(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2008-0087994
(43) 공개일자 2008년10월02일

(51) Int. Cl.
G06F 17/27 (2006.01) *G06F 17/30* (2006.01)
(21) 출원번호 10-2007-0030277
(22) 출원일자 2007년03월28일
심사청구일자 2007년03월28일

(71) 출원인
인하대학교 산학협력단
인천 남구 용현동 253 인하대학교
(72) 발명자
이주홍
경기 고양시 일산구 마두동 강촌마을 711동 603호
김태순
서울 서초구 서초4동 서초래미안아파트 113동 1501호
(뒷면에 계속)
(74) 대리인
이은철

전체 청구항 수 : 총 10 항

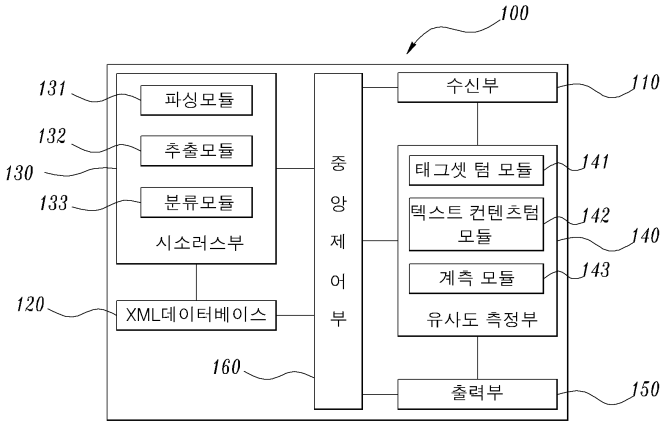
(54) 엑스엠엘 문서 유사도 측정 장치 및 그 방법

(57) 요 약

본 발명은 엑스엠엘 문서 유사도 측정 장치 및 그 방법에 관한 것으로서, 사용자로부터 입력된 질의를 수신하는 수신부; 엑스엠엘 문서를 데이터베이스로 구축하는 엑스엠엘 데이터베이스; 엑스엠엘 데이터베이스로부터 태그 및 단어를 추출해 같은 의미를 가지고 있는 태그 및 단어들을 분류하는 시소러스부; 엑스엠엘 문서의 태그 구조 정보와 의미정보와 텍스트 컨텐츠정보를 이용하여 유사도를 계산하는 유사도 측정부; 시소러스부와 유사도 측정부를 통해 계산된 유사도를 출력하는 출력부; 및 수신부, 엑스엠엘 데이터베이스, 시소러스부, 유사도 측정부 및 출력부를 중앙제어하는 중앙제어부;를 포함한다.

상기와 같은 본 발명은, 유사한 의미를 가지는 태그와 텍스트 컨텐츠 정보에 시소러스를 적용하여 유사 태그와 텍스트 내의 단어로 분류하며, 엑스엠엘 문서에서 상위태그의 의미정보를 하위태그에 반영하는 태그 셋 텀과 엑스엠엘 문서의 태그에 연결된 텍스트 길이에 따른 중요도를 고려한 텍스트 컨텐츠텀을 적용하여 유사도를 계산함으로써 정확한 유사도를 계산할 수 있는 효과가 있다.

대 표 도 - 도1



(72) 발명자

송재원

인천 남동구 간석1동 498-28

김덕환

서울 양천구 목동 940번지 아이파크 아파트 103동 1004호

특허청구의 범위

청구항 1

엑스엠엘 문서 유사도 측정장치에 있어서,

사용자로부터 입력된 질의를 수신하는 수신부(110);

엑스엠엘 문서를 데이터베이스로 구축하는 엑스엠엘 데이터베이스(120);

상기 엑스엠엘 데이터베이스로부터 태그 및 단어를 추출해 같은 의미를 가지고 있는 태그 및 단어를 분류하는 시소러스부(130);

상기 엑스엠엘 문서의 태그 구조정보, 의미정보 및 텍스트 컨텐츠정보를 이용하여 유사도를 계산하는 유사도 측정부(140);

상기 시소러스부(130)와 유사도 측정부(140)를 통해 계산된 유사도를 출력하는 출력부(150); 및

상기 수신부, 엑스엠엘 데이터베이스, 시소러스부, 유사도 측정부 및 출력부를 중앙제어하는 중앙제어부(160); 를 포함하는 것을 특징으로 하는 엑스엠엘 문서 유사도 측정 장치.

청구항 2

제 1 항에 있어서,

상기 시소러스부(130)는,

엑스엠엘 문서를 파싱하여 엑스엠엘 트리를 생성하는 파싱모듈(131);

상기 파싱모듈에서 생성한 엑스엠엘 트리로부터 태그 및 단어의 원형을 추출하는 추출모듈(132); 및

상기 추출모듈에서 추출한 원형 태그 및 단어를 의미가 같은 태그 및 단어로 분류하는 분류모듈(133); 를 포함하는 것을 특징으로 하는 엑스엠엘 문서 유사도 측정 장치.

청구항 3

제 1 항에 있어서,

상기 유사도 측정부(140)는,

엑스엠엘 문서 태그의 구조정보와 의미정보를 이용하여 유사도를 계산하는 태그셋 텀 모듈(141);

엑스엠엘 문서의 텍스트 컨텐츠정보를 이용하여 유사도를 계산하는 텍스트 컨텐츠텀 모듈(142); 및

상기 태그셋 텀 모듈의 태그셋 텀 유사도와 상기 텍스트 컨텐츠텀 모듈의 텍스트 컨텐츠텀 유사도에 대한 가중치를 이용하여 전체 유사도를 계산하는 계측 모듈(143); 을 포함하는 것을 특징으로 하는 엑스엠엘 문서 유사도 측정 장치.

청구항 4

제 3 항에 있어서,

상기 태그셋 텀 모듈(141)은,

엑스엠엘 문서에 대한 태그의 구조정보와 의미정보를 포함하는 태그 셋을 텀으로 하는 것을 특징으로 하는 엑스엠엘 문서 유사도 측정 장치.

청구항 5

제 3 항에 있어서,

상기 태그셋 텀 모듈(141)은,

태그셋 텀 빈도와 태그셋 텀 역문서빈도를 이용하여 태그셋 텀의 가중치를 계산한 후, 상기 태그셋 텀의 가중치를 이용한 태그셋 텀 유사도를 벡터의 상관도로 정량화하여 계산하는 것을 특징으로 하는 엑스엠엘 문서 유사도

측정 장치.

청구항 6

제 3 항에 있어서,

상기 텍스트 컨텐츠텀 모듈(142)은,

엑스엠엘 문서에 대한 텍스트 컨텐츠를 텀으로 하는 것을 특징으로 하는 엑스엠엘 문서 유사도 측정 장치.

청구항 7

제 3 항에 있어서,

상기 텍스트 컨텐츠텀 모듈(142)은,

텍스트 컨텐츠텀 가중빈도와 텍스트 컨텐츠텀 역문서빈도를 이용하여 텍스트 컨텐츠텀의 가중치를 계산한 후, 상기 텍스트 컨텐츠텀의 가중치를 이용한 텍스트 컨텐츠텀 유사도를 벡터의 상관도로 정량화하여 계산하는 것을 특징으로 하는 엑스엠엘 문서 유사도 측정 장치.

청구항 8

엑스엠엘 문서 유사도를 측정하는 방법에 있어서,

(a) 사용자로부터 입력된 질의를 수신하는 단계;

(b) 상기 수신한 질의와 관련된 엑스엠엘 문서를 파싱하고 엑스엠엘 트리를 생성하는 단계;

(c) 상기 엑스엠엘 트리로부터 태그와 단어의 원형을 추출하는 단계;

(d) 상기 추출한 원형 태그 및 단어에서 의미가 같은 태그 및 단어를 분류하는 단계;

(e) 상기 엑스엠엘 문서의 태그에 대한 구조정보 및 의미정보를 이용하여 태그셋 텀 유사도를 계산하는 단계;

(f) 상기 엑스엠엘 문서의 태그에 대한 텍스트 컨텐츠정보를 반영하여 텍스트 컨텐츠텀 유사도를 계산하는 단계;

(g) 상기 태그셋 텀 유사도와 상기 텍스트 컨텐츠텀 유사도에 대한 가중치를 이용하여 전체 유사도를 계산하는 단계; 및

(h) 상기 제 (g) 단계의 전체 유사도를 출력하는 단계; 를 포함하는 것을 특징으로 하는 유사도 측정 방법.

청구항 9

제 8 항에 있어서,

상기 태그셋 텀 유사도를 계산하는 단계는,

(e-1) 엑스엠엘 문서로부터 상위태그의 의미정보를 하위태그에 반영하기 위한 태그 셋 텀을 설정하는 단계;

(e-2) 태그셋 텀 빈도와 태그셋 텀 역문서빈도를 이용하여 태그셋 텀 가중치를 계산하는 단계; 및

(e-3) 상기 태그셋 텀 가중치를 이용하여 태그셋 텀 유사도를 계산하는 단계; 를 포함하는 것을 특징으로 하는 유사도 측정 방법.

청구항 10

제 8 항에 있어서,

상기 태그셋 텀 유사도를 계산하는 단계는,

(f-1) 엑스엠엘 문서로부터 텍스트 컨텐츠를 포함하는 텍스트 컨텐츠텀을 설정하는 단계;

(f-2) 텍스트 컨텐츠텀에 가중빈도와 텍스트 컨텐츠텀 역문서빈도를 이용하여 텍스트 컨텐츠텀의 가중치를 계산하는 단계; 및

(f-3) 상기 텍스트 컨텐츠텀의 가중치를 이용하여 텍스트 컨텐츠텀 유사도를 계산하는 단계; 를 포함하는 것을

특징으로 하는 유사도 측정 방법.

명 세 서

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

<14> 본 발명은 엑스엠엘 문서 유사도 측정 장치 및 그 방법에 관한 것으로서, 특히, 엑스엠엘 문서의 태그 구조정보, 의미정보 및 텍스트 컨텐츠정보를 고려하여 유사도의 정확도를 개선한 엑스엠엘 문서 유사도 측정 장치 및 그 방법에 관한 것이다.

<15> 종래, 유사문서를 검색하는 기술은 대한한국 공개특허 제1999-0048714호(인터넷 정보검색시 유사문서 우선순위 판별방법)가 이미 공개된 상태이다. 상기 인터넷 정보검색시 유사문서 우선순위 판별방법은 인터넷에서 정보 검색시 유사문서를 판별하는 방법에 있어서, 사용자로부터 키워드를 입력받으면 상기 키워드에 따른 관련 문서들을 상기 인터넷에 연결된 각 사이트의 데이터 색인으로부터 검색하는 제1과정과, 상기 검색된 관련 문서들 중에서 상기 키워드의 빈도수를 체크하여 상기 키워드의 빈도수가 높은 문서에 높은 점수를 부여하는 제2과정과, 상기 키워드의 빈도수에 따라 점수가 부여된 상기 검색된 문서들 중에서 상기 키워드의 위치를 체크하여 상기 키워드가 앞쪽에 위치한 문서에 높은 점수를 부여하는 제3과정과, 상기 키워드의 문서내 위치에 따라 점수가 부여된 상기 검색된 문서들 중에서 상위 디렉토리에 존재하는 문서에 높은 점수를 부여하는 제4과정과, 상기 검색된 문서들을 상기 제2과정과 제3과정 및 제4과정에서 부여된 점수가 높은 순으로 상기 웹 브라우저 화면에 디스플레이시키는 제5과정을 구비함을 특징으로 한다.

<16> 하지만, 문서의 상위태그의 의미정보를 하위태그에 반영하지 못해 유사문서 계산에 정확도가 떨어지는 문제점이 있다. 그리고, 태그에 연결된 텍스트 컨텐츠정보도 문장 길이에 따른 중요도가 유사도 계산에 반영되지 않아 유사문서 계산에 정확도가 떨어지는 문제점이 있다.

발명이 이루고자 하는 기술적 과제

<17> 본 발명의 목적은, 엑스엠엘 문서의 태그 구조정보, 의미정보 및 텍스트 컨텐츠정보를 고려한 엑스엠엘 문서 유사도 측정 장치 및 그 방법을 제공함에 있다.

<18> 본 발명의 다른 목적은, 시소러스를 적용하여 유사한 의미를 가지는 태그와 텍스트 내의 단어를 분류함으로써 유사도를 계산하도록 하는 엑스엠엘 문서 유사도 측정 장치 및 그 방법을 제공함에도 있다.

발명의 구성 및 작용

<19> 본 발명에 따른 엑스엠엘 문서 유사도 측정장치는, 사용자로부터 입력된 질의를 수신하는 수신부; 엑스엠엘 문서를 데이터베이스로 구축하는 엑스엠엘 데이터베이스; 상기 엑스엠엘 데이터베이스로부터 태그 및 단어를 추출해 같은 의미를 가지고 있는 태그 및 단어를 분류하는 시소러스부; 상기 엑스엠엘 문서의 태그 구조정보, 의미정보 및 텍스트 컨텐츠정보를 이용하여 유사도를 계산하는 유사도 측정부; 상기 시소러스부와 유사도 측정부를 통해 계산된 유사도를 출력하는 출력부; 및 상기 수신부, 엑스엠엘 데이터베이스, 시소러스부, 유사도 측정부 및 출력부를 중앙제어하는 중앙제어부; 를 포함한다.

<20> 바람직하게, 상기 시소러스부는 엑스엠엘 문서를 파싱하여 엑스엠엘 트리를 생성하는 파싱모듈; 상기 파싱모듈에서 생성한 엑스엠엘 트리로부터 태그 및 단어의 원형을 추출하는 추출모듈; 및 상기 추출모듈에서 추출한 원형 태그 및 단어를 의미가 같은 태그 및 단어로 분류하는 분류모듈; 를 포함하는 것을 특징으로 한다.

<21> 그리고 바람직하게, 상기 유사도 측정부는 엑스엠엘 문서 태그의 구조정보와 의미정보를 이용하여 유사도를 계산하는 태그셋 텀 모듈; 엑스엠엘 문서의 텍스트 컨텐츠정보를 이용하여 유사도를 계산하는 텍스트 컨텐츠텀 모듈; 및 상기 태그셋 텀 모듈의 태그셋 텀 유사도와 상기 텍스트 컨텐츠텀 모듈의 텍스트 컨텐츠텀 유사도에 대한 가중치를 이용하여 전체 유사도를 계산하는 계측 모듈; 을 포함하는 것을 특징으로 한다.

<22> 더욱 바람직하게, 상기 태그셋 텀 모듈은 엑스엠엘 문서에 대한 태그의 구조정보와 의미정보를 포함하는 태그셋을 텀으로 하는 것을 특징으로 한다.

<23> 또한 더욱 바람직하게, 상기 태그셋 텀 모듈은 태그셋 텀 빈도와 태그셋 텀 역문서빈도를 이용하여 태그셋 텀의 가중치를 계산한 후, 상기 태그셋 텀의 가중치를 이용한 태그셋 텀 유사도를 벡터의 상관도로 정량화하여 계산하는 것을 특징으로 한다.

<24> 또한 더욱 바람직하게, 상기 텍스트 컨텐츠텀 모듈은 엑스엠엘 문서에 대한 텍스트 컨텐츠를 텀으로 하는 것을 특징으로 한다.

<25> 그리고 더욱 바람직하게, 상기 텍스트 컨텐츠텀 모듈은 텍스트 컨텐츠텀 가중빈도와 텍스트 컨텐츠텀 역문서빈도를 이용하여 텍스트 컨텐츠텀의 가중치를 계산한 후, 상기 텍스트 컨텐츠텀의 가중치를 이용한 텍스트 컨텐츠텀 유사도를 벡터의 상관도로 정량화하여 계산하는 것을 특징으로 한다.

<26> 한편, 엑스엠엘 문서 유사도를 측정하는 방법에 있어서, (a) 사용자로부터 입력된 질의를 수신하는 단계; (b) 엑스엠엘 데이터베이스(120)로부터 상기 수신한 질의와 관련된 엑스엠엘 문서를 파싱하고 엑스엠엘 트리를 생성하는 단계; (c) 상기 엑스엠엘 트리로부터 태그와 단어의 원형을 추출하는 단계; (d) 상기 추출한 원형 태그 및 단어에서 의미가 같은 태그 및 단어를 분류하는 단계; (e) 상기 엑스엠엘 문서의 태그에 대한 구조정보 및 의미정보를 이용하여 태그셋 텀 유사도를 계산하는 단계; (f) 상기 엑스엠엘 문서의 태그에 대한 텍스트 컨텐츠정보를 반영하여 텍스트 컨텐츠텀 유사도를 계산하는 단계; (g) 상기 태그셋 텀 유사도와 상기 텍스트 컨텐츠텀 유사도에 대한 가중치를 이용하여 전체 유사도를 계산하는 단계; 및 (h) 상기 제 (g) 단계의 전체 유사도를 출력하는 단계; 를 포함하는 것을 특징으로 한다.

<27> 바람직하게, 상기 태그셋 텀 유사도를 계산하는 단계는 (e-1) 엑스엠엘 문서로부터 상위태그의 의미정보를 하위태그에 반영하기 위한 태그 셋 텀을 설정하는 단계; (e-2) 태그셋 텀 빈도와 태그셋 텀 역문서빈도를 이용하여 태그셋 텀 가중치를 계산하는 단계; 및 (e-3) 상기 태그셋 텀 가중치를 이용하여 태그셋 텀 유사도를 계산하는 단계; 를 포함하는 것을 특징으로 한다.

<28> 그리고 바람직하게, 상기 태그셋 텀 유사도를 계산하는 단계는 (f-1) 엑스엠엘 문서로부터 텍스트 컨텐츠를 포함하는 텍스트 컨텐츠텀을 설정하는 단계; (f-2) 텍스트 컨텐츠텀에 가중빈도와 텍스트 컨텐츠텀 역문서빈도를 이용하여 텍스트 컨텐츠텀의 가중치를 계산하는 단계; 및 (f-3) 상기 텍스트 컨텐츠텀의 가중치를 이용하여 텍스트 컨텐츠텀 유사도를 계산하는 단계; 를 포함하는 것을 특징으로 한다.

<29> 본 발명의 일실시예에 따른 엑스엠엘 문서 유사도 측정 장치에 관하여 도 1을 참조하여 설명하면 다음과 같다.

<30> 도 1은 본 발명의 일실시예에 따른 엑스엠엘 문서 유사도 측정 장치에 관한 블록도이며, 도 2는 본 발명의 일실시예에 따른 엑스엠엘 트리를 나타낸 도면이다.

<31> 본 발명의 일실시예에 따른 엑스엠엘 문서 유사도 측정 장치(100)는 도 1에 도시된 바와 같이, 수신부(110), 엑스엠엘 데이터베이스(120), 시소러스부(130), 유사도 측정부(140), 출력부(150) 및 중앙제어부(160)를 포함한다.

<32> 수신부(110)는 사용자로부터 입력된 질의를 수신하는 기능을 수행한다.

<33> 또한, 엑스엠엘 데이터베이스(120)는 엑스엠엘 문서를 저장하는 기능을 수행한다.

<34> 또한, 시소러스부(130)는 엑스엠엘 데이터베이스(120)의 엑스엠엘 문서로부터 태그 및 단어를 추출해 같은 의미를 가지고 있는 태그 및 단어를 분류하는 기능을 수행한다. 즉, picture와 image 또는 author와 writer는 서로 다른 태그와 단어로 표현하였지만 같은 의미를 가진 한 개의 태그와 단어로 일치시켜 분류한다.

<35> 여기서, 시소러스부(130)는 파싱모듈(131), 추출모듈(132) 및 분류모듈(133)을 포함한다.

<36> 파싱모듈(131)은 엑스엠엘 문서를 파싱하여 도 2에 도시된 바와 같은 엑스엠엘 트리를 생성하는 기능을 수행한다.

<37> 또한 추출모듈(132)은 파싱모듈(131)에서 생성한 엑스엠엘 트리의 태그 및 단어의미를 파악하기 위해 태그 및 단어의 원형을 추출하는 기능을 수행한다. 즉, 추출모듈(132)은 엑스엠엘 트리의 태그 및 단어의미를 파악하기 위해 공백, 하이픈 등이 걸러지고, "an apple"은 "apple"로, "cars"는 "car"로 원형을 추출한다.

<38> 그리고 분류모듈(133)은 oracle의 wordnet2.0을 사용하여 추출모듈(132)에서 추출한 원형 태그 및 단어에서 의미가 같은 태그 및 단어를 분류하는 기능을 수행한다. 이때, 대표 태그 및 단어는 분류된 태그 및 단어를 알파벳 순서로 나열한 중에서 첫 번째 위치한 것으로 한다. 본 실시예에서, 태그 및 단어의 의미를 분류하기 위해

oracle의 wordnet2.0을 사용하는 것으로 설정하였으나, 본 발명이 이에 한정되는 것은 아니다.

<39> 그리고, 유사도 측정부(140)는 엑스엠엘 문서의 태그 구조정보, 의미정보 및 텍스트 컨텐츠정보에 벡터모델을 적용함으로써, 유사도를 계산하는 기능을 수행한다.

<40> 유사도 측정부(140)는 이러한 기능을 수행하기 위해 태그셋 텀 모듈(141), 텍스트 컨텐츠텀 모듈(142) 및 계측 모듈(143)을 포함한다.

<41> 태그셋 텀 모듈(141)은 태그의 구조정보 및 의미정보를 이용하여 유사도를 계산한다.

<42> 벡터모델을 적용하기 위해서는 텀(term)을 정의하여야 한다. 태그를 텀으로 간주하면 태그 일부의 의미정보만을 반영하므로, 태그 구조정보와 의미정보를 동시에 반영할 수 없다. 예를 들면, movie/actor/name인 제1 트리와 zoo/animal/name인 제2 트리가 있다고 가정하면, 제1 트리의 하위 태그인 name과 제2 트리의 하위 태그인 name은 태그 이름이 동일하지만, 의미정보가 다르다. 제1 트리의 하위태그 name은 어떤 영화에 출연한 배우의 이름이고, 제2 트리의 하위태그 name은 동물원의 어떤 동물 이름이다. 그러므로, 벡터모델을 적용해 유사도의 정확한 계산을 하기 위해서는 태그의 구조정보와 의미정보를 포함하는 태그 패스(path)를 반영한 태그 셋(tag set)을 텀으로 사용해야 한다.

<43> 태그 셋을 적용하는 것은 패스 구조가 다르지만 의미정보가 같은 패스를 반영하기 위해서이다. 예를 들면, toy/shop인 제1 패스와 shop/toy인 제2 패스가 있다고 가정하면, 제1 패스와 제2 패스의 패스구조는 다르지만 의미정보는 같다. 즉, 단독 태그만을 고려한 shop과 toy는 서로 다른 의미정보를 가지고 있어 다른 태그로 분류된다.

<44> 하지만, 제1 패스의 하위태그 shop은 상위태그 toy의 의미정보를 포함한 장난감을 취급하는 shop의 의미정보이고, 제2 패스의 하위태그 toy도 상위태그 shop의 의미정보를 포함한 shop중 toy를 취급하는 같은 의미정보가 된다. 즉, 패스구조에서 하위태그는 상위태그의 의미를 포함하는 포괄적 의미정보를 지닌다. 그러므로 태그셋 텀 모듈(141)은 벡터모델을 적용한 유사도의 정확한 계산을 하기 위해 포괄적 의미정보를 지니고 있는 태그 셋을 텀으로 표현한다. 여기서, 태그 셋은 패스의 순서가 없는 태그의 집합이다.

표 1

<45>

tag set term	tag set term	tag set term	tag set term
{page}	{shop,page}	{book,location}	{book,shop,location}
{lname}	{book,page}	{book,heading}	{book,shop,heading}
{fname}	{shop,writer}	{book,story}	{book,shop,story}
{shop}	{book,shop}	{shop,book,page}	{music,shop,heading}
{writer}	{music,shop}	{shop,writer,lname}	{music,shop,story}
{music}	{shop,image}	{shop,writer,fname}	{book,shop,writer,lname}
{image}	{shop,location}	{book,shop,writer}	{book,shop,writer,fname}
{location}	{shop,heading}	{music,shop,writer}	{music,shop,writer,lname}
{heading}	{shop,story}	{music,shop,location}	{music,shop,writer,fname}
{story}	{writer,lname}	{book,writer,lname}	
{book}	{writer,fname}	{book,writer,fname}	
{book,writer}	{book,image}	{book,shop,image}	

<46> 표 1에 도시된 바와 같이, 이때 사용되는 텀을 태그 셋 텀(tag set term)이라 한다. 이러한 태그 셋 텀의 가중치(TSTW:Tag Set Term Weight)는 태그셋 텀 빈도(TSTF:Tag Set Term Frequency)와 태그셋 텀 역문서빈도(TSTIDF:Tag Set Term Inverse Document Frequency)의 곱이며, 다음 수학식 1과 같다.

수학식 1

<47>
$$TSTW_{ij} = TSTF_{ij} \times TSTIDF_{j}$$

<48> $TSTF_{ij}$는 i^{th} 문서에서 j^{th} 태그 셋 텀의 빈도를 나타내고, $TSTIDF_{j}$는 다음 수학식 2와 같다.

수학식 2

<49>
$$TSTIDF_{j} = \log \frac{N}{DF_{j}}$$

<50> 이때, N은 전체 문서의 개수를 나타내고, DF_j는 문서빈도수로서 N 개의 문서들 중에서 j^{th}태그 셋 텀이 존재하는 문서 수를 나타낸다

<51> 태그 셋 텀의 가중치(TSTW)를 사용한 유사도(TSTS:Tag Set Term Similarity)는 벡터모델에서 문서d_i 와 질의 d_j의 유사도 측정은 두 벡터 D_i와 D_j의 상관도로 구할 수 있으며, 이 상관도는 두 벡터 간 사이각의 코사인 값으로 다음 수학식 3과 같이 정량화될 수 있다.

수학식 3

<52>
$$TSTS(d_i, d_j) = \frac{D_i \cdot D_j}{|D_i| \times |D_j|} = \frac{\sum_{k=1}^{t} TSTW_{ki} \times TSTW_{kj}}{\sqrt{\sum_{k=1}^{t} TSTW_{ki}^2} \times \sqrt{\sum_{k=1}^{t} TSTW_{kj}^2}}$$

<53> 여기서, t는 d_i, d_j 내에 전체 텀의 수를 나타내며, 두 벡터D_i와 D_j는 다음 수학식 4와 같이 표현된다.

수학식 4

<54>
$$D_i = (TSTW_{1i}, TSTW_{2i}, \ldots TSTW_{ti}), D_j = (TSTW_{1j}, TSTW_{2j}, \ldots TSTW_{tj})$$

<55> 엑스엠엘 문서의 태그 구조만으로 유사도를 계산하면, 태그 구조정보는 유사하지만 의미정보가 다른 문서가 함께 유사문서로 분류되는 경우가 발생한다. 예를 들어, 도 2에 도시된 바와 같이, 문서1(doc1)과 문서2(doc2)는 책의 목록문서를 문서3(doc3)은 음악정보문서를 나타내는 엑스엠엘 트리에서, 문서1과 문서3의 구조정보는 유사하지만 전체적인 의미는 다르다. 즉, 문서1과 문서3의 book, page, location 및 music은 서로 다르지만, shop, writer, lname, fname, story 및 title은 같다.

<56> book/shop와 music/shop의 shop은 같은 태그명을 가지지만, book/shop은 책을 파는 가게를 가리키는 의미정보를 가지며, music/shop은 음반을 파는 가게를 가리키는 의미정보를 가지고 있다. 반면, 문서1과 문서2는 shop, story, writer, lname, fname, title, page 및 book 태그는 서로 같고, 그 외의 picture, location 및 image 태그들은 서로 다르다. 여기서, picture와 image는 서로 다른 구조지만 같은 의미정보를 담고 있고, book/shop 패스와 shop/book패스도 다른 패스를 가지고 있지만 같은 의미정보를 담고 있다.

표 2

<57>

	distance	similarity=1-distance
doc1,doc2	0.77868	0.22132
doc1,doc3	0.34335	0.65665
doc2,doc3	0.74678	0.25322

<58> 따라서, 태그의 구조 정보만을 사용하면 표 2에 도시된 바와 같이, 문서1과 문서3 사이의 유사도가 문서1과 문서2 사이보다 유사도가 더 높게 된다. 그러나, 태그 구조정보와 의미정보를 반영하면 표 3에 도시된 바와 같이, 문서1과 문서2 사이의 유사도가 문서1과 문서3 사이보다 유사도가 더 높게 나타낸다.

표 3

<59>

	distance	similarity=1-distance
doc1,doc2	0.3680	0.63199
doc1,doc3	0.5436	0.45642
doc2,doc3	0.7499	0.25013

<60> 태그셋 텀 빈도, 태그셋 텀 역문서빈도 및 태그 셋 텀의 가중치를 계산한 결과는 표 4에 도시된 바와 같다.

표 4

tag set term	TSTF			TSIDF			TSTW			tag set term	TSTF			TSIDF			TSTW		
	doc1	doc2	doc3	doc1	doc2	doc3	doc1	doc2	doc3		doc1	doc2	doc3	doc1	doc2	doc3	doc1	doc2	doc3
page	1	1	0	1.176	1.176	0	1.176	1.176	0	music,shop,heading	0	0	1	0	0	1.477	0	0	1.477
shop,page	1	0	0	1.477		0	1.477	0	0	music,shop,story	0	0	1	0	0	1.477	0	0	1.477
shop,book,page	1	1	0	1.176	1.176	0	1.176	1.176	0	shop,image	1	0	0	1.477	0	0	1.477	0	0
book,page	0	1	0		1.477	0	0	1.477	0	shop,location	1	0	1	1.176	0	1.176	1.176	0	1.176
lname	2	2	1	1	1	1	2	2	1	shop,heading	1	0	1	1.176	0	1.176	1.176	0	1.176
fname	2	2	1	1	1	1	2	2	1	shop,story	1	0	1	1.176	0	1.176	1.176	0	1.176
shop	1	1	1	1	1	1	1	1	1	writer	2	2	1	1	1	1	2	2	1
shop,writer	2	0	1	1.176	0	1.176	2.352	0	1.176	writer,lname	2	2	1	1	1	1	2	2	1
shop,writer,lname	2	0	1	1.176	0	1.176	2.352	0	1.176	writer,fname	2	2	1	1	1	1	2	2	1
shop,writer,fname	2	0	1	1.176	0	1.176	2.352	0	1.176	book,writer	0	2	0	0	1.477	0	0	2.954	0
book,shop,writer	2	2	0	1.176	1.176	0	2.352	2.352	0	book,writer,lname	0	2	0	0	1.477	0	0	2.954	0
book,shop,writer,lname	2	2	0	1.176	1.176	0	2.352	2.352	0	book,writer,fname	0	2	0	0	1.477	0	0	2.954	0
book,shop,writer,fname	2	2	0	1.176	1.176	0	2.352	2.352	0	book	1	1	0	1.176	1.176	0	1.176	1.176	0
music,shop,writer	0	0	1	0	0	1.477	0	0	1.477	book,image	0	1	0	0	1.477	0	0	1.477	0
music,shop,writer,lname	0	0	1	0	0	1.477	0	0	1.477	book,location	0	1	0	0	1.477	0	0	1.477	0
music,shop,writer,fname	0	0	1	0	0	1.477	0	0	1.477	book,heading	0	1	0	0	1.477	0	0	1.477	0
book,shop	1	1	0	1.176	1.176	0	1.176	1.176	0	book,story	0	1	0	0	1.477	0	0	1.477	0
book,shop,image	1	1	0	1.176	1.176	0	1.176	1.176	0	music	0	0	1	0	0	1.477	0	0	1.477
book,shop,location	1	1	0	1.176	1.176	0	1.176	1.176	0	image	1	1	0	1.176	1.176	0	1.176	1.176	0
book,shop,heading	1	1	0	1.176	1.176	0	1.176	1.176	0	location	1	1	1	1	1	1	1	1	1
book,shop,story	1	1	0	1.176	1.176	0	1.176	1.176	0	heading	1	1	1	1	1	1	1	1	1
music,shop	0	0	1	0	0	1.477	0	0	1.477	story	1	1	1	1	1	1	1	1	1
music,shop,location	0	0	1	0	0	1.477	0	0	1.477										

<61>

<62> 또한, 텍스트 컨텐츠텀 모듈(142)은 텍스트 컨텐츠정보를 이용하여 유사도를 계산한다.

<63> 엑스엠엘 문서의 텍스트 컨텐츠정보는 일반문서의 텍스트 컨텐츠정보와는 다르게 태그 구조에 연결되어 포함되어 있다. 동일한 출현빈도를 갖는 텍스트 컨텐츠라도 태그의 위치에 따라 중요도가 달리 표현되므로 태그 구조에 포함된 텍스트 컨텐츠의 가중치를 다르게 반영해야 한다. 즉, 만약 논문 보고서 형식의 엑스엠엘 문서를 예로 든다면, 제목, 요약과 같은 태그들은 그 안의 텍스트 컨텐츠의 길이가 대체로 짧지만, 텍스트 컨텐츠 단어들의 중요도는 높다고 할 수 있다. 소개글과 같은 태그는 대체로 그 안의 텍스트의 길이가 길며 그 안의 단어들의 중요도는 제목, 요약 안의 단어들에 비해서 그 중요도가 비교적 낮다고 할 수 있다.

<64> 여기서, 텍스트 컨텐츠텀은 표 5에 도시된 엑스엠엘 문서들로부터 표 6에 도시된 바와 같은 텍스트 컨텐츠로 작성된 텀을 의미한다.

표 5

<book> <shop> <location>seoul</location> <writer> <lname>Michal</lname><fname>cruze</fname> </writer> <writer> <lname>Michal</lname><fname>cruze</fname> </writer> <title>cook book using flower dynamic</title> <story>The first time in book is dynamic history. The wide modern man and wide modern woman read a newspaper.</story> <picture>b1 book,b2 book</picture> <page>67</page> </shop> </book>	<shop> <book> <location>pusan</location> <writer> <lname>Michal</lname><fname>cruze</fname> </writer> <writer> <lname>Michal</lname><fname>cruze</fname> </writer> <title>comic book and flower book</title> <story>The first time in modern man is dynamics. And modern people is dynamic flower book.</story> <image>c1 book</image> <page>120</page> </book> </shop>	<music> <shop> <location>seoul</location> <writer> <lname>Michal</lname><fname>cruze</fname> </writer> <story>The first time in music, from modern people</story> <title>instrumental music</title> </shop> </music>
[doc 1]	[doc 1]	[doc 1]

<65>

표 6

text content term	text content term	text content term
book	instrument	read
comic	michal	seoul
cook	modern	time
cruze	music	us
dynam	newspap	wide
flower	peopl	woman
histori	pusan	

<66>

<67> 이러한 텍스트 컨텐츠텀의 가중치(TCTW:Text Content Term Weight)는 텍스트 컨텐츠텀 가중빈도(TCTWF:Text Content Term Weighted Frequency)와 텍스트 컨텐츠텀 역문서빈도(TCTIDF:Text Content Term Inverse Document Frequency)의 곱이며, 다음 수학식 5와 같다.

수학식 5

<68> $$TCTW_{ij} = TCTWF_{ij} \times TCTIDF_{j}$$

<69> 텍스트 컨텐츠텀 가중빈도(TCTWF)는 다음 수학식 6과 같다.

수학식 6

$$TCTWF_{ij} = \sum_{k=1}^{n} (cfreq_{ikj} \times (1 + \log(\frac{\#\,text\ content\ term\ in\ i^{th}\ document}{\#\,text\ content\ term\ in\ k^{th}\ element})))$$

<70>

<71> 이때, k는 문서의 모든 태그 인덱스이고, $cfreq_{ikj}$는 i^{th}문서의 k^{th}태그에 포함되어 있는 j^{th}텍스트 컨텐츠텀의 빈발수이다. (1+log(#text content term in i^{th} document/#text content term in k^{th} element))는 문장 길이에 따른 태그의 중요도를 계산하는 식이다. 여기서, log함수는 완만한 값이 나오도록 하기 위해서 사용된다.

<72> 텍스트 컨텐츠텀 역문서빈도(TCTIDF)는 다음 수학식 7과 같다.

수학식 7

$$TCTIDF_j = \log \frac{N}{DF_j}$$

<73>

<74> 이때, N은 전체 문서의 개수를 나타내고, DF_j는 문서빈도수로서 N개의 문서들 중에서 j^{th}텍스트 컨텐츠텀이 존재하는 문서 수를 나타낸다.

<75> 텍스트 컨텐츠텀 가중치(TCTW)를 사용한 유사도를 텍스트 컨텐츠텀 유사도(TCTS:Text Content Term Similarity)라고 하며, 벡터모델에서 문서d_i 와 질의 d_j의 유사도 측정은 두 벡터 D_i와 D_j의 상관도로 구할 수 있으며, 이 상관도는 두 벡터 간 사이각의 코사인 값으로 다음 수학식 8과 같이 정량화될 수 있다.

수학식 8

$$TCTS(d_i, d_j) = \frac{D_i \cdot D_j}{|D_i| \times |D_j|} = \frac{\sum_{k=1}^{t} TCTW_{ki} \times TCTW_{kj}}{\sqrt{\sum_{k=1}^{t} TCTW_{ki}^2} \times \sqrt{\sum_{k=1}^{t} TCTW_{kj}^2}}$$

<76>

<77> 여기서, t는 d_i, d_j내에 전체 텀의 수를 나타내며, 두 벡터 D_i와 D_j는 다음 수학식 9와 같이 표현된다.

수학식 9

<78> $$D_i = (TCTW_{1i}, TCTW_{2i}, \ldots TCTW_{ti}), D_j = (TCTW_{1j}, TCTW_{2j}, \ldots TCTW_{tj})$$

표 7

text content	TCTWF			TCTIDF			TCTW		
	doc1	doc2	doc3	doc1	doc2	doc3	doc1	doc2	doc3
book	2,9	4,58	0	1,18	1,18	0	3,42	5,39	0
comic	0	1,63	0	0	1,48	0	0	2,41	0
cook	1,62	0	0	1,48	0	0	2,4	0	0
cruze	4,64	4,46	1,95	1	1	1	4,64	4,46	1,95
dynam	2,9	2,65	0	1,18	1,18	0	3,42	3,12	0
flower	1,62	2,96	0	1,18	1,18	0	1,91	3,48	0
histori	1,28	0	0	1,48	0	0	1,89	0	0
instrument	0	0	1,65	0	0	1,48	0	0	2,44
michal	4,64	4,46	1,95	1	1	1	4,64	4,46	1,95
modern	2,56	2,65	1,35	1	1	1	2,56	2,65	1,35
music	0	0	3,01	0	0	1,48	0	0	4,44
newspap	1,28	0	0	1,48	0	0	1,89	0	0
peopl	0	1,33	1,35	0	1,18	1,18	0	1,56	1,59
pusan	0	2,23	0	0	1,48	0	0	3,29	0
read	1,28	0	0	1,48	0	0	1,89	0	0
seoul	2,32	0	1,95	1,18	0	1,18	2,73	0	2,3
time	1,28	1,33	1,35	1	1	1	1,28	1,33	1,35
us	1,62	0	0	1,48	0	0	2,4	0	0
wide	2,56	0	0	1,48	0	0	3,78	0	0
woman	1,28	0	0	1,48	0	0	1,89	0	0

<79>

<80> 엑스엠엘 문서는 일반문서와는 다른 구조를 지니고 있다. 따라서, 엑스엠엘 문서의 텍스트 컨텐츠정보의 유사도

를 측정할 경우 태그를 고려한 방법을 사용해야 한다. 이때, 시소러스를 이용하여 텍스트 내의 서로 유사한 단어들을 하나의 단어로 일치시켜 분류한다. 일반 문서 방법으로 유사도 측정을 계산하면 문서1과 문서3 사이의 유사도가 문서1과 문서2 사이의 유사도 보다 높다. 그러나 텍스트 컨텐츠정보를 포함한 태그를 고려한 방법으로 유사도를 측정하면 문서1과 문서2 사이의 유사도가 문서1과 문서3 사이의 유사도보다 더 높다. 그러므로, 엑스엠엘 문서의 컨텐츠정보와 함께 태그 구조도 동시에 고려해야 좀더 신중한 유사도 계산이 되는 것을 알 수 있다.

<81> 텍스트 컨텐츠텀 가중빈도(TCTWF), 텍스트 컨텐츠텀 역문서빈도(TCTIDF) 및 텍스트 컨텐츠텀 가중치(TCTW)를 계산한 결과는 표 7에 도시된 바와 같다.

표 8

<82>

	distance	similarity=1-distance
doc1,doc2	0.29905	0.70095
doc1,doc3	0.60545	0.39455
doc2,doc3	0.65277	0.34723

<83> 엑스엠엘 문서1, 엑스엠엘 문서2 및 엑스엠엘 문서3의 유사도 측정결과는 표 8에 도시된 바와 같다.

<84> 그리고 계측 모듈(143)은 태그셋 텀 모듈(141)의 태그셋 텀 유사도와 텍스트 컨텐츠텀 모듈(142)의 텍스트 컨텐츠텀 유사도에 대한 가중치를 이용하여 다음 수학식 10과 같이 전체 유사도를 계산한다.

수학식 10

<85>
$$\begin{cases} sim(d_1,d_2) = tsw \times TSTS(d_1,d_2) + tcw \times TCTS(d_1,d_2) \\ tsw + tcw = 1 \end{cases}$$

<86> 여기서, tsw는 태그셋 텀 유사도의 가중치이고, tcw는 텍스트 컨텐츠텀 유사도의 가중치이다.

<87> 또한, 출력부(150)는 상술한 시소러스부(130)와 유사도 측정부(140)를 통해 계산된 전체 유사도를 출력하는 기능을 수행한다.

<88> 그리고, 중앙제어부(160)는 수신부(110), 엑스엠엘 데이터베이스(120), 시소러스부(130), 유사도 측정부(140) 및 출력부(150)를 중앙제어하는 기능을 수행한다.

<89> 상술한 구성을 가지는 본 발명의 일실시예에 따른 엑스엠엘 문서 유사도 측정 장치에 대한 정확도 실험 결과에 관하여 표 9를 참조하여 설명하면 다음과 같다.

표 9

<90>

	윈스콘데이터	오아시스데이터
일반문서에 벡터모델을 적용한 유사도	0.75329	0.81538
XML문서 구조 정보만을 반영한 유사도	0.79934	0.83154
XML문서의 태그 셋 텀과 텍스트 컨텐츠텀을 반영한 유사도	0.83553	0.87692
시소러스를 사용한 XML문서의 태그 셋 텀과 텍스트 컨텐츠텀을 반영한 유사도	0.92846	0.94955

<91> 표 9는 일반 문서에 벡터모델을 적용한 유사도, 엑스엠엘 문서의 구조 정보만을 반영한 유사도, 엑스엠엘 문서의 태그 셋 텀과 텍스트 컨텐츠텀을 반영한 유사도, 시소러스를 사용한 엑스엠엘 문서의 태그 셋 텀과 텍스트 컨텐츠텀을 반영한 유사도를 비교한 것이다. 여기서, 실험데이터는 윈스콘데이터 및 오아시스데이터로 실험하였다.

<92> 정확도 실험은 퓨리티(purity)를 사용해 정확도를 측정한다. 이러한 퓨리티는 각 클러스터가 기본적으로 하나의 클래스로부터 데이터들을 포함하게 되는 범위를 측정하고, 개별 클러스터 퓨리티의 가중치 합계로서 다음 수학

식 11과 같다. 여기서 클러스터는 파일을 저장하는 논리적 단위이다.

수학식 11

$$purity = \sum_{i=1}^{K} \frac{n_i}{n} P(S_i)$$

<93>

<94> 여기에서, S_i는 크기가 n_i인 i^{th} 클러스터이며, K는 클러스터의 수, 그리고 n은 문서의 총수를 나타낸다. 클러스터 S_i의 개별 클러스터 퓨리티인 P(S_i)는 다음 수학식 12와 같다.

수학식 12

$$P(S_i) = \frac{1}{n_i} \max_j (n_i^j)$$

<95>

<96> 여기에서, n_i^j 는 j번째 클러스터에 지정된 i번째 입력 클래스의 문서들의 수를 나타낸다. 이때, 퓨리티 값이 높을수록 정확도가 더 높다.

<97> 이러한 실험 결과는 표 9에 도시된 바와 같이, 엑스엠엘 문서의 구조정보와 태그 구조정보와 컨텐츠정보에 시소러스를 적용시킨 유사도가 더 높은 정확도를 나타내는 것을 알 수 있다.

<98> 한편, 상술한 구성을 가지는 본 발명의 일실시예에 따른 엑스엠엘 문서 유사도 측정 장치를 이용한 방법(이하, 유사도 측정 방법)에 관하여 도 3 내지 도 5를 참조하여 설명하면 다음과 같다.

<99> 도 3은 본 발명의 일실시예에 따른 유사도 측정 방법에 관한 전체흐름도이고, 도 4는 본 발명의 일실시예에 따른 태그셋 텀 유사도 계산에 관한 상세흐름도이며, 도 5는 본 발명의 일실시예에 따른 텍스트 컨텐츠텀 유사도 계산에 관한 상세흐름도이다.

<100> 본 발명의 일실시예에 따른 유사도 측정 방법에 관한 전체적인 흐름을 도 3을 참조하여 설명하면 다음과 같다.

<101> 도 3에 도시된 바와 같이, 엑스엠엘 문서 유사도 측정 장치(100)의 수신부(110)는 사용자로부터 입력된 질의를 수신한다(S2).

<102> 엑스엠엘 문서 유사도 측정 장치(100)의 시소러스부(130)는 엑스엠엘 데이터베이스(120)로부터 수신한 질의와 관련된 엑스엠엘 문서를 파싱하고 엑스엠엘 트리를 생성한 후(S4), 엑스엠엘 트리로부터 태그와 단어의 원형을 추출하고(S6), 추출한 원형 태그와 단어에서 의미가 같은 태그와 단어를 분류한다(S8).

<103> 엑스엠엘 문서 유사도 측정 장치(100)의 유사도 측정부(140)는 태그의 구조정보 및 의미정보를 이용하여 태그셋 텀 유사도를 계산한다(S10).

<104> 엑스엠엘 문서 유사도 측정 장치(100)의 유사도 측정부(140)는 텍스트 컨텐츠정보를 이용하여 텍스트 컨텐츠텀 유사도를 계산한다(S12).

<105> 엑스엠엘 문서 유사도 측정 장치(100)의 유사도 측정부(140)는 제 S10 단계의 태그셋 텀 유사도와 제 S12 단계의 텍스트 컨텐츠텀 유사도에 대한 가중치를 이용하여 전체 유사도를 계산한다(S14).

<106> 엑스엠엘 문서 유사도 측정 장치(100)의 출력부(150)는 계산한 전체 유사도를 출력한다(S16).

<107> 다음으로 상술한 제 S10 단계의 태그셋 텀 유사도 계산에 대하여 도 4를 참조하여 상세히 살펴본다.

<108> 도 4에 도시된 바와 같이, 엑스엠엘 문서 유사도 측정 장치(100)의 유사도 측정부(140)는 엑스엠엘 문서로부터 상위태그의 의미정보를 하위태그에 반영하기 위한 태그 셋 텀을 설정한다(S22).

<109> 엑스엠엘 문서 유사도 측정 장치(100)의 유사도 측정부(140)는 태그셋 텀 빈도와 태그셋 텀 역문서빈도를 이용하여 태그셋 텀 가중치를 계산한다(S24).

<110> 엑스엠엘 문서 유사도 측정 장치(100)의 유사도 측정부(140)는 제 S24 단계에서 계산한 태그셋 텀 가중치를 이용한 태그셋 텀 유사도를 두 벡터의 상관도로 정량화하여 계산한다(S26).

<111> 다음으로 상술한 제 S12 단계의 텍스트 컨텐츠텀 유사도 계산에 대하여 도 5를 참조하여 상세히 살펴본다.

<112> 도 5에 도시된 바와 같이, 엑스엠엘 문서 유사도 측정 장치(100)의 유사도 측정부(140)는 엑스엠엘 문서로부터 텍스트 컨텐츠로 작성된 텍스트 컨텐츠텀을 설정한다(S32).

<113> 엑스엠엘 문서 유사도 측정 장치(100)의 유사도 측정부(140)는 텍스트 컨텐츠텀 가중빈도와 텍스트 컨텐츠텀 역문서빈도를 이용하여 텍스트 컨텐츠텀의 가중치를 계산한다(S34).

<114> 엑스엠엘 문서 유사도 측정 장치(100)의 유사도 측정부(140)는 제 S34 단계에서 계산한 텍스트 컨텐츠텀의 가중치를 이용한 텍스트 컨텐츠텀 유사도를 두 벡터의 상관도로 정량화하여 계산한다(S36).

<115> 이상으로 본 발명의 기술적 사상을 예시하기 위한 바람직한 실시예와 관련하여 설명하고 도시하였지만, 본 발명은 이와 같이 도시되고 설명된 그대로의 구성 및 작용에만 국한되는 것이 아니며, 기술적 사상의 범주를 일탈함이 없이 본 발명에 대해 다수의 변경 및 수정이 가능함을 당업자들은 잘 이해할 수 있을 것이다. 따라서, 그러한 모든 적절한 변경 및 수정과 균등물들도 본 발명의 범위에 속하는 것으로 간주되어야 할 것이다.

발명의 효과

<116> 상기와 같은 본 발명은, 유사한 의미를 가지는 태그와 텍스트 컨텐츠 정보에 시소러스를 적용하여 유사 태그와 텍스트 내의 단어로 분류하며, 엑스엠엘 문서에서 상위태그의 의미정보를 하위태그에 반영하는 태그 셋 텀과 엑스엠엘 문서의 태그에 연결된 텍스트 길이에 따른 중요도를 고려한 텍스트 컨텐츠텀을 적용하여 유사도를 계산함으로써 정확한 유사도를 계산할 수 있는 효과가 있다.

도면의 간단한 설명

<1> 도 1은 본 발명의 일실시예에 따른 엑스엠엘 문서 유사도 측정 장치에 관한 블록도.

<2> 도 2는 본 발명의 일실시예에 따른 엑스엠엘 트리를 나타낸 도면.

<3> 도 3은 본 발명의 일실시예에 따른 유사도 측정 방법에 관한 전체흐름도.

<4> 도 4는 본 발명의 일실시예에 따른 태그셋 텀 유사도 계산에 관한 상세흐름도.

<5> 도 5는 본 발명의 일실시예에 따른 텍스트 컨텐츠텀 유사도 계산에 관한 상세흐름도.

<6> <도면의 주요 부분에 대한 부호의 설명>

<7> 100 : 엑스엠엘 문서 유사도 측정 장치

<8> 110 : 수신부 120 : 엑스엠엘 데이터베이스

<9> 130 : 시소러스부 131 : 파싱모듈

<10> 132 : 추출모듈 133 : 분류모듈

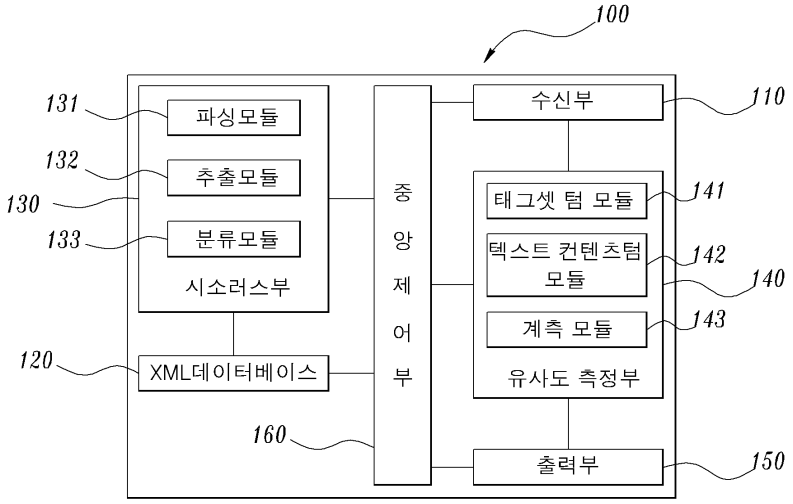
<11> 140 : 유사도 측정부 141 : 태그셋 텀 모듈

<12> 142 : 텍스트 컨텐츠텀 모듈 143 : 계측 모듈

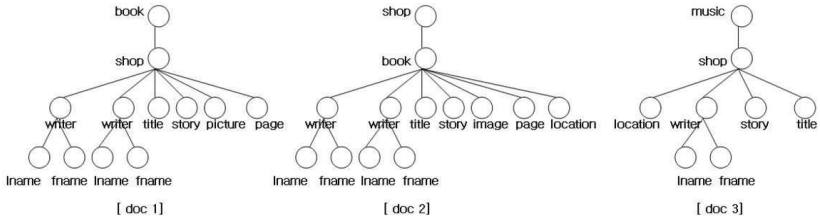
<13> 150 : 출력부 160 : 중앙제어부

도면

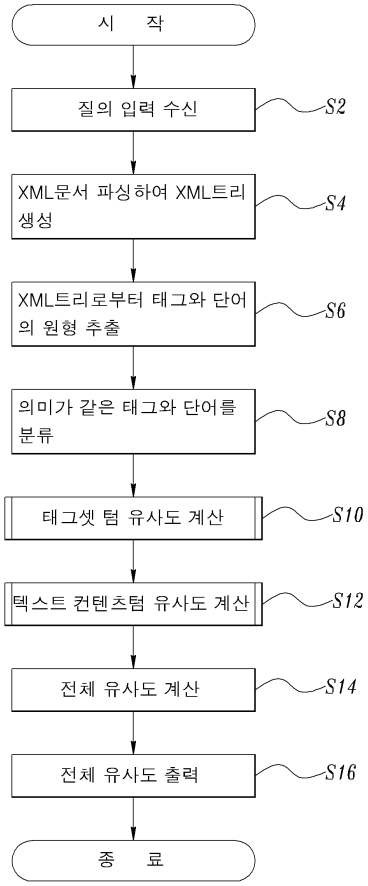
도면1


100
131
파싱모듈
132
추출모듈
130
133
분류모듈
시소러스부
120
XML데이터베이스
160
중앙제어부
수신부
110
태그셋 템 모듈
141
텍스트 컨텐츠템 모듈
142
140
계측 모듈
143
유사도 측정부
출력부
150

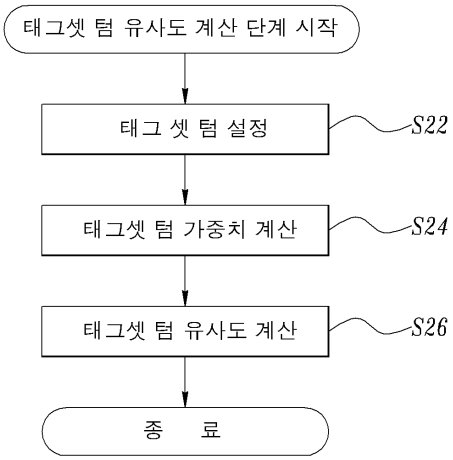
도면2


book
shop
writer
writer
title
story
picture
page
lname
fname
lname
fname
[doc 1]
shop
book
writer
writer
title
story
image
page
location
lname
fname
lname
fname
[doc 2]
music
shop
location
writer
story
title
lname
fname
[doc 3]

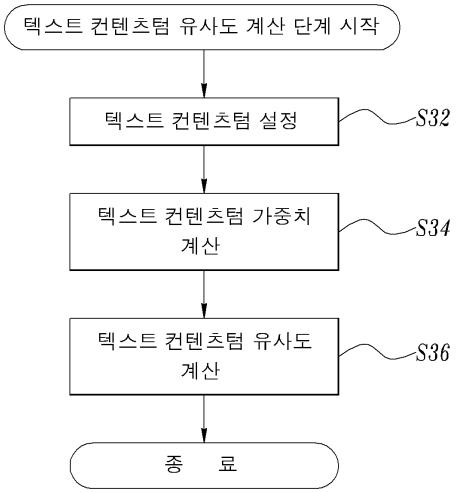
도면3


시 작
질의 입력 수신
S2
XML문서 파싱하여 XML트리 생성
S4
XML트리로부터 태그와 단어의 원형 추출
S6
의미가 같은 태그와 단어를 분류
S8
태그셋 텀 유사도 계산
S10
텍스트 컨텐츠텀 유사도 계산
S12
전체 유사도 계산
S14
전체 유사도 출력
S16
종 료

도면4


태그셋 텀 유사도 계산 단계 시작
태그 셋 텀 설정
S22
태그셋 텀 가중치 계산
S24
태그셋 텀 유사도 계산
S26
종 료

도면5


텍스트 컨텐츠텀 유사도 계산 단계 시작
텍스트 컨텐츠텀 설정
S32
텍스트 컨텐츠텀 가중치 계산
S34
텍스트 컨텐츠텀 유사도 계산
S36
종 료