 한글표제어 정보, 상기 영어표제어 정보, 상기 한글뜻풀이 정보를 분석하여, 각 정보가 개념 속성 또는 인스턴스 속성을 갖는지의 여부를 상기 각 용어별로 판별하여, 상기 개념 속성인 경우 개념 설정코드를 부가하고 상기 인스턴스 속성인 경우 인스턴스 설정코드를 부가하는 판별부; 및

상기 각 용어별 상기 각 정보에 관하여 상기 개념 설정코드가 부가된 갯수와 상기 인스턴스 설정코드가 부가된 갯수를 비교하는 것에 의해 해당 용어가 개념 또는 인스턴스인지를 판단하여, 상기 각 용어를 개념 또는 인스턴스인 것으로 분류하는 분류부를 포함하는 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템.

청구항 2

제 1항에 있어서, 상기 판별부는,

상기 용어에 관한 개념 속성의 판단을 위한 세 가지 규칙으로서,

상기 용어에 관한 상기 한글표제어 정보가 기 구축된 온톨로지 정보 상의 개념 군에 속하는 경우, 상기 한글표제어 정보에 관하여 상기 개념 설정코드를 부가하고,

상기 용어의 대역어인 상기 영어표제어 정보의 첫째 알파벳이 소문자인 경우 상기 한글표제어 정보에 관하여 상기 개념 설정코드를 부가하고,

상기 용어의 한글뜻풀이 정보에 포함된 한글뜻풀이의 어구에 '총칭', '전체', '개념', '집합', '등이', '아우르는' 또는 '단위' 중 선택된 하나 또는 복수 개의 어휘를 포함하는 경우, 상기 한글뜻풀이 정보에 관하여 상기 선택된 하나 또는 복수 개의 각 어휘에 대응되도록 각 어휘별로 상기 개념 설정코드를 개별적으로 부가하는 것을 특징으로 하는 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템.

청구항 3

제 1항 또는 제 2항에 있어서, 상기 판별부는,

상기 용어에 관한 인스턴스 속성의 판단을 위한 세 가지 규칙으로서,

상기 용어에 관한 상기 한글표제어 정보가 기 구축된 온톨로지 정보 상의 인스턴스 군에 속하는 경우, 상기 한글표제어 정보에 숫자가 포함된 경우, 상기 한글표제어 정보에 영문자가 포함된 경우, 또는 상기 한글표제어 정보가 기 구축된 온톨로지 정보 상의 파생어 또는 합성어 군에 속하는 경우, 해당되는 각 경우 별로 인스턴스 설정코드를 개별적으로 부가하고,

상기 용어의 대역어인 상기 영어표제어 정보의 첫째 알파벳이 대문자인 경우 상기 한글표제어 정보에 관하여 상기 인스턴스 설정코드를 부가하고,

상기 용어의 한글뜻풀이 정보에 포함된 한글뜻풀이의 어구에 고유명사가 포함된 경우, 상기 어구에 숫자가 포함된 경우, 또는 상기 어구에 영문자가 포함된 경우, 해당되는 각 경우별로 상기 인스턴스 설정코드를 개별적으로 부가하는 것을 특징으로 하는 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템.

청구항 4

제 3항에 있어서, 상기 분류부는,

해당 용어의 각 정보에 관하여 상기 개념 설정코드가 부가되는 경우 음(-)의 가중치를 부가하고, 상기 인스턴스 설정코드가 부가되는 경우 양(+)의 가중치를 부가하여, 상기 음(-)의 가중치와 양(+)의 가중치를 합산하여 음수가 나오는 경우에는 해당 용어는 개념이고, 양수가 나오는 경우에는 해당 용어는 인스턴스인 것으로 분류하는 것을 특징으로 하는 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템.

청구항 5

제 4항에 있어서,

상기 각 용어별로 상기 한글표제어 정보, 상기 영어표제어 정보 및 상기 한글뜻풀이 정보를 표시하고, 상기 분류부에 의해 상기 각 용어가 상기 개념 또는 상기 인스턴스로 분류되었는지의 여부가 확인되도록 상기 각 용어별로 개념 또는 인스턴스 분류 내역을 표시 제공하는 표시부를 더 포함하는 것을 특징으로 하는 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템.

명 세 서

발명의 상세한 설명

기 술 분 야

<1> 본 발명은 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템에 관한 것으로서, 보다 상세하게는 개념과 인스턴스에 관한 분류가 명확히 정의된 온톨로지가 구축될 수 있도록, 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템에 관한 것이다.

배 경 기 술

<2> 컴퓨터를 똑똑하게 만들기 위해서는 사람의 개념체계를 컴퓨터상에 심어주고 개념들 간의 관계나 속성을 설정해 주어야 하고 각 개념에 관한 예를 알려주어야 한다.

<3> 이러한 일련의 과정을 위해서 사람이 가진 개념체계를 컴퓨터가 인식할 수 있는 형태로 바꾸어 줄 필요가 있는데, 그 형태가 온톨로지로 나타나게 된다. 또한 이러한 배경을 기반으로 현재 온톨로지에 대한 연구가 활발히 이루어지고 있다.

<4> 지금까지 온톨로지에 대한 연구는 크게 온톨로지 구축, 온톨로지 인스턴스 생성, 온톨로지를 이용한 추론, 온톨로지를 이용한 응용서비스 등이 있다.

<5> 이러한 상기 온톨로지를 구축하기 위해서는 많은 시간과 인력이 필요하다.

<6> 우선, 기반이 되는 말뭉치나 사전 및 시소러스 또는 이미 구축된 온톨로지 등의 자원이 필요하며 말뭉치(문서 등)에서 텍스트 정제, 용어 추출 및 검증, 용어의 정의문 대역어용례 추출, 추출된 용어의 개념 및 인스턴스 검증을 통한 개념화, 개념 간의 관계속성 추출, OWL(Web Ontology Language;웹 온톨로지 언어)로 변환, 분류체계 구축, 관계 정의 및 설정, 인스턴스 부착, 온톨로지 통합 등 많은 단계의 작업을 거쳐야 한다.

<7> 이 모든 공정을 사람이 수행을 하려면 많은 인력과 시간과 자원이 소모된다.

<8> 시간이 지날수록 개념 및 인스턴스가 기하급수적으로 늘어날 것인데 비하여 그때마다 새로운 개념 및 인스턴스 구축을 위해 대규모의 인력과 자원을 투입할 수는 없을 것이다.

<9> 또한, 추출한 용어가 인스턴스임에도 불구하고 온톨로지 자동 구축 과정을 통해 개념으로 할당된다면, 온톨로지의 용량 증가 및 신뢰성 저하 등의 문제가 발생될 수 있다. 따라서, 보다 정확한 온톨로지 구축을 위해 자동 구축 시스템들 중 하나인 개념과 인스턴스를 분류하는 시스템이 구축되어야 한다.

발명의 내용

해결 하고자하는 과제

<10> 본 발명은 상술한 문제점을 해결하기 위하여 창출된 것으로서, 보다 정확한 온톨로지 구축을 위해 해당 용어가 개념인지 또는 인스턴스인지를 정확성 있게 자동 분류하여, 구축된 온톨로지에 관한 신뢰성을 향상시킬 수 있는 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템을 제공하는 데 그 목적이 있다.

과제 해결수단

<11> 상기와 같은 목적을 달성하기 위한 본 발명의 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템은, 개념 및 인스턴스 분류 대상에 해당되는 말뭉치 내의 각 용어에 관하여, 상기 용어에 관한 한글표제어 정보, 상기

용어에 대응되는 대역어인 영어표제어 정보 및 상기 용어에 관한 한글뜻풀이 정보를 상기 각 용어별로 입력받는 입력부; 상기 입력부로부터 입력받은 상기 한글표제어 정보, 상기 영어표제어 정보, 상기 한글뜻풀이 정보를 분석하여, 각 정보가 개념 속성 또는 인스턴스 속성을 갖는지의 여부를 상기 각 용어별로 판별하여, 상기 개념 속성인 경우 개념 설정코드를 부가하고 상기 인스턴스 속성인 경우 인스턴스 설정코드를 부가하는 판별부; 및 상기 각 용어별 상기 각 정보에 관하여 상기 개념 설정코드가 부가된 갯수와 상기 인스턴스 설정코드가 부가된 갯수를 비교하는 것에 의해 해당 용어가 개념 또는 인스턴스인지를 판단하여, 상기 각 용어를 개념 또는 인스턴스인 것으로 분류하는 분류부를 포함한다.

<12> 또한, 상기 판별부는, 상기 용어에 관한 개념 속성의 판단을 위한 세 가지 규칙으로서, 상기 용어에 관한 상기 한글표제어 정보가 기 구축된 온톨로지 정보 상의 개념 군에 속하는 경우, 상기 한글표제어 정보에 관하여 상기 개념 설정코드를 부가하고, 상기 용어의 대역어인 상기 영어표제어 정보의 첫째 알파벳이 소문자인 경우 상기 한글표제어 정보에 관하여 상기 개념 설정코드를 부가하고, 상기 용어의 한글뜻풀이 정보에 포함된 한글뜻풀이의 어구에 '총칭', '전체', '개념', '집합', '등이', '아우르는' 또는 '단위' 중 선택된 하나 또는 복수 개의 어휘를 포함하는 경우, 상기 한글뜻풀이 정보에 관하여 상기 선택된 하나 또는 복수 개의 각 어휘에 대응되도록 각 어휘별로 상기 개념 설정코드를 개별적으로 부가할 수 있다.

<13> 또한, 상기 판별부는, 상기 용어에 관한 인스턴스 속성의 판단을 위한 세 가지 규칙으로서, 상기 용어에 관한 상기 한글표제어 정보가 기 구축된 온톨로지 정보 상의 인스턴스 군에 속하는 경우, 상기 한글표제어 정보에 숫자가 포함된 경우, 상기 한글표제어 정보에 영문자가 포함된 경우, 또는 상기 한글표제어 정보가 기 구축된 온톨로지 정보 상의 파생어 또는 합성어 군에 속하는 경우, 해당되는 각 경우 별로 인스턴스 설정코드를 개별적으로 부가하고, 상기 용어의 대역어인 상기 영어표제어 정보의 첫째 알파벳이 대문자인 경우 상기 한글표제어 정보에 관하여 상기 인스턴스 설정코드를 부가하고, 상기 용어의 한글뜻풀이 정보에 포함된 한글뜻풀이의 어구에 고유명사가 포함된 경우, 상기 어구에 숫자가 포함된 경우, 또는 상기 어구에 영문자가 포함된 경우, 해당되는 각 경우별로 상기 인스턴스 설정코드를 개별적으로 부가할 수 있다.

<14> 그리고, 상기 분류부는, 해당 용어의 각 정보에 관하여 상기 개념 설정코드가 부가되는 경우 음(-)의 가중치를 부가하고, 상기 인스턴스 설정코드가 부가되는 경우 양(+)의 가중치를 부가하여, 상기 음(-)의 가중치와 양(+)의 가중치를 합산하여 음수가 나오는 경우에는 해당 용어는 개념이고, 양수가 나오는 경우에는 해당 용어는 인스턴스인 것으로 분류할 수 있다.

효 과

<15> 개시된 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템에 따르면, 온톨로지 구축을 위해 해당 용어가 개념인지 또는 인스턴스인지를 정확성 있게 자동 분류함에 따라, 추출한 용어가 인스턴스임에도 불구하고 온톨로지 자동 구축 과정을 통해 개념으로 할당되는 일이 방지되어 온톨로지의 용량 효율 및 신뢰성이 향상되는 효과가 있다.

발명의 실시를 위한 구체적인 내용

<16> 이하 첨부된 도면을 참조로 본 발명의 바람직한 실시예들을 상세히 설명하기로 한다.

<17> 이에 앞서, 본 명세서 및 청구범위에 사용된 용어나 단어는 통상적이거나 사전적인 의미로 한정해서 해석되어서는 아니되며, 발명자는 그 자신의 발명을 가장 최선의 방법으로 설명하기 위해 용어의 개념을 적절하게 정의할 수 있다는 원칙에 입각하여 본 발명의 기술적인 사상에 부합하는 의미와 개념으로 해석되어야만 한다.

<18> 따라서, 본 명세서에 기재된 실시예와 도면에 도시된 구성은 본 발명의 가장 바람직한 일 실시예에 불과할 뿐이고 본 발명의 기술적 사상을 모두 대변하는 것은 아니므로, 본 출원시점에 있어서 이들을 대체할 수 있는 다양한 균등물과 변형예들이 있을 수 있음을 이해하여야 한다.

<19> 도 1은 본 발명의 실시예에 따른 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템의 구성도, 도 2는 도 1의 판별부에서 특정 용어의 한글표제어 정보, 영어표제어 정보, 한글뜻풀이 정보에 관하여 개념 설정코드 또는 인스턴스 설정코드를 부가하기 위한 조건을 나타내는 구성도이다.

<20> 또한, 도 3은 도 1의 판별부에서 특정 용어의 한글뜻풀이 정보에 관하여 개념 설정코드 또는 인스턴스 설정코드가 부가되는 예시도, 도 4는 도 1의 표시부에 관한 실시예를 나타내는 화면구성도, 도 5는 도 4에서 특정 용어가 위치한 행을 클릭하는 경우 표시되는 상기 특정 용어에 관한 상세 정보를 나타내는 화면구성도이다.

<21> 그리고, 도 6은 도 4의 '폴더열기' 버튼을 통해 입력되는 '한글표제어 정보, 영어표제어 정보, 한글뜻풀이 정보가 용어별로 저장된 텍스트 파일'의 예시도이다.

<22> 도 1에 도시된 바와 같이, 본 발명의 실시예에 따른 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템(100)은, 입력부(110), 판별부(120), 분류부(130)를 포함한다.

<23> 상기 입력부(110)는, 개념 및 인스턴스 분류 대상에 해당되는 말뭉치 내의 각 용어에 관하여, 상기 용어에 관한 한글표제어 정보, 상기 용어에 대응되는 대역어인 영어표제어 정보 및 상기 용어에 관한 한글뜻풀이 정보를 상기 각 용어별로 입력받는 부분이다.

<24> 이러한 상기 입력부(110)의 입력 동작은 도 4의 '폴더열기' 버튼에 의해 실행 가능할 수 있다.

<25> 예를 들면, 상기 입력부(110)는 해당 용어에 관한 한글표제어 정보, 상기 용어의 영어표제어 정보 및 상기 용어의 한글뜻풀이 정보가 상기 용어별로 각각 정리된 도 6과 같은 텍스트 파일을 상기 도 4의 '폴더열기' 버튼을 통해 입력받을 수 있다.

<26> 참고로, 도 6에서 특정 용어에 관한 한글뜻풀이 정보, 영어표제어 정보, 한글뜻풀이 정보는 띄어쓰기에 의해 순차적으로 구분되어 있으며, 용어별 상호 구분은 문단나눔에 의해 설정되어 있다.

<27> 한편, 본 발명에 의해 개념 및 인스턴스로 분류되기 위한 도 6에 개시된 용어들은 대부분 전자 통신분야의 용어들을 나타내고 있으나, 도 6의 경우는 단지 본 발명의 실시예에 불과하며 상기 전자분야 이외의 다양한 분야에 관한 용어들 또한 본 발명에 의해 개념 및 인스턴스로 분류될 수 있음은 물론이다.

<28> 상기 판별부(120)는 상기 입력부(110)로부터 입력받은 상기 한글표제어 정보, 상기 영어표제어 정보, 상기 한글뜻풀이 정보를 각각 분석하여, 각 정보가 개념 속성 또는 인스턴스 속성을 갖는지의 여부를 상기 각 용어별로 판별하여, 상기 개념 속성인 경우 개념 설정코드를 부가하고 상기 인스턴스 속성인 경우 인스턴스 설정코드를 부가하는 부분이다.

<29> 즉, 상기 판별부(120)는 상기 입력부(110)로부터 입력받은 도 6의 텍스트 파일 상에 존재하는 해당 용어에 관한 한글표제어 정보, 영어표제어 정보, 한글뜻풀이 정보 각각에 관하여 개념 속성을 갖는지 또는 인스턴스 속성을 갖는지를 분석하여 개념 설정코드 또는 인스턴스 설정코드를 부가하게 된다.

<30> 상기 분류부(130)는 상기 각 용어별 상기 각 정보에 관하여 상기 개념 설정코드가 부가된 갯수와 상기 인스턴스 설정코드가 부가된 갯수를 비교하는 것에 의해 해당 용어가 개념 또는 인스턴스인지를 판단하여, 상기 각 용어를 개념 또는 인스턴스인 것으로 분류하는 부분이다.

<31> 즉, 상기 분류부(130)에 의한 분류는, 특정 용어에 관한 상기 한글표제어 정보, 영어표제어 정보, 한글뜻풀이 정보 각각에 관해 부가된 개념 설정코드 및 인스턴스 설정코드의 갯수를 비교하여 그 갯수가 큰 어느 한쪽을 기준으로 하여, 상기 개념 설정코드가 부가된 갯수가 많은 경우는 해당 용어를 개념으로 분류하고, 상기 인스턴스 설정코드가 부가된 갯수가 많은 경우에는 해당 용어를 인스턴스인 것으로 분류하게 된다.

<32> 여기서, 상기 판별부(120) 및 분류부(130)의 동작은 도 4의 '실행' 버튼에 의해 가능할 수 있다.

<33> 한편, 본 발명에 따르면, 상기 분류부(130)에 의해 개념 또는 인스턴스로 분류된 결과를 해당 용어별로 표시하는 도 4와 같은 표시부(140)를 더 포함할 수 있다.

<34> 이러한 상기 표시부(140)는, 상기 각 용어별로 상기 한글표제어 정보, 상기 영어표제어 정보 및 상기 한글뜻풀이 정보를 표시하고, 상기 분류부(130)에 의해 상기 각 용어가 상기 개념 또는 상기 인스턴스로 분류되었는지의 여부가 확인되도록 상기 각 용어별로 개념 또는 인스턴스 분류 내역을 표시 제공하는 부분이다.

<35> 물론, 상기 표시부(140)는 도 6의 좌측에 위치한 표시창과 같이 상기 입력부(110)로 부터 입력받은 도 6과 같은 해당 텍스트 파일에 관한 파일명이 텍스트 파일별로 함께 표시되며, 특정 파일명이 더블 클릭(또는 클릭)되는 경우, 오른쪽 표시창과 같이 상기 특정 파일명에 해당하는 텍스트 파일에 포함된 용어별로, 해당 용어에 관한 한글명(한글표제어 정보), 영어명(영어표제어 정보)뿐만 아니라 한글뜻풀이 정보까지 포함하여 표시 가능하다.

<36> 또한, 상기 분류부(130)에 의해 개념 또는 인스턴스로 분류된 결과는, 도 4의 표시부에서 '추천결과'란에 의해 개념인지 인스턴스인지가 표시 가능하다.

<37> 더욱이, 도 4와 같이 상기 '추천결과'에 관한 '추천이유' 또한 표시 가능한데, 상기 '추천이유'에 관한 보다 상세한 이유를 알고자 한다면 해당 용어가 있는 행을 선택하여 더블클릭하는 것 또는 좌측 상단의 '자세히 보기'

버튼 클릭에 의해 가능할 수 있다. 이에 관한 보다 상세한 설명은 추후 소개하기로 한다.

<38> 이하에서는, 이상과 같은 본 발명의의 실시예에 관하여 도 1 내지 도 6을 참고로 하여 보다 상세히 설명하고자 한다.

<39> 먼저, 본 발명의 시스템을 구축하기 위해 가장 기본 바탕이 되는 '개념'과 '인스턴스'에 대한 정의 및 분류 기준을 국어사전, 백과사전, 영영사전, 위키피디아(wikipedia), 온톨로지 관련 논문 등을 참고하여 설정하였다.

<40> 상기 '개념'의 정의는 '추상적', '일반적', '보편적', '공통적인 요소', '집합'이라는 속성들을 기반으로 하고 있다.

<41> 이에 반해, '인스턴스'의 정의는 '구체적', '특정한', '개별적인 요소', '사례', '예'와 같은 속성들을 기반으로 하고 있다.

<42> 이러한 특성들은 우리가 일반적으로 사용하는 명사들의 분류에서도 나타나는데, 상기 '개념'은 보통명사, 일반명사를 포함하며, 상기 '인스턴스'는 고유명사를 포함한다고 볼 수 있다.

<43> 상기 일반명사의 특성은 '개념'과 같이 일반적이고 포괄적인 의미의 특징들을 가지고 있으며, 이에 반해 상기 고유명사의 특성은 특정하고 개별적이고 유일한 특징을 가지며 특히 특정 명칭이 보통 이에 속한다.

<44> 이러한 각종 용어들이나 패턴들을 추출하고 분석한 결과는, 본 발명에서 분류하고자 하는 해당 용어에 관한 '개념' 및 '인스턴스'에 대한 설정 규칙으로 사용될 수 있다.

<45> 이상과 같이 해당 용어에 관한 '개념' 또는 '인스턴스'의 분류를 위해, 말뭉치(예를 들면, 특정 문서) 내에서 상기 분류의 대상인 특정 용어(한글표제어 정보)와 상기 용어의 대역어(영어표제어 정보), 그리고 상기 용어의 정의문 또는 용어의 설명(한글뜻풀이 정보)이 도 6과 같이 필요하다.

<46> 그리고 더욱 세밀한 '개념' 및 '인스턴스' 규칙 설정을 위해 기구축된 시소러스나 온톨로지 상의 개념 집합, 그와 관련된 동의어 및 유의어 자원과, 국가명, 지역명, 인명, 기관명, 단체명, 사건명, 책명 등의 특정한 명칭을 나타내는 고유명사들을 확보하여야 한다.

<47> 본 발명에서는 해당 용어에 관하여 '개념 설정 규칙' 또는 '인스턴스 설정 규칙'을 적용하고, 이를 더욱 세부적으로 한다면 아래의 표 1과 같이 '한글표제어 규칙', '영어표제어 규칙', '한글뜻풀이 규칙'마다 '개념 설정 규칙' 또는 '인스턴스 설정 규칙'을 적용한 총 6가지 규칙이 이용될 수 있고 해당 규칙마다 코드가 부가될 수 있다.

<48> [표 1]

규칙명	규칙 세부 설명
CKT	한글 표제어에 대한 개념 규칙
CKX	한글 뜻풀이에 대한 개념 규칙
CET	영어 표제어에 대한 개념 규칙
IKT	한글 표제어에 대한 인스턴스 규칙
IKX	한글 뜻풀이에 대한 인스턴스 규칙
IET	영어 표제어에 대한 인스턴스 규칙

<49>

<50> 표 1을 참고로 하면, 개념 규칙에 관한 코드 및 인스턴스 규칙에 관한 코드가 각각 3개로 존재하며, 각 코드는 세 자릿수로 이루어진다.

<51> 코드의 첫째 자리는 개념에 대한 규칙(C)인지, 인스턴스에 대한 규칙(I)인지를 설정하는 자리이다.

<52> 코드의 둘째 자리는 한국어(K)인지 영어(E)인지를 구분하고, 마지막 셋째 자리는 표제어(T)인지 뜻풀이(X)인지를 구분하는 자리이다.

<53> 일단, 본 발명에서는 개념 및 인스턴스의 분류대상 용어에 해당되는 각 용어별로 한글표제어 정보, 영어표제어 정보, 한글뜻풀이 정보가 정리된 도 6과 같은 텍스트 문서를 상기 입력부(110)에 의해 입력받으며 상기한 입력은 도 4의 '폴더열기' 버튼을 통해 가능할 수 있다.

<54> 상기 판별부(120)는 상기 입력받은 용어에 포함된 각 정보별로, 해당 정보가 개념 속성을 갖는 경우 표 1의 해당되는 개념 설정코드를 부여하고, 인스턴스 속성을 갖는 경우 해당되는 인스턴스 설정코드를 부가하게 된다.

<55> 먼저 상기 판별부(120)에 의해 해당 용어의 각 정보에 '개념 설정코드'가 부가되는 실시예는 아래와 같다.

<56> 상기 용어에 관한 개념 속성의 판단을 위한 상술한 세 가지 규칙 중 첫 번째 규칙으로서, 상기 판별부(120)는 상기 용어에 관한 상기 한글표제어 정보가 기 구축된 온톨로지 정보 상의 개념 군에 속하는 경우, 상기 한글표제어 정보에 관하여 상기 표 1의 개념 설정코드(CKT)를 부가하게 된다.

<57> 여기서 상기 기 구축된 온톨로지 정보란, 앞서 상술한 바와 같이 본 시스템의 구축을 위해 가장 기본 바탕이 되는 '개념' 또는 '인스턴스'에 대한 정의 및 분류 기준을 국어사전, 백과사전, 영영사전, 위키피디아(wikipedia), 온톨로지 관련 논문 등을 참고하여 설정되어 구축된 온톨로지 상의 정보일 수 있다.

<58> 즉, 앞서 상술한 바와 같이, 상기 한글표제어 정보가 기 구축된 온톨로지 정보 상의 '개념' 군에 해당되는 것, 즉 '일반적', '보편적', '공통적인 요소', '집합'의 속성을 갖는 흔히 부르는 보통명사, 일반명사에 해당되는 경우, 해당 한글표제어 정보에 대해 개념 설정코드(CKT)가 부가된다.

<59> 다시 말해서, 한글표제어가 일반적, 공통적, 집합적인 용어를 포함하거나 그러한 속성을 갖는 경우 그 한글표제어는 개념이 될 가능성이 크다.

<60> 여기서 본 발명은 도 1과 같이 상기 한글표제어에 관한 개념 규칙을 설정하기 위하여 기 구축된 온톨로지 중에서 상위 3 depth 즉, 상위 3 계층(hierarchy) 까지의 개념과 상기 한글표제어 정보 간을 비교하여 해당 한글표제어 정보가 상기 개념 속성에 포함되는 지의 여부를 가릴 수 있으며, 이러한 본 발명은 3 계층 이상을 포함하는 온톨로지 계층을 아우르는 정보를 이용하는 것도 가능함은 물론이다.

<61> 상기 개념 속성의 판단을 위한 두 번째 규칙으로서, 상기 판별부(120)는 상기 용어의 대역어인 상기 영어표제어 정보의 첫째 알파벳이 소문자인 경우 상기 한글표제어 정보에 관하여 상기 표 1의 개념 설정코드(CET)를 부가한다.

<62> 이는, 일반적으로 영어표제어가 고유명사인 경우(인스턴스 속성을 갖는 경우), 영어 표제어의 제일 앞 글자가 대문자로 표기되므로, 표제어가 개념 속성을 갖는 경우에는 영어 표제어의 첫글자는 소문자로 표기되는 이유에서이다.

<63> 즉, 상기 판별부(120)는 상기 입력부(110)로부터 입력받은 용어에 관한 영어표제어의 첫 자리 알파벳이 소문자인 경우, 이 영어표제어는 개념 속성인 것으로 판단하여 상기 표 1의 개념 설정코드(CET)를 부가할 수 있다.

<64> 다음으로, 상기 개념 속성의 판단을 위한 세 번째 규칙으로서, 상기 표 1 또는 아래의 표 2와 같이 상기 판별부(120)는 상기 용어의 한글뜻풀이 정보에 포함된 한글뜻풀이의 각 어구에 '총칭', '전체', '개념', '집합', '등이', '아우르는' 또는 '단위' 중 선택된 하나 또는 복수 개의 어휘를 포함하는 경우, 상기 한글뜻풀이 정보에 관하여 상기 선택된 하나 또는 복수 개의 각 어휘에 대응되도록 각 어휘별로 상기 해당되는 개념 설정코드(CKX)를 개별적으로 부가하게 된다.

<65> 보통, 해당 용어에 관한 한글뜻풀이는 뜻풀이를 위한 문장으로 이루어져 있으며, 그 문장 또한 개별 어구로 이루어져 있으므로, 상기 판별부(120)는 각각의 어구별로 해당 어구가 앞서 나열한 개념 속성을 갖는 어구들과 일치되는가를 판단하게 된다.

<66> 상기 '총칭', '전체', '개념', '집합', '등이', '아우르는' 또는 '단위'는 앞서 상술한 '일반적', '보편적', '공통적인 요소', '집합'의 속성을 갖는 부분이므로 이러한 어휘가 해당 한글뜻풀이 정보에 포함되어 있는 경우, 이에 관한 개념 설정코드(CKX)를 부가되게 된다.

<67> [표 2]

어휘	구체적 어휘	뜻풀이에 나타난 예
총칭	총칭, 총칭함 총칭이다, 총칭한다, 총칭하는 용어	<휴대용 컴퓨터>:간단히 이동할 수 있도록 설계된 컴퓨터의 **총칭**.
		<형식 의미론>:~ 파악하려는 연구를 **총칭하는 용어**.
전체	전체적인 전체를, 전체의 전체적으로 전체	<데이터 기술 자료부>:~ 데이터의 기술이나 **전체적인** 데이터베이스의~
		<유즈 웨어>:개인용 컴퓨터(PC) 관련 서비스 **전체를** 말하는 것
개념	개념,~을,~으로 개념이다	<유틸리티 컴퓨팅>:~비용을 지불한다는 **개념**.
집합 집합체	~의 집합 ~의 집합으로 ~의 집합체 ~의 집합체로	<시스템 문서>:~유지 보수에 관한 사항을 기술한 문서들**의 집합**.
		<오류 수정 보조기>:~도와주기 위해 마련된 루틴들**의 집합체**.
등이	등이 있~ 등이다	<비교 정렬>:~버블 정렬, 셀 정렬, 퀵 정렬, 힙 정렬 **등이 있다**.
아우르는	아우르는	<유비쿼터스 정보 기술>:~기술을 **아우르는** 폭 넓은 개념이다.
단위	단위, 단위로 단위이다.	<토큰>:~분할할 수 있는 **단위로**, 컴파일러나 ~ 어휘 분석 **단위**.

<68>

<69> 즉, 표 2의 예와 같이, 해당 용어에 관한 한글뜻풀이 정보 상에 '총칭'이라는 어휘가 있는 경우, 이에 대해 개념 설정코드(CKX)를 부가하고, 만약에 '총칭'과 '아우르는'이 함께 포함된 표 2의 6번째 예인 <유비쿼터스 정보 기술>인 용어의 경우, 상기 개념 설정코드(CKX)가 해당 어휘에 대응되어 각각 부가됨에 따라 총 2개가 부가되어 개념 설정코드에 관한 가중치가 더욱 증가될 수 있다.

<70> 이상과 같은 해당 용어에 관한 상기 개념 속성의 판단을 규칙과는 달리, 본 발명에서는 상기 용어에 관한 인스턴스 속성의 판단을 위한 세 가지 규칙을 더 포함한다.

<71> 상기 인스턴스 규칙에 관한 상술한 세 가지 규칙 중 첫 번째 규칙으로서, 표 3을 참고하면 상기 판별부(120)는 a)상기 용어에 관한 상기 한글표제어 정보가 기 구축된 온톨로지 정보 상의 인스턴스 군에 속하는 경우, b)상기 한글표제어 정보에 숫자가 포함된 경우, c)상기 한글표제어 정보에 영문자가 포함된 경우, 또는 d)상기 한글표제어 정보가 기 구축된 온톨로지 정보 상의 파생어 또는 합성어 군에 속하는 경우, 해당되는 각 경우(a,b,c,d) 별로 표 1과 같은 인스턴스 설정코드(IKT)를 개별적으로 부가하게 된다.

<72> [표 3]

패턴	구체적 패턴	표제어 예제
이름	1.제조사명, 기관명, 단체명 2.국가 및 지역명 3.인명 4.책명 5.사건명	삼성,엘지,정보통신연구진흥원,MBC
		한국,서울,보스톤,도쿄
		이순신,노무현
		성호사설(星湖僿說),과학혁명의 구조
숫자 및 영문자	1.숫자+용어 2.영문자+용어 3.숫자+영문자	LG-LV4200
		KIPONTO2006
		넷데이 96
		윈도즈 CE
파생어 및 합성어	1.접두어+인스턴스 2.인스턴스+접미어 3.접두어+인스턴스+접미어	<(주),(유),주식회사가 붙은 표제어> (주)시소러스,듀폰(유)
	4.인스턴스+인스턴스 5.(일반어)+인스턴스+(일반어)	<~호~사 접미사 > 독일TV 위성 1호

<73>

<74> 여기서, a) 또는 d)의 경우에 있어서 상기 기 구축된 온톨로지 정보란, 앞서 상술한 바와 같이 본 시스템의 구축을 위해 가장 기본 바탕이 되는 '개념' 또는 '인스턴스'에 대한 정의 및 분류 기준을 국어사전, 백과사전, 위키피디아, 온톨로지 관련 논문 등을 참고하여 구축된 온톨로지 정보를 의미할 수 있다.

<75> 즉, 앞서 상술한 바와 같이 상기 한글표제어 정보가 상기 기 구축된 온톨로지 상의 '인스턴스' 군에 해당되는 것, 즉 '구체적', '특정한', '개별적인 요소', '사례', '예'와 같은 속성을 갖는 흔히 부르는 고유명사에 해당되는 경우, 해당 한글표제어 정보에 인스턴스 설정코드(IKT)가 부가된다.

<76> 한편, 이러한 각 a), b), c), d)의 경우는, 도 2의 한글뜻풀이에 관한 인스턴스 규칙 적용 부분에 기재된 '명칭들(기 구축된 온톨로지 상의 정보)과의 비교', '숫자 확인', '영문자 확인', '파생어 및 합성어 조사' 부분에 각각 대응된다.

<77> 이러한 상기 첫 번째 규칙에 관하여 상기 표 3을 이용하여 보다 상세히 설명하자면, 한글표제어가 국가명, 지역명, 인명, 기관명, 단체명 등의 특정한 명칭을 나타내는 용어를 포함할 때 그 한글표제어는 인스턴스가 될 가능성이 많다.

<78> 또한, 한글표제어가 상술한 바와 같이 특정한 명칭을 나타내는 고유명사인 경우 이외에도 표제어에 포함되어 파생어 및 합성어로 사용되는 경우에도 그 한글표제어는 인스턴스일 가능성이 크다.

<79> 또한, 상기 한글표제어가 특정 숫자(예를 들면, 연도) 및 영문자를 포함(숫자, 영문자, 숫자+영문자로 이루어진 경우 모두 포괄)인 경우 그 한글표제어는 인스턴스일 가능성이 크다.

<80> 예를 들면, 상기 한글표제어가 숫자 및 영문자가 조합된 경우인 표 3의 'KIPONT2006'인 경우, 상기 인스턴스 설정코드(IKT)가 상기 영문자와 숫자인 경우 각각에 부가됨에 따라 총 2개가 부가되어 인스턴스 설정코드에 관한 가중치가 더욱 증가될 수 있다.

<81> 상기 인스턴스 속성의 판단을 위한 두 번째 규칙으로서, 상기 판별부(120)는 상기 용어의 대역어인 상기 영어표제어 정보의 첫째 알파벳이 대문자인 경우 상기 한글표제어 정보에 관하여 상기 표 1의 인스턴스 설정코드(IET)를 부가할 수 있다.

<82> 이는, 앞서 상술한 개념 설정코드(CET) 부가시와는 반대되는 경우로서, 영어표제어 정보가 고유명사인 경우, 표제어의 제일 앞 알파벳을 대문자로 표기하는 일반적인 특징에 기반한 것이다.

<83> 다음으로, 상기 인스턴스 속성 판단을 위한 세 번째 규칙으로서, 상기 표 1 또는 아래의 표 4와 같이 상기 판별부(120)는 상기 용어의 한글뜻풀이 정보에 포함된 1)한글뜻풀이의 어구에 고유명사가 포함된 경우, 2)상기 어구에 숫자가 포함된 경우, 또는 3)상기 어구에 영문자가 포함된 경우, 해당되는 각 경우별로 상기 인스턴스 설정코드(IKX)를 개별적으로 부가하게 된다.

<84> [표 4]

패턴	구체적 패턴	예제
이름	뜻풀이에 다음의 명칭들 포함 여부. 1.제조사명,기관명,집단명,단체명 2.국가 및 지역명 3.인명 4.책명 5.사건명	<오라클>:**미국 오라클사**의 관계 ~ 관리 시스템의 이름.
		<APCN>:**한국** 및 **일본**을 비롯한 **동아시아** 지역과 **동남아시아** 5개~
		<한국과학기술사>:**전상운**(全相運)이 ~ 한국 과학기술사의 쓴 **책**.
		<타스>:~**소련** 공산당 기관지 《 **프라우다 Pravda**》의 논설을 ~
숫자 및 영문자	특정 연도의 포함 여부	<영사 관계에 관한 빈 협약>: **1963년 4월** ~ **1977년 4월** 발효된 조약.
		<아이보>:**1999년** 소니에서 발표한 세계 최초~ 완구 로봇 애완견.
		<3DAV>:**MPEG** 표준화에 포함될 3차원~기존의 **MPEG-2**, **MPEG-4**
특정 용어	1.'명칭','상품명','이름' 2.'개발한','설립한','제작한' 3.인스턴스+'동의어','유의어','은어'	<알니코>:강자성(强磁性)의 영구 자석으로 사용되는 재료의 **상품명**.
		<챔프스>:SCI (System Coor dinati on,Inc.)사가 **개발한** 원자력~
		<응용 지향 언어 (AOL)>:~사용자 ~POL (Problem-Orien ted Langu age)과 **동의어**이다.

<85>

<86> 즉, 1)의 경우는 상기 한글뜻풀이의 문구를 분리한 각 어구에 표 4와 같이 국가명, 지역명, 인명, 기관명, 단체명 등의 특정한 명칭을 나타내는 용어가 포함된 경우 즉, 해당 어구가 고유명사인 경우를 의미한다.

<87> 상기 표 4의 예와 같이, 해당 용어에 관한 한글뜻풀이 정보 상에 '국가명'이 기재된 <오라클> 용어의 경우, 상기 용어에 관하여 상기 인스턴스 설정코드(IKX)가 부가될 수 있다. 이에 부가적으로 상기 <오라클> 용어의 뜻풀이에 '제조사명'이 더 포함되어 있으므로 상기 인스턴스 설정코드(IKT)는 총 2개 부가될 수 있음은 물론이다.

<88> 한편, 이러한 각 1), 2), 3)의 경우는, 도 2의 한글뜻풀이에 관한 인스턴스 규칙 적용 부분에 기재된 '잘라낸 어절과 명칭들과의 비교', '숫자 확인', '영문자 확인', '파생어 및 합성어 조사' 부분에 각각 대응된다.

<89> 물론, 본 발명에서 상기 판별부(120)는 표 4의 마지막 부분에 개시된 바와 같이, 상기 한글뜻풀이 정보에 인스턴스 속성을 갖는 '특정 어구'(예를 들어, '개발한', '설립한', '제작한', '인스턴스와 동의어, 유의어, 은어 간의 결합')가 포함된 경우에도 상기 인스턴스 설정코드(IKX)를 부가할 수 있음은 물론이다.

<90> 도 3은 본 발명에 따른 판별부(120)에 의해 한글뜻풀이 정보 상에서 상술한 개념 설정코드(CKX) 또는 인스턴스 설정코드(IKX)가 부가되는 예시를 나타낸 것이다.

<91> 한글표제어가 <알니코>, <삼성애니콜>인 용어의 경우, 그 한글뜻풀이 상에 상품명 또는 기관명이 포함되어 있으므로 이는 인스턴스 속성을 갖는 것으로 판단되어 인스턴스 설정코드가 부가될 수 있다.

<92> 그리고, 한글표제어가 <영상 단국 장치>, <초전도 원소>인 용어의 경우, 그 한글뜻풀이 상에 '총칭'을 포함하고 있어, 이는 개념 속성을 갖는 것으로 판단되어 개념 설정코드가 부가될 수 있다.

<93> 한편, 상기 분류부(130)는, 해당 용어의 각 정보에 관하여 상기 개념 설정코드가 부가되는 경우 음(-)의 가중치를 부가하고, 상기 인스턴스 설정코드가 부가되는 경우 양(+)의 가중치를 부가하여, 상기 음(-)의 가중치와 양(+)의 가중치를 합산하여 음수가 나오는 경우에는 해당 용어는 개념이고, 양수가 나오는 경우에는 해당 용어는 인스턴스인 것으로 최종 분류할 수 있다.

<94> 여기서, 상기 분류부(130)는 상기 합산이 단순히 양수 또는 음수가 아닌, 특정 임계치를 갖는 양수나 특정 임계치를 갖는 음수를 기준으로 하여 그 임계치에 따라 인스턴스인지 개념인지를 분류하는 것으로 변경 가능함은 물론이다.

<95> 도 4의 마우스 선택 부분은, '디지털디바이스 LD-4200E'라는 용어가 인스턴스로 분류된 경우에 관한 예시를 나타낸 것으로서, 이를 더블클릭하거나 도 4의 화면상의 좌측 상단에 구비된 '자세히보기' 버튼을 클릭하면, 상기

'디지털디바이스 LD-4200E' 용어가 '인스턴스인 것으로 분류된 이유가 도 5와 같이 표시될 수 있다.

<96> 도 5를 살펴보면, 상기 '디지털디바이스 LD-4200E' 용어에 관한 인스턴스 설정코드가 모두 6개가 부가되어 양(+)의 가중치가 6개 발생되고, 개념 설정코드는 0개 부가되어 음(-)의 가중치는 발생되지 않음으로 하여 상기 각 가중치를 합산한 경우가 양수이므로 상기 '디지털디바이스 LD-4200E'는 인스턴스로 분류된다.

<97> 도 5를 참고로 하여 이를 좀 더 상세히 살펴보면, 상기 '디지털디바이스 LD-2400E' 용어의 한글표제어 정보는 '디지털디바이스 LD-2400E', 영어표제어 정보는 'DigitalDevice LD-4200E', 한글뜻풀이 정보는 '디지털디바이스 사(社)에서 제조한 42인치형 디지털텔레비젼이다.2007년 01월에~'이다.

<98> 여기서, 상기 영어표제어 정보의 첫 번째 알파벳이 대문자로서 고유명사 특징을 가짐에 따라 상기 영어표제어 정보에 관한 1개의 인스턴스 설정코드(IET)가 부가되었다.

<99> 또한, 상기 한글뜻풀이 정보에 숫자(2007년 01월), 회사명과 같은 집단명(디지털디바이스)이 각각 포함되어 있어, 이러한 각각에 대응되도록 2개의 인스턴스 설정코드(IKX)가 부가되었다.

<100> 이에 더하여, 상기 한글표제어 정보에 상기 회사명과 같은 집합체(디지털디바이스), 영문자(LD 등), 숫자(4200)가 포함되어 있어, 이러한 각각에 대응되도록 3개의 인스턴스 설정코드(IKT)가 부가된다.

<101> 이에 따라, 결과적으로 총 6개의 인스턴스 설정코드가 부가되고 개념 설정코드는 전혀 부가되지 않음으로 인하여 상기 '디지털디바이스 LD-2400E'는 인스턴스인 것으로 최종적으로 분류될 수 있다.

<102> 도시되어 있지는 않으나, 특정 용어에 관하여 인스턴스와 개념 규칙이 같이 나타나는 경우의 일례는, 상기 특정 용어의 영어표제어가 소문자이고 한글뜻풀이 상에는 숫자(연도 등)나 영문자 등의 인스턴스 규칙이 나오는 경우로서, 각각의 경우에 부가된 개념 설정코드와 인스턴스 설정코드 간의 가중치 합산에 의해, 상기 해당 용어가 개념인지 인스턴스인지 분류 및 판단이 가능하게 된다.

<103> 이상과 같이, 본 발명에 따른 개념 분류 규칙 및 인스턴스 분류 규칙에 있어서, 용어의 한글표제어와 영어표제어 상의 첫 알파벳의 대문자 표기 여부, 숫자 포함 여부 등의 형태정보를 이용하고, 이에 부가적으로 한글뜻풀이의 문맥정보를 주로 분석하여 규칙 설정에 이용하고 있다.

<104> 이러한 본 발명에 따른 상기 개념 및 인스턴스 분류 시스템의 성능 평가를 위해, 그 실험 대상 분야로서 IT839 정책과 9대 신성장동력 분야 중 이동통신, 디지털TV, 지능형로봇 3가지 분야로 설정하고, 실험 대상 말뭉치는 백과사전과 TTA용어사전의 한글 및 영어 표제어, 뜻풀이를 가진 용어중 2,000개를 대상으로 실험을 해보았다.

<105> 또한, 본 발명에 따른 개념 및 인스턴스의 분류 결과에 관한 정확률 측정을 위해, 상술한 실험에 앞서 상기 2,000개의 용어에 대해 미리 개념(1,000개) 및 인스턴스(1,000개)를 분류해 놓았다.

<106> 개념과 인스턴스 분류 실험결과 그 정확률은 상기 개념의 경우 97.7%, 인스턴스의 경우 86.4%로 측정되었고, 전체 정확률인 그 평균값은 92.1%로 산출되었다.

<107> 실험 결과를 살펴보면, 개념에 비해 인스턴의 분류 정확률이 다소 낮은 것을 볼 수 있는데, 이는 해당 용어에 관한 한글뜻풀이 정보가 부정확하거나 한글뜻풀이나 영어표제어 정보에 관한 데이터가 누락되어 있어서 제대로 분석되지 못한 경우이거나, 상기 판별부(120)에서 적절한 인스턴스 설정 규칙이 부족한 경우로 볼 수 있다.

<108> 그러나, 향후 기구축된 온톨로지를 참고하는 것에 의해 상기 인스턴스 설정 규칙을 더욱 세밀화하고, 문법 규칙의 확장 및 통계적 또는 기계학습 방법을 이용한다면, 상기 개념 및 인스턴스 분류에 관한 정확률을 크게 향상시킬 수 있을 것으로 기대된다.

<109> 또한, 이상과 같은 본 발명에 따르면, 추출한 용어가 인스턴스임에도 불구하고 온톨로지 자동 구축 과정을 통해 개념으로 할당되는 경우, 온톨로지의 용량 증가 및 신뢰성 저하 등의 문제가 발생될 수 있는 기존의 문제점이 해결될 수 있는 이점이 있다.

<110> 한편, 국내에서 구축된 몇 가지 온톨로지는, 코넷(Cornet), 유윈(U-Win), 에트리(ETRI) 어휘개념망, 세종 의미부류체계, 국가과학 기술 기반 정보 온톨로지, 스마트메신저 영화정보 온톨로지 등이 있다.

<111> 상기 구축된 온톨로지들은 서로 분야가 다르며 각 온톨로지마다 개념명이나 관계명, 또는 매핑된 관계, 속성의 정도 등이 동일한 것도 있지만 서로 상이한 것들 또한 많으므로, 기구축된 온톨로지들 간의 통합에 많은 어려움이 있다.

<112> 또한, 각 온톨로지에 맞게 개발된 시스템 및 응용서비스는 해당 온톨로지에 종속적이므로, 다른 온톨로지에 그

시스템 및 응용서비스를 재사용 또는 재적용하기가 곤란한 문제점이 있으나 본 발명에 따른 개념 및 인스턴스 분류 시스템(100)을 이용한다면 상술한 적용성, 호환성 등의 문제가 해결될 수 있는 이점이 있다.

<113> 이상과 같이 본 발명은 비록 한정된 실시예와 도면에 의해 설명되었으나, 본 발명은 이것에 의해 한정되지 않으며 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 본 발명의 기술 사상과 아래에 기재될 특허청구범위의 균등범위 내에서 다양한 수정 및 변형이 가능함은 물론이다.

도면의 간단한 설명

<114> 도 1은 본 발명의 실시예에 따른 온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템의 구성도,

<115> 도 2는 도 1의 판별부에서 특정 용어의 한글표제어 정보, 영어표제어 정보, 한글뜻풀이 정보에 관하여 개념 설정코드 또는 인스턴스 설정코드를 부가하기 위한 조건을 나타내는 구성도,

<116> 도 3은 도 1의 판별부에서 특정 용어의 한글뜻풀이 정보에 관하여 개념 설정코드 또는 인스턴스 설정코드가 부가되는 예시도,

<117> 도 4는 도 1의 표시부에 관한 실시예를 나타내는 화면구성도,

<118> 도 5는 도 4에서 특정 용어가 위치한 행을 클릭하는 경우 표시되는 상기 특정 용어에 관한 상세 정보를 나타내는 화면구성도,

<119> 도 6은 도 4의 '폴더열기' 버튼을 통해 입력되는 '한글표제어 정보, 영어표제어 정보, 한글뜻풀이 정보가 용어별로 저장된 텍스트 파일'에 관한 예시도이다.

<120> <도면의 주요 부분에 대한 부호의 설명>

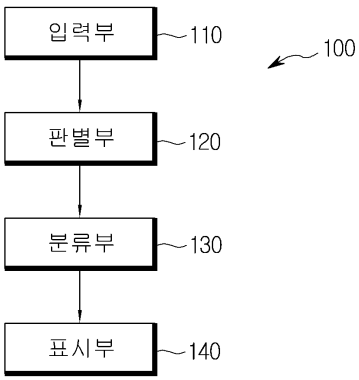
<121> 100...온톨로지 자동 구축을 위한 개념 및 인스턴스 분류 시스템

<122> 110...입력부 120...판별부

<123> 130...분류부 140...표시부

도면

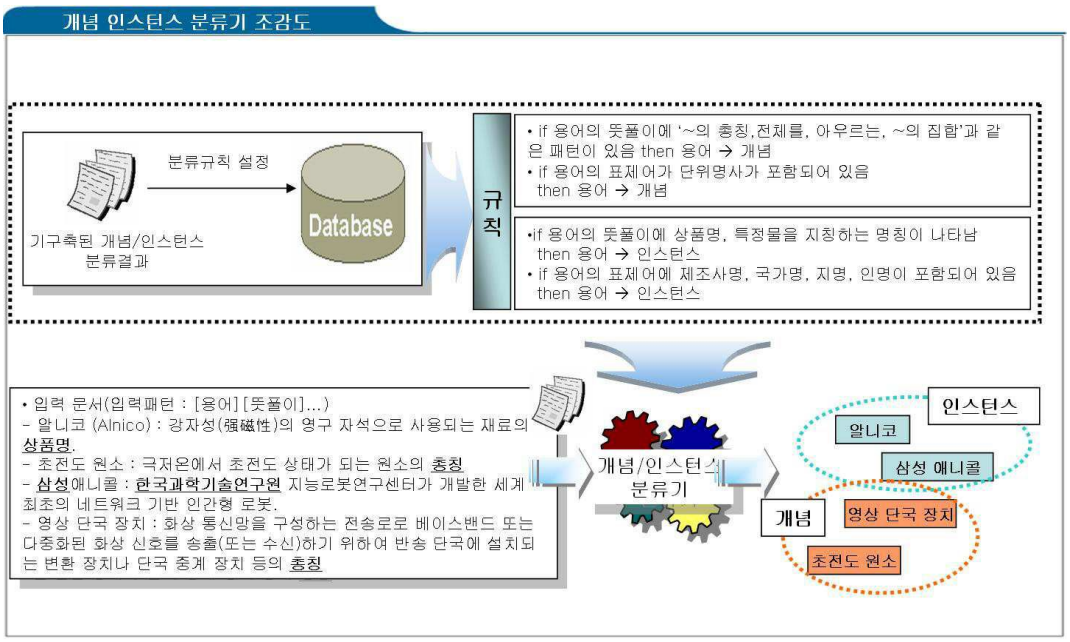
도면1



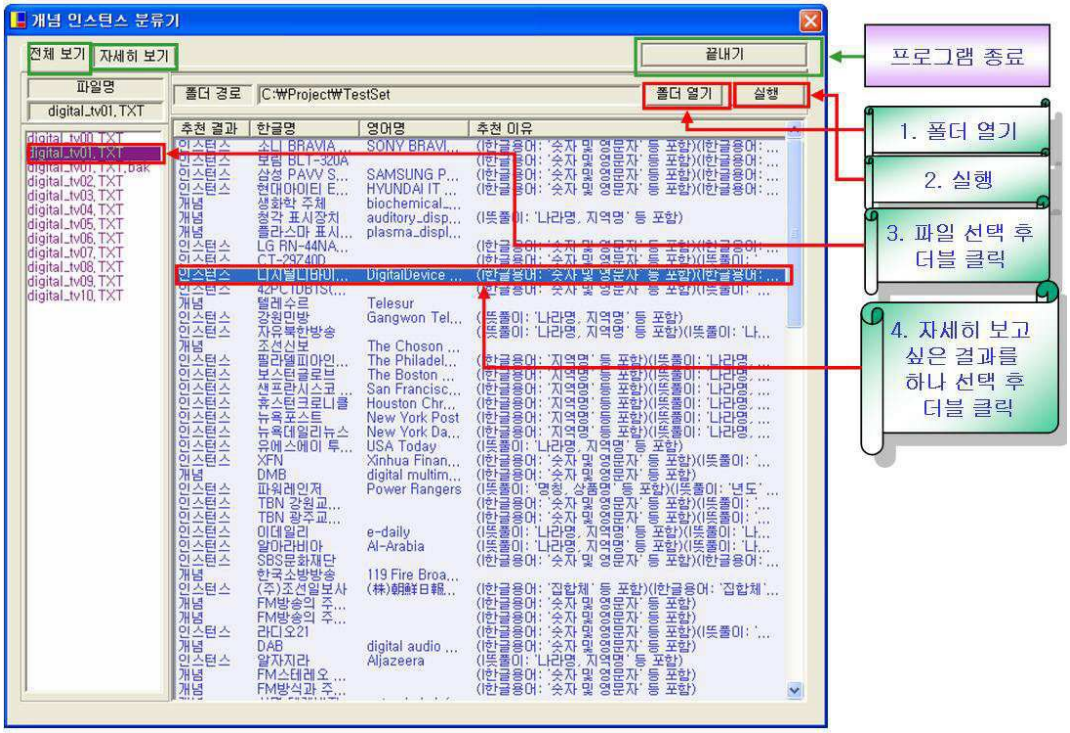
도면2


개념 규칙
인스턴스 규칙
한글 표제어
온톨로지 상위 3 depth까지의 개념과 비교
명칭들과 비교
숫자 및 영문자 확인
파생어 및 합성어 조사
영어 표제어
제일 앞의 글자 대 / 소 문자 확인
어절 별 자르기
뜻풀이
잘라낸 어절과 특정 어구 비교
잘라낸 어절과 명칭들과 비교
잘라낸 어절과 특정 어구 비교
숫자 및 영문자 확인
조사 및 어미 제거 후 비교

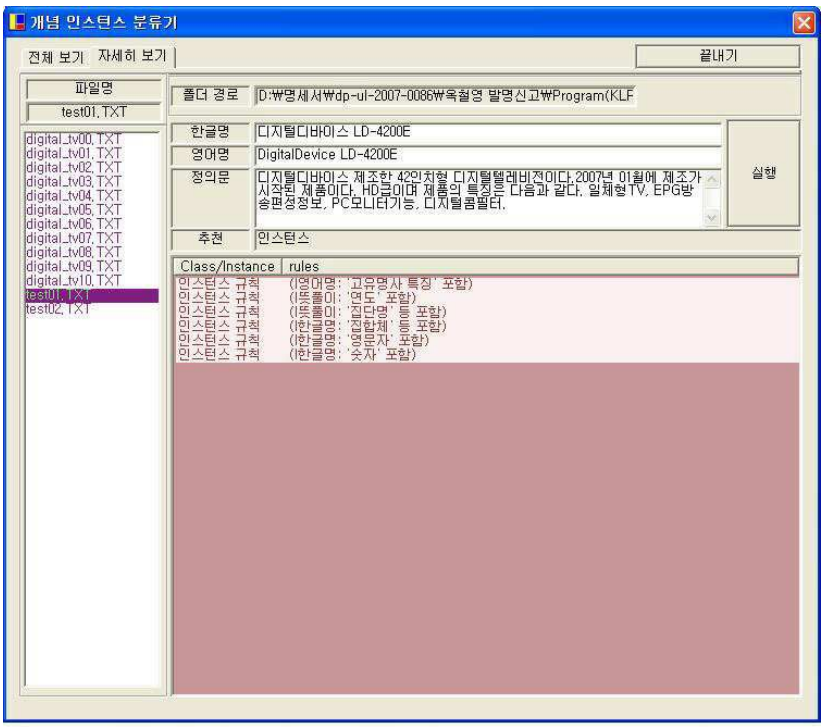
도면3


개념 인스턴스 분류기 조감도
분류규칙 설정
Database
기구축된 개념/인스턴스 분류결과
규칙
• if 용어의 뜻풀이에 '~의 총칭,전체를, 아우르는, ~의 집합'과 같은 패턴이 있음 then 용어 → 개념
• if 용어의 표제어가 단위명사가 포함되어 있음 then 용어 → 개념
•if 용어의 뜻풀이에 상품명, 특정물을 지칭하는 명칭이 나타남 then 용어 → 인스턴스
• if 용어의 표제어에 제조사명, 국가명, 지명, 인명이 포함되어 있음 then 용어 → 인스턴스
• 입력 문서(입력패턴 : [용어][뜻풀이]...)
- 알니코 (Alnico) : 강자성(强磁性)의 영구 자석으로 사용되는 재료의 상품명.
- 초전도 원소 : 극저온에서 초전도 상태가 되는 원소의 총칭
- 삼성애니콜 : 한국과학기술연구원 지능로봇연구센터가 개발한 세계 최초의 네트워크 기반 인간형 로봇.
- 영상 단국 장치 : 화상 통신망을 구성하는 전송로로 베이스밴드 또는 다중화된 화상 신호를 송출(또는 수신)하기 위하여 반송 단국에 설치되는 변환 장치나 단국 중계 장치 등의 총칭
개념/인스턴스 분류기
인스턴스
알니코
삼성 애니콜
개념
영상 단국 장치
초전도 원소

도면4


개념 인스턴스 분류기
전체 보기
자세히 보기
끝내기
프로그램 종료
파일명
digital_tv01.TXT
폴더 경로
C:₩Project₩TestSet
폴더 열기
실행
1. 폴더 열기
2. 실행
3. 파일 선택 후 더블 클릭
4. 자세히 보고 싶은 결과를 하나 선택 후 더블 클릭

도면5


개념 인스턴스 분류기
전체 보기
자세히 보기
끝내기
파일명
test01.TXT
폴더 경로
D:₩명세서₩dp-ul-2007-0086₩옥철영 발명신고₩Program(KLP
한글명
디지털디바이스 LD-4200E
영어명
DigitalDevice LD-4200E
정의문
디지털디바이스 제조한 42인치형 디지털텔레비전이다.2007년 이월에 제조가 시작된 제품이다. HD급이며 제품의 특징은 다음과 같다. 일체형TV, EPG방송편성정보, PC모니터기능, 디지털콤필터.
실행
추천
인스턴스
Class/Instance
rules

도면6


test01 - 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)
BSAC BSAC 기존의 고급 오디오 부호화(AAC:Advanced Audio Coding) 압축 방식의 성능을 일부 개선한 MPEG 4 오디오 압축 방식. 삼성전자와 삼성종합기술원 등이 개발해 제안한 압축 방식으로 뮤직캠(MUSICAM)과 함께 국내 지상파 디지털 멀티미디어 방송(DMB)의 오디오 압축 기술의 복수 표준이다.
디지털 케이블 텔레비전 Digital cable television,Digital CATV 디지털 방식의 종합 유선 방송(CATV). 고품질,다채널은 물론 고도의 양방향 서비스가 가능한 뉴미디어로서 HD급 이상의 고화질과 CD 수준의 고음질을 제공하고,실시간 주문형 비디오(RVOD:Real Video On Demand),데이터 방송,전자 상거래(T-Commerce),TV를 통한 전자 정부(T-Government),광대역 초고속 인터넷 등의 다양한 양방향 서비스를 가입자에게 제공할 수 있다.
서비스 영역 service area 라디오 방송의 청취나 텔레비전 방송의 시청이 가능한 지역 범위. 보통 전계 강도계(電界强度計)로 실측하여 중파 라디오 방송에서는 250㎶/m,텔레비전에서는 500㎶/m 이상의 전계 강도가 얻어지는 지역을 서비스 영역이라고 한다.
각광 foot light 연기자의 발 아래 쪽에서 위쪽을 향하여 투사되는 조명. 보통 대상물을 강조하거나 긴장감이나 공포감을 일으키기 위해 사용한다.
비주얼 라디오 visual radio 라디오 수신의 기본적 기능 외에 터치스크린을 통해 영상과 문자를 동시에 전송하는 양방향 라디오. 이동 전화망을 이용해 라디오 서비스를 제공하며,이메일,휴대 전화,개인 일정 관리,달력,비디오 플레이어,게임,디지털 카메라,이미지 저장,웹 검색 등 다양한 부가 기능을 겸비하고 있고,청취자와 방송국 간 양방향 커뮤니케이션이 가능하여 투표,퀴즈쇼 등 프로그램에서 청취자의 적극적인 참여를 이끌어 낼 수 있는 장점이 있다.
멀티미디어 객체 전송 Multimedia Object Transfer 지상파 디지털 멀티미디어 방송(DMB)의 표준 프로토콜의 하나. 텍스트,정지 영상,동영상,오디오 콘텐츠 등을 파일 형태로 전송하면 지상파 DMB 수신기에서 이를 수신해 재생할 수 있도록 한 프로토콜이다. 하나의 DMB 채널로는 노래를 틀어주면서 이와 연동된 다른 데이터 채널을 통해서는 방송 중인 노래와 관련된 그림이나 사진 등을 슬라이드처럼 순차적으로 제공하거나 곡명,작곡자,가수,연주자,가사,콘서트 일정 등을 음악과 함께 그림과 문자로 제공할 수 있다.
방송 웹 서비스 Broadcast Website Service 웹 사이트의 내용을 디지털 멀티미디어 방송으로 제공하는 서비스. 방송과 인터넷이 결합된 서비스를 다양한 형태로 방송에서 구현한다. 예를 들어 증권사 인터넷 사이트에서 실시간으로 제공하는 증권 시세를 그대로 디지털 멀티미디어 방송(DMB)망을 통해 송출하는 서비스이다.
브로드캐스트 멀티캐스팅 서비스 Broadcast Multicasting Service 부호 분할 다중 접속(CDMA)의 PCS (EVDO) 주파수 대역에서 멀티미디어 방송 서비스가 가능하도록 채널을 정의한 기술 규격. 위성 디지털 미디어 방송 서비스(DMB)의 대안으로 고려되고 있으며,최근 3GPP2 등 표준화 기구에서 표준화 작업 중이다.
제2 오디오 프로그램 Second Audio Program 공중파나 케이블 텔레비전(TV) 방송에서 방송하는 부가적인 제2의 음성 프로그램. 보통 제2 외국어나 프로그램 정보를 방송하기 위해 사용한다. TV 수상기에는 스테레오용으로 제2 오디오 채널이 있어 그것을 다른 목적으로 사용할 수도 있는데,이것이 좌우 스테레오