ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2016 December: 22(4): 193~215 http://dx.doi.org/10.13088/jiis.2016.22.4.193

텍스트 분석을 통한 이종 매체 카테고리 다중 매핑 방법론*

김다솜
국민대학교 비즈니스IT전문대학원
(dskim1225@kookmin.ac.kr)

김남규
국민대학교 비즈니스IT전문대학원
(ngkim@kookmin.ac.kr)

최근 다양한 소셜 네트워크 서비스의 증가로 인해 사용자들은 각자의 목적 및 취향에 따라 여러 매체를 동시에 이용하는 경향을 보이고 있다. 또한 특정 주제에 대한 정보를 수집할 때에도 소셜 네트워크 서비스, 인터넷 뉴스, 블로그 등 여러 매체를 동시에 활용하는 것이 일반적이다. 하지만 다양한 매체를 통해 유통되는 문서들은 서로 유사한 주제, 심지어는 동일한 내용을 다루더라도 각 매체 별 정책 및 기준에 따라 각기 다른 카테고리로 관리되고 있으며, 이는 이종 매체를 아우르는 범위에서 특정 카테고리에 대한 탐색을 수행하고자 하는 시도에 걸림돌로 작용하고 있다. 이러한 제약을 극복하기 위해, 본 연구에서는 기존 매체 고유의 카테고리 체계는 그대로 유지하면서 이종 매체 간 카테고리 매핑을 수행하는 방법을 제시한다. 즉, 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 제안한다. 본 논문에서는 국내 인터넷 뉴스 포털 사이트 두 곳의 뉴스 기사 6,000건에 대해 제안 방법론을 적용한 실험을 통해 각 기사에 매체와 카테고리 정보로 구성된 2차원 레이블을 부여하였으며, 매체 간, 지도 학습과 준지도 학습 간, 동질 학습 데이터와 이질학습 데이터 간의 정확도 비교 실험을 수행하였다. 특히 매우 흥미롭게도, 일부 카테고리에서 이질 학습 데이터를 사용한 준지도 학습의 분류 정확도가 지도 학습 및 동질 학습 데이터를 사용한 준지도 학습의 분류 정확도보다 높게 나타나는 현상을 발견하였다.

주제어 : Category Mapping, Document Classification, Text Mining, Topic Modeling

논문접수일 : 2016년 8월 17일 논문수정일 : 2016년 12월 10일 게재확정일 : 2016년 12월 28일
원고유형 : 일반논문 교신저자 : 김남규

1. 서론

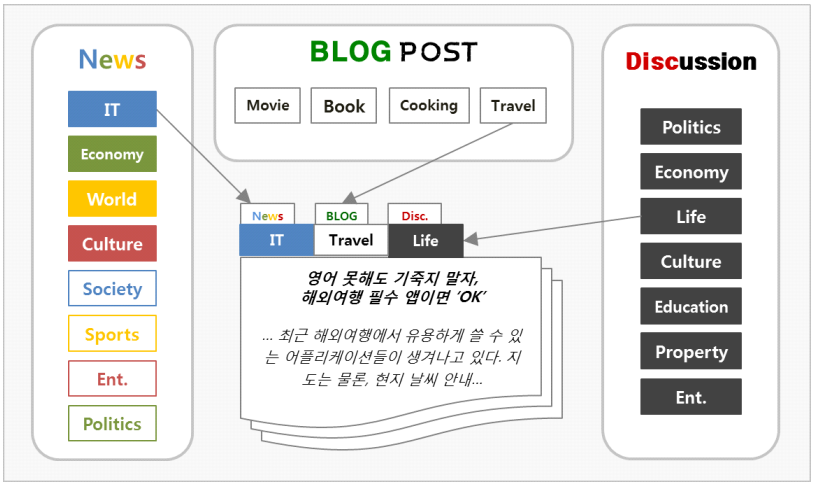
최근 스마트 기기의 발달과 인터넷의 보급화로 인해 사용자들은 소셜 네트워크 서비스(Social Network Service), 인터넷 뉴스, 웹 커뮤니티 등 다양한 매체를 시간과 장소에 제약 받지 않고 사용할 수 있게 되었다. 이에 부응하여 다양한 기능과 목적을 지닌 매체들 또한 꾸준히 개발되고 있으며, 사용자들은 각자의 목적 및 취향에 따라 일반적으로 여러 매체들을 동시에 이용하고 있다. 다양한 매체 가운데 특히 인스타그램(Instargram), 트위터(Twitter), 페이스북(Facebook) 등의 사용이 두드러지며, 2013년 기준 국내 소셜 네트워크 서비스 사용자는 평균 2.09개의 매체를

* 이 논문은 2016년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2016S1A5A2A01022228).

동시에 이용하고 있는 것으로 나타났다(한국직업능력개발원, 2013). 이처럼 사용자들은 본인의 의견 또는 정보를 다양한 매체를 통해 공유함을 물론, 반대로 특정 주제에 대한 정보를 수집할 때에도 여러 매체를 동시에 활용하고 있다. 이렇듯 다양한 매체를 통해 공유되고 수집되는 디지털 정보의 양은 2020년에 35제타바이트(ZB)를 훨씬 넘을 것으로 전망되고 있다(한국인터넷진흥원, 2014).

이처럼 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 <Figure 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.

이러한 매체에 따른 분류 체계의 상이함은 전체 매체를 아우르는 분석을 통해 새로운 지식을 창출하기 위한 시도에 걸림돌로 작용할 수 있다. 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 찾고자 하는 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, 이처럼 이종 매체를 아우르는 범위에서



〈Figure 1〉 Diverse category names in heterogeneous sources

특정 카테고리에 대한 탐색이 이루어지기란 매우 어렵다.

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. 따라서 본 연구에서는 기존 매체 고유의 카테고리 체계는 그대로 유지하면서, 위에서 언급한 한계를 극복하기 위해 이종 매체간 카테고리 매핑을 수행하는 방안을 제시하고자 한다. 즉 개별 문서를 다양한 매체의 관점에서 재분류하고 이러한 분류 결과를 문서에 2차원 레이블(Label)로 저장함으로써, 이종 매체에 속한 다양한 문서들을 마치 한 매체에 속한 것과 같이 동일한 카테고리 기준으로 탐색할 수 있는 논리적 장치를 제안하고자 한다.

본 논문의 이후 부분은 다음과 같이 구성된다. 다음 장인 2장에서는 본 연구와 관련된 선행 연구들을 요약하고, 3장에서는 본 연구의 전체적인 개요와 방법론을 소개한다. 4장에서는 3장에서 제시한 방법론을 실제 데이터에 적용한 실험 결과를 살펴보고 마지막 5장에서는 본 연구의 기여, 한계, 그리고 후속 연구의 방향을 제시한다.

2. 관련 연구

최근 사회 전반에 걸쳐 정형 또는 비정형 데이터가 실시간으로 생성되고 있으며, 이렇게 발생되는 방대한 양의 데이터를 분석하는 기법을 빅데이터 분석이라 한다(McKinsey, 2011). 또한 최근 스마트 기기와 인터넷의 발달로 인해 필요한 정보를 수집할 수 있는 매체들이 더욱 다양해졌으며, 이를 통해 수집한 방대한 양의 텍스트를 대상으로 텍스트 마이닝(Text Mining) 분야의 연구가 활발히 진행되고 있다. 텍스트 마이닝은 다량의 텍스트 문서 또는 문장에 대한 분석을 통해 의미 있는 정보를 추출하는 과정으로 정의될 수 있으며(Hearst, 1999), 기본적으로 비정형 텍스트에 대한 구조화 이후 기존의 일반적인 데이터 마이닝 기법을 적용하는 형태로 분석이 이루어진다. 비정형 텍스트의 구조화에는 행렬, 계층, 벡터 등이 사용될 수 있으며, 일반적으로 벡터공간 모델(Vector Space Model)(Salton et al., 1975)에 기반을 둔 구조화가 수행된다.

텍스트 마이닝의 세부 분야 중 가장 활발한 응용이 이루어지고 있는 대표적인 분야로 토픽 모델링(Topic Modeling)과 문서 분류(Document Classification)을 들 수 있다. 우선 토픽 모델링은 대상 문서를 유사도에 따라 그룹화하고 각 그룹을 대표하는 주요 용어를 제시함으로써, 방대한 양의 문서로부터 주요 토픽을 추출하는 일련의 과정을 나타낸다. 문서 간 유사도는 기본적으로 코사인 유사도(Salton and McGill, 1986)로 측정되며, 최근에는 LSA(Latent Semantic Analysis)(Deer wester et al., 1990), PLSA(Probabilistic LSA)(Hofmann, 1999), LDA(Latent Dirichlet Allocation)(Blei et al., 2003) 등 다양한 기법이 토픽 모델링에 사용되고 있다. 기본적으로 토픽 모델링은 유사성에 따라 객체들을 그룹화한다는 점에서 전통적인 군집화(Clustering)의 한 영역이라고 볼 수 있다. 하지만 토픽 모델링은 경성 군집화(Hard Clustering)가 아닌 연성 군집화(Soft Clustering)을 적용한다는 점, 그리고 결과 군집에 단순 식별자를 부여하는 대신 용어 집합을 통한 의미를 부여한다는 점에서 전통적인 군집화

와는 다른 특징을 갖는다. 이러한 토픽 모델링 기법을 활용하여 국가 R&D 과제와 과학기술정보를 융합하여 다차원 연계 지식 맵 서비스 구축을 시도한 연구(Jeong, 2015), 병사들의 SNS와 군 내부 데이터를 활용하여 군 내 병사 사고 예측방법론을 제안한 연구(Yoon, 2015) 등 국내에서도 여러 연구가 진행되고 있다.

텍스트 마이닝의 또 다른 대표적 응용인 문서 분류는 주어진 미문서가 포함하고 있는 용어를 분석하여 해당 문서를 특정 분류로 구분하는 과정을 나타내며, 기본적으로 이미 분류가 이루어진 다수의 문서에 대한 기계학습(Machine Learning)을 통해 분류기를 생성하고 이를 미분류 문서의 구분에 적용하는 형태로 수행된다. 이 분야는 Joachims(1998)가 SVM(Vapnik, 1995) 기법을 문서 분류에 적용하면서 많은 후속 연구가 이루어졌으며, 특히 최근 다양한 매체를 통해 실시간으로 생성되는 방대한 양의 텍스트 데이터를 체계적으로 관리할 필요성이 증가함에 따라 더욱 주목을 받고 있다. 문서 분류에 대한 연구는 주로 지도학습(Supervised Learning) 기반으로 이루어졌으며, SVM을 포함한 다양한 분류기의 특성 선택(Feature Selection) 성능을 비교한 연구(Rogati and Yang, 2002), 분류의 수가 많을 때 분류의 정확도가 낮아지는 부작용을 개선한 연구(Rubin et al., 2012), 단문의 주요 키워드를 질의어로 사용하여 단문을 분류한 연구(Sun, 2012) 등을 대표적인 예로 들 수 있다. 문서 분류에 대한 연구는 국내에서도 많은 연구자들에 의해 활발하게 이루어지고 있으며, 대표적인 예로 차원 축소를 통해 한글 문서 분류기의 성능을 개선한 연구(Li et al., 2010), 특허 문서에 대해 다양한 문서 분류 알고리즘의 성능을 비교한 연구(Kang et al., 2016), 단일 분류를 갖고 있는 문서의 기준을 확장하여 다중 분류를 배정하는 연구(Hong et al., 2014) 등 문서 분류의 성능을 개선하거나 그 활용성을 증대하고자 하는 노력이 계속되고 있다.

하지만 지도학습 기반의 전통적인 문서 분류기는 학습 과정에 상당히 많은 수의 기분류 문서(Labeled Documents)를 요구하기 때문에, 기분류 문서를 충분히 확보하고 있지 못한 매체의 경우 전통적인 분류기를 통해 문서 분류를 수행하는 것은 현실적으로 불가능하다는 한계를 갖는다. 이미 알려진 바와 같이 기분류 문서의 수가 현저히 적을 때에는 분류의 편중(Bias), 과적합(Overfitting), 과소적합(Underfitting) 등으로 인해 모형의 정확성 및 신뢰성이 낮아지는 부작용이 발생한다. 특히 문서 분류기를 서로 다른 매체에 반복적으로 상호 적용함으로써 각 문서가 다양한 매체에 대응될 수 있는 카테고리를 식별하는 것이 본 연구의 핵심임을 감안할 때, 기존 지도학습 기반의 전통적인 문서 분류기는 제안 방법론에 적용되기 어려운 측면이 있다. 최근에는 지도학습 방식의 한계를 극복하기 위해 고안된 준지도 학습(Semi-Supervised Learning)에 대한 관심이 높아지고 있다. 준지도 학습은 기분류 문서만을 학습 데이터로 사용하는 기존 방식과는 달리, 미분류 문서와 기분류 문서를 함께 학습 데이터로 활용함으로써 충분한 기분류 문서를 확보하지 못한 환경에서의 문서 분류에 적합한 기법으로 평가받고 있다. 이러한 준지도 학습 방식의 장점으로 인해 국내외에서 준지도 방식의 문서 분류에 대한 많은 연구(Ko and Seo, 2008; Kim and Lee, 2007; Liu et al., 2003; Lu et al., 2013)가 활발하게 수행되었으며, 준지도 문서 분류기의 정확도를 평가(Silva and Ribeiro, 2004)하고 향상시키기 위한 연구(Lee et al., 2015; Nigam

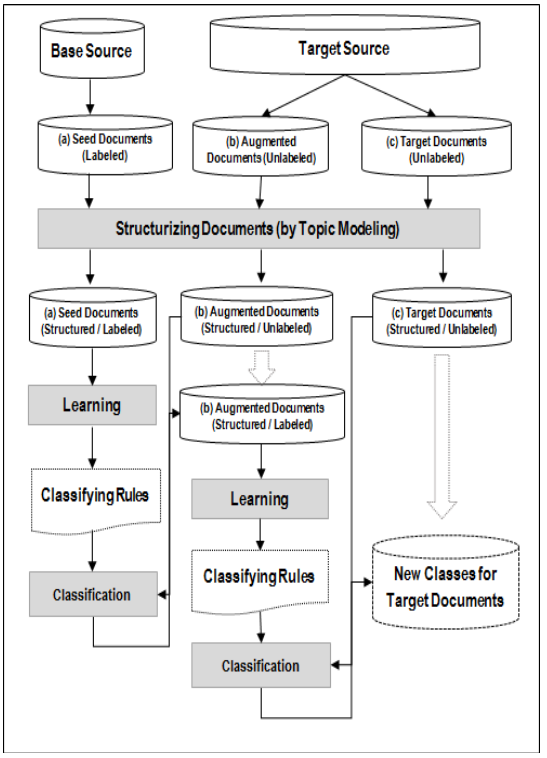
et al., 2006)도 다수 이루어지고 있다. 이러한 준지도 학습에 대한 기존의 연구 성과를 활용하여, 본 연구에서는 각 매체의 매우 소량의 기분류 문서만을 기준 문서(Seed Documents)로 사용하여 각 문서의 매체별 대응 카테고리를 2차원 레이블로 식별하여 관리할 수 있는 방안을 제시한다.

3. 제안 방법론

3.1 연구모형

본 장에서는 이종 매체에서 추출된 문서들에 대한 토픽 모델링을 실시하고 각 매체 별 토픽과 문서 간의 관계를 분석하여 문서 분류기를 생성한 뒤, 생성된 분류기를 활용하여 각 문서를 다양한 매체의 기준으로 레이블링하는 방안을 제시한다. 제안하는 방법론은 둘 이상의 다양한 매체에 대해 적용되며, 하나의 매체를 기준 소스(Base Source)로, 다른 매체들을 대상 소스(Target Source)로 간주한다. 이러한 과정은 각 매체에 대해 반복적으로 수행된다. 예를 들어 총 N개의 매체가 있는 경우 각 매체는 한 번은 기준 소스, (N-1)번은 대상 소스의 역할로 참여한다. 이렇듯 제안 방법론은 이론상 개수의 제한이 없는 복수개의 매체에 반복 적용 가능하지만, 본 부절에서는 설명의 편의를 위해 단 두 개의 매체만 존재하는 경우를 가정한다. 제안 방법론의 개요는 <Figure 2>에 제시되어 있다.

<Figure 2>에서 기준 소스는 이미 고유의 카테고리 체계를 갖고 있으며, 기준 소스로부터 추출한 문서들은 모두 기분류 문서(Labeled Documents)이다. 이들 문서의 집합을 시드 문서(Seed Documents)라고 하며, 분류 학습의 원천으로 사용된다. 한편 분류의 대상이 되는 대상 소스의 경우 기준 소스의 관점에서 새로운 카테고리를 부여받게 되므로, 고유의 카테고리 체계는 모두 무시된다. 따라서 대상 소스로부터 추출한 문서들은 모두 미분류 문서(Unlabeled Documents)로 간주된다. 또한 기준 소스의 규모가 매우 작아서 시드 문서의 수가 현저히 부족한 경우 학습을 위해 필요한 기분류 문서를 보강할 필요가 있으므로, 제안 방법론은 대상 소스로부터 미분류 문서 일부를 추출하여 학습에 활용하는 준지도 학습 방법을 채택한다. 그림에서 (b)는 기분류 문서 보강을 위한 문서, (c)는 최종적으로 분류될 문서를 나타낸다. 제안 방법론의 프로세스는 크게 문서의 구조화를 위한 토픽 모델링, 1차 학습 및 분



〈Figure 2〉 Research overview

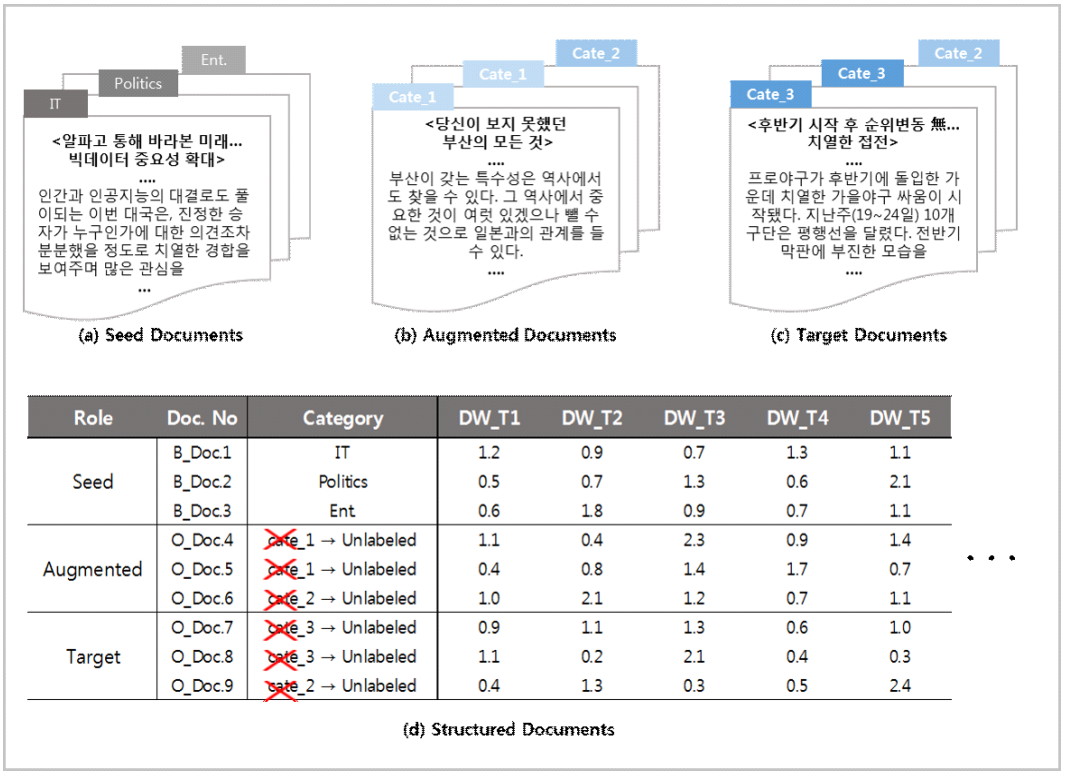
류, 그리고 2차 학습 및 분류로 구성되며, 각각 <Figure 2>의 상단, 하단 좌측, 그리고 하단 중앙에 나타나있다. 본 방법론의 단계별 설명은 3.2절과 3.3절에서 자세히 다루고, 전체 과정을 실제 데이터에 적용한 실험 결과는 4장에서 소개한다.

3.2 토픽 모델링을 통한 비정형 문서의 구조화

본 절에서는 토픽 모델링을 통해 문서와 토픽 간의 대응도를 산출하고, 이를 활용하여 각 문서를 구조화하는 과정을 소개한다. 우선 웹 상에 존재하는 이종 매체들로부터 문서를 수집한 후, 수집된 문서들을 모두 통합하여 토픽 모델링을 수행한다. 토픽 모델링은 이미 기존의 많은 연구들을 통해 충분히 설명되었으므로, 본 연구에서는 토픽 모델링의 과정에 대한 자세한 소개 대신 주요 개념만을 간략히 소개한다.

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. 이 때 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 값을 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 연구에서는 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. 토픽 모델링을 통한 문서 구조화의 예가 <Figure 3>에 나타나있다.

<Figure 3(a)>는 기준 소스로부터 도출된 시드 문서를 나타내며 "IT", "Politics", "Ent." 등의 카테고리로 구분되어 있다. 한편 <Figure 3(b)>와 <Figure 3(c)>의 경우 대상 소스로부터 도출된 문서로 "Cate1", "Cate2", "Cate3" 등의 카테고리로 구분되어 있다. 하지만 본 분석에서는 기준 소스의 카테고리만이 유효하게 작용하기 때문에, 대상 소스로부터 도출된 문서의 기존 카테고리는 모두 무시된다. 따라서 <Figure 3>에서 "B_Doc1" ~ "B_Doc3"은 기분류 문서로, "O_Doc4" ~ "O_Doc9"는 미분류 문서로 사용된다. <Figure 3>에서 (a), (b), 그리고 (c)의 문서는 모두 한꺼번에 토픽 모델링의 입력으로 사용되며, 그 분석 결과가 <Figure 3(d)>에 나타나있다. <Figure 3(d)>의 우측 부분은 각 토픽에 대한 각 문서의 문서 가중치를 나타내며, 향후 분석에서 입력 변수로 사용된다. 예를 들어 "B_Doc1" 문서의 경우 ("IT", 1.2, 0.9, 0.7, 1.3, 1.1, ...)의 벡터로 구조



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

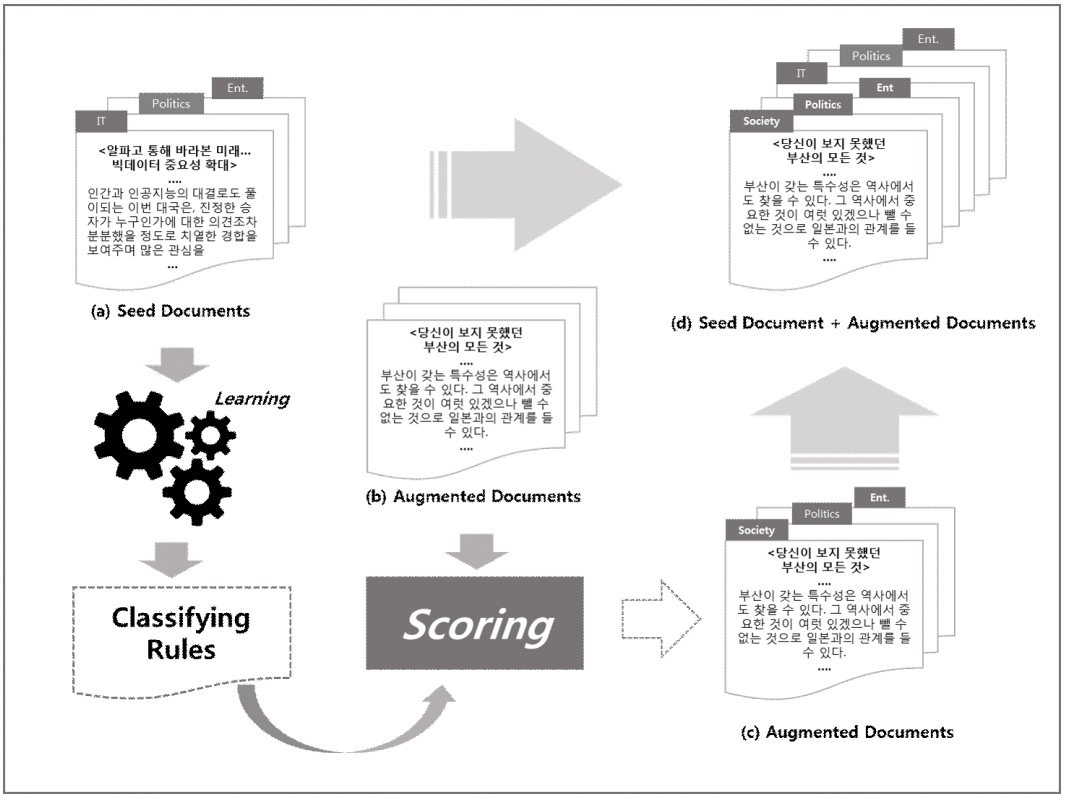
(d) Structured Documents

〈Figure 3〉 Example of structured documents using topic modeling

화되며, 가장 첫 요소는 목적 변수, 그리고 나머지 요소들은 입력 변수로 구분된다. 이렇게 구조화된 문서에 대한 분류 과정은 본 장의 이후 부분에서 소개한다.

3.3 1차 학습 및 분류

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 연구에서는 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 연구에서 핵심적으로 다루는 내용이 아니므로, 본 연구에서는 준지도 학습 과정에서 가장 직관적이고 단순한 방법을 사용한다. 즉, 소량의 기분류 문서를 학습 집합으로 활용하여 미분류 문서의 일부를 분류한 후, 이렇게 분류된 미분류 문서를 기존의 기분류 문서와 통합하여 새로운 학습 집합으로 사용한다. 준지도 학습 기반의 문



〈Figure 4〉 Learning stage of semi-supervised learning

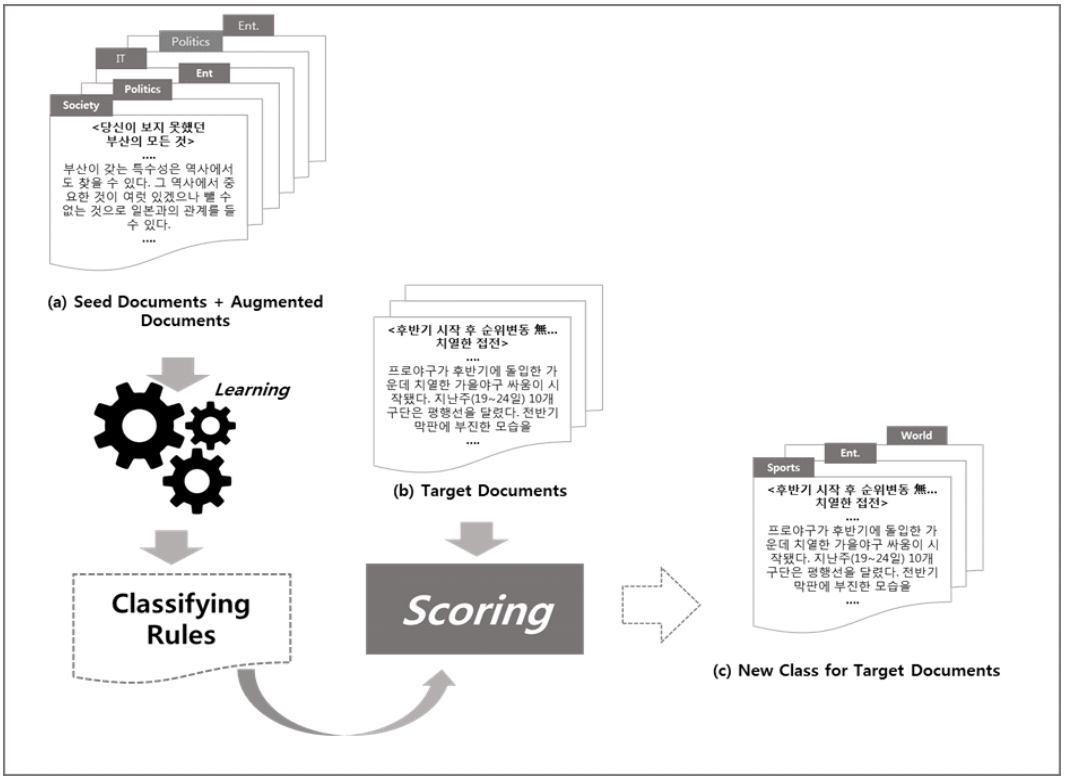
서 분류 중 1차 학습 및 분류 단계에 관한 설명이 <Figure 4>에 나타나있다.

<Figure 4(a)>는 소량의 기분류 문서를 나타내며, 이들 문서에 대한 학습을 통해 분류 알고리즘을 생성한다. 이렇게 생성된 분류 알고리즘을 통해 <Figure 4(b)>의 미분류 문서를 분류함으로써 <Figure 4(c)>와 같은 기분류 문서를 추가로 획득할 수 있으며, 이렇게 추가 분류된 문서가 기존의 기분류 문서와 함께 추후 학습에 활용된다. 이 과정에서 분류된 문서 전부가 아닌 일부, 즉 분류된 문서의 분류 확률(Probability or Score)이 특정 임계값 이상인 문서만을 추후 학습에 활용할 수도 있다.

3.4 2차 학습 및 분류

1차 학습 및 분류를 통해 새로 분류된 문서는 기존의 기분류 문서와 통합되어 2차 분류의 학습 데이터로 활용된다. 2차 분류는 앞에서 소개한 1차 분류와 매우 유사한 형태로 수행되며, 이 과정은 <Figure 5>에 나타나있다.

<Figure 5(a)>는 <Figure 4(d)>에 해당하는 문서 집합으로, 2차 분류의 학습 데이터로 사용된다. 이러한 과정을 통해 최종적으로 타겟 미분류



〈Figure 5〉 Scoring stage of semi-supervised learning

문서(<Figure 5(b)>)의 카테고리를 식별하게 되며, 그 결과가 <Figure 5(c)>에 나타나있다.

본 장에서는 편의를 위해 두 매체로부터 문서가 도출된 경우만을 예로 들어 설명하였지만, 제안하는 방법론은 유사한 과정의 반복 적용을 통해 둘 이상의 매체에 확장 적용될 수 있다. 이 경우 최종 결과물은 <Figure 6>과 같은 형태로 나타나게 되며, 각 문서는 원 소속 매체의 카테고리 뿐 아니라 서로 상이한 구조를 가진 다른 매체의 카테고리 정보 또한 동시에 갖게 된다. 예

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	
. . .	. . .		. . .			

〈Figure 6〉 Category mapping example

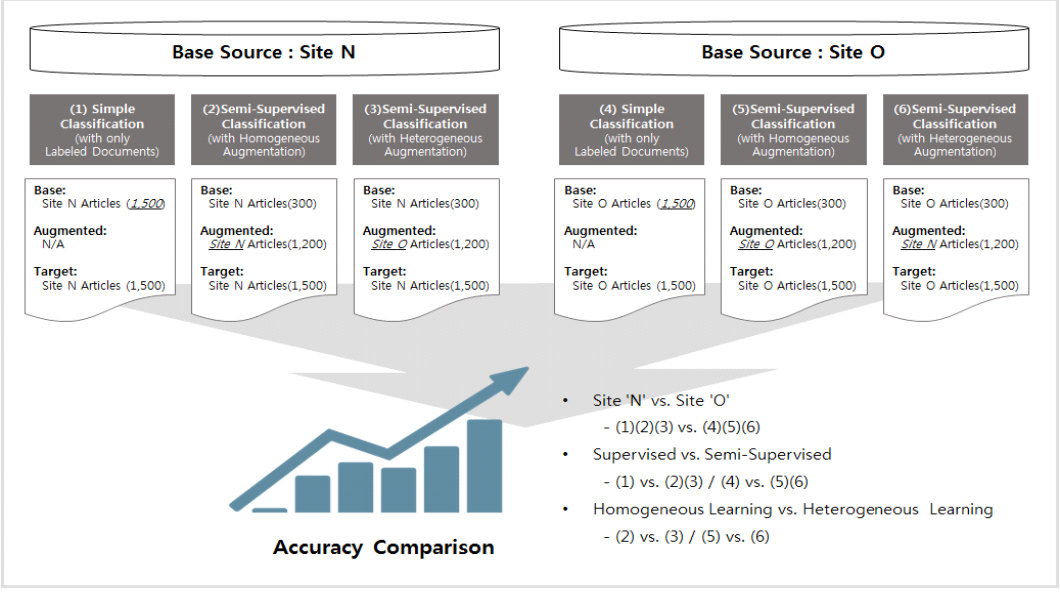
를 들어 <Figure 6>에서 1번 문서의 경우 원래 매체 “D News”의 카테고리 “IT”에 속한 문서이며, 제안 방법론의 적용을 통해 매체 “N Blog”의 카테고리 “Travel”과 매체 “A Discussion”의 카테고리 “Life”에도 추가로 연결되었음을 알 수 있다.

본 장에서는 제안 방법론의 개념 및 과정을 간략한 예시를 통해 설명하였다. 다음 장에서는 실제 데이터를 통해 본 방법론을 적용한 실험 과정 및 결과를 소개한다.

4. 실험

4.1 실험 개요

본 장에서는 실제 수집된 이종 매체 문서에 대하여 제안 방법론을 적용한 실험 과정 및 분석 결과를 소개한다. 실험 대상 매체로는 인터넷 뉴스 포털인 ‘N’ 사이트와 ‘O’ 사이트를 선정하였으며, 각 사이트로부터 뉴스 기사 3,000건씩 총 6,000건의 뉴스 기사 수집하였다. 기사 원문은 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 ‘IT과학’, ‘정치’, ‘경제’, ‘사회’, ‘생활문화’, ‘세계’, ‘스포츠’, 그리고 ‘연예’ 등 총 8개의 카테고리 분류 체계를 갖고 있었으며, 사이트 ‘O’는 ‘경제’, ‘교육’, ‘미디어’, ‘민족/국제’, ‘사회’, ‘정치’, 그리고 ‘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 연구에



〈Figure 7〉 Experiment Design

서 수행한 실험의 개요가 <Figure 7>에 제시되어 있다.

본 실험의 목적은 제안 방법론을 통해 실제 문서에 대해 다양한 매체의 기준에 따른 카테고리 식별 과정을 직접 수행함으로써 제안 방법론의 실무 적용 가능성을 파악함과 동시에, 제안 방법론에 따른 문서 분류의 정확성을 평가하는 것이다. 하지만 기준 문서와 대상 문서의 카테고리 체계 및 명칭이 상이하기 때문에, 제안 방법론에 대한 직접적인 정확도 평가는 불가능한 것으로 판단한다. 예를 들어 사이트 'O'의 '교육' 카테고리에 속하는 기사가 사이트 'N'의 '사회' 카테고리로 분류되었을 때, 이 분류에 대한 참/거짓을 판단할 수 없다. 따라서 제안 방법론의 정확성을 직접 평가하는 대신, 본 장에서는 여러 상황에 따른 제안 방법론의 정확도 차이를 비교하는 실험을 수행하고 그 결과를 소개한다.

우선 각 매체의 특성이 정확도에 미치는 영향을 파악하기 위해 전체 실험을 사이트 'N'의 문서를 분류하는 실험과 사이트 'O'의 문서를 분류하는 실험의 두 가지로 나누어 수행하였다. <Figure 7>에서 (1) ~ (3)은 사이트 'N'에 대한 실험을, (4) ~ (6)은 사이트 'O'에 대한 실험을 나타낸다. 다음으로 학습을 위한 기분류 문서의 수가 충분한 경우의 지도 학습과 기분류 문서의 수가 충분하지 않은 경우의 준지도 학습의 정확도 차이를 비교하기 위한 실험을 수행하였다. 사이트 'N'의 문서 분류 실험을 예로 들면, <Figure 7>에서 (1)의 경우 사이트 'N'의 기분류 문서 1,500개에 대한 학습을 수행하고 이를 통해 사이트 'N'의 미분류 문서 1,500개를 분류하는 실험을 의미한다. 한편 (2)와 (3)의 경우 사이트 'N'의 기분류 문서 300개가 주어져있을 때 사이트 'N'의 미분류 문서 1,500개를 분류하는 상황을 가정한 실험이다. 마지막으로 1차 분류 과정에서 학습 데이터의 이질성이 최종 분류의 정확도에 미치는 영향을 분석하였다. 이를 위해 <Figure 7>에서 (2)의 경우 기준 문서와 동일하게 사이트 'N'으로부터 추출한 미분류 문서 1,200개를 1차 분류에 사용하였으며, (3)의 경우 기준 문서는 다르게 사이트 'O'로부터 추출한 미분류 문서 1,200개를 1차 분류에 사용하였다. 사이트 'O'를 대상으로 한 실험 (4) ~ (6)의 경우 사이트 'N'을 대상으로 한 실험 (1) ~ (3)에 대응되므로, 사이트 'O'의 실험에 대한 자세한 설명은 생략한다.

다음 절에서는 3장에서 제시한 방법론에 따른 실험 과정 및 결과를 <Figure 7>의 실험 (2)를 기준으로 소개하며, 사이트 'N'과 사이트 'O', 지도 학습과 준지도 학습, 그리고 동질 학습 데이터 사용과 이질 학습 데이터 사용으로 인한 정확도의 차이에 대한 분석은 4.3절에서 다룬다.

4.2 단계별 실험 과정 및 중간 산출물

4.2.1 토픽 모델링을 통한 데이터 구조화

본 단계에서는 수집된 문서의 토픽 모델링 진행 과정 및 결과를 소개한다. <Figure 8>에 나타난 바와 같이 데이터 마이닝 상용 도구 중 하나인 SAS Enterprise Miner 12.1의 Text Miner 모듈을 사용하여 파싱, 필터링, 토픽 분석 순으로 문서 3,000건에 대해 토픽 모델링을 수행하였으며, 이 중 이메일, URL, 기타 무의미한 단어 등 불필요한 어휘를 제거하기 위해 총 68,822개의 어휘를 수록한 불용어 사전(Stop List)을 적용하였다. 본 연구에서는 문서에 포함된 여러 용어들 중 핵심 용어만을 문서의 구조화에 활용하기 위해 명사만을 활용하여 토픽 모델링을 수행하였으며, 구조화를 위한 차원의 수, 즉 전체 토픽의 수는

〈Figure 8〉 Flow Diagram for topic modeling

	_DOC	TextTopi	TextTopi	TextTopic	TextTopic	T_CATE	T_BODY
1	1	0.141	0.186	0.126	0.067	IT과학	최근 급증하고 있는 스마트폰에서 최고의 콘텐…
2	2	0.159	0.205	0.135	0.051	IT과학	킨들 보이지(Voyage)/아마존 세계 최대 온라인…
3	3	0.181	0.193	0.181	0.096	IT과학	나노기업 74% 제품생산… 매출은 턱없이 적어…
4	4	0.122	0.131	0.099	0.048	IT과학	유선 결합 상품에 가입하면 통신사 별로 각각 38…
5	5	0.191	0.233	0.217	0.197	IT과학	대형사고 발생때마다 사고 수습에 기존예산 유…
6	6	0.133	0.227	0.130	0.140	IT과학	의무화 폐지… 자율에 맡겨 멤점 PG에 결제정보…
7	7	0.148	0.197	0.087	0.028	IT과학	어느덧 바츠해방전쟁 10주년이 되었다. 바츠해…
8	8	0.176	0.224	0.122	0.150	IT과학	보조금 상한액 규모 & 주체별 분리 공시 여부 (…
9	9	0.127	0.143	0.175	0.182	IT과학	기업이 정보보호를 위해 법률로 강제한 '정보보…
10	10	0.177	0.209	0.173	0.001	IT과학	연간 인건비 7천억↓…내부 동요·경쟁력 회복 과…
11	11	0.245	0.152	0.193	0.109	IT과학	정책 · 연구개발 기능 통합 'ICT특별법' 시행……
12	12	0.200	0.278	0.054	0.084	IT과학	4월 첫째 주, 구글코리아가 구글 플레이 스토어…
13	13	0.140	0.225	0.139	0.151	IT과학	60년 만에 찾아온다는 청마의 해, 2014년이 밝…
14	14	0.240	0.237	0.260	0.218	IT과학	▲서울 남산 북측 순환로에 물든 단풍, 상상력의…
15	15	0.131	0.199	0.068	0.141	IT과학	네이버, 라인 성장 덕에 2분기 영업익 38.5% 증…

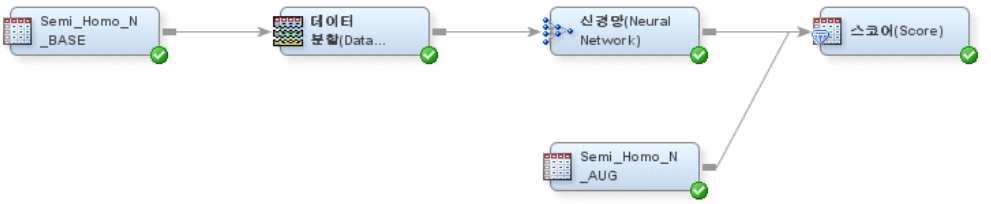
〈Figure 9〉 Flow diagram for first learning and scoring

50개로 한정하였다.

<Figure 8>의 과정을 통해 토픽 모델링을 수행한 결과물의 일부가 <Figure 9>에 소개되어 있다. 첫째 열은 문서의 고유 식별 번호를 나타내며, 이후 총 50개의 열을 통해 각 토픽에 대해 문서가 갖는 대응도인 문서 가중치가 나타난다. 마지막 두 열에는 문서가 원래 속해있던 카테고리의 정보와 문서의 본문이 나타나 있다. <Figure 9>의 테이블에서 문서 가중치 50개는 문서 분류 학습의 입력 변수로, 카테고리는 목적 변수로 활용되며, 문서의 본문은 구조화 이후의 단계에서는 더 이상 활용되지 않는다.

4.2.2 1차 학습 및 분류 실험

본 단계에서는 3.3절에서 소개한 1차 학습 및 분류 실험의 과정 및 결과를 소개한다. 이 과정 역시 앞에서 사용한 SAS Enterprise Miner 12.1을 사용하여 수행하였으며, 학습 및 분류 과정은 <Figure 10>에 나타난 프로세스에 따라 진행된다. 앞에서 언급한 바와 같이 본 절에서는 <Figure 7>의 6가지 실험 중 실험 (2), 즉 시드 문서와 추가 문서, 그리고 목적 문서로 모두 사이트 'N'의 기사를 사용한 실험을 예로 들어 소개한다. 구체적으로 <Figure 10>에서 학습용 데이터는 'N' 사의 인터넷 뉴스 300 건을 사용하였으며, 학습을 위한 분석용과 평가용 데이터의 비율을 7 : 3으로 설정하였다. 또한 예측 모형의 선정을 위해 의사결정나무, 회귀분석, 신경망의 세 가지 모형 모두를 실험에 사용하였으며, 실험 결과 신경망의 분류 정확도가 가장 높은 것으로 나타나 이후 모든 실험에서는 동일하게 신경망 모형만을 분류에 활용하였다. 마지막으로 신경망 분석을 통해 도출된 규칙을 'N'사의 문서 1,200건의 분류(Classification or Scoring)에 적용한 결



〈Figure 10〉 Flow diagram for first learning and scoring

	DOCNO	Topic_1	Topic_2	...	Topic_49	Topic_50	Category_1	...	Category_8	EM_PROBABILITY	EM_CLASSIFICATION
1	301	0,029	0,056		0,038	0,023	0,0000135719		0,0009321368	0,9058543551	연예
2	302	0,138	0,153		0,093	0,134	0,0001830807		0,2083916732	0,6391619778	사회
3	303	0,168	0,114		0,099	0,143	0,2767856809		0,3059881333	0,3059881333	IT과학
4	304	0,152	0,207		0,097	0,098	0,1434991194		0,4093612536	0,4093612536	IT과학
5	305	0,102	0,177		0,151	-0,037	0,603770243		0,004184547	0,603770243	정치
6	306	0,106	0,148		0,246	-0,015	0,0124014995		0,0217596816	0,6551055109	세계
7	307	0,131	0,176		0,080	0,131	0,5555104394		0,0850914935	0,5555104394	정치
8	308	0,105	0,283		0,141	-0,054	0,0002564824		0,2211718201	0,6213591938	사회
9	309	0,063	0,342		0,085	0,005	0,0000191518		0,0032671423	0,853103324	연예
10	310	0,116	0,261		0,095	0,067	0,1053965507		0,0097519105	0,8779737194	생활문화
11	311	0,117	0,180		0,101	0,082	0,0217590561		0,0014511526	0,9629948882	생활문화
12	312	0,232	0,205		0,071	0,076	0,0001018794		0,0176722236	0,88317812	경제
13	313	0,101	0,834		0,076	0,041	1,3221296E-6		0,0010690927	0,9527491226	스포츠
14	314	0,158	0,140		0,123	-0,023	0,0576909577		0,4559486859	0,4559486859	IT과학
15	315	0,200	0,265		0,123	-0,019	0,0000727068		0,0113751922	0,9020912078	경제

〈Figure 11〉 Result of first stage (part)

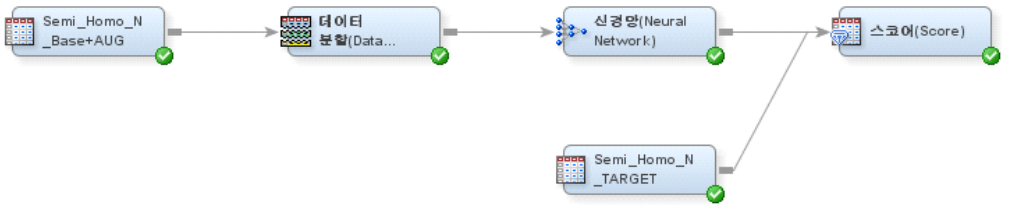
과가 <Figure 11>에 나타나있다.

<Figure 11>의 맨 좌측 열부터 “Topic_50”까지의 열은 <Figure 9>에서 소개한 바와 동일하므로 설명을 생략한다. 이후 “Category_1” ~ “Category_8”까지의 8개 열은 각 문서가 해당 카테고리로 분류될 확률을 나타내며, 8개 값 중 가장 큰 값이 “EM_Probability” 열에, 그리고 이 값에 대응되는 카테고리의 명칭이 “EM_Classification” 열에 나타나있다. 이렇게 카테고리가 식별된 미분류 문서는 시드 문서로 사용된 기분류 문서와 통합되며, 2차 학습 및 분류 과정의 학습 데이터로 활용된다.

한편 1차 분류에 사용된 미분류 문서 1,200건의 경우 시드 문서와 마찬가지로 사이트 ‘N’으로부터 추출되었으며 카테고리 정보를 갖고 있으므로, 1차 분류를 통해 식별된 카테고리와 원 소속 카테고리와의 비교를 통해 분류 정확도를 측정할 수 있다. <Figure 10>을 통해 나타난 실험 (2)의 경우 분류 정확도가 65.08%로 나타났으며, 사이트 ‘O’에 대해 수행한 실험 (4)의 경우 분류 정확도가 63.5%로 나타났다. 전술한 바와 같이 준지도 학습을 위한 최근 알고리즘의 적용을 통해 1차 분류에서의 오분류로 인한 성능 저하를 줄일 수 있지만, 준지도 학습의 성능 개선은 본 연구에서 핵심적으로 다루는 내용이 아니므로, 본 실험에서는 <Figure 11>의 결과를 이후 분석에 그대로 사용하였다.

4.2.3 2차 학습 및 분류 실험

본 단계는 실험의 마지막 단계로서 1차 학습 및 분류 실험을 통해 보강된 학습 데이터를 활용하여 최종적인 목적 문서에 카테고리를 부여하는 과정이다. 이 과정 역시 SAS Enterprise Miner



〈Figure 12〉 Flow diagram for second learning and scoring

	DOCNO	Topic_1	Topic_2	...	Topic_49	Topic_50	Category_1	...	Category_8	EM_PROBABILITY	EM_CLASSIFICATION
1	1501	0,175	0,194		0,103	0,018	0,0160873544		0,020884551	0,869575279	생활문화
2	1502	0,039	0,068		0,017	-0,001	0,0021610781		0,0002531598	0,9788089593	연예
3	1503	0,095	0,138		0,067	0,008	0,805327476		0,0008589457	0,805327476	정치
4	1504	0,218	0,187		0,143	0,156	0,0095828915		0,0184625394	0,8941868896	생활문화
5	1505	0,116	0,146		0,084	0,082	0,8641528324		0,0002722905	0,8641528324	정치
6	1506	0,163	0,253		0,152	0,095	0,0101175551		0,0169170733	0,9018849866	생활문화
7	1507	0,094	0,687		0,160	0,006	0,0000264142		0,0003556364	0,9880195718	스포츠
8	1508	0,069	0,120		0,167	-0,162	0,5253995071		0,0034990385	0,5253995071	정치
9	1509	0,066	0,086		0,051	0,055	0,4913030755		0,0031320141	0,4913030755	정치
10	1510	0,147	0,164		0,066	0,091	0,0006373994		0,0371860978	0,5082754645	경제
11	1511	0,139	0,161		0,140	-0,188	0,0692148781		0,0244323989	0,6505476112	생활문화
12	1512	0,020	0,060		0,005	-0,034	0,0021593677		0,000252162	0,9788966744	연예
13	1513	0,140	0,209		0,117	0,073	0,0067339517		0,0454540292	0,6732881686	생활문화
14	1514	0,157	0,138		0,108	0,107	0,0005014816		0,0339482624	0,5377437698	경제
15	1515	0,157	0,154		0,087	0,134	0,1275468334		0,0210114265	0,5359096286	사회

〈Figure 13〉 Result of second stage (part)

12.1을 활용하였으며, 실험은 1차 학습 및 분류 실험과 매우 유사한 형태로 진행된다. <Figure 12>는 2차 학습 및 분류 과정을 나타내며, 최종 목적 문서로는 사이트 'N'의 문서 1,500건이 사용되었다. 또한 보강된 학습 데이터 1,500건을 통해 도출한 규칙을 목적 데이터의 1,500건의 분류에 적용한 결과가 <Figure 13>에 나타나있다.

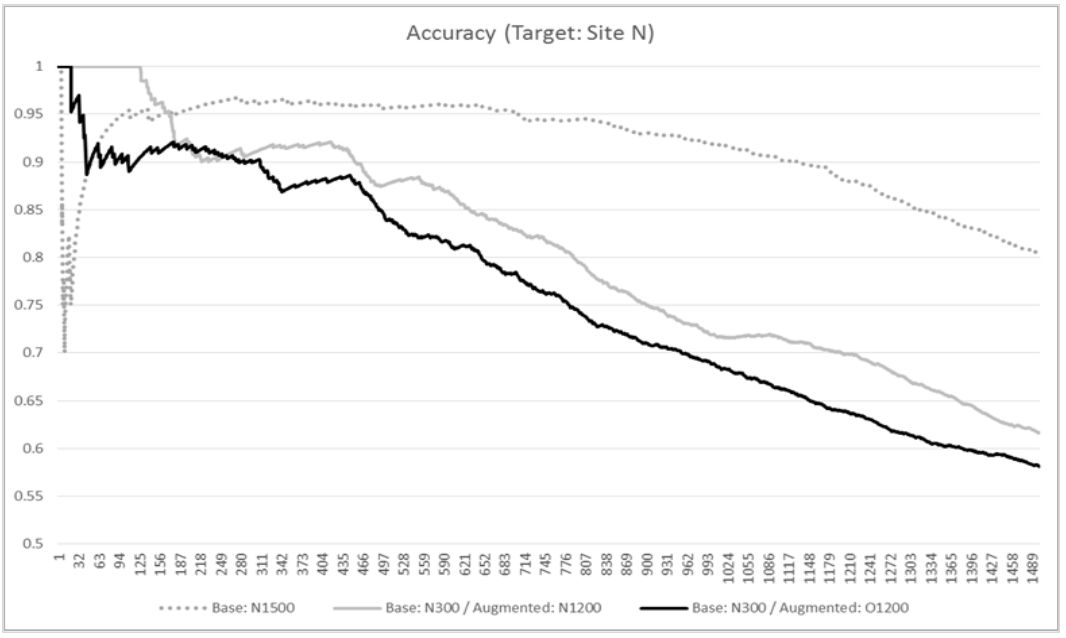
4.3 실험 결과 분석

본 절에서는 제안 방법론의 성능을 간접적으로 파악하기 위해, 여러 상황에 따른 제안 방법론의 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 <Table 1>과 같다.

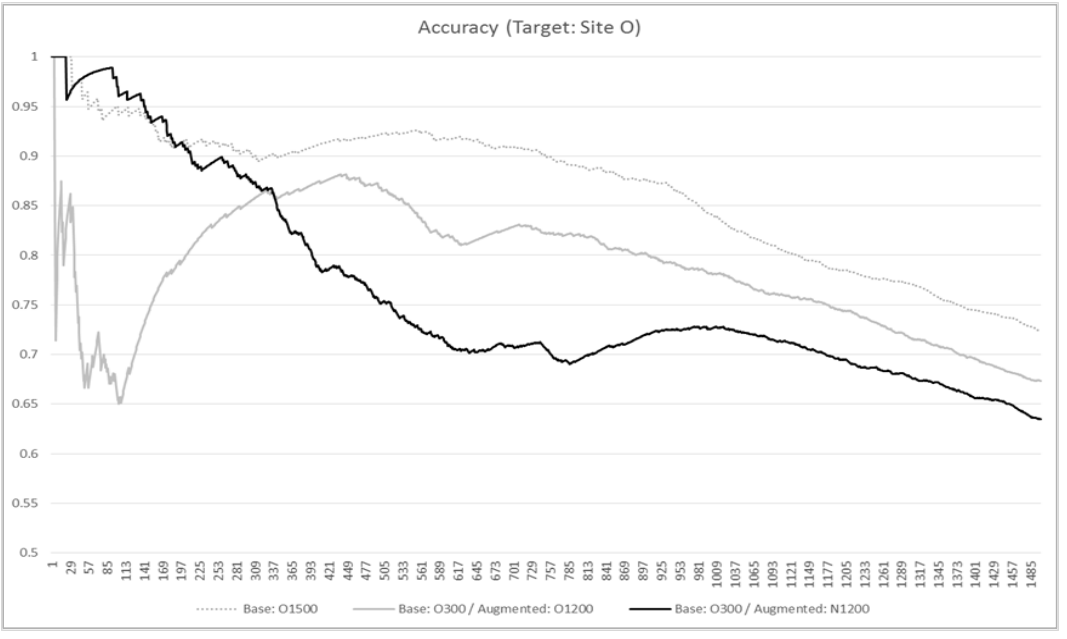
<Table 1>의 결과에 따르면 지도 학습의 경우 사이트 'N'이 사이트 'O'에 비해 분류 정확도가 높게 나타났으며, 준지도 학습의 경우 반대로 사이트 'O'가 사이트 'N'에 비해 분류 정확도가 높게 나타났다. 한편 두 사이트 모두에 대해, 학습

〈Table 1〉 Accuracy Comparison

(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



〈Figure 14〉 Accuracy of experiment for site 'N'

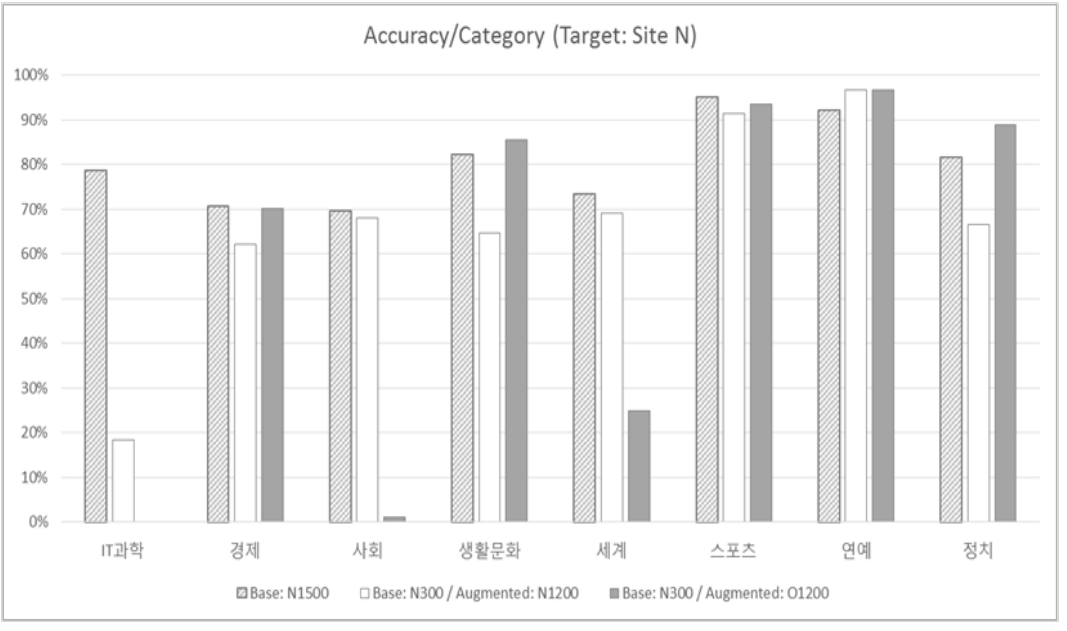


〈Figure 15〉 Accuracy of experiment for site 'O'

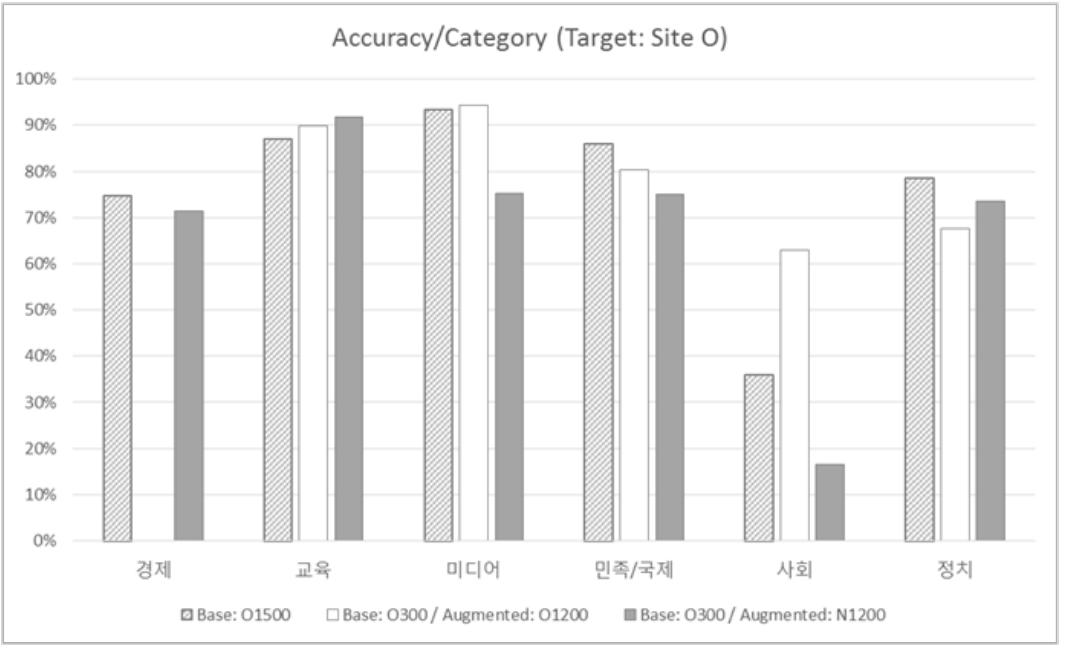
데이터가 충분한 경우 수행 가능한 지도 학습이 준지도 학습에 비해 분류 정확도가 높게 나타났다. 마지막으로 학습 데이터의 이질성 비교 실험의 경우, 동일 소스로부터 학습 데이터를 보강한 경우가 이질 소스로부터 학습 데이터를 보강한 경우에 비해 분류 정확도가 높게 나타났다. 각 실험에 대한 보다 자세한 분석은 다음과 같다.

우선 <Figure 14>는 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. 본 그래프는 각 실험에서 나타난 각 문서의 분류 확률의 내림차순으로 문서를 정렬한 뒤, 정렬 순서에 따른 각 문서의 분류 정확도를 누적으로 측정한 것이다. 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. 이러한 현상은 <Figure 15>의 사이트 'O'에 대한 문서 분류 실험에서도 동일하게 나타난다.

위의 <Figure 14>와 <Figure 15>를 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. 본 연구에서는 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 <Figure 16>에, 사이트 'O'에



〈Figure 16〉 Accuracy of experiment by category (Site 'N')



〈Figure 17〉 Accuracy of experiment by category (Site 'O')

대한 실험 결과는 <Figure 17>에 제시되어 있으며, 그림에서 빗금으로 나타난 막대(Bar)는 지도 학습, 흰색으로 나타난 막대는 동질 준지도 학습, 그리고 회색으로 나타난 막대는 이질 준지도 학습의 분류 정확도를 나타낸다.

매우 흥미롭게도, '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 <Figure 17>의 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.

요약하면, 본 장에서 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.

5. 결론

다양한 매체를 통해 유통되는 문서들은 서로 유사한 주제, 심지어는 동일한 내용을 다루더라도, 각 매체 별 정책 및 기준에 따라 각기 다른 카테고리로 관리되고 있으며, 이는 이종 매체를 아우르는 범위에서 특정 카테고리에 대한 탐색을 수행하고자 하는 시도에 걸림돌로 작용하고 있다. 이러한 제약을 극복하기 위해, 본 연구에서는 기존 매체 고유의 카테고리 체계는 그대로 유지하면서 이종 매체간 카테고리 매핑을 수행하는 방안을 제시하였다. 구체적으로 개별 문서를 다양한 매체의 관점에서 재분류하고 이러한 분류 결과를 문서의 추가 속성으로 저장함으로써, 이종 매체에 속한 다양한 문서들을 마치 한 매체에 속한 것과 같이 동일한 카테고리 기준으로 탐색할 수 있는 논리적 장치를 제안하였다. 또한 제안 방법론의 성능 평가를 위해 실제 인터넷 뉴스 포털 사이트 두 곳으로부터 뉴스 기사 6,000건을 수집하여 매체 간, 지도 학습과 준지도 학습 간, 동질 학습 데이터와 이질 학습 데이터 간의 정확도 비교 실험을 수행하였다. 특히 매우 흥미롭게도, 일부 카테고리에서 이질 학습 데이터를 사용한 준지도 학습의 분류 정확도가 지도 학습 및 동질 학습 데이터를 사용한 준지도 학습의 분류 정확도보다도 높게 나타나는 현상을 발견하였다.

제안 방법론은 다음의 측면에서 학술적, 실무적 차원의 기여를 갖는 것으로 판단한다. 우선 학술적 측면에서 제안 방법론은 상이한 분류 체계를 갖는 이질적인 매체에 대해, 기존의 물리적 분류 체계를 그대로 유지하면서도 전체를 통합 관리할 수 있는 논리적인 체계 구축 방안을 제안했다는 점에서 의의를 갖는다. 또한 준지도 학습 과정에서 기준 문서와 동일한 소스의 문서뿐 아니라, 별도의 소스로부터 문서를 추출하여 이를 학습 데이터로 보강하는 시도를 했다는 점은 매우 새로운 시도로 인정받을 수 있다. 특히 이질적인 학습 데이터의 사용으로 인한 분류 정확도의 차이가 각 카테고리 별로 매우 상이하게 나타나는 현상을 보임으로써, 카테고리 별 특징 분석을 통해 제안 방법론의 성능을 향상시키기 위한 후속 연구가 활발하게 수행될 수 있을 것으로 기대한다.

이와 같은 학술적 기여 외에 본 연구의 기여는 실무적 측면에서 더욱 크게 나타날 것으로 기대한다. 최근 인터넷을 통한 탐색의 범위는 하나의 매체에만 국한되지 않으며, 다양한 매체의 문서에 대한 탐색, 수집 및 분석에 대한 수요가 증가하고 있는 추세이다. 하지만 각 매체마다 서로 상이한 카테고리 구조 및 명칭을 갖기 때문에, 이종 매체를 아우르는 범위에서 특정 카테고리에 대한 탐색이 이루어지기란 매우 어렵다. 이러한 상황에서 제안 방법론은 기존 사이트의 특성 및 구조를 그대로 유지하면서, 사용자가 원하는 사이트를 선택하여 해당 사이트의 카테고리 분류를 기준으로 전체 문서를 조회할 수 있는 방안을 제시하였다는 점에서 의의가 있다. 더욱이 제안 방법론을 통해 사용자가 자신만의 카테고리 구조를 설계하고, 여러 매체에 수록된 다양한 문서를 자신이 설계한 카테고리 구조에 맞게 관리할 수도 있다는 점에서 본 연구의 실무 활용도는 더욱 높을 것으로 기대한다.

이와 같은 기여에도 불구하고, 본 연구에서 제안하는 방법론은 향후 다음과 같은 측면에서 보완될 필요가 있다. 우선 본 연구에서는 제안 방법론의 성능에 대한 간접적인 비교 평가만이 이루어졌지만, 향후 연구에서는 제안 방법론의 정

확도에 대한 보다 직접적인 검증이 이루어질 필요가 있다. 즉 대상 소스의 문서를 기준 소스의 카테고리 체계에 따라 재분류한 뒤, 이러한 분류가 정확하게 이루어졌는지 여부를 실제 사용자 평가를 통해 확인할 필요가 있다. 또한 방법론의 정교화를 통해 분류 정확도를 향상시킬 필요가 있다. 예를 들어 준지도 학습을 활용한 1차 분류 과정에서도 기존의 연구 성과를 활용하여 보다 양질의 학습 데이터를 보강할 필요가 있으며, 1차와 2차 분류과정 모두에서 더욱 다양한 분류 모형을 적용하여 정확도를 비교할 필요가 있다. 마지막으로 본 실험의 마지막 결과로 제시한 카테고리 별 분류 정확도 비교에 대한 깊은 고찰이 이루어져야 한다. 구체적으로 이질 준지도 학습의 분류 정확도가 오히려 지도 학습의 분류 정확도에 비해 더욱 높게 나타난 일부 카테고리의 특성을 파악함으로써, 다양한 매체로부터 이질적인 문서를 획득하고 이를 활용하여 문서 분류의 정확도를 향상시킬 수 있는 방안에 대한 모색이 이루어져야 한다.

참고문헌(References)

Blei, D. M., Ng, A. Y., and Jordan, M. I., "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3(2003), 993~1022.

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas and R. A. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, Vol. 41, No. 6(1990), 391~407.

Hearst, M. A., "Untangling Text Data Mining," *Proceedings of the 37th ACL*, 1999.

Hofmann, T., "Probabilistic Latent Semantic Indexing," *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 1999, 50~57.

Hong, J. S., N. Kim, and S. Lee. "A Methodology for Automatic Multi-Categorization of Single-Categorized Documents," *Journal of Intelligence and Information Systems,* Vol. 20, No. 3(2014), 77~92.

Jeong, H., "A Study on Ontology and Topic Modeling-based Multi-dimensional Knowledge Map Services," *Journal of Intelligence and Information Systems,* Vol. 21, No. 4(2015), 79~92.

Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Proceedings of the 10th European Conference on Machine Learning,* 1998, 137~142.

Kang, J. H., J. C. Kim, J. H. Lee, S. S. Park and D. S. Jang, "A Comparative Study on Patent Document Classification Algorithms," *Proceedings of KIIS Spring Conference*, Vol. 26, No. 1(2016), 9~10.

Kim, P. J. and J. Y. Lee, "Utilizing Unlabeled Documents in Automatic Classification with Inter-document Similarities," *Journal of the Korean Society for Information Management*, Vol. 24, No. 1(2007), 251~271.

Ko, Y. and J. Seo, "Automatic Text Categorization based on Semi-Supervised Learning," *Journal of KIISE: Software and Applications*, Vol. 35, No. 5(2008), 325~334.

Korea Internet Security Agency, *2014 Korea Internet White Paper,* Korea Internet Security Agency, 2014.

Korea Research Institute for Vocational Education & Training , *THE HRD*, Vol. 16, No. 6(2013), 136~151.

Lee, S., J. Kim and S. H. Myaeng, "An Extension of Topic Models for Text Classification: A Term Weighting Approach", *Proceedings of the 2015 International Conference on Big Data and Smart Computing(BigComp)*, 2015, 217~224.

Li, C., D. R. Byun, and S. C. Park "BPNN Algorithm with SVD Technique for Korean Document Categorization", *Journal of the Korea Industrial Information System Society*, Vol. 15, No. 2(2010), 49~57.

Liu, B., Y. Dai, X. Li, W. S. Lee and P. S. Yu, "Building Text Classifiers Using Positive and Unlabeled Examples", *Proceedings of the 3rd IEEE International Conference on Data Mining*, 2003, 179~188.

Lu, Y., S. Okada and K. Nitta, "Semi-supervised Latent Dirichlet Allocation for Multi-label Text Classification", *Proceedings of 26th IEA/AIE, 2013,* 351~360.

McKinsey Global Institute, Big Data : The next Frontier for Innovation, Competition, and Productivity, McKinsey and Company, 2011.

Nigam, K., A. K. McCallum, S. Thrun and T. Mitchell, "Learning to Classify Text from Labeled and Unlabeled Documents", *Proceedings of 15th national conference on artificial intelligence*, 1998, 792~799.

Nigam, K., A. K. McCallum, S. Thrun and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents Using EM", *Machine Learning*, Vol. 39, No. 2(2000), 103~134.

Nigam, K., A. McCallum, and T. Mitchell, "Semi-Supervised Text Classification Using EM", *Supervised Learning*, MIT Press, 2006.

Rogati, M. and Y. Yang, "High-Performing Feature Selection for Text Classification", *Proceedings of the International Conference on Information and Knowledge Management*, 2002, 659~661.

Rubin, T. N., A. Chambers, P. Smyth and M. Steyvers, "Statistical Topic Models for Multi-label Document Classification", *Machine learning*, Vol. 88, No. 1(2012), 157~208.

Salton, G. and M. J. McGill, *Introduction to Modern Information Retrieval,* McGraw-Hill, 1986.

Salton, G., A. Wong and C. S. Yang, "A Vector Space Model for Automatic Indexing", *Communications of the ACM,* Vol. 18, No. 11(1975), 613~620.

Silva, C. and B. Ribeiro, "Labeled and Unlabeled Data in Text Categorization", *Proceedings of the IEEE International Joint Conference on Neural Networks*, 2004, 2971~2976.

Sun, A., "Short Text Classification Using Very Few Words", *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012, 1145~1146.

Vapnik, V. N., *The Nature of Statistical Learning Theory*, Springer, 1995.

Yoon, S, S. Kim, and K. Shin, "Development of the Accident Prediction Model for Enlisted Men through an Integrated Approach to Datamining and Textmining," *Journal of Intelligence and Information Systems,* Vol. 21, No.3(2015), 1~17.

Abstract

Mapping Categories of Heterogeneous Sources Using Text Analytics

Dasom Kim* · Namgyu Kim**

In recent years, the proliferation of diverse social networking services has led users to use many mediums simultaneously depending on their individual purpose and taste. Besides, while collecting information about particular themes, they usually employ various mediums such as social networking services, Internet news, and blogs. However, in terms of management, each document circulated through diverse mediums is placed in different categories on the basis of each source's policy and standards, hindering any attempt to conduct research on a specific category across different kinds of sources. For example, documents containing content on "Application for a foreign travel" can be classified into "Information Technology," "Travel," or "Life and Culture" according to the peculiar standard of each source. Likewise, with different viewpoints of definition and levels of specification for each source, similar categories can be named and structured differently in accordance with each source.

To overcome these limitations, this study proposes a plan for conducting category mapping between different sources with various mediums while maintaining the existing category system of the medium as it is. Specifically, by re-classifying individual documents from the viewpoint of diverse sources and storing the result of such a classification as extra attributes, this study proposes a logical layer by which users can search for a specific document from multiple heterogeneous sources with different category names as if they belong to the same source. Besides, by collecting 6,000 articles of news from two Internet news portals, experiments were conducted to compare accuracy among sources, supervised learning and semi-supervised learning, and homogeneous and heterogeneous learning data. It is particularly interesting that in some categories, classifying accuracy of semi-supervised learning using heterogeneous learning data proved to be higher than that of supervised learning and semi-supervised learning, which used homogeneous learning data.

* Graduate School of Business IT, Kookmin University
** Corresponding Author: Namgyu Kim
School of Management Information Systems, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Korea
Tel: +82-2-910-5425, Fax: +82-2-910-4017, E-mail: ngkim@kookmin.ac.kr

This study has the following significances. First, it proposes a logical plan for establishing a system to integrate and manage all the heterogeneous mediums in different classifying systems while maintaining the existing physical classifying system as it is. This study's results particularly exhibit very different classifying accuracies in accordance with the heterogeneity of learning data; this is expected to spur further studies for enhancing the performance of the proposed methodology through the analysis of characteristics by category. In addition, with an increasing demand for search, collection, and analysis of documents from diverse mediums, the scope of the Internet search is not restricted to one medium. However, since each medium has a different categorical structure and name, it is actually very difficult to search for a specific category insofar as encompassing heterogeneous mediums. The proposed methodology is also significant for presenting a plan that enquires into all the documents regarding the standards of the relevant sites' categorical classification when the users select the desired site, while maintaining the existing site's characteristics and structure as it is.

This study's proposed methodology needs to be further complemented in the following aspects. First, though only an indirect comparison and evaluation was made on the performance of this proposed methodology, future studies would need to conduct more direct tests on its accuracy. That is, after re-classifying documents of the object source on the basis of the categorical system of the existing source, the extent to which the classification was accurate needs to be verified through evaluation by actual users. In addition, the accuracy in classification needs to be increased by making the methodology more sophisticated. Furthermore, an understanding is required that the characteristics of some categories that showed a rather higher classifying accuracy of heterogeneous semi-supervised learning than that of supervised learning might assist in obtaining heterogeneous documents from diverse mediums and seeking plans that enhance the accuracy of document classification through its usage.

Key Words : Category Mapping, Document Classification, Text Mining, Topic Modeling

Received : August 17, 2016 Revised : December 10, 2016 Accepted : December 28, 2016
Publication Type : Regular Paper Corresponding Author : Namgyu Kim

저 자 소 개

김 다 솜
현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학 중이며, 평생교육진흥원 경영학 학사 학위를 취득하였다. 주요 관심분야는 데이터 마이닝 및 텍스트 마이닝이다.

김 남 규
현재 국민대학교 경영정보학부에서 부교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국지능정보시스템학회 부회장, 한국정보기술응용학회 부회장, 한국경영정보학회 이사, 한국CRM학회 이사, 한국IT서비스학회지 편집위원, JITAM 편집위원, 한국인터넷정보학회논문지 편집위원을 역임하였으며, 한국생산성본부 TOPCIT 개발사업 자문위원으로 활동 중이다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 데이터베이스 설계 등이다.