(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2010-0122298
(43) 공개일자 2010년11월22일

(51) Int. Cl.
G06F 17/21 (2006.01) *G06F 17/27* (2006.01)
(21) 출원번호 10-2009-0041272
(22) 출원일자 2009년05월12일
심사청구일자 2009년05월12일

(71) 출원인
인하대학교 산학협력단
인천 남구 용현동 253 인하대학교
(72) 발명자
조태호
인천광역시 남구 용현동 230번지 인하대학교
(74) 대리인
특허법인무한

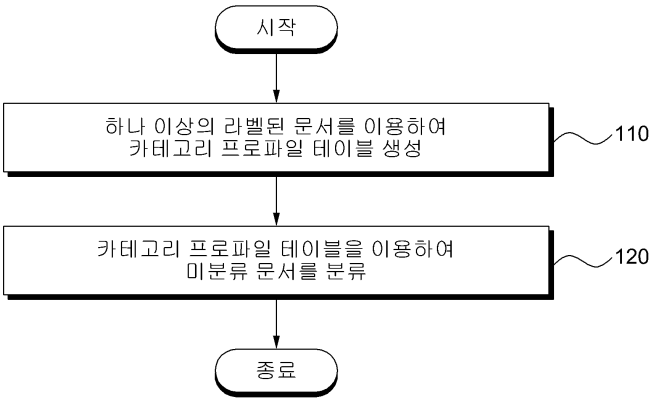
전체 청구항 수 : 총 12 항

(54) 전자문서 자동 분류 방법 및 장치

(57) 요 약

전자문서 자동 분류 방법 및 장치가 개시된다. 전자문서를 자동으로 분류하는 방법은, 하나 이상의 라벨된 문서를 이용하여 카테고리 프로파일 테이블을 생성하는 단계 및 상기 카테고리 프로파일 테이블을 이용하여 미분류 문서를 분류하는 단계를 포함한다.

대 표 도 - 도1



특허청구의 범위

청구항 1

하나 이상의 라벨된(labeled) 문서를 이용하여 카테고리 프로파일 테이블을 생성하는 단계; 및

상기 카테고리 프로파일 테이블을 이용하여 미분류 문서를 분류하는 단계

를 포함하는, 전자문서 자동 분류 방법.

청구항 2

제1항에 있어서,

상기 카테고리 프로파일 테이블을 생성하는 단계는,

상기 하나 이상의 라벨된 문서에 포함된 텍스트를 하나의 텍스트로 통합하는 단계; 및

상기 하나의 텍스트를 하나 이상의 단어 및 상기 하나 이상의 단어에 대한 가중치를 포함하는 카테고리 프로파일 테이블로 인코딩하는 단계

를 포함하는, 전자문서 자동 분류 방법.

청구항 3

제1항에 있어서,

상기 분류하는 단계는,

상기 미분류 문서를 하나 이상의 단어 및 상기 하나 이상의 단어에 대한 가중치를 포함하는 미분류 문서 테이블로 인코딩하는 단계;

상기 미분류 문서 테이블과 하나 이상의 카테고리 프로파일 테이블의 유사도를 각각 판단하는 단계; 및

상기 각 유사도에 기초하여 상기 미분류 문서를 분류하는 단계

를 포함하는, 전자문서 자동 분류 방법.

청구항 4

제2항 또는 제3항에 있어서,

상기 인코딩하는 단계는,

문서 내의 텍스트를 서브스트링으로 분리하여 토큰(token)을 생성하는 단계;

상기 토큰을 원형으로 변환하여 하나 이상의 원형 단어를 생성하는 단계;

상기 문서 내의 접속사, 관사, 전치사, 대명사 중 적어도 하나를 포함하는 불사용어를 제거하는 단계; 및

상기 원형 단어에 대응하는 가중치를 할당하는 단계

를 포함하는, 전자문서 자동 분류 방법.

청구항 5

제3항에 있어서,

상기 유사도를 판단하는 단계는,

하나의 카테고리 프로파일 테이블을 선택하는 단계;

상기 미분류 문서 테이블과 상기 선택된 카테고리 프로파일 테이블의 공통 단어를 추출하는 단계; 및

상기 미분류 문서 테이블로부터 상기 공통 단어에 대응하는 가중치를 합산하여 유사도 스코어를 산출하는 단계

를 포함하는, 전자문서 자동 분류 방법.

청구항 6

제3항에 있어서,

상기 유사도를 판단하는 단계는,

하나의 카테고리 프로파일 테이블을 선택하는 단계;

상기 미분류 문서 테이블과 상기 선택된 카테고리 프로파일 테이블의 공통 단어를 추출하는 단계; 및

상기 카테고리 프로파일 테이블로부터 상기 공통 단어에 대응하는 가중치를 합산하여 유사도 스코어를 산출하는 단계

를 포함하는, 전자문서 자동 분류 방법.

청구항 7

제3항에 있어서,

상기 유사도를 판단하는 단계는,

하나의 카테고리 프로파일 테이블을 선택하는 단계;

상기 미분류 문서 테이블과 상기 선택된 카테고리 프로파일 테이블의 공통 단어를 추출하는 단계;

상기 공통 단어 별로 상기 미분류 문서 테이블 및 상기 선택된 카테고리 프로파일 테이블의 가중치 값을 곱하여 곱셈 스코어를 생성하는 단계; 및

상기 공통 단어 별 곱셈 스코어를 합산하여 유사도 스코어를 산출하는 단계

를 포함하는, 전자문서 자동 분류 방법.

청구항 8

제5항 내지 제7항 중 어느 하나의 항에 있어서,

상기 각 유사도에 기초하여 상기 미분류 문서를 분류하는 단계는,

상기 유사도 스코어가 가장 큰 카테고리 프로파일 테이블에 대응하는 카테고리를 상기 미분류 문서의 카테고리로 분류하는 것을 특징으로 하는 전자문서 자동 분류 방법.

청구항 9

하나 이상의 라벨된(labeled) 문서를 이용하여 카테고리 프로파일 테이블을 생성하는 카테고리 프로파일 생성부; 및

상기 카테고리 프로파일 테이블을 이용하여 미분류 문서를 분류하는 문서 분류부

를 포함하는, 전자문서 자동 분류 장치.

청구항 10

제9항에 있어서,

상기 카테고리 프로파일 생성부는,

상기 하나 이상의 라벨된 문서에 포함된 텍스트를 하나의 텍스트로 통합하는 텍스트 통합부; 및

상기 하나의 텍스트를 하나 이상의 단어 및 상기 하나 이상의 단어에 대한 가중치를 포함하는 카테고리 프로파일 테이블로 인코딩하는 인코딩부

를 포함하는, 전자문서 자동 분류 장치.

청구항 11

제9항에 있어서,

상기 문서 분류부는,

상기 미분류 문서를 하나 이상의 단어 및 상기 하나 이상의 단어에 대한 가중치를 포함하는 미분류 문서 테이블로 인코딩하는 인코딩부;

상기 미분류 문서 테이블과 하나 이상의 카테고리 프로파일 테이블의 유사도를 각각 판단하는 유사도 판단부; 및

상기 각 유사도에 기초하여 상기 미분류 문서를 분류하는 분류부

를 포함하는, 전자문서 자동 분류 장치.

청구항 12

제10항 또는 제11항에 있어서,

상기 인코딩부는,

텍스트를 서브스트링으로 분리하여 토큰(token)을 생성하는 토큰 생성부;

상기 토큰을 원형(root)으로 변환하여 원형 단어를 생성하는 원형 생성부;

상기 문서 내의 접속사, 관사, 전치사, 대명사 중 적어도 하나를 포함하는 불사용어를 제거하는 불사용어 제거부; 및

상기 원형 단어에 대응하는 가중치를 할당하는 가중치 할당부

를 포함하는, 전자문서 자동 분류 장치.

명 세 서

발명의 상세한 설명

기 술 분 야

[0001] 본 발명은 전자문서 자동 분류 방법 및 장치에 관한 것으로 특히, 문서를 테이블로 인코딩하고, 카테고리 프로파일을 이용하여 전자문서를 자동으로 분류하는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 최근, 문서를 자동으로 분류하는 기술은 문서 관리 시스템, 텍스트 정보 시스템 등의 발전을 위해 중요한 기술로 인식되고 있다. 문서 분류는 미분류된 문서를 미리 설정된 카테고리 중 하나의 카테고리로 할당하여 수행할 수 있다.

[0003] 종래에는 문서 분류를 위해 문서 내의 텍스트를 수치벡터(numerical vector)로 표현해야 하며, 수치벡터로 표현해야 하는 벡터가 고차원화되고, 수치벡터가 대부분 0값으로 채워질 수 있었다.

[0004] 따라서, 수치벡터를 사용하지 않고 문서 자동 분류의 효율을 높일 수 있는 문서 자동 분류 방법 및 장치가 요구된다.

발명의 내용

해결 하고자하는 과제

[0005] 본 발명은 문서를 텍스트로 인코딩하고 미리 분류된 카테고리 프로파일을 사용하여 문서를 자동 분류함으로써, 문서 자동 분류를 위한 연산량을 줄일 수 있는 문서 자동 분류 방법 및 장치를 제공한다.

[0006] 본 발명은 문서를 텍스트로 인코딩하고 미리 분류된 카테고리 프로파일을 사용하여 문서를 자동 분류함으로써, 문서 자동 분류의 성능을 향상시키고, 분류의 정확성을 향상시키는 문서 자동 분류 방법 및 장치를 제공한다.

과제 해결수단

[0007] 본 발명의 일실시예에 따른 전자문서 자동 분류 방법은, 하나 이상의 라벨된(labeled) 문서를 이용하여 카테고

리 프로파일 테이블을 생성하는 단계 및 상기 카테고리 프로파일 테이블을 이용하여 미분류 문서를 분류하는 단계를 포함한다

[0008] 본 발명의 일측면에 따르면, 상기 카테고리 프로파일 테이블을 생성하는 단계는, 상기 하나 이상의 라벨된 문서에 포함된 텍스트를 하나의 텍스트로 통합하는 단계; 및 상기 하나의 텍스트를 하나 이상의 단어 및 상기 하나 이상의 단어에 대한 가중치를 포함하는 카테고리 프로파일 테이블로 인코딩하는 단계를 포함할 수 있다.

[0009] 본 발명의 일측면에 따르면, 상기 분류하는 단계는, 상기 미분류 문서를 하나 이상의 단어 및 상기 하나 이상의 단어에 대한 가중치를 포함하는 미분류 문서 테이블로 인코딩하는 단계, 상기 미분류 문서 테이블과 하나 이상의 카테고리 프로파일 테이블의 유사도를 각각 판단하는 단계 및 상기 각 유사도에 기초하여 상기 미분류 문서를 분류하는 단계를 포함할 수 있다.

[0010] 본 발명의 일측면에 따르면, 상기 인코딩하는 단계는, 문서 내의 텍스트를 서브스트링으로 분리하여 토큰(token)을 생성하는 단계, 상기 토큰을 원형으로 변환하여 하나 이상의 원형 단어를 생성하는 단계, 상기 문서 내의 접속사, 관사, 전치사, 대명사 중 적어도 하나를 포함하는 불사용어를 제거하는 단계 및 상기 원형 단어에 대응하는 가중치를 할당하는 단계를 포함할 수 있다.

[0011] 본 발명의 일실시예에 따른 전자문서 자동 분류 장치는, 하나 이상의 라벨된(labeled) 문서를 이용하여 카테고리 프로파일 테이블을 생성하는 카테고리 프로파일 생성부 및 상기 카테고리 프로파일 테이블을 이용하여 미분류 문서를 분류하는 문서 분류부를 포함한다

[0012] 본 발명의 일측면에 따르면, 상기 문서 분류부는, 상기 미분류 문서를 하나 이상의 단어 및 상기 하나 이상의 단어에 대한 가중치를 포함하는 미분류 문서 테이블로 인코딩하는 인코딩부, 상기 미분류 문서 테이블과 하나 이상의 카테고리 프로파일 테이블의 유사도를 각각 판단하는 유사도 판단부 및 상기 각 유사도에 기초하여 상기 미분류 문서를 분류하는 분류부를 포함할 수 있다.

[0013] 본 발명의 일측면에 따르면, 상기 유사도 판단부는, 하나의 카테고리 프로파일 테이블을 선택하는 카테고리 프로파일 선택부, 상기 미분류 문서 테이블과 상기 선택된 카테고리 프로파일 테이블의 공통 단어를 추출하는 공통단어 추출부 및 상기 미분류 문서 테이블로부터 상기 공통 단어에 대응하는 가중치를 합산하여 유사도 스코어를 산출하는 스코어 산출부를 포함할 수 있다.

[0014] 본 발명의 일측면에 따르면, 상기 유사도 판단부는, 하나의 카테고리 프로파일 테이블을 선택하는 카테고리 프로파일 선택부, 상기 미분류 문서 테이블과 상기 선택된 카테고리 프로파일 테이블의 공통 단어를 추출하는 공통단어 추출부 및 상기 카테고리 프로파일 테이블로부터 상기 공통 단어에 대응하는 가중치를 합산하여 유사도 스코어를 산출하는 스코어 산출부를 포함할 수 있다.

[0015] 본 발명의 일측면에 따르면, 상기 유사도 판단부는, 하나의 카테고리 프로파일 테이블을 선택하는 카테고리 프로파일 선택부, 상기 미분류 문서 테이블과 상기 선택된 카테고리 프로파일 테이블의 공통 단어를 추출하는 공통단어 추출부, 상기 공통 단어 별로 상기 미분류 문서 테이블 및 상기 선택된 카테고리 프로파일 테이블의 가중치 값을 곱하여 곱셈 스코어를 생성하는 곱셈 스코어 산출부 및 상기 공통 단어 별 곱셈 스코어를 합산하여 유사도 스코어를 산출하는 유사도 스코어 산출부를 포함할 수 있다.

[0016] 본 발명의 일측면에 따르면, 상기 분류부는, 상기 유사도 스코어가 가장 큰 카테고리 프로파일 테이블에 대응하는 카테고리를 상기 미분류 문서의 카테고리로 분류할 수 있다.

효 과

[0017] 본 발명의 일실시예에 따르면, 문서를 텍스트로 인코딩하고 미리 분류된 카테고리 프로파일을 사용하여 문서를 자동 분류함으로써, 문서 자동 분류를 위한 연산량을 줄일 수 있는 문서 자동 분류 방법 및 장치가 제공된다.

[0018] 본 발명의 일실시예에 따르면, 문서를 텍스트로 인코딩하고 미리 분류된 카테고리 프로파일을 사용하여 문서를 자동 분류함으로써, 문서 자동 분류의 성능을 향상시키고, 분류의 정확성을 향상시키는 문서 자동 분류 방법 및 장치가 제공된다.

발명의 실시를 위한 구체적인 내용

[0019] 이하, 첨부된 도면들에 기재된 내용들을 참조하여 본 발명의 실시예들을 상세하게 설명한다. 다만, 본 발명이 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조부호는 동일한 부재를 나타

낸다. 전자문서 자동 분류 방법은 전자문서 자동 분류 장치를 구성하는 구성 요소에 의해 수행될 수 있다.

[0020] 도 1은 본 발명의 일실시예에 있어서, 전자문서 자동 분류 방법을 도시한 흐름도이다.

[0021] 도 1을 참고하면, 단계(110)에서는 하나 이상의 라벨된(labeled) 문서를 이용하여 카테고리 프로파일 테이블을 생성할 수 있다. 즉, 라벨된 문서를 이용하여 수치벡터를 생성하지 않고, 카테고리별 테이블을 생성할 수 있다. 여기서, 단계(110)는 도 2 및 도 3을 참고하여 이하에서 더욱 상세하게 설명한다.

[0022] 도 2는 도 1에 도시된 카테고리 프로파일 테이블을 생성하는 과정을 나타낸 동작 흐름도이다.

[0023] 도 2를 참고하면, 카테고리 프로파일을 생성하기 위해 단계(210)에서는 상기 하나 이상의 라벨된 문서에 포함된 텍스트를 하나의 텍스트로 통합할 수 있다. 예를 들어, 동일한 카테고리에 속하는 문서1, 문서2, 및 문서3이 존재하는 경우, 상기 문서1, 문서2, 및 문서3에 포함된 텍스트를 하나의 텍스트로 통합할 수 있다.

[0024] 단계(220)에서는 상기 하나의 텍스트를 하나 이상의 단어 및 상기 하나 이상의 단어에 대한 가중치를 포함하는 카테고리 프로파일 테이블로 인코딩할 수 있다. 즉, 상기 통합된 하나의 텍스트에서 단어들을 추출하고, 추출된 단어들에 대응하는 가중치를 할당하여 카테고리 프로파일 테이블로 인코딩할 수 있다. 여기서, 카테고리 프로파일 테이블 인코딩을 위해 텍스트를 서브스트링으로 분리하여 토큰(token)을 생성하는 토큰화(tokenization), 상기 토큰을 원형으로 변환하여 하나 이상의 원형 단어를 생성하는 스테밍(stemming), 상기 문서 내의 접속사, 관사, 전치사, 대명사 중 적어도 하나를 포함하는 불사용어를 제거하는 불사용어 제거(stop word removal), 및 상기 원형 단어에 대응하는 가중치를 할당하는 가중치 할당(term weighting) 중 적어도 하나를 수행할 수 있다.

[0025] 카테고리 프로파일 테이블 인코딩의 일예로, 도 3에 도시된 바와 같이, 카테고리1에 속하는 하나 이상의 문서(310)에 포함된 텍스트를 통합하고, 상기 통합된 텍스트로부터 카테고리 프로파일 테이블을 생성할 수 있다. 따라서, 각 카테고리별로 카테고리에 속하는 문서의 텍스트를 통합하고, 통합된 텍스트로부터 각각의 카테고리 프로파일 테이블을 생성할 수 있다.

[0026] 다시 도 1을 참고하면, 단계(120)에서는 상기 카테고리 프로파일 테이블을 이용하여 미분류 문서를 분류할 수 있다. 따라서, 수치벡터를 이용하지 않고, 테이블을 이용하여 미분류 문서를 분류할 수 있다. 여기서, 단계(120)은 도 4 및 도 5를 참고하여 이하에서 더욱 상세하게 설명한다.

[0027] 도 4는 도 1에 도시된 미분류 문서를 분류하는 과정을 나타낸 동작 흐름도이다.

[0028] 도 4를 참고하면, 단계(410)에서는 상기 미분류 문서를 하나 이상의 단어 및 상기 하나 이상의 단어에 대한 가중치를 포함하는 미분류 문서 테이블로 인코딩할 수 있다. 여기서, 인코딩하는 과정은 도 5를 참고하여 더욱 상세하게 설명한다.

[0029] 도 5를 참고하면, 단계(510)에서는 문서 내의 텍스트를 서브스트링으로 분리하여 토큰을 생성하는 토큰화를 수행할 수 있다. 여기서, 상기 토큰화는 텍스트를 토큰이라는 서브스트링(substring)으로 분리(segmenting)하는 과정이며, 공백(white spaces), 구두점(punctuation marks), 줄바꿈(carridge returns) 등에 의해 토큰화될 수 있다. 이때, 특수문자 및 수치값 등은 제거될 수 있다.

[0030] 단계(520)에서는 상기 토큰을 원형으로 변환하여 하나 이상의 원형 단어를 생성할 수 있다. 예를 들어, 과거, 현재, 미래 시제를 가지는 동사, 복수 형태를 가지는 명사 등이 원형으로 변환될 수 있다.

[0031] 단계(530)에서는 상기 문서 내의 접속사, 관사, 전치사, 대명사 중 적어도 하나를 포함하는 불사용어를 제거하는 불사용어 제거가 수행될 수 있다. 여기서, 상기 불사용어 제거는 문서의 내용(content)과 관련이 없고 문법적으로만 사용되는 단어를 제거하는 과정일 수 있다. 또한, 불사용어는 접속사, 관사, 전치사, 대명사 중 적어도 하나를 포함할 수 있다. 불사용어를 제거는 카테고리 분류를 위해 필요없는 단어를 제거하여 카테고리 분류를 보다 용이하게 수행하기 위해 수행될 수 있다.

[0032] 단계(540)에서는 상기 원형 단어에 대응하는 가중치를 할당할 수 있다. 여기서, 가중치를 할당을 통해 각 단어가 문서의 내용과 관련하여 얼마나 중요한 단어인지를 나타낼 수 있다. 여기서, 가중치 할당은 아래 [수학식 1]에 따라 할당할 수 있다.

수학식 1

[0033] $$weight(w_k) = tf(w_k)(log_2 D - log_2 df(w_k) + 1)$$

[0034] 여기서, Wk는 특정단어, tf(Wk)는 주어진 코퍼스(corpus)에서 Wk의 전체 빈도, D는 코퍼스의 사이즈(코퍼스의 전체 문서수), df(Wk)는 Wk를 포함하는 문서의 수를 각각 의미할 수 있다.

[0035] 또한, 가중치 할당 시, 중복되는 단어를 제거한 후, 각 단어에 가중치 할당할 수 있다.

[0036] 다시 도 4를 참고하면, 단계(420)에서는 상기 미분류 문서 테이블과 하나 이상의 카테고리 프로파일 테이블의 유사도를 각각 판단할 수 있다. 즉, 상기 미분류 문서 테이블과 하나 이상의 카테고리 프로파일 테이블을 순차적으로 비교하여 유사한 정도를 수치화하여 나타낼 수 있다. 유사도 판단은 도 6 및 도 7을 참고하여 이하에서 더욱 상세하게 설명한다.

[0037] 도 6은 본 발명의 일실시예에 있어서, 미분류 문서를 분류하는 과정을 설명하기 위한 도면이다.

[0038] 도 6을 참고하면, 미분류 문서(610)는 인코딩 과정을 통해 하나 이상의 단어 및 상기 하나 이상의 단어의 가중치를 포함하는 테이블(620)로 생성될 수 있다. 이후, 상기 테이블은 제1 카테고리 프로파일 테이블(630), 제2 카테고리 프로파일 테이블(640) 등과 순차적으로 유사도를 판단할 수 있다. 이때, 제1 카테고리 프로파일 테이블(630), 제2 카테고리 프로파일 테이블(640) 등과의 유사도는 유사도 스코어를 통해 수치화하여 나타낼 수 있다. 예를 들어, 제1 카테고리 프로파일 테이블(630)과 미분류 문서(610)의 유사도 스코어가 1.5이고, 제2 카테고리 프로파일 테이블(640)과 미분류 문서(610)의 유사도 스코어가 1.2인 것으로 수치화될 수 있다. 유사도 스코어를 산출하는 방법의 실시예는 도 7을 참고하여 이하에서 더욱 상세하게 설명한다.

[0039] 도 7은 본 발명의 일실시예에 있어서, 미분류 문서를 분류하기 위해 유사도를 판단하는 과정을 도시한 도면이다.

[0040] 도 7을 참고하면, 유사도 판단을 위해 복수의 카테고리 프로파일 테이블 중 유사도를 판단하여 유사도 스코어를 산출하기 위한 카테고리 프로파일 테이블(720)을 선택할 수 있다. 이때, 카테고리를 분류하고자 하는 미분류 문서 테이블(710)과 카테고리 프로파일 테이블(720)은 단어와 상기 단어에 대한 가중치를 포함할 수 있다.

[0041] 이후, 미분류 문서 테이블(710)과 카테고리 프로파일 테이블(720)에 공통적으로 포함된 단어를 추출하고, 상기 공통적으로 포함된 단어의 가중치 중 카테고리 프로파일 테이블(720)에 포함된 가중치를 합산하여 유사도 스코어를 산출할 수 있다. 예를 들어, 미분류 문서 테이블(710)이 단어 및 상기 단어의 가중치로 (computer, 0.4), (information, 0.7), (data, 0.3), (internet, 0.6)를 가지고, 카테고리 프로파일 테이블(720)이 (computer, 0.2), (information, 0.1), (hardware, 0.3), (communication, 0.9)을 가지는 경우, 공통적으로 포함되는 단어는 'computer' 및 'information'이다. 따라서, 상기 공통적으로 포함되는 단어, 상기 단어에 대응하는 각각의 가중치를 별도의 테이블(730)로 생성할 수 있다. 따라서, 생성된 별도의 테이블(730)은 (computer, 0.4, 0.2), (infomration, 0.7, 0.1)을 포함할 수 있다. 한편, 카테고리 프로파일 테이블(720)로부터 추출된 상기 'computer' 및 'information'에 대응하는 가중치를 추출하면, 각각 '0.2' 및 '0.1'이므로 두 값을 합산한 0.3이 카테고리 프로파일 테이블(720)에 대한 유사도 스코어로 결정될 수 있다.

[0042] 유사도 판단을 위한 다른 일예로, 미분류 문서 테이블(710)과 카테고리 프로파일 테이블(720)에 공통적으로 포함된 단어를 추출하고, 상기 공통적으로 포함된 단어의 가중치 중 미분류 문서 테이블(710)에 포함된 가중치를 합산하여 유사도 스코어를 산출할 수 있다. 예를 들어, 미분류 문서 테이블(710)이 단어 및 상기 단어의 가중치로 (computer, 0.4), (information, 0.7), (data, 0.3), (internet, 0.6)를 가지고, 카테고리 프로파일 테이블(720)이 (computer, 0.2), (information, 0.1), (hardware, 0.3), (communication, 0.9)을 가지는 경우, 공통적으로 포함되는 단어는 'computer' 및 'information'이다. 따라서, 상기와 같이, (computer, 0.4, 0.2), (infomration, 0.7, 0.1)를 포함하는 별도의 테이블(730)을 생성할 수 있다. 여기서, 상기 'computer'에 대하여 미분류 문서 테이블(710)로부터 추출된 가중치는 '0.4'이고, 'information'에 미분류 문서 테이블(710)로부터 추출된 가중치는 '0.7'이므로 두 값을 합산한 1.1이 카테고리 프로파일 테이블(720)에 대한 유사도 스코어로 결정될 수 있다.

[0043] 또다른 일예로, 유사도 판단을 위해 미분류 문서 테이블(710)과 카테고리 프로파일 테이블(720)에 공통적으로 포함된 단어를 추출하고, 상기 공통적으로 포함된 단어 별로 미분류 문서 테이블(710) 및 카테고리 프로파일 테이블(720)의 가중치 값을 곱하여 곱셈 스코어를 생성할 수 있다. 이후, 상기 공통적으로 포함된 단어 별 곱셈

스코어를 합산하여 유사도 스코어를 산출할 수 있다. 예를 들어, 미분류 문서 테이블(710)이 단어 및 상기 단어의 가중치로 (computer, 0.4), (information, 0.7), (data, 0.3), (internet, 0.6)를 가지고, 카테고리 프로파일 테이블(720)이 (computer, 0.2), (information, 0.1), (hardware, 0.3), (communication, 0.9)을 가지는 경우, 공통적으로 포함되는 단어는 'computer' 및 'information'이다. 따라서, 상기와 같이, (computer, 0.4, 0.2), (infomration, 0.7, 0.1)를 포함하는 별도의 테이블(730)을 생성할 수 있다. 이때, 상기 'computer'에 대응하는 미분류 문서 테이블(710)로부터 추출된 가중치(ex. 0.4)와 카테고리 프로파일 테이블(720)로부터 추출된 가중치(ex. 0.2)를 곱하여 단어 'computer'에 대한 곱셈 스코어 0.08을 산출할 수 있다. 또한, 'information'에 대응하는 미분류 문서 테이블(710)로부터 추출된 가중치(ex. 0.7)와 카테고리 프로파일 테이블(720)로부터 추출된 가중치(ex. 0.1)를 곱하여 단어 'information'에 대한 곱셈 스코어 0.07을 산출할 수 있다. 이후, 각 단어별 곱셈 스코어를 합산한 0.15가 카테고리 프로파일 테이블(720)에 대한 유사도 스코어로 결정될 수 있다.

[0044] 다시 도 4를 참고하면, 단계(430)에서는 상기 각 유사도에 기초하여 상기 미분류 문서를 분류할 수 있다. 즉, 각 카테고리 프로파일 테이블에 대한 유사도 스코어 중 가장 큰 유사도 스코어를 결정하고, 상기 유사도 스코어가 가장 큰 카테고리 프로파일 테이블에 대응하는 카테고리를 상기 미분류 문서의 카테고리로 분류할 수 있다. 도 6에 도시된 바와 같이, 상기 수치화된 유사도 스코어 중 최대값을 갖는 카테고리 프로파일 테이블이 미분류 문서(610)와 가장 유사한 테이블인 것으로 결정할 수 있고, 상기 예의 경우, 유사도 스코어가 1.5인 제1 카테고리 프로파일 테이블(630)이 미분류 문서(610)와 가장 유사한 테이블로 결정될 수 있다. 따라서, 미분류 문서(610)는 제1 카테고리 프로파일 테이블(630)의 카테고리와 동일한 카테고리로 분류될 수 있다.

[0045] 도 8은 본 발명의 일실시예에 있어서, 전자문서 자동 분류 장치를 도시한 블록도이다.

[0046] 도 8을 참고하면, 전자문서 자동 분류 장치(800)는 프로파일 생성부(810) 및 문서 분류부(820)를 포함할 수 있다.

[0047] 프로파일 생성부(810)는 하나 이상의 라벨된 문서를 이용하여 카테고리 프로파일 테이블을 생성할 수 있다. 여기서, 프로파일 생성부(810)는 상기 하나 이상의 라벨된 문서에 포함된 텍스트를 하나의 텍스트로 통합하는 텍스트 통합부(811) 및 상기 하나의 텍스트를 하나 이상의 단어 및 상기 하나 이상의 단어에 대한 가중치를 포함하는 카테고리 프로파일 테이블로 인코딩하는 인코딩부(812)를 포함할 수 있다.

[0048] 문서 분류부(820)는 상기 카테고리 프로파일 테이블을 이용하여 미분류 문서를 분류할 수 있다. 여기서, 문서 분류부(820)는 인코딩부(821), 유사도 판단부(822), 및 분류부(823)를 포함할 수 있다.

[0049] 인코딩부(821)는 상기 미분류 문서를 하나 이상의 단어 및 상기 하나 이상의 단어에 대한 가중치를 포함하는 미분류 문서 테이블로 인코딩할 수 있다. 여기서, 가중치는 미분류 문서에 포함된 단어가 문서의 내용과 관련하여 얼마나 중요한지를 나타내는 수치일 수 있다. 또한, 인코딩부(821)는 토큰 생성부(미도시), 원형 생성부(미도시), 불사용어 제거부(미도시), 및 가중치 할당부(미도시)를 포함할 수 있다. 토큰 생성부(미도시)는 텍스트를 서브스트링으로 분리하여 토큰을 생성하고, 원형 생성부(미도시)는 상기 토큰을 원형으로 변환하여 원형 단어를 생성할 수 있다. 불사용어 제거부(미도시)는 상기 문서 내의 접속사, 관사, 전치사, 대명사 중 적어도 하나를 포함하는 불사용어를 제거하고, 가중치 할당부(미도시)는 상기 원형 단어에 대응하는 가중치를 할당할 수 있다.

[0050] 유사도 판단부(822)는 상기 미분류 문서 테이블과 하나 이상의 카테고리 프로파일 테이블의 유사도를 각각 판단할 수 있다. 또한, 유사도 판단부(822)는 일예로, 카테고리 프로파일 선택부(미도시), 공통단어 추출부(미도시), 및 스코어 산출부(미도시)를 포함할 수 있다. 프로파일 선택부(미도시)는 하나의 카테고리 프로파일 테이블을 선택하고, 공통단어 추출부(미도시)는 상기 미분류 문서 테이블과 상기 선택된 카테고리 프로파일 테이블의 공통 단어를 추출할 수 있다. 스코어 산출부(미도시)는 상기 미분류 문서 테이블로부터 상기 공통 단어에 대응하는 가중치를 합산하여 유사도 스코어를 산출할 수 있다. 또한, 스코어 산출부(미도시)는 상기 카테고리 프로파일 테이블로부터 상기 공통 단어에 대응하는 가중치를 합산하여 유사도 스코어를 산출할 수 있다.

[0051] 또한, 유사도 판단부(822)는 다른 일예로, 카테고리 프로파일 선택부(미도시), 공통단어 추출부(미도시), 곱셈 스코어 산출부(미도시), 및 유사도 스코어 산출부(미도시)를 포함할 수 있다.

[0052] 이때, 프로파일 선택부(미도시)는 하나의 카테고리 프로파일 테이블을 선택하고, 공통단어 추출부(미도시)는 상기 미분류 문서 테이블과 상기 선택된 카테고리 프로파일 테이블의 공통 단어를 추출할 수 있다. 곱셈 스코어

산출부(미도시)는 상기 공통 단어 별로 상기 미분류 문서 테이블 및 상기 선택된 카테고리 프로파일 테이블의 가중치 값을 곱하여 곱셈 스코어를 생성하고, 유사도 스코어 산출부(미도시)는 상기 공통 단어 별 곱셈 스코어를 합산하여 유사도 스코어를 산출할 수 있다.

[0053] 분류부(823)는 상기 각 유사도에 기초하여 상기 미분류 문서를 분류할 수 있다. 즉, 유사도 스코어가 가장 큰 카테고리 프로파일 테이블에 대응하는 카테고리를 상기 미분류 문서의 카테고리로 분류할 수 있다.

[0054] 한편, 도 8에서 설명되지 않은 부분은 도 1 내지 도 7의 설명을 참고할 수 있다.

[0055] 상기와 같이, 문서를 텍스트로 인코딩하고 미리 분류된 카테고리 프로파일을 사용하여 문서를 자동 분류함으로써, 문서 자동 분류를 위한 연산량을 줄이고, 분류의 정확성을 향상시킬 수 있다.

[0056] 이상과 같이 본 발명의 일실시예는 비록 한정된 실시예와 도면에 의해 설명되었으나, 본 발명의 일실시예는 상기 설명된 실시예에 한정되는 것은 아니며, 이는 본 발명이 속하는 분야에서 통상의 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형이 가능하다. 따라서, 본 발명의 일실시예는 아래에 기재된 특허청구범위에 의해서만 파악되어야 하고, 이의 균등 또는 등가적 변형 모두는 본 발명 사상의 범주에 속한다고 할 것이다.

도면의 간단한 설명

[0057] 도 1은 본 발명의 일실시예에 있어서, 전자문서 자동 분류 방법을 도시한 흐름도이다.

[0058] 도 2는 도 1에 도시된 카테고리 프로파일 테이블을 생성하는 과정을 나타낸 동작 흐름도이다.

[0059] 도 3은 본 발명의 일실시예에 있어서, 라벨된 문서를 이용하여 카테고리 프로파일 테이블을 생성하는 과정을 설명하기 위한 도면이다.

[0060] 도 4는 도 1에 도시된 미분류 문서를 분류하는 과정을 나타낸 동작 흐름도이다.

[0061] 도 5는 도 4에 도시된 인코딩 과정을 나타낸 동작 흐름도이다.

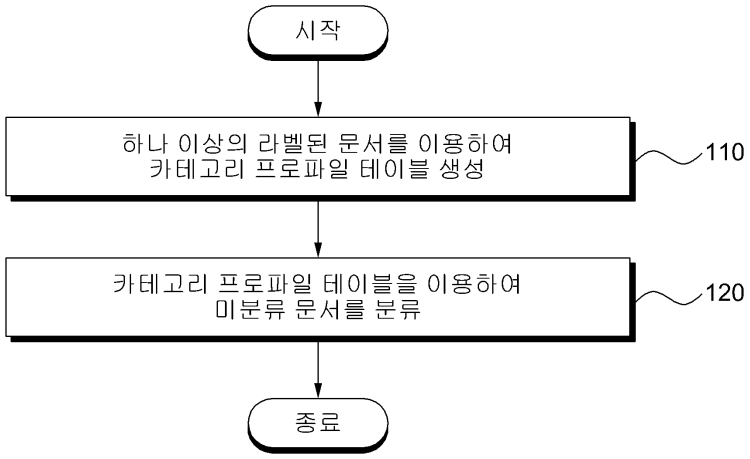
[0062] 도 6은 본 발명의 일실시예에 있어서, 미분류 문서를 분류하는 과정을 설명하기 위한 도면이다.

[0063] 도 7은 본 발명의 일실시예에 있어서, 미분류 문서를 분류하기 위해 유사도를 판단하는 과정을 도시한 도면이다.

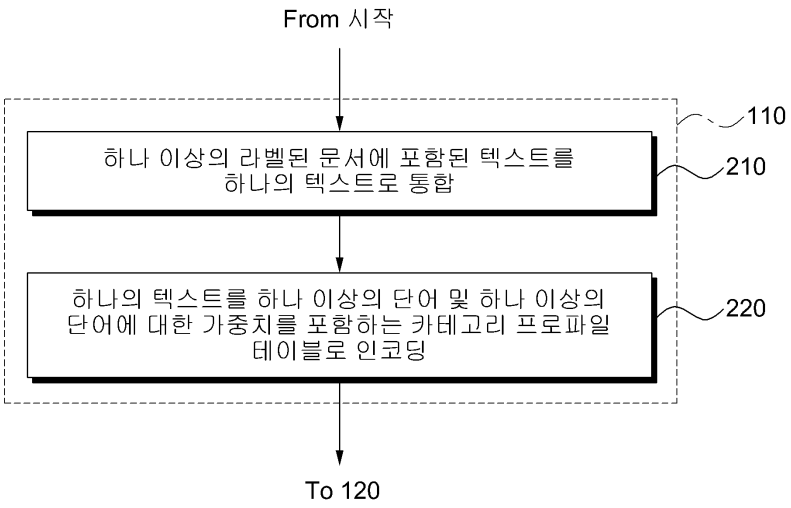
[0064] 도 8은 본 발명의 일실시예에 있어서, 전자문서 자동 분류 장치를 도시한 블록도이다.

도면

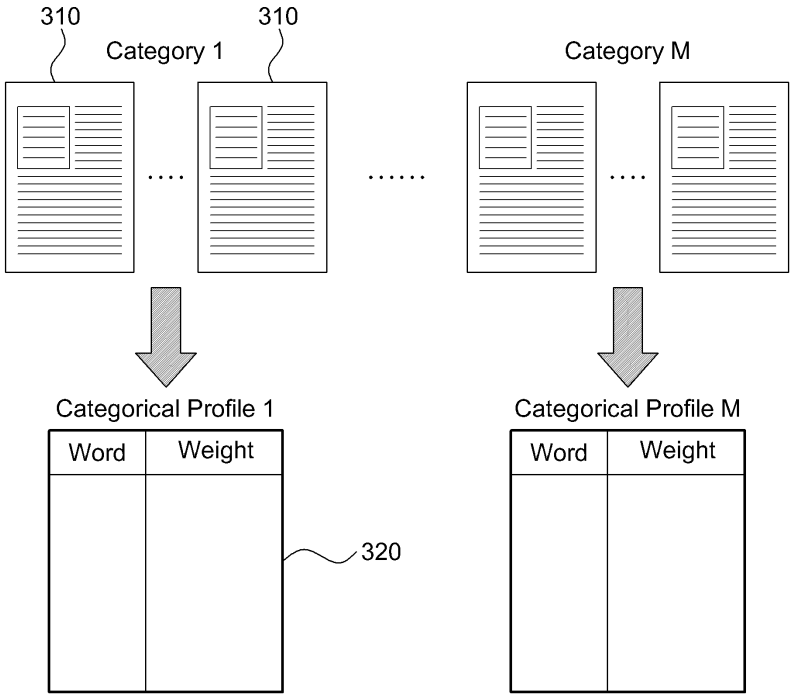
도면1



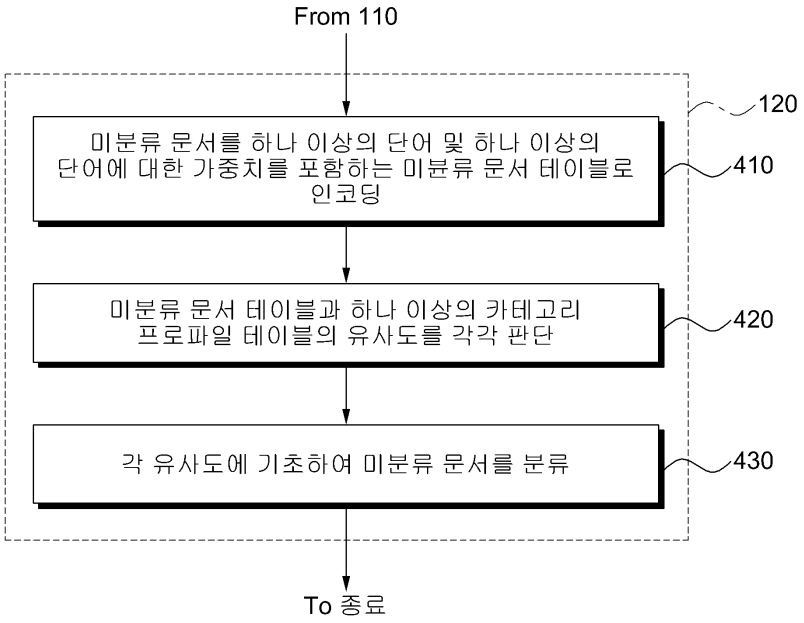
도면2


From 시작
110
하나 이상의 라벨된 문서에 포함된 텍스트를 하나의 텍스트로 통합
210
하나의 텍스트를 하나 이상의 단어 및 하나 이상의 단어에 대한 가중치를 포함하는 카테고리 프로파일 데이블로 인코딩
220
To 120

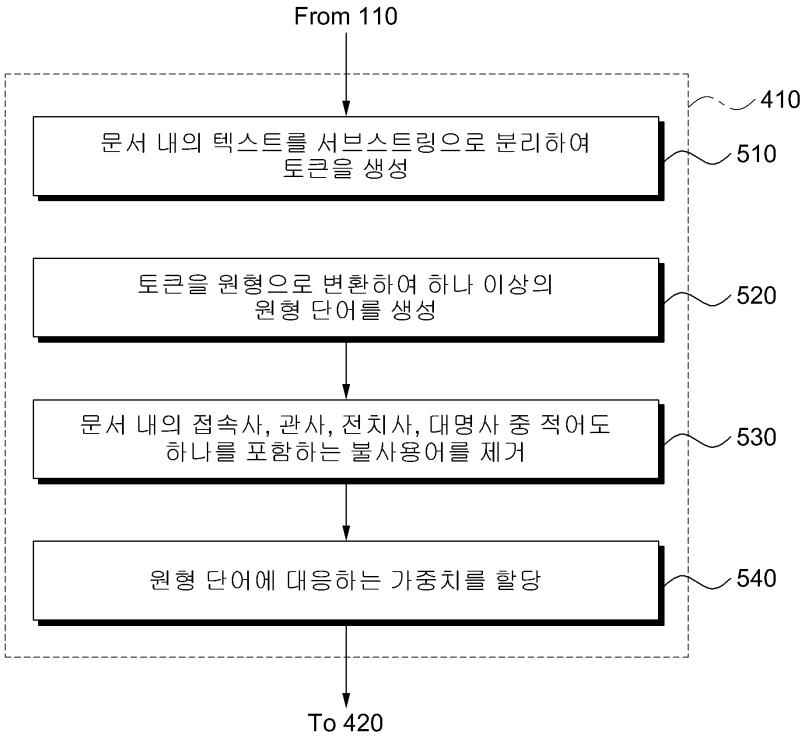
도면3


310
310
Category 1
Category M
Categorical Profile 1
Word
Weight
320
Categorical Profile M
Word
Weight

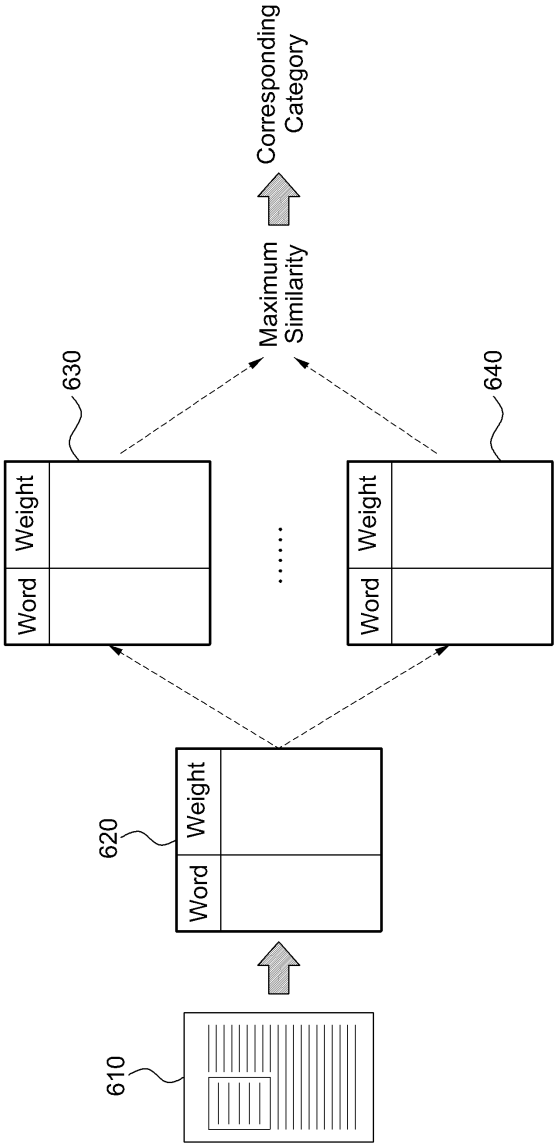
도면4


From 110
120
미분류 문서를 하나 이상의 단어 및 하나 이상의 단어에 대한 가중치를 포함하는 미분류 문서 데이블로 인코딩
410
미분류 문서 데이블과 하나 이상의 카테고리 프로파일 데이블의 유사도를 각각 판단
420
각 유사도에 기초하여 미분류 문서를 분류
430
To 종료

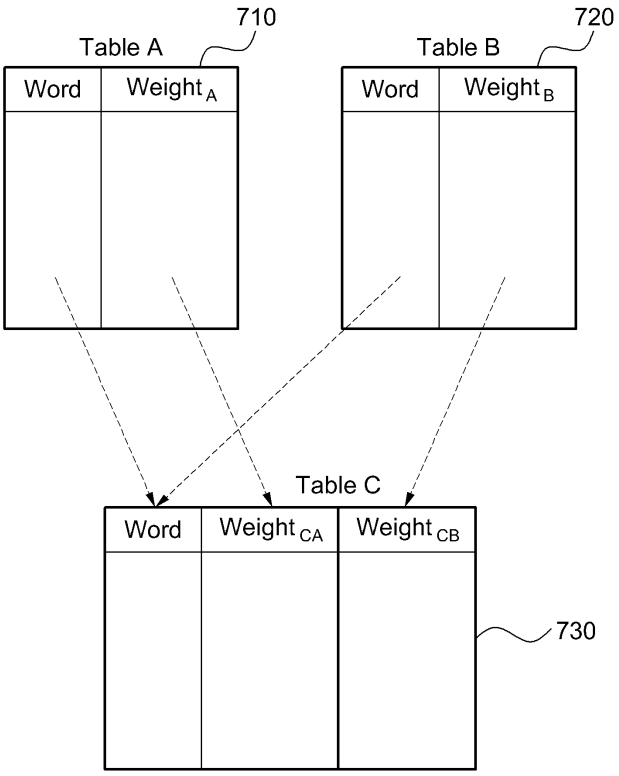
도면5


From 110
410
문서 내의 텍스트를 서브스트링으로 분리하여 토큰을 생성
510
토큰을 원형으로 변환하여 하나 이상의 원형 단어를 생성
520
문서 내의 접속사, 관사, 전치사, 대명사 중 적어도 하나를 포함하는 불사용어를 제거
530
원형 단어에 대응하는 가중치를 할당
540
To 420

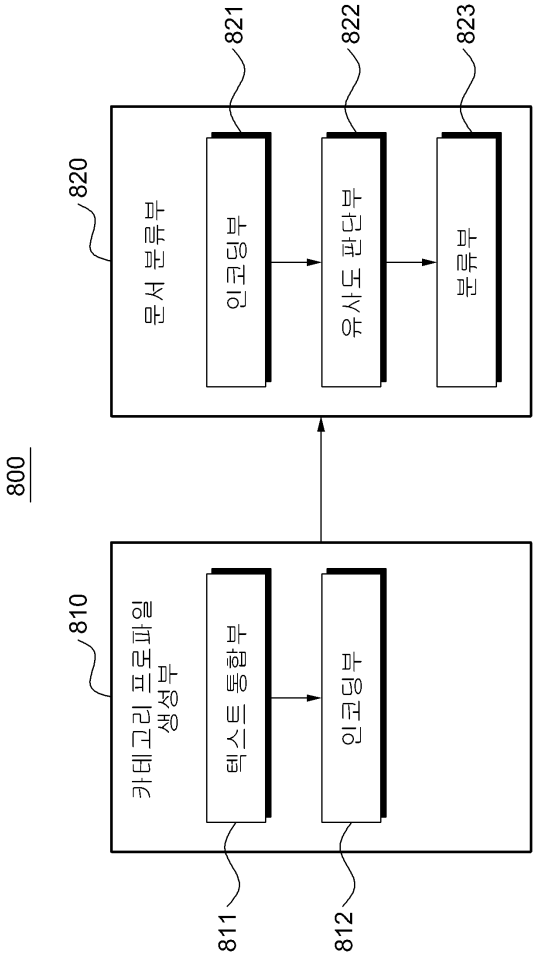
도면6


Corresponding Category
Maximum Similarity
630
640
Weight
Word
Weight
Word
.....
620
Weight
Word
610

도면7


710
Table A
Word
Weight A
720
Table B
Word
Weight B
Table C
Word
Weight CA
Weight CB
730

도면8


800
810
카테고리 프로파일 생성부
텍스트 통합부
인코딩부
811
812
820
문서 분류부
인코딩부
유사도 판단부
분류부
821
822
823