국민대학교
KOOKMIN UNIVERSITY

DB 매핑 시스템

- 이름 : 김남규
- 소속 : 경영정보학부
- 연구분야 : 시멘틱 웹, DB설계

텍스트 분류 정확도가 우수한 다중 매핑 시스템

기술개요

- 본 기술은 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템 기술이다.
- 기존 매체 고유의 카테고리 체계는 그대로 유지하면서 이종 매체 간 카테고리 매칭을 수행하는 시스템이다.
- 또한, 개별 문서를 2차원 레이블로 저장하여 다양한 문서의 분류 정확도를 향상시킬 수 있다.

Role	Doc. No	Category	DW_T1	DW_T2	DW_T3	DW_T4	DW_T5
Seed	B_Doc1	IT	1.2	0.9	0.7	1.3	1.1
	B_Doc2	Politics	0.5	0.7	1.3	0.6	2.1
	B_Doc3	Ent.	0.6	1.8	0.9	0.7	1.1
Augmented	O_Doc4	~~Cate_1~~ → Unlabeled	1.1	0.4	2.3	0.9	1.4
	O_Doc5	~~Cate_1~~ → Unlabeled	0.4	0.8	1.4	1.7	0.7
	O_Doc6	~~Cate_2~~ → Unlabeled	1.0	2.1	1.2	0.7	1.1
Target	O_Doc7	~~Cate_3~~ → Unlabeled	0.9	1.1	1.3	0.6	1.0
	O_Doc8	~~Cate_3~~ → Unlabeled	1.1	0.2	2.1	0.4	0.3
	O_Doc9	~~Cate_2~~ → Unlabeled	0.4	1.3	0.3	0.5	2.4

(d) 구조화된 문서

기술성

- 개별 문서를 다양한 매체의 관점으로 2차원 레이블로 저장하여 분류 정확도 향상
- 이종 매체에 속한 문서를 한 매체에 속한 것과 같이 구성 가능
- 고유 카테고리는 유지하며 이종 매체 간 카테고리 매칭 수행 가능

대표청구항

- 이종 매체들로부터 문서를 수집한 후, 문서와 토픽 간의 대응도를 산출하여 수집된 문서를 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화하는 토픽 모델링부와 준지도 학습 기반의 문서 분류를 활용하여 상기 토픽 모델링부에서 구조화된 문서 중 기본류 문서에 대한 학습을 통해 생성된 분류 알고리즘을 이용하여 미분류 문서를 분류하여 기존의 기분류 문서와 통합된 1차 학습데이터를 생성하는 1차 학습 및 분류부와 상기 1차 학습 및 분류부를 통해 보강된 1차 학습 데이터를 활용하여 최종적인 대상 미분류 문서에 카테고리를 부여하여 2차 분류된 2차 학습 데이터를 생성하는 2차 학습...

지식재산권

- 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템 및 방법 (비공개)

(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2018년12월13일
(11) 등록번호 10-1928732
(24) 등록일자 2018년12월07일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01) *G06F 15/18* (2018.01)
(52) CPC특허분류
G06F 17/3002 (2013.01)
G06F 15/18 (2018.05)
(21) 출원번호 10-2017-0031217
(22) 출원일자 2017년03월13일
심사청구일자 2017년03월13일
(65) 공개번호 10-2018-0104446
(43) 공개일자 2018년09월21일
(56) 선행기술조사문헌
US20090030862 A1*
JP2009122851 A
US20090125463 A1
홍진성 외, 단일 카테고리 문서의 다중 카테고리 자동 확장 방법론, 한국지능정보시스템학회 학술대회 논문집, pp.332-339 (2014.05)*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
국민대학교산학협력단
서울특별시 성북구 정릉로 77 (정릉동, 국민대학교)
(72) 발명자
김남규
서울특별시 중랑구 신내로7나길 24, 209동 1701호(상봉동, 건영2차아파트)
김다솜
서울특별시 성북구 솔샘로11길 39, 103호 (정릉동)
(74) 대리인
정부연

전체 청구항 수 : 총 14 항 심사관 : 경연정

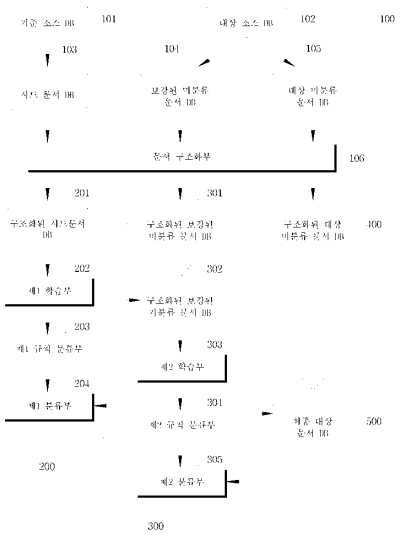
(54) 발명의 명칭 **텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템 및 방법**

(57) 요 약

본 발명은 개별 문서를 다양한 매체의 관점에서 재분류하고 이러한 결과를 문서에 2차원 레이블로 저장함으로써, 이종 매체에 속한 다양한 문서들을 마치 한 매체에 속한 것과 같이 동일한 카테고리 기준으로 탐색할 수 있는 논리적 장치를 제안하여 기존 매체 고유의 카테고리 체계는 그대로 유지하면서 이종 매체 간 카테고리 매핑을 수행

(뒷면에 계속)

대 표 도 - 도2



하는 시스템 및 방법을 제공하기 위한 것으로서, 이종 매체들로부터 문서를 수집한 후, 문서와 토픽 간의 대응도를 산출하여 수집된 문서들을 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화하는 토픽 모델링부와, 준지도 학습 기반의 문서 분류를 활용하여 상기 토픽 모델링부에서 구조화된 문서 중 기분류 문서에 대한 학습을 통해 생성된 분류 알고리즘을 이용하여 미분류 문서를 분류하여 기존의 기분류 문서와 통합된 1차 학습 데이터를 생성하는 1차 학습 및 분류부와, 상기 1차 학습 및 분류부를 통해 보강된 1차 학습 데이터를 활용하여 최종적인 대상 미분류 문서에 카테고리를 부여하여 2차 분류된 2차 학습 데이터를 생성하는 2차 학습 및 분류부를 포함하여 구성되는데 있다.

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

자체적 혹은 고유적으로 운영하는 카테고리가 존재하는 분류의 기준이 되는 기준매체와 기준매체로부터 카테고리를 부여받는 분류의 대상이 되는 대상매체로 구성되는 이종 매체들로부터 문서를 수집한 후, 문서와 토픽 간의 대응도를 산출하여 수집된 문서들을 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화하는 토픽 모델링부와,

준지도 학습 기반의 문서 분류를 활용하여 상기 토픽 모델링부에서 구조화된 문서 중 기분류 문서에 대한 학습을 통해 생성된 분류 알고리즘을 이용하여 미분류 문서를 분류하여 기존의 기분류 문서와 통합된 1차 학습 데이터를 생성하는 1차 학습 및 분류부와,

상기 1차 학습 및 분류부를 통해 보강된 1차 학습 데이터를 활용하여 최종적인 대상 미분류 문서에 카테고리를 부여하여 2차 분류된 2차 학습 데이터를 생성하는 2차 학습 및 분류부를 포함하여 구성되는 것을 특징으로 하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템.

청구항 2

제 1 항에 있어서,

상기 이종 매체는 일반 웹 사이트, 포털 사이트, 언론, 소셜 미디어를 포함하는 문서를 제공하는 웹 플랫폼인 것을 특징으로 하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템.

청구항 3

제 2 항에 있어서,

상기 이종 매체 중 기준매체는 분류에 사용될 카테고리 제공을 위하여 하나의 매체만 선정 가능하고,

상기 이종 매체 중 대상매체는 고유의 카테고리 체계는 존재하거나 존재하지 않아도 되며, 복수 개의 매체가 선정 가능한 것을 특징으로 하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템.

청구항 4

제 1 항에 있어서, 상기 토픽 모델링부는

기준매체로부터 입력되는 기준소스(Base Source)를 저장하는 기준소스 DB와,

대상매체로부터 입력되는 대상소스(Target Source)를 저장하는 대상소스 DB와,

상기 기준소스 DB에 저장된 기준소스로부터 추출한 모든 기분류 문서(Labeleed Documents)의 집합인 시드문서(Seed Documents)를 저장하는 시드문서 DB와,

대상소스 DB에 저장된 대상소스로부터 추출한 미분류 문서(Unlabeled Document) 중 준지도 학습방법을 채택하기 위한 일부를 추출하여 보강된 미분류 문서로 저장하는 보강된 미분류 문서 DB와,

대상소스 DB에 저장된 대상소스로부터 추출한 미분류 문서(Unlabeled Document) 중 상기 보강된 미분류 문서로 추출된 문서를 제외한 나머지 미분류 문서인 대상 미분류 문서를 저장하는 대상 미분류 문서 DB와,

상기 시드문서 DB, 보강된 미분류 문서 DB 및 대상 미분류 문서 DB에 각각 수집되어 저장된 문서들(시드문서, 보강된 미분류 문서, 대상 미분류 문서)을 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화하는 문서 구조화부를 포함하여 구성되는 것을 특징하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템.

청구항 5

제 1 항에 있어서, 상기 1 차 학습 및 분류부는

토픽 모델링부에서 구조화된 문서 중 시드문서 및 보강된 미분류 문서를 입력으로 각각 저장하는 구조화된 시드문서 DB와,

상기 구조화된 시드문서 DB에 저장된 소량의 기분류 시드문서에 대한 학습을 통해 분류 알고리즘을 생성하는 제 1 학습부와,

상기 제 1 학습부에서 생성된 분류 알고리즘을 통해 분류 규칙을 생성하는 제 1 규칙 분류부와,

상기 제 1 규칙 분류부에서 생성된 규칙을 적용하여 보강된 미분류 문서를 분류하는 제 1 분류부를 포함하여 구성되는 것을 특징으로 하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템.

청구항 6

제 5 항에 있어서,

상기 제 1 학습부는 카테고리 분류 과정에 준지도 학습을 활용하여 학습하는 것을 특징으로 하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템.

청구항 7

제 1 항에 있어서, 상기 2차 학습 및 분류부는

토픽 모델링부에서 구조화된 문서 중 보강된 미분류 문서와 상기 1차 학습 및 분류부에서 분류된 문서를 저장하는 구조화된 보강된 미분류 문서 DB와,

상기 보강된 미분류 문서와 분류된 문서를 통합하여 저장하는 구조화된 보강된 기분류 문서 DB와,

상기 구조화된 보강된 기분류 문서 DB에 저장된 통합 문서에 대한 학습을 통해 분류 알고리즘을 생성하는 제 2 학습부와,

상기 제 2 학습부에서 생성된 분류 알고리즘을 통해 분류 규칙을 생성하는 제 2 규칙 분류부와,

토픽 모델링부에서 구조화된 문서 중 구조화된 대상 미분류 문서 DB에 저장된 대상 미분류 문서의 구조화된 문서를 상기 제 2 규칙 분류부에서 생성된 분류 규칙을 적용하여 분류하여 최종 대상문서 DB에 저장하는 제 2 분류부를 포함하여 구성되는 것을 특징으로 하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템.

청구항 8

(A) 토픽 모델링부를 통해 자체적 혹은 고유적으로 운영하는 카테고리가 존재하는 분류의 기준이 되는 기준매체와 기준매체로부터 카테고리를 부여받는 분류의 대상이 되는 대상매체로 구성되는 이종 매체들로부터 문서를 수집한 후, 문서와 토픽 간의 대응도를 산출하여 수집된 문서들을 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화하는 단계와,

(B) 1차 학습 및 분류부를 통해 준지도 학습 기반의 문서 분류를 활용하여 상기 토픽 모델링부에서 구조화된 문서 중 기분류 문서에 대한 학습을 통해 생성된 분류 알고리즘을 이용하여 미분류 문서를 분류하여 기존의 기분류 문서와 통합된 1차 학습 데이터를 생성하는 단계와,

(C) 2차 학습 및 분류부를 통해 상기 생성된 1차 학습 데이터를 활용하여 최종적인 대상 미분류 문서에 카테고리를 부여하여 2차 분류된 2차 학습 데이터를 생성하는 단계를 포함하여 이루어지는 것을 특징으로 하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 방법.

청구항 9

제 8 항에 있어서, 상기 (A) 단계는

(A1) 이종매체 중 기준매체로부터 기준소스를, 대상매체로부터 대상소스를 각각 입력받아 추출한 문서의 집합인 시드문서 및 미분류 문서를 저장하는 단계와,

(A2) 상기 저장된 미분류 문서의 일부를 추출하여 보강된 미분류 문서로 저장하는 단계와,

(A3) 문서 구조화부를 통해 저장된 시드문서, 보강된 미분류 문서, 대상 미분류 문서를 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화하는 단계를 포함하여 이루어지는 것을 특징으로 하는 텍스트 분석을 통한 이

종 매체 카테고리 다중 매핑 방법.

청구항 10

제 9 항에 있어서,

상기 기준소스는 이미 고유의 카테고리 체계를 갖고 있는 기분류 문서이며,

상기 대상소스는 상기 기준소스의 관점에서 새로운 카테고리를 부여받게 되는 미분류 문서인 것을 특징으로 하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 방법.

청구항 11

제 9 항에 있어서, 상기 (A3) 단계는

문서 구조화부를 통해 저장된 시드문서, 보강된 미분류 문서, 대상 미분류 문서를 모두 통합하는 단계와,

상기 통합된 문서를 용어의 빈도에 기반을 두어 용어에 대한 차원의 수(토픽의 수)를 축소하는 단계와,

상기 축소된 각 용어가 토픽에 대응되는 정도인 용어 가중치를 산출하는 단계와,

상기 산출된 용어 가중치가 미리 정해진 용어 임계값 이상인 경우 해당 토픽을 나타내는 용어로 설정하는 단계와,

각 문서를 각 토픽에 대응되는 정도인 문서 가중치의 벡터로 나타냄으로서, 문서를 구조적 형태로 표현하는 단계를 포함하여 이루어지는 것을 특징으로 하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 방법.

청구항 12

제 11 항에 있어서,

상기 (B) 단계는 문서 가중치 벡터를 입력 변수로 카테고리를 목적 변수로 설정하여 구조적 형태로 표현된 문서의 분류를 위한 학습 및 분류를 수행하는 것을 특징으로 하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 방법.

청구항 13

제 8 항에 있어서, 상기 (B) 단계는

제 1 학습부를 통해 구조화된 문서 중 일부가 추출된 보강된 미분류 문서에 대한 준지도 학습을 활용하여 분류 알고리즘을 생성하는 단계와,

상기 생성된 분류 알고리즘을 통해 제 1 규칙 분류부에서 분류 규칙을 생성하는 단계와,

상기 생성된 분류 규칙을 적용하여 제 1 분류부에서 보강된 미분류 문서를 분류하여 기존의 기분류 문서와 통합된 1차 학습 데이터를 생성하는 단계를 포함하여 이루어지는 것을 특징으로 하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 방법.

청구항 14

제 8 항에 있어서, 상기 (C) 단계는

제 2 학습부를 통해 상기 생성된 1차 학습 데이터에 대한 준지도 학습을 활용하여 분류 알고리즘을 생성하는 단계와,

상기 생성된 분류 알고리즘을 통해 제 2 규칙 분류부에서 분류 규칙을 생성하는 단계와,

상기 구조화된 문서 중 대상소스를 입력받아 추출한 미분류 문서에서 (B) 단계에서 보강된 미분류 문서로 추출된 문서를 제외한 나머지 미분류 문서를 상기 생성된 분류 규칙을 적용하여 분류하여 최종 대상문서인 2차 학습 데이터를 생성하는 단계를 포함하여 이루어지는 것을 특징으로 하는 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 이종 매체 간 카테고리 매핑을 수행하는 시스템 및 방법에 관한 것으로, 특히 기존 매체 고유의 카테고리 체계는 그대로 유지하면서 이종 매체 간 카테고리 매핑을 수행하는 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 최근 스마트 기기의 발달과 인터넷의 보급화로 인해 사용자들은 소셜 네트워크 서비스(Social Network Service), 인터넷 뉴스, 웹 커뮤니티 등 다양한 매체를 시간과 장소에 제약 받지 않고 사용할 수 있게 되었다. 이에 부응하여 다양한 기능과 목적을 지닌 매체들 또한 꾸준히 개발되고 있으며, 사용자들은 각자의 목적 및 취향에 따라 일반적으로 여러 매체들을 동시에 이용하고 있다.

[0003] 다양한 매체 가운데 특히 인스타그램(Instargram), 트위터(Twitter), 페이스북(Facebook) 등의 사용이 두드러지며, 2013년 기준 국내 소셜 네트워크 서비스 사용자는 평균 2.09개의 매체를 동시에 이용하고 있는 것으로 나타났다(한국직업능력개발원, 2013). 이처럼 사용자들은 본인의 의견 또는 정보를 다양한 매체를 통해 공유함은 물론, 반대로 특정 주제에 대한 정보를 수집할 때에도 여러 매체를 동시에 활용하고 있다. 그리고 다양한 매체를 통해 공유되고 수집되는 디지털 정보의 양은 2020년에 35제타바이트(ZB)를 훨씬 넘을 것으로 전망되고 있다(한국인터넷진흥원, 2014).

[0004] 다양한 매체를 통해 유통되는 문서들은 서로 유사한 주제, 심지어는 동일한 내용을 다루더라도 각 매체 별 정책 및 기준에 따라 각기 다른 카테고리(Category)로 관리될 수 있다. 예를 들어 도 1 은 "해외여행용 어플리케이션"에 대한 내용을 다루는 문서가 각 매체 고유의 카테고리 기준에 따라 "IT" "Travel", "Life"등으로 상이하게 분류될 수 있는 상황을 보여준다. 이렇듯 각 매체마다 카테고리를 정의하는 관점과 세분화 수준이 다르기 때문에, 유사 카테고리가 매체마다 서로 다른 명칭과 구조로 관리될 수 있다.

[0005] 이러한 매체에 따른 분류 체계의 상이함은 전체 매체를 아우르는 분석을 통해 새로운 지식을 창출하기 위한 시도에 걸림돌로 작용할 수 있다.

[0006] 일반적으로 정보의 조회는 크게 키워드를 통한 검색과 카테고리를 통한 탐색으로 구분된다. 전자의 경우 획득하고자 하는 정보의 주제가 비교적 구체적인 경우 사용되며, 문서가 속한 카테고리의 명칭 및 구조와 무관하게 내용에 기반하여 결과가 도출된다는 특징이 있다. 하지만 찾고자하는 문서의 주제가 키워드 수준으로 명확하지 않고 분야 수준에 머무는 초기 탐색의 경우, 후자와 같이 특정 카테고리를 선택하여 해당 카테고리 내의 문서를 조회하는 것이 일반적이다. 또한 이러한 탐색의 범위는 하나의 매체에만 국한되지 않으며, 점차 다양한 매체의 문서에 대한 탐색, 수집 및 분석에 대한 수요가 증가하고 있는 추세이다. 하지만 전술한 바와 같이 각 매체마다 서로 상이한 카테고리 구조 및 명칭을 갖기 때문에, 이종 매체를 아우르는 범위에서 특정 카테고리에 대한 탐색이 이루어지기란 매우 어렵다.

[0007] 이러한 한계점을 극복하기 위한 가장 직접적인 방법으로 모든 매체의 카테고리 체계를 표준화하는 방안을 생각할 수 있다. 하지만 각 매체들은 고유의 목적과 관점을 갖고 있기 때문에, 모든 매체의 카테고리 체계를 통일하는 것은 바람직하지도 않으며 가능하지도 않다.

[0008] 즉, 다양한 매체를 통해 유통되는 문서들은 서로 유사한 주제, 심지어는 동일한 내용을 다루더라도 각 매체별 정책 및 기준에 따라 각기 다른 카테고리로 관리되고 있으며, 이는 이종 매체를 아우르는 범위에서 특정 카테고리에 대한 탐색을 수행하고자 하는 시도에 걸림돌로 작용하고 있다.

선행기술문헌

특허문헌

[0009] (특허문헌 0001) 등록특허공보 제10-1088483호 (등록일자 2011.11.24)

(특허문헌 0002) 등록특허공보 제10-1478348호 (등록일자 2014.12.24)

발명의 내용

해결하려는 과제

[0010] 따라서 본 발명은 상기와 같은 문제점을 해결하기 위해 안출한 것으로서, 개별 문서를 다양한 매체의 관점에서 재분류하고 이러한 결과를 문서에 2차원 레이블로 저장함으로써, 이종 매체에 속한 다양한 문서들을 마치 한 매체에 속한 것과 같이 동일한 카테고리 기준으로 탐색할 수 있는 논리적 장치를 제안하여 기존 매체 고유의 카테고리 체계는 그대로 유지하면서 이종 매체 간 카테고리 매핑을 수행하는 시스템 및 방법을 제공하는데 그 목적이 있다.

[0011] 본 발명의 다른 목적들은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 또 다른 목적들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0012] 상기와 같은 목적을 달성하기 위한 본 발명에 따른 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템의 특징은 이종 매체들로부터 문서를 수집한 후, 문서와 토픽 간의 대응도를 산출하여 수집된 문서들을 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화하는 토픽 모델링부와, 준지도 학습 기반의 문서 분류를 활용하여 상기 토픽 모델링부에서 구조화된 문서 중 기분류 문서에 대한 학습을 통해 생성된 분류 알고리즘을 이용하여 미분류 문서를 분류하여 기존의 기분류 문서와 통합된 1차 학습 데이터를 생성하는 1차 학습 및 분류부와, 상기 1차 학습 및 분류부를 통해 보강된 1차 학습 데이터를 활용하여 최종적인 대상 미분류 문서에 카테고리를 부여하여 2차 분류된 2차 학습 데이터를 생성하는 2차 학습 및 분류부를 포함하여 구성되는데 있다.

[0013] 바람직하게 상기 이종 매체는 일반 웹 사이트, 포털 사이트, 언론, 소셜 미디어를 포함하는 문서를 제공하는 웹 플랫폼인 것을 특징으로 한다.

[0014] 바람직하게 상기 이종 매체는 분류의 기준이 되는 매체로, 자체적 혹은 고유적으로 운영하는 카테고리가 존재하며, 분류에 사용될 카테고리를 제공하기 때문에 하나의 매체만 선정 가능한 기준매체와, 분류의 대상이 되는 매체로, 상기 기준매체로부터 카테고리를 부여받아 고유의 카테고리 체계는 존재하거나 존재하지 않아도 되며, 복수 개의 매체가 선정 가능한 대상매체로 구성되는 것을 특징으로 한다.

[0015] 바람직하게 상기 토픽 모델링부는 기준매체로부터 입력되는 기준소스(Base Source)를 저장하는 기준소스 DB와, 대상매체로부터 입력되는 대상소스(Target Source)를 저장하는 대상소스 DB와, 상기 기준소스 DB에 저장된 기준소스로부터 추출한 모든 기분류 문서(Labeled Documents)의 집합인 시드문서(Seed Documents)를 저장하는 시드문서 DB와, 대상소스 DB에 저장된 대상소스로부터 추출한 미분류 문서(Unlabeled Document) 중 준지도 학습방법을 채택하기 위한 일부를 추출하여 보강된 미분류 문서로 저장하는 보강된 미분류 문서 DB와, 대상소스 DB에 저장된 대상소스로부터 추출한 미분류 문서(Unlabeled Document) 중 상기 보강된 미분류 문서로 추출된 문서를 제외한 나머지 미분류 문서인 대상 미분류 문서를 저장하는 대상 미분류 문서 DB와, 상기 시드문서 DB, 보강된 미분류 문서 DB 및 대상 미분류 문서 DB에 각각 수집되어 저장된 문서들(시드문서, 보강된 미분류 문서, 대상 미분류 문서)을 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화하는 문서 구조화부를 포함하여 구성되는 것을 특징하는 한다.

[0016] 바람직하게 상기 1 차 학습 및 분류부는 토픽 모델링부에서 구조화된 문서 중 시드문서 및 보강된 미분류 문서를 입력으로 각각 저장하는 구조화된 시드문서 DB와, 상기 구조화된 시드문서 DB에 저장된 소량의 기분류 시드문서에 대한 학습을 통해 분류 알고리즘을 생성하는 제 1 학습부와, 상기 제 1 학습부에서 생성된 분류 알고리즘을 통해 분류 규칙을 생성하는 제 1 규칙 분류부와, 상기 제 1 규칙 분류부에서 생성된 규칙을 적용하여 보강된 미분류 문서를 분류하는 제 1 분류부를 포함하여 구성되는 것을 특징으로 한다.

[0017] 바람직하게 상기 제 1 학습부는 카테고리 분류 과정에 준지도 학습을 활용하여 학습하는 것을 특징으로 한다.

[0018] 바람직하게 상기 2차 학습 및 분류부는 토픽 모델링부에서 구조화된 문서 중 보강된 미분류 문서와 상기 1차 학습 및 분류부에서 분류된 문서를 저장하는 구조화된 보강된 미분류 문서 DB와, 상기 보강된 미분류 문서와 분류된 문서를 통합하여 저장하는 구조화된 보강된 기분류 문서 DB와, 상기 구조화된 보강된 기분류 문서 DB에 저장된 통합 문서에 대한 학습을 통해 분류 알고리즘을 생성하는 제 2 학습부와, 상기 제 2 학습부에서 생성된 분류 알고리즘을 통해 분류 규칙을 생성하는 제 2 규칙 분류부와, 토픽 모델링부에서 구조화된 문서 중 구조화된 대상 미분류 문서 DB에 저장된 대상 미분류 문서의 구조화된 문서를 상기 제 2 규칙 분류부에서 생성된 분류 규칙을 적용하여 분류하여 최종 대상문서 DB에 저장하는 제 2 분류부를 포함하여 구성되는 것을 특징으로 한다.

[0019] 상기와 같은 목적을 달성하기 위한 본 발명에 따른 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 방법의 특징은 (A) 토픽 모델링부를 통해 이종 매체들로부터 문서를 수집한 후, 문서와 토픽 간의 대응도를 산출하여 수집된 문서들을 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화하는 단계와, (B) 1차 학습 및 분류부를 통해 준지도 학습 기반의 문서 분류를 활용하여 상기 토픽 모델링부에서 구조화된 문서 중 기분류 문서에 대한 학습을 통해 생성된 분류 알고리즘을 이용하여 미분류 문서를 분류하여 기존의 기분류 문서와 통합된 1차 학습 데이터를 생성하는 단계와, (C) 2차 학습 및 분류부를 통해 상기 생성된 1차 학습 데이터를 활용하여 최종적인 대상 미분류 문서에 카테고리를 부여하여 2차 분류된 2차 학습 데이터를 생성하는 단계를 포함하여 이루어지는데 있다.

[0020] 바람직하게 상기 (A) 단계는 (A1) 이종매체로부터 기준소스 및 대상소스를 각각 입력받아 추출한 문서의 집합인 시드문서 및 미분류 문서를 저장하는 단계와, (A2) 상기 저장된 미분류 문서의 일부를 추출하여 보강된 미분류 문서로 저장하는 단계와, (A3) 문서 구조화부를 통해 저장된 시드문서, 보강된 미분류 문서, 대상 미분류 문서를 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화하는 단계를 포함하여 이루어지는 것을 특징으로 한다.

[0021] 바람직하게 상기 기준소스는 이미 고유의 카테고리 체계를 갖고 있는 기분류 문서이며, 상기 대상소스는 상기 기준소스의 관점에서 새로운 카테고리를 부여받게 되는 미분류 문서인 것을 특징으로 한다.

[0022] 바람직하게 상기 (A3) 단계는 문서 구조화부를 통해 저장된 시드문서, 보강된 미분류 문서, 대상 미분류 문서를 모두 통합하는 단계와, 상기 통합된 문서를 용어의 빈도에 기반을 두어 용어에 대한 차원의 수(토픽의 수)를 축소하는 단계와, 상기 축소된 각 용어가 토픽에 대응되는 정도인 용어 가중치를 산출하는 단계와, 상기 산출된 용어 가중치가 미리 정해진 용어 임계값 이상인 경우 해당 토픽을 나타내는 용어로 설정하는 단계와, 각 문서를 각 토픽에 대응되는 정도인 문서 가중치의 벡터로 나타냄으로서, 문서를 구조적 형태로 표현하는 단계를 포함하여 이루어지는 것을 특징으로 한다.

[0023] 바람직하게 상기 (B) 단계는 문서 가중치 벡터를 입력 변수로 카테고리를 목적 변수로 설정하여 구조적 형태로 표현된 문서의 분류를 위한 학습 및 분류를 수행하는 것을 특징으로 한다.

[0024] 바람직하게 상기 (B) 단계는 제 1 학습부를 통해 구조화된 문서 중 일부가 추출된 보강된 미분류 문서에 대한 준지도 학습을 활용하여 분류 알고리즘을 생성하는 단계와, 상기 생성된 분류 알고리즘을 통해 제 1 규칙 분류부에서 분류 규칙을 생성하는 단계와, 상기 생성된 분류 규칙을 적용하여 제 1 분류부에서 보강된 미분류 문서를 분류하여 기존의 기분류 문서와 통합된 1차 학습 데이터를 생성하는 단계를 포함하여 이루어지는 것을 특징으로 한다.

[0025] 바람직하게 상기 (C) 단계는 제 2 학습부를 통해 상기 생성된 1차 학습 데이터에 대한 준지도 학습을 활용하여 분류 알고리즘을 생성하는 단계와, 상기 생성된 분류 알고리즘을 통해 제 2 규칙 분류부에서 분류 규칙을 생성하는 단계와, 상기 구조화된 문서 중 대상소스를 입력받아 추출한 미분류 문서에서 (B) 단계에서 보강된 미분류 문서로 추출된 문서를 제외한 나머지 미분류 문서를 상기 생성된 분류 규칙을 적용하여 분류하여 최종 대상문서인 2차 학습 데이터를 생성하는 단계를 포함하여 이루어지는 것을 특징으로 한다.

발명의 효과

[0026] 이상에서 설명한 바와 같은 본 발명에 따른 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템 및 방법은 다음과 같은 효과가 있다.

[0027] 첫째, 개별 문서를 다양한 매체의 관점에서 재분류하고 이러한 결과를 문서에 2차원 레이블로 저장함으로써, 이종 매체에 속한 다양한 문서들의 분류를 높은 정확도를 나타나는 효과가 있다.

[0028] 둘째, 이종 매체에 속한 다양한 문서들을 마치 한 매체에 속한 것과 같이 동일한 카테고리 기준으로 탐색할 수 있는 논리적 장치를 제안하여 기존 매체 고유의 카테고리 체계는 그대로 유지하면서 이종 매체 간 카테고리 매핑을 수행할 수 있는 효과가 있다.

도면의 간단한 설명

[0029] 도 1 은 종래의 "해외여행용 어플리케이션"에 대한 내용을 다루는 문서가 각 매체 고유의 카테고리 기준에 따라 "IT" "Travel", "Life"등으로 상이하게 분류될 수 있는 상황을 나타낸 구성도

도 2 는 본 발명의 실시예에 따른 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템의 구성을 나타낸 블록도

도 3 은 도 2에서 토픽 모델링부의 구성을 설명하기 위한 도면

도 4 는 도 2에서 1차 학습 및 분류부의 구성을 설명하기 위한 도면

도 5 는 도 2에서 2차 학습 및 분류부의 구성을 설명하기 위한 도면

도 6 은 도 5에서 도출된 최종 결과물을 나타낸 도면

도 7 은 본 발명의 일 실시예를 위해 수행한 실험의 개요를 나타낸 도면

도 8 은 본 발명의 다중 매핑 방법을 통해 사이트 "N"의 문서에 대한 분류 실험 세 가지의 누적 반응 검출률(Cumulative Response)을 나타낸 도면

도 9 는 본 발명의 다중 매핑 방법을 통해 사이트 "O"의 문서에 대한 분류 실험 세 가지의 누적 반응 검출률(Cumulative Response)을 나타낸 도면

도 10 은 본 발명의 다중 매핑 방법을 통한 실험에서 각각에 대해 각 카테고리 별 분류 정확도를 측정하는 사이트 "N"에 대한 실험 결과를 나타낸 도면

도 11 은 본 발명의 다중 매핑 방법을 통한 실험에서 각각에 대해 각 카테고리 별 분류 정확도를 측정하는 사이트 "O"에 대한 실험 결과를 나타낸 도면

발명을 실시하기 위한 구체적인 내용

[0030] 본 발명의 다른 목적, 특성 및 이점들은 첨부한 도면을 참조한 실시예들의 상세한 설명을 통해 명백해질 것이다.

[0031] 본 발명에 따른 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템 및 방법의 바람직한 실시예에 대하여 첨부한 도면을 참조하여 설명하면 다음과 같다. 그러나 본 발명은 이하에서 개시되는 실시예에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예는 본 발명의 개시가 완전하도록하며 통상의 지식을 가진자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이다. 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.

[0032] 도 2 는 본 발명의 실시예에 따른 텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 시스템의 구성을 나타낸 블록도이다.

[0033] 도 2에서 도시하고 있는 것과 같이, 이종 매체들로부터 문서를 수집한 후, 문서와 토픽 간의 대응도를 산출하여 수집된 문서들을 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화하는 토픽 모델링부(100)와, 준지도 학습 기반의 문서 분류를 활용하여 상기 토픽 모델링부(100)에서 구조화된 문서 중 기분류 문서에 대한 학습을 통해 생성된 분류 알고리즘을 이용하여 미분류 문서를 분류하여 기존의 기분류 문서와 통합된 1차 학습 데이터를 생성하는 1차 학습 및 분류부(200)와, 상기 1차 학습 및 분류부(200)를 통해 보강된 1차 학습 데이터를 활용하여 최종적인 대상 미분류 문서에 카테고리를 부여하여 2차 분류된 2차 학습 데이터를 생성하는 2차 학습 및 분류부(300)로 구성된다.

[0034] 이때, 문서를 수집하는 이종 매체는 문서를 제공하는 웹 플랫폼을 의미하며 구체적으로는 일반 웹 사이트, 포털 사이트, 언론, 소셜 미디어 등 웹 문서를 수집 가능한 모든 매체를 포함하며, 기준매체와 대상매체로 구성된다.

[0035] 상기 기준매체는 분류의 기준이 되는 매체를 뜻하며, 이는 매체에서 자체적 혹은 고유적으로 운영하는 카테고리가 존재해야 한다. 또한 기준매체는 분류에 사용될 카테고리를 제공하기 때문에 하나의 매체만 선정이 가능하다. 실제 사용 시 본 발명을 사용하는 사용자는 상기 조건을 만족시키는 매체에 한해 사용자가 자율적으로 기준매체를 선정할 수 있다. 또한 상기 대상매체는 분류의 대상이 되는 매체를 뜻하며, 기준매체로부터 카테고리를 부여받기에 대상매체 고유의 카테고리 체계는 존재하거나 존재하지 않아도 된다. 이때, 기준매체는 하나의 매체로 선정이 가능하나 대상매체는 복수 개의 매체를 선정하여 사용가능하다.

[0036] 상기 토픽 모델링부(100)는 기준매체로부터 입력되는 기준소스(Base Source)를 저장하는 기준소스 DB(101)와, 대상매체로부터 입력되는 대상소스(Target Source)를 저장하는 대상소스 DB(102)와, 상기 기준소스 DB(101)에 저장된 기준소스로부터 추출한 모든 기분류 문서(Labeleed Documents)의 집합인 시드문서(Seed Documents)를 저장하는 시드문서 DB(103)와, 대상소스 DB(102)에 저장된 대상소스로부터 추출한 미분류 문서(Unlabeled Document) 중 준지도 학습방법을 채택하기 위한 일부를 추출하여 보강된 미분류 문서로 저장하는 보강된 미분류 문서 DB(104)와, 대상소스 DB(102)에 저장된 대상소스로부터 추출한 미분류 문서(Unlabeled Document) 중 상기 보강된 미분류 문서로 추출된 문서를 제외한 나머지 미분류 문서인 대상 미분류 문서를 저장하는 대상 미분류 문서 DB(105)와, 상기 시드문서 DB(103), 보강된 미분류 문서 DB(104) 및 대상 미분류 문서 DB(105)에 각각 수집되어 저장된 문서들(시드문서, 보강된 미분류 문서, 대상 미분류 문서)을 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화하는 문서 구조화부(106)로 구성된다.

[0037] 이와 같이 구성되는 상기 토픽 모델링부(100)를 통한 비정형 문서를 구조화하는 구성을 상세히 설명하면 다음과 같다.

[0038] 상기 토픽 모델링부(100)는 둘 이상의 다양한 매체에 대해 적용되며, 하나의 매체는 기준소스(Base Source)로 기존소스 DB(101)에 저장되고, 다른 매체들은 대상소스(Target Source)로 대상소스 DB(102)에 저장된다. 이러한 과정은 각 매체에 대해 반복적으로 수행된다. 예를 들어 총 N개의 매체가 있는 경우 각 매체는 한 번은 기준소스, (N-1)번은 대상소스의 역할로 참여한다. 이렇듯 이론상 개수의 제한이 없는 복수개의 매체에 반복 적용 가능하지만, 본 발명에서는 설명의 편의를 위해 단 두 개의 매체만 존재하는 경우를 가정한다.

[0039] 상기 기준소스는 이미 고유의 카테고리 체계를 갖고 있으며, 기준소스로부터 추출한 문서들은 모두 기분류 문서(Labeleed Documents)로서, 이들 문서의 집합을 시드문서(Seed Documents)라고 하며, 시드문서 DB(103)에 저장된다. 상기 시드문서가 분류 학습의 원천으로 사용된다.

[0040] 한편, 분류의 대상이 되는 대상소스는 상기 기준소스의 관점에서 새로운 카테고리를 부여받게 되므로, 고유의 카테고리 체계는 모두 무시된다. 따라서 대상소스로부터 추출한 문서들은 모두 미분류 문서(Unlabeled Document)로 간주된다. 또한 기준소스의 규모가 매우 작아서 시드문서의 수가 현저히 부족한 경우 학습을 위해 필요한 기분류 문서를 보강할 필요가 있음으로, 학습에 활용하는 준지도 학습방법을 채택하기 위해 대상소스로부터 미분류 문서의 일부를 추출하여 보강된 미분류 문서 DB(104)에 저장되고, 나머지 미분류 문서는 대상 미분류 문서 DB(105)에 저장된다.

[0041] 그리고 문서 구조화부(106)를 통해 상기 시드문서 DB(103), 보강된 미분류 문서 DB(104) 및 대상 미분류 문서 DB(105)에 각각 저장된 문서들(시드문서, 보강된 미분류 문서, 대상 미분류 문서)을 모두 통합하고, 토픽 모델링을 수행하여 각 문서를 구조화한다. 이때, 상기 토픽 모델링은 이미 기존의 많은 연구들을 통해 충분히 설명되었으므로, 본 발명에서는 토픽 모델링의 과정에 대한 자세한 설명 대신 주요 개념만을 간략하게 소개한다.

[0042] 즉, 분석의 대상이 되는 문서가 포함하고 있는 용어의 수는 일반적으로 매우 방대하기 때문에, 문서를 용어의 빈도에 기반을 두어 구조화하는 과정에서 용어에 대한 차원 축소가 반드시 필요하다. 이때 사용된 차원의 수가 일반적인 토픽 모델링에서의 토픽의 수를 나타낸다. 이후 각 용어가 토픽에 대응되는 정도인 용어 가중치(Term Topic Weight)를 산출할 수 있으며, 용어 가중치는 정해진 용어 임계값(Term Cutoff) 이상인 경우, 해당 토픽을 나타내는 용어로 간주된다. 임계값으로는 주로 각 토픽의 모든 용어 가중치의 "평균 + 1 σ (Sigma, 표준편차)"가 사용된다. 유사한 방식으로 각 문서의 문서 가중치(Document Topic Weight) 또한 산출할 수 있는데, 이는 TF-IDF(Term Frequency - Inverse Document Frequency)와 용어 가중치의 곱의 표준합(Normalized Sum)으로 계산된다. 문서 가중치 또한 임계값 이상의 값의 갖는 경우 해당 문서가 해당 토픽에 속하는 것으로 분류되며, 임계값으로는 주로 각 토픽의 모든 문서 가중치의 "평균 + 1 σ"가 사용된다. 이러한 방식을 통해 방대한 문서로부터 주요 토픽을 추출할 수 있지만, 본 발명에서는 토픽의 추출보다는 토픽 모델링 과정에서 산출되는 문서 가중치에 주목한다. 즉, 각 문서를 각 토픽에 대응되는 정도인 문서 가중치의 벡터로 나타냄으로써 문서를 구조적 형태로 표현할 수 있으며, 이후 문서 가중치 벡터를 입력 변수로, 카테고리를 목적 변수로 설정하여 문서 분류를 위한 학습 및 분류를 수행하게 된다.

[0043] 도 3(a)는 기준소스로부터 도출된 시드문서를 나타내며, "IT", "Politics", "Ent." 등의 카테고리로 구분되어 있다. 한편, 도 3(b)와 도 3(c)의 경우 대상소스로부터 도출된 문서로 "Cate1", "Cate2", "Cate3" 등의 카테고리로 구분되어 있다. 하지만 본 분석에서는 기준소스의 카테고리만이 유효하게 작용하기 때문에 대상소스로부터 도출된 문서의 기존 카테고리는 모두 무시된다. 따라서 도 3에서 기분류 문서는 "B_Doc1"~"B_Doc3"로, 미분류

문서는 "O_Doc4"~"O_Doc9"로 사용된다.

[0044] 문서 구조화부(106)는 도 3(a), 도 3(b) 및 도 3(c)의 문서를 모두 한꺼번에 토픽 모델링의 입력으로 사용되며, 그 분석 결과가 도 3(d)에 나타나 있다.

[0045] 도 3(d)의 우측 부분은 각 토픽에 대한 각 문서의 문서 가중치를 나타내며, 향후 분석에서 입력변수로 사용된다. 예를 들어 "B_Doc1" 문서의 경우("IT", 1.2, 0.9, 0.7, 1.3, 1.1, ...)의 벡터로 구조화되며, 가장 첫 요소는 목적 변수, 그리고 나머지 요소들은 입력 변수로 구분된다.

[0046] 그리고 상기 1 차 학습 및 분류부(200)는 토픽 모델링부(100)에서 구조화된 문서 중 시드문서 및 보강된 미분류 문서를 입력으로 각각 저장하는 구조화된 시드문서 DB(103)와, 상기 구조화된 시드문서 DB(103)에 저장된 소량의 기분류 시드문서에 대한 학습을 통해 분류 알고리즘을 생성하는 제 1 학습부(202)와, 상기 제 1 학습부(202)에서 생성된 분류 알고리즘을 통해 분류 규칙을 생성하는 제 1 규칙 분류부(203)와, 상기 제 1 규칙 분류부(203)에서 생성된 분류 규칙을 적용하여 보강된 미분류 문서를 분류하는 제 1 분류부(204)로 구성된다.

[0047] 이와 같이 구성되는 1차 학습 및 분류부(200)를 통한 1차 학습 데이터를 생성하는 방법을 상세히 설명하면 다음과 같다.

[0048] 일반적인 문서 분류기가 채택하는 방식인 지도 학습의 경우, 소량의 기분류 문서만을 학습 데이터로 사용할 경우에는 편중, 과적합, 과소적합 등의 현상으로 인해 분류기의 성능이 낮게 나타남이 이미 알려진 바 있다. 본 발명에서는 이를 극복하기 위해 카테고리 분류 과정에 지도 학습이 아닌 준지도 학습을 활용한다. 이미 준지도 학습의 성능 향상을 위해 최대기대(Expectation Maximization) 기반, 그래프(Graph) 기반, co-training 기반 알고리즘 등 다양한 방법이 고안되어 왔다. 하지만 준지도 학습의 성능 개선은 본 발명의 핵심적으로 다루는 내용이 아니므로, 본 발명에서는 준지도 학습 과정에서 가장 직관적이고 단순한 방법으로 사용한다. 즉, 소량의 기분류 문서를 학습 집합으로 활용하여 미분류 문서를 학습 집합으로 활용하여 미분류 문서의 일부를 분류한 후, 이후 2차 학습 및 분류부(300)에서 상기 분류된 미분류 문서를 기존의 기분류 문서와 통합하여 새로운 학습 집합으로 사용한다.

[0049] 상기 준지도 학습 기반의 문서 분류 중 1차 학습 및 분류 과정에 관한 설명이 도 4에서 나타내고 있다.

[0050] 도 4(a)는 소량의 기분류 문서를 나타내며, 이들 문서에 대한 학습을 통해 분류 알고리즘을 생성한다. 이렇게 생성된 분류 알고리즘을 통해 도 4(b)의 미분류 문서를 분류함으로써, 도 4(c)와 같은 기분류 문서를 추가로 획득할 수 있으며, 이렇게 추가 분류된 문서가 기존의 기분류 문서와 함께 추후 학습에 활용된다.

[0051] 이 과정에서 분류된 문서가 전부가 아닌 일부, 즉 분류된 문서의 분류 확률(Probability or Score)이 특정 임계값 이상인 문서만을 추후 학습에 활용할 수도 있다.

[0052] 마지막으로 상기 2차 학습 및 분류부(300)는 토픽 모델링부(100)에서 구조화된 문서 중 보강된 미분류 문서와 상기 1차 학습 및 분류부(200)에서 분류된 문서를 저장하는 구조화된 보강된 미분류 문서 DB(301)와, 상기 보강된 미분류 문서와 분류된 문서를 통합하여 저장하는 구조화된 보강된 기분류 문서 DB(302)와, 상기 구조화된 보강된 기분류 문서 DB(302)에 저장된 통합 문서에 대한 학습을 통해 분류 알고리즘을 생성하는 제 2 학습부(303)와, 상기 제 2 학습부(303)에서 생성된 분류 알고리즘을 통해 분류 규칙을 생성하는 제 2 규칙 분류부(304)와, 토픽 모델링부(100)에서 구조화된 문서 중 구조화된 대상 미분류 문서 DB(400)에 저장된 대상 미분류 문서의 구조화된 문서를 상기 제 2 규칙 분류부(304)에서 생성된 분류 규칙을 적용하여 분류하여 최종 대상문서 DB(500)에 저장하는 제 2 분류부(305)로 구성된다.

[0053] 이와 같이 구성되는 2차 학습 및 분류부(300)를 통한 2차 학습 데이터를 생성하는 구성을 상세히 설명하면 다음과 같다.

[0054] 2차 학습 및 분류부(200)를 통해 새로 분류된 문서는 기존의 기분류 문서와 통합되어 2차 분류의 학습 데이터로 활용된다. 2차 분류는 상기 1차 학습 및 분류부(200)의 구성과 매우 유사한 형태로 수행되며, 도 5에서 나타내고 있다.

[0055] 도 5(a)는 도 4(d)에 해당하는 문서 집합으로, 2차 분류의 학습 데이터로 사용된다. 이러한 과정을 통해 최종적으로 도 5(b)에서 나타내고 있는 대상 미분류 문서의 카테고리를 식별하게 되며, 그 결과가 도 5(c)에 나타내고 있다.

[0056] 본 발명에서는 편의를 위해 두 매체로부터 문서가 도출된 경우만을 예로 들어 설명하였지만, 본 발명에 따른 방

법은 유사한 과정의 반복 적용을 통해 둘 이상의 매체에 확장 적용될 수 있다. 이 경우 최종 결과물은 도 6과 같은 형태로 나타나게 되며, 각 문서는 원 소속 매체의 카테고리 뿐 아니라 서로 상이한 구조를 가진 다른 매체의 카테고리 정보 또한 동시에 갖게 된다.

[0057] 예를 들어 도 6에서 1번 문서의 경우 원래 매체 "D News"의 카테고리 "IT"에 속한 문서이며, 본 발명의 방법을 통해 매체 "N Blog"의 카테고리 "Travel"과 매체 "A Discussion"의 카테고리 "Life"에도 추가로 연결되었음을 알 수 있다.

[0058] 실시예

[0059] 본 발명의 방법을 적용하여 실제 수집된 이종 매체 문서에 대하여 실험 및 분석하였다. 실험 대상 매체로는 인터넷 뉴스 포털인 "N" 사이트와 "O" 사이트를 선정하였으며, 각 사이트로부터 뉴스 기사 3,000건씩 총 6,000건의 뉴스 기사를 수집하였다. 기사 원본은 JAVA 기반의 크롤러를 직접 제작하여 수집하였으며, 제안 방법론은 시간의 흐름에 따른 변화나 추이 등의 영향을 받지 않으므로 데이터 수집 기간에 대해서는 별도의 제약을 두지 않았다. 사이트 "N"은 "IT 과학", "정치", "사회", "생활문화", "세계", "스포츠" 그리고 "연애" 등 총 8개의 카테고리 분류 체계를 갖고 있었으며, 사이트 "O"는 "경제", "교육", "미디어", "민족/국제", "사회", "정치" 및 "여성" 등 총 7개의 카테고리를 관리하고 있었다. 하지만 "여성" 카테고리의 경우 보유하고 있는 문서 수가 극히 적어 본 실험에서 제외하였으며, 사이트 "N"의 8개 카테고리와 사이트 "O"의 6개의 카테고리에 포함된 기사만을 대상으로 실험을 진행하였다. 본 연구에서 수행한 실험의 개요가 도 7에 제시되어 있다.

[0060] 본 발명의 방법에 따른 성능을 간접적으로 파악하기 위해 여러 상황에 따른 제안 방법의 정확도를 비교 분석한다. 정확도 비교는 매체 간 비교, 지도 학습과 준지도 학습 비교, 학습 데이터의 이질성 비교의 세 가지 관점에서 이루어졌다. 이상 전체 6가지 실험의 정확도를 요약한 결과는 다음 표 1과 같다.

표 1

[0061]

(1) Simple_N	(2) Semi_Homo_N	(3) Semi_Hetero_N	(4) Simple_O	(5) Semi_Homo_O	(6) Semi_Hetero_O
0.8033	0.6167	0.581	0.7227	0.6733	0.634

[0062] 표 1의 결과에 따르면 지도 학습의 경우 사이트 "N"이 사이트 "O"에 비해 분류 정확도가 높게 나타났으며, 준지도 학습의 경우 반대로 사이트 "O"이 사이트가 "N"에 비해 분류 정확도가 높게 나타났다. 한편, 두 사이트 모두에 대해, 학습 데이터가 충분한 경우 수행 가능한 지도 학습이 준지도 학습에 비해 분류 정확도가 높게 나타났다. 마지막으로 학습 데이터의 이질성 비교 실험의 경우, 동일 소스로부터 학습 데이터를 보강한 경우에 비해 분류 정확도가 높게 나타났다.

[0063] 각 실험에 대한 보다 자세한 분석은 다음과 같다.

[0064] 우선 도 8 은 사이트 "N"의 문서에 대한 분류 실험 세 가지의 누적 반응 검출률(Cumulative Response)을 나타낸다. 실험 (1), (2), 그리고 (3)의 결과는 그래프에서 각각 점선, 흐린 실선, 그리고 짙은 실선으로 나타나있다.

[0065] 본 그래프는 각 실험에서 나타난 각 문서의 분류 확률의 내림차순으로 문서를 정렬한 뒤, 정렬 순서에 따른 각 문서의 분류 정확도를 누적으로 측정한 것이다.

[0066] 실험 결과 (1)번 실험 최상위 문서의 일부 구간을 제외하면, 세 가지 실험 모두 가파른 우하향 형태를 나타냄을 알 수 있다. 즉 세 가지 경우 모두 상위 스코어를 갖는 문서의 예측 정확도가 하위 스코어를 갖는 문서에 비해 비교적 높게 나타나는 바람직한 특징을 가짐을 알 수 있다.

[0067] 이러한 현상은 도 9의 사이트 "O"에 대한 문서 분류 실험에서도 동일하게 나타난다.

[0068] 위의 도 8과 도 9를 통해, 두 개의 사이트 모두에서 기준 소스와 동일한 매체의 문서를 학습 데이터의 보강에 사용하는 경우가 기준 소스와 다른 매체의 문서를 사용하는 경우에 비해 분류 정확도가 높게 나타남을 알 수 있었다.

[0069] 본 발명에서는 이러한 현상이 각 매체의 모든 카테고리에 대해 일반적으로 나타나는 현상인지 여부를 살펴보기 위해, 위의 6가지 실험에 각각에 대해 각 카테고리 별 분류 정확도를 측정하는 추가 실험을 수행하였다. 사이트 "N"에 대한 실험 결과는 도 10에, 사이트 "O"에 대한 실험 결과는 도 11에 제시되어 있으며, 도면에서 빗금으로 나타난 막대(Bar)는 지도학습, 흰색으로 나타난 막대는 동질 준지도 학습, 그리고 회색으로 나타난 막대는 이질 준지도 학습의 분류 정확도를 나타낸다.

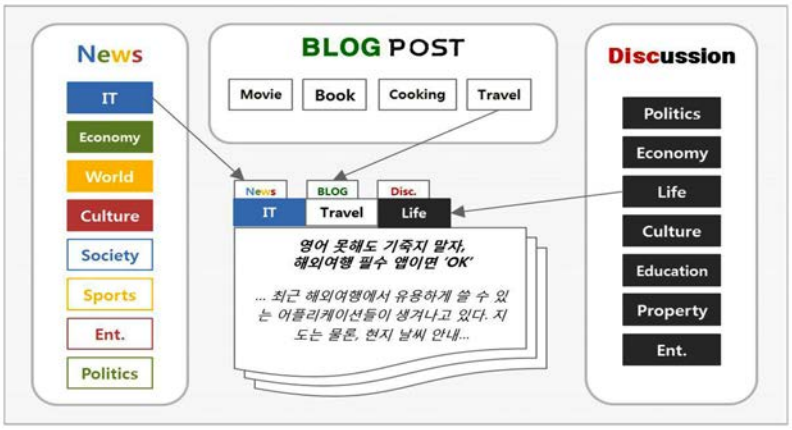
[0070] 도 10에서 나타낸 그래프를 보면, "경제", "생활문화", "스포츠", "연예" 그리고 "정치" 등의 카테고리의 경우 이질 학습 데이터를 사용한 준지도 학습의 정확도가 동질 학습 데이터를 사용한 경우에 비해 더욱 높은 것으로 나타났다. 특히 "생활문화"와 "정치" 카테고리의 경우 이질 학습 데이터를 사용한 준지도 학습의 정확도가 지도 학습의 경우보다도 높게 나타났다. 이러한 현상은 도 11의 사이트 "O"에 대한 실험에서도 마찬가지로 나타나서, "경제", "교육" 그리고 "정치" 등의 카테고리에서 이질 학습 데이터를 사용한 준지도 학습의 정확도가 동질 학습 데이터를 사용한 경우에 비해 더욱 높게 나타났으며, 이들 중 "교육" 카테고리의 경우는 이질 학습 데이터를 사용한 준지도 학습의 정확도가 지도 학습의 경우보다도 높게 나타났다.

[0071] 위 실험과 같이, 수행한 6가지 성능 비교 실험의 결과는 각 카테고리에 따라 매우 상이한 형태로 나타나며, 특히 일부 카테고리의 경우 이질 준지도 학습의 분류 정확도가 동질 준지도 학습 뿐 아니라 지도 학습의 분류 정확도보다도 오히려 높게 나타남을 알 수 있었다. 향후 분류 정확도가 카테고리 별로 상이하게 나타나는 원인 및 이질적인 문서가 학습 데이터 보강에 어떤 영향을 주는지에 대한 보다 엄밀한 분석을 통해, 제안 방법론의 정확도와 활용성을 더욱 높일 수 있을 것으로 기대한다.

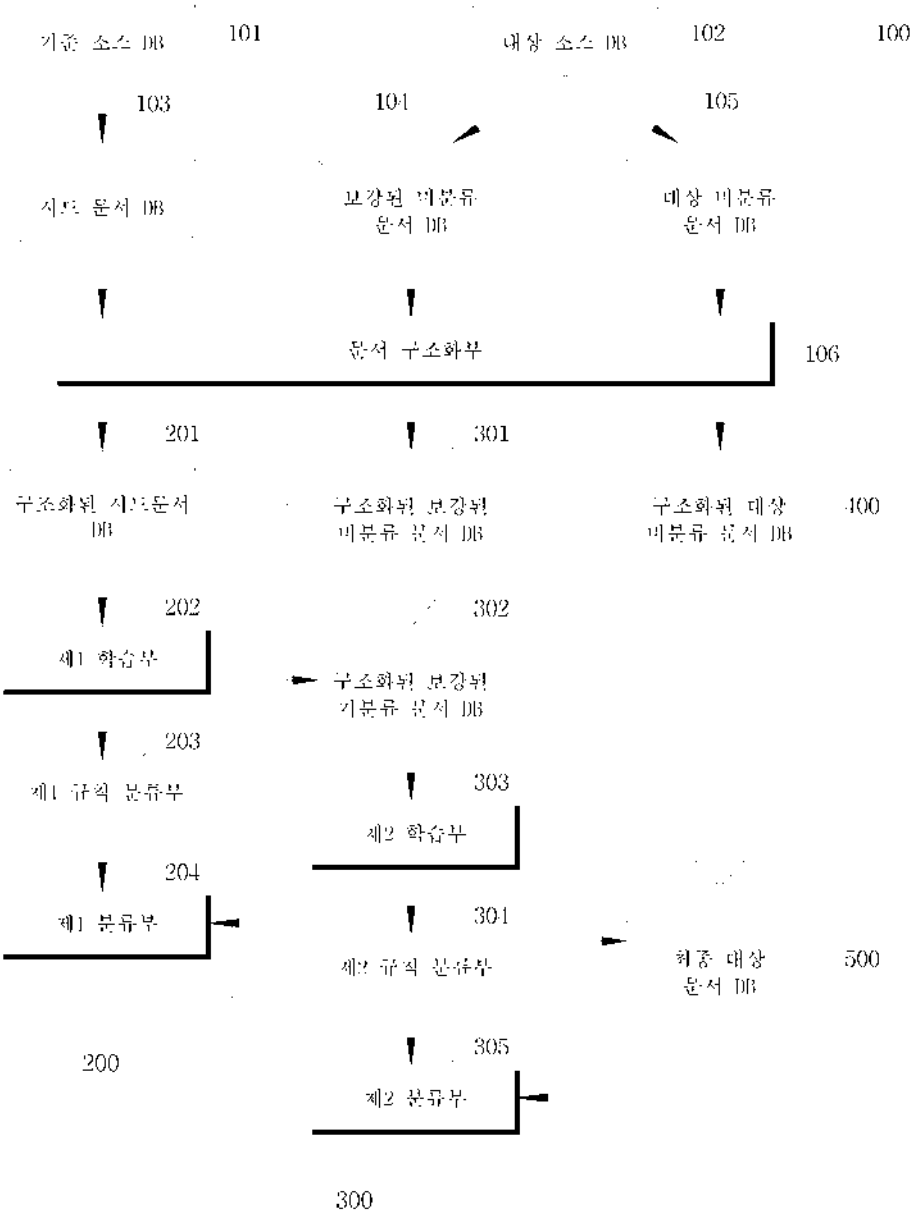
[0072] 상기에서 설명한 본 발명의 기술적 사상은 바람직한 실시예에서 구체적으로 기술되었으나, 상기한 실시예는 그 설명을 위한 것이며 그 제한을 위한 것이 아님을 주의하여야 한다. 또한, 본 발명의 기술적 분야의 통상의 지식을 가진자라면 본 발명의 기술적 사상의 범위 내에서 다양한 실시예가 가능함을 이해할 수 있을 것이다. 따라서 본 발명의 진정한 기술적 보호 범위는 첨부된 특허청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

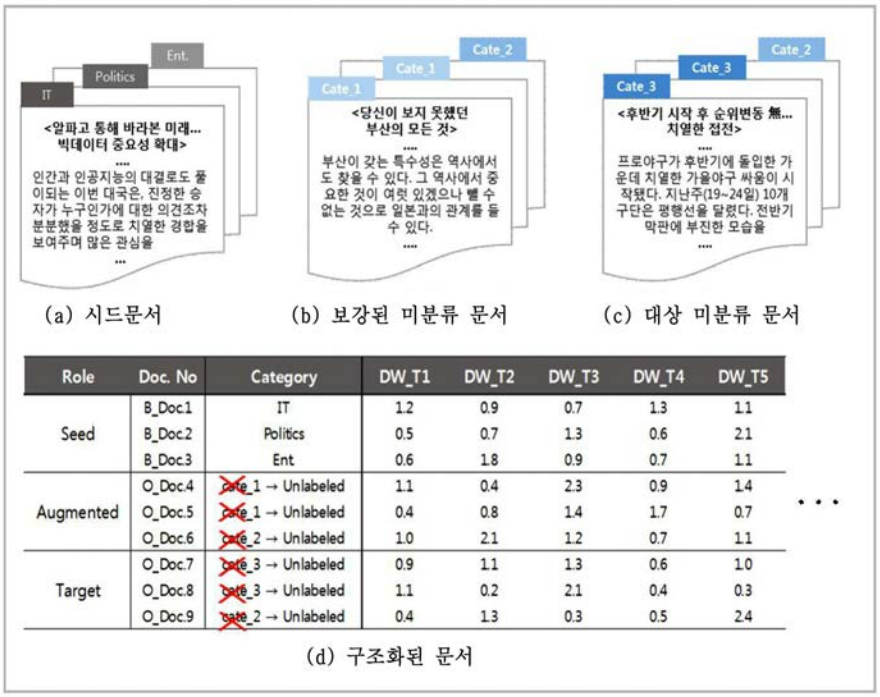
도면1



도면2


기준 소스 DB
101
대상 소스 DB
102
100
103
104
105
시드 문서 DB
보강된 비분류 문서 DB
대상 비분류 문서 DB
문서 구조화부
106
201
301
구조화된 시드문서 DB
구조화된 보강된 비분류 문서 DB
구조화된 대상 비분류 문서 DB
400
202
302
제1 학습부
구조화된 보강된 비분류 문서 DB
203
303
제1 규칙 분류부
제2 학습부
204
301
제1 분류부
제2 규칙 분류부
최종 대상 문서 DB
500
200
305
제2 분류부
300

도면3

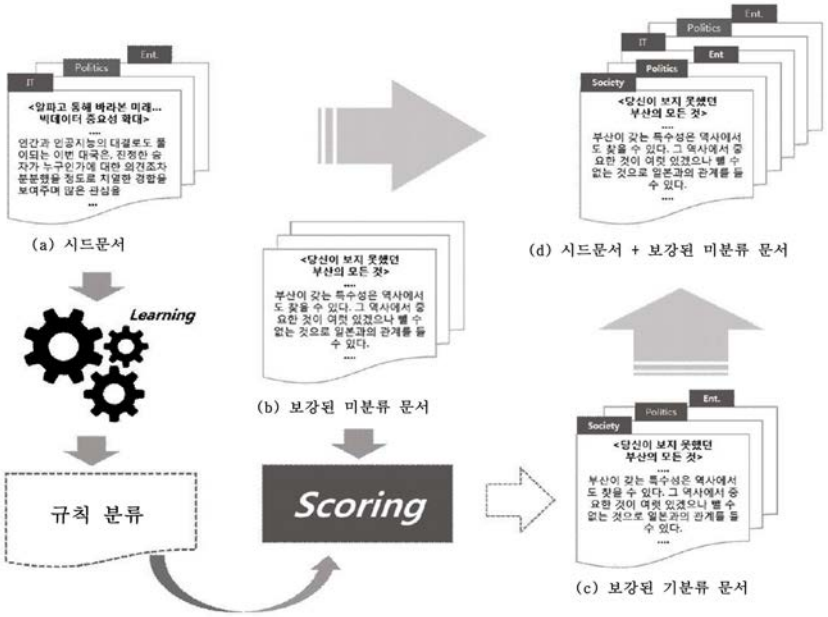


(a) 시드문서 (b) 보강된 미분류 문서 (c) 대상 미분류 문서

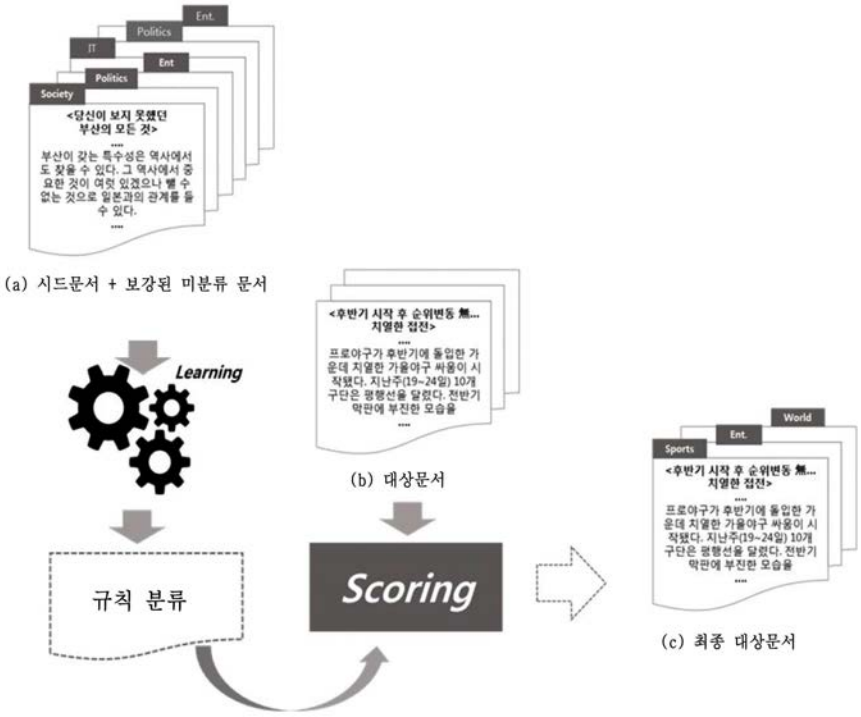
Role	Doc. No	Category	DW_T1	DW_T2	DW_T3	DW_T4	DW_T5
Seed	B_Doc.1	IT	1.2	0.9	0.7	1.3	1.1
	B_Doc.2	Politics	0.5	0.7	1.3	0.6	2.1
	B_Doc.3	Ent	0.6	1.8	0.9	0.7	1.1
Augmented	O_Doc.4	~~Cate_1~~ → Unlabeled	1.1	0.4	2.3	0.9	1.4
	O_Doc.5	~~Cate_1~~ → Unlabeled	0.4	0.8	1.4	1.7	0.7
	O_Doc.6	~~Cate_2~~ → Unlabeled	1.0	2.1	1.2	0.7	1.1
Target	O_Doc.7	~~Cate_3~~ → Unlabeled	0.9	1.1	1.3	0.6	1.0
	O_Doc.8	~~Cate_3~~ → Unlabeled	1.1	0.2	2.1	0.4	0.3
	O_Doc.9	~~Cate_2~~ → Unlabeled	0.4	1.3	0.3	0.5	2.4

(d) 구조화된 문서

도면4



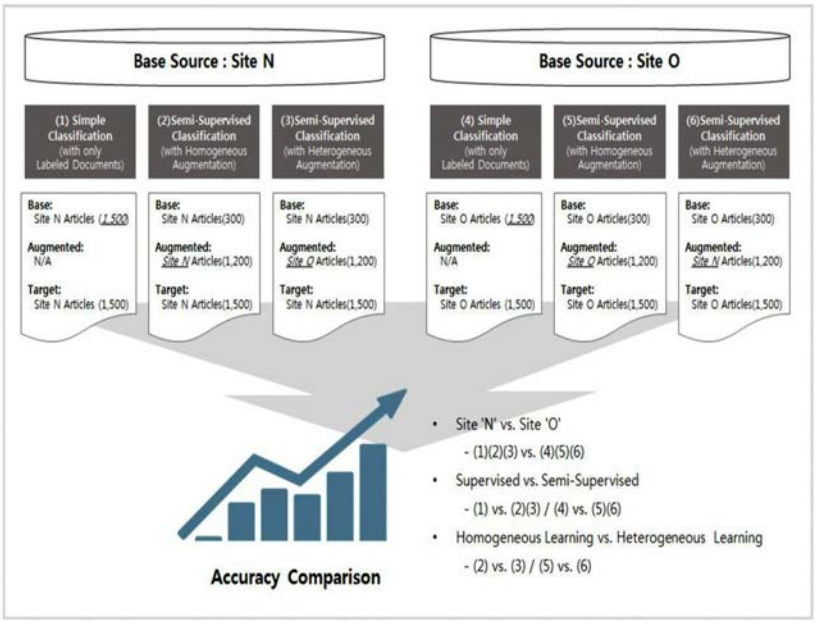
도면5



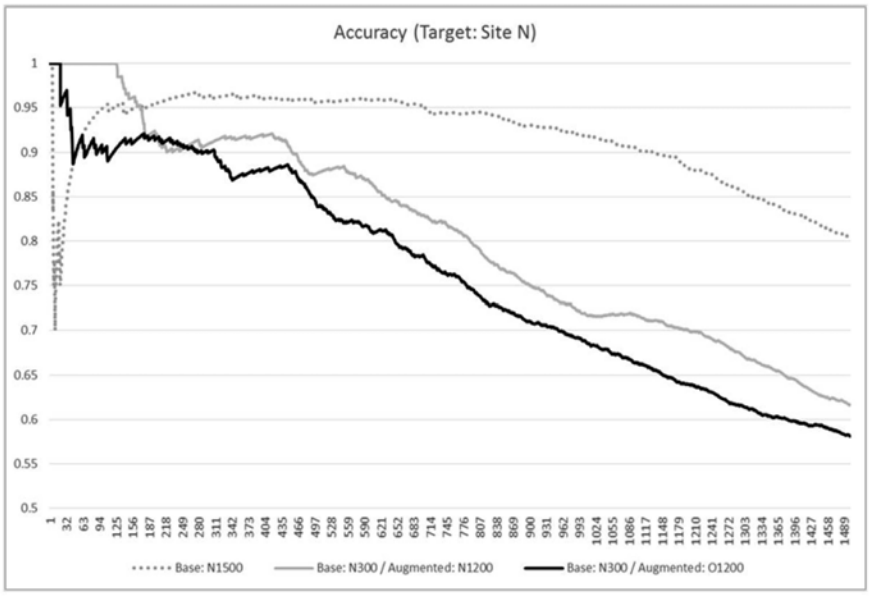
도면6

Doc.No	Original		Extended			
	Source	Category	Cat1 (D News)	Cat2 (N Blog)	Cat3(A Discussion)	. . .
1	D News	IT	IT	Travel	Life	
2	N Blog	Movie	Culture	Movie	Culture	. . .
3	A Discussion	Life	Culture	Cooking	Life	
. . .	. . .			. . .		

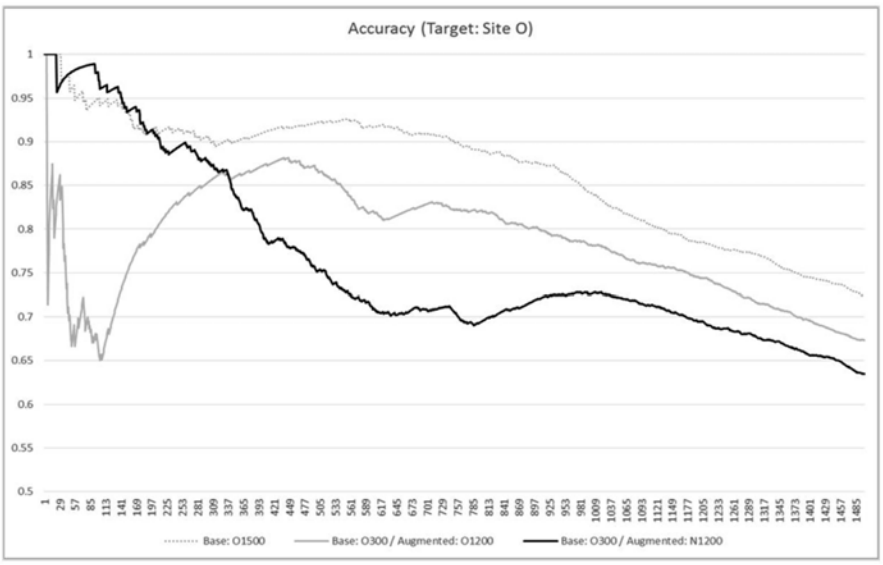
도면7


Base Source : Site N
Base Source : Site O
(1) Simple Classification (with only Labeled Documents)
(2)Semi-Supervised Classification (with Homogeneous Augmentation)
(3)Semi-Supervised Classification (with Heterogeneous Augmentation)
(4) Simple Classification (with only Labeled Documents)
(5)Semi-Supervised Classification (with Homogeneous Augmentation)
(6)Semi-Supervised Classification (with Heterogeneous Augmentation)
Base: Site N Articles (1,500)
Augmented: N/A
Target: Site N Articles (1,500)
Base: Site N Articles(300)
Augmented: Site N Articles(1,200)
Target: Site N Articles(1,500)
Base: Site N Articles(300)
Augmented: Site O Articles(1,200)
Target: Site N Articles(1,500)
Base: Site O Articles (1,500)
Augmented: N/A
Target: Site O Articles (1,500)
Base: Site O Articles(300)
Augmented: Site O Articles(1,200)
Target: Site O Articles(1,500)
Base: Site O Articles(300)
Augmented: Site N Articles(1,200)
Target: Site O Articles(1,500)
Accuracy Comparison
• Site 'N' vs. Site 'O'
- (1)(2)(3) vs. (4)(5)(6)
• Supervised vs. Semi-Supervised
- (1) vs. (2)(3) / (4) vs. (5)(6)
• Homogeneous Learning vs. Heterogeneous Learning
- (2) vs. (3) / (5) vs. (6)

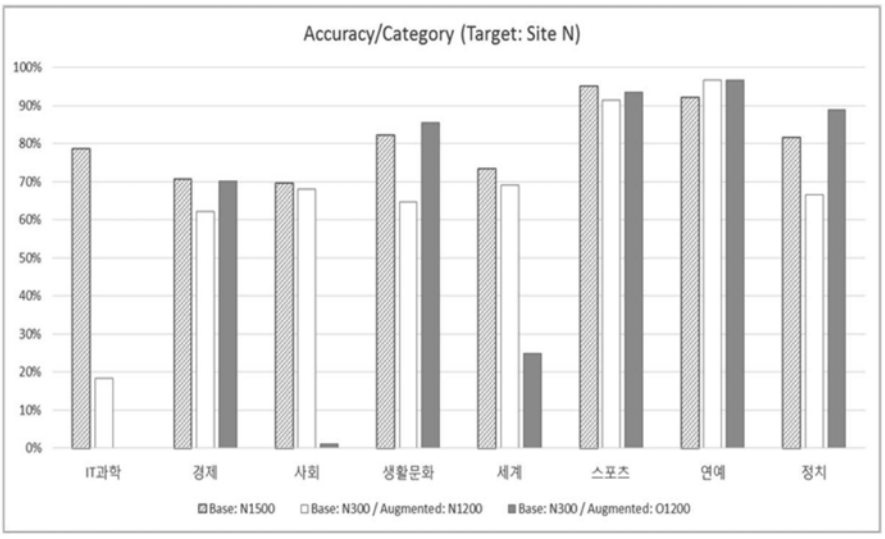
도면8


Accuracy (Target: Site N)
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Base: N1500
Base: N300 / Augmented: N1200
Base: N300 / Augmented: O1200

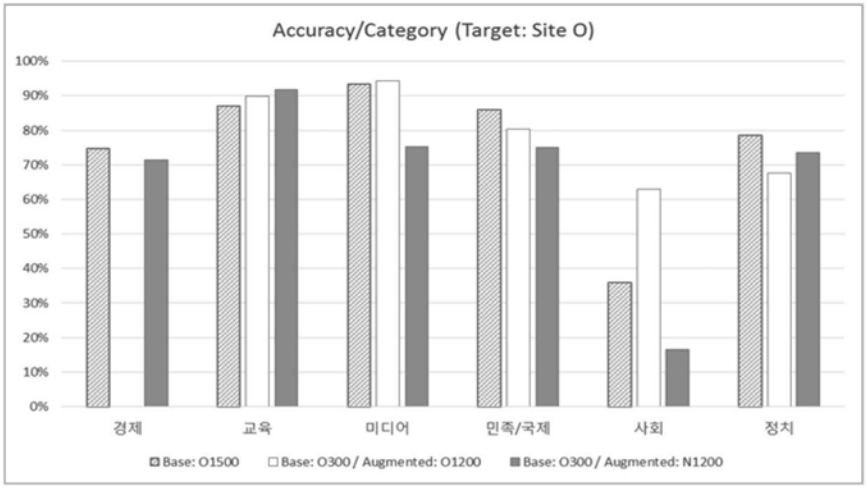
도면9


Accuracy (Target: Site O)
Base: O1500
Base: O300 / Augmented: O1200
Base: O300 / Augmented: N1200

도면10


Accuracy/Category (Target: Site N)
IT과학
경제
사회
생활문화
세계
스포츠
연예
정치
Base: N1500
Base: N300 / Augmented: N1200
Base: N300 / Augmented: O1200

도면11


Accuracy/Category (Target: Site O)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
경제
교육
미디어
민족/국제
사회
정치
Base: O1500
Base: O300 / Augmented: O1200
Base: O300 / Augmented: N1200