대한경영정보학회
「경영과 정보연구」 제37권 제4호 2018년 12월
http://dx.doi.org/10.29214/damis.2018.37.4.003

용어 사전의 특성이 문서 분류 정확도에 미치는 영향 연구†

정해강* · 김남규**

〈요 약〉

다양한 소셜 미디어 활동과 인터넷 뉴스 기사, 블로그 등을 통해 유통되는 비정형 데이터의 양이 급증함에 따라 비정형 데이터를 분석하고 활용하기 위한 연구가 활발히 진행되고 있다. 텍스트 분석은 주로 특정 도메인 또는 특정 주제에 대해 수행되므로, 도메인별 용어 사전의 구축과 적용에 대한 중요성이 더욱 강조되고 있다. 용어 사전의 품질은 비정형 데이터 분석 결과의 품질에 직접적인 영향을 미치게 되며, 분석 과정에서 정제의 역할을 수행함으로써 분석의 관점을 정의한다는 측면에서 그 중요성이 더욱 강조된다. 이렇듯 용어 사전의 중요성은 기존의 많은 연구에서도 강조되어 왔으나, 용어 사전이 분석 결과의 품질에 어떤 방식으로 어떤 영향을 미치는지에 대한 엄밀한 분석은 충분히 이루어지지 않았다. 따라서 본 연구에서는 전체 문서에서의 용어 빈도수에 기반을 두어 사전을 구축하는 일괄 구축 방식, 카테고리별 주요 용어를 추출하여 통합하는 용어 통합 방식, 그리고 카테고리별 주요 특질(Feature)을 추출하여 통합하는 특질 통합 방식의 세 가지 방식으로 사전을 구축하고 각 사전의 품질을 비교한다. 품질을 간접적으로 평가하기 위해 각 사전을 적용한 문서 분류의 정확도를 비교하고, 각 사전에 고유율의 개념을 도입하여 정확도의 차이가 나타나는 원인을 심층 분석한다. 본 연구의 실험에서는 5개 카테고리의 뉴스 기사 총 39,800건을 분석하였다. 실험 결과를 심층 분석한 결과 문서 분류의 정확도가 높게 나타나는 사전의 고유율이 높게 나타남을 확인하였으며, 이를 통해 사전의 고유율을 높임으로써 분류의 정확도를 더욱 향상시킬 수 있는 가능성을 발견하였다.

핵심주제어: 텍스트 분석, 용어 사전, 문서 분류, 토픽 모델링, 고유율

논문접수일: 2018년 09월 04일 수정일: 2018년 10월 22일 게재확정일: 2018년 11월 02일
† 이 논문은 2017년 대한민국교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A5A2A03067632).
* 국민대학교 비즈니스IT전문대학원 석사과정, k-kang@kookmin.ac.kr
** 국민대학교 경영정보학부 부교수, ngkim@kookmin.ac.kr

Ⅰ. 서 론

최근 스마트 기기의 대중화로 페이스북(Facebook), 트위터(Twitter), 인스타그램(Instagram) 등과 같은 소셜 네트워크 서비스(SNS, Social Network Service)를 이용한 실시간 정보 생산과 의견 공유가 활발해지고 있다. 이렇듯 다양한 소셜 미디어 활동과 인터넷 뉴스 기사, 블로그 등을 통해 유통되는 비정형 데이터의 양이 급증함에 따라, 각 산업 분야에서 비정형 데이터를 분석하고 활용하기 위한 관심과 연구가 활발히 진행되고 있다. 특히 텍스트 마이닝(Text Mining)은 다량의 텍스트 문서 또는 문장에 대한 분석을 통해 의미 있는 정보를 추출하는 과정으로(Hearst, 1999), 개별 데이터의 기밀성(Confidentiality)이 정형 데이터에 비해 상대적으로 낮을 뿐 아니라 크롤링(Crawling)을 통해 상대적으로 용이하게 대량의 데이터를 수집할 수 있다는 특징으로 인해 다양한 텍스트 데이터에 대한 다양한 분석이 여러 도메인에서 이루어지고 있다.

텍스트 분석은 어휘의 출현 여부 및 출현 빈도에 기반을 두어 텍스트를 수치로 표현하는 구조화 단계를 거치게 된다. 다만 텍스트에 포함된 방대한 용어를 모두 분석에 사용하는 대신, 분석 주제에 집중하고 분석 결과의 품질을 향상시키기 위해 용어 사전(Start List) 또는 불용어 사전(Stop List)을 사용하는 것이 일반적이다. 용어 사전은 분석에 사용될 어휘를 정의한 목록으로 특정 분야의 문서 추출 및 분석에 주로 사용된다. 한편 불용어 사전은 분석에서 배제되는 용어의 목록으로, 주로 문서의 내용 파악이나 주제 식별에 영향을 주지는 않지만 자주 출현하는 용어들로 구성된다.

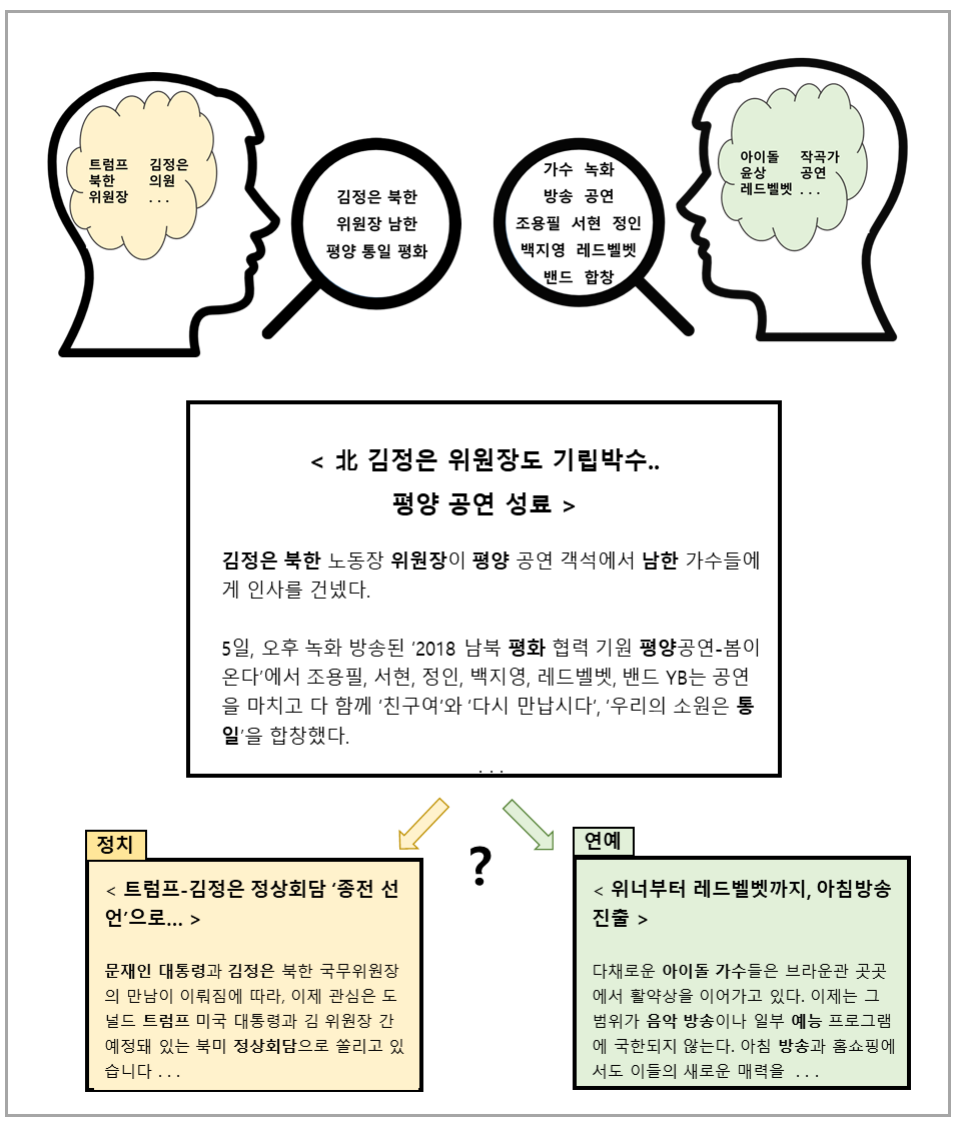
일반적으로 토픽 모델링과 같이 최종 분석 결과가 용어의 집합으로 표현되어 상대적으로 높은 수준의 정제가 필요한 경우는 용어 사전이 사용되며, 버즈(Buzz) 분석과 같이 정제 수준은 낮더라도 다양한 용어를 포함한 결과를 도출하고자 하는 경우는 불용어 사전을 사용한 분석이 수행된다. 이렇듯 텍스트를 대상으로 하는 분석은 사전을 활용하게 되며, 분석 목적에 따라 용어 사전과 불용어 사전이 적용된다. 최근 많은 분석이 특정 도메인 또는 특정 주제에 대해 수행되므로, 불용어 사전보다는 도메인별 용어 사전의 구축, 그리고 이러한 용어 사전의 적용에 대한 중요성이 점차 강조되고 있다.

용어 사전의 품질은 비정형 데이터 분석 결과의 품질에 직접적인 영향을 미치게 된다. 예를 들어 용어 사전이 충분히 많은 어휘를 포함하지 않고 어휘가 과도하게 정제된 경우, 분석 주제와 관련된 용어, 문장, 문서가 분석에서 누락되어 결과를 왜곡하게 된다. 반대로 정제 수준이 지나치게 낮은 경우는 분석 결과에 다듬어지지 않은 어휘가 그대로 포함되어 나타나거나 분석 주제와 무관한 문서가 포함되어 분석의 신뢰성을 떨어뜨리게 된다. 따라서 주제에 대한 보다 정확한 분석을 위해, 주요 용어의 누락이 없으면서도 불필요한 용어를 포함하지 않는 용어 사전이 구축되어야 한다.

용어 사전은 분석 과정 및 결과에서 정제의 역할을 수행할 뿐 아니라, 이를 통해 분석의 관점을 정의한다는 측면에서 그 중요성이 더욱 강조된다. 적용된 용어 사전에 따라 동일한 문서에 대해서도 분석 관점이 상이하게 나타나는 예는 <그림 1>을 통해 설명할 수 있다.

<그림 1>은 우리나라 가수들의 평양 공연을 보도한 기사에 대한 문서 분류 예를 나타낸다. 구체적으로는 해당 기사를 '정치' 또는 '연예'의 카테고리로 분류하고자 하며, 해당 문서의 하단에는 이미 '정치'와 '연예'로 각각 분류된 문서가

비교를 위해 제시되어 있다. 그림에서 상단 좌측은 극단적으로 정치에 편향된 용어 사전을 갖고 있는 경우를, 그리고 상단 우측은 연예에 편향된 용어 사전을 갖고 있는 경우를 나타낸다. 즉 좌측 분석에서는 주어진 문서의 용어 중 '김정은, 북한, 위원장' 등의 용어를 기준으로 문서의 내용을 판단하게 되며, 그 결과 해당 문서를 이들과 유사한 용어를 포함한 문서로 구성된 '정치' 카테고리로 분류하게 된다. 한편 우측 분석은 주어진 문서의 용어 중 '가수, 방송, 공연' 등의 용어를 기준으로 문서의 내용을 판단하게 되며, 그 결과 해당 문서를 '연예' 카테고리로 분류하게 된다.



<그림 1> 용어 사전이 문서 분류에 미치는 영향

<그림 1>은 사전을 구성하고 있는 용어에 따라 동일한 문서를 바라보는 관점이 달라짐을 나타내며, 이는 곧 용어 사전의 내용에 따라 문서 분류의 결과가 달라짐을 의미한다. 실제로 기존의 많은 연구에서 용어 사전의 품질에 따라 문서 분류의 정확도가 영향을 받을 것이라는 주장이 있어 왔으나(최성이, 2014; 김민철, 2013; 홍진성, 2014), 사전의 구성 내용 및 구성 과정이 구체적으로 분류 정확도에 어떻게 영향을 주는지에 대한 엄밀한 검증은 이루어지지 않았다. 따라서 본 연구에서는 용어 사전을 다양하게 구축하고 각 사전의 단어 구성 비율을 고유율 관점에서 비교 및 분석함으로써 융합된 분야를 분석하는 경우 용어 사전의 단어 구성 비율과 사전 구축 방법이 문서 분류의 정확도에 미치는 영향을 살펴보고자 한다.

구체적으로 본 연구에서는 전체 문서에서의 용어 빈도수에 기반을 두어 사전을 구축하는 일괄 구축 방식과 카테고리별 용어 수의 균형을 맞춘 개별 구축 방식의 두 가지 방식으로 용어 사전을 구축하여 결과를 분석하고자 한다. 또한 개별 구축 방식은 카테고리별 주요 용어를 추출하여 통합하는 용어 통합 방식과 카테고리별 주요 특질(Feature)을 추출하여 통합하는 특질 통합 방식의 두 가지 방식으로 다시 세분화하여 성능을 비교한다. 성능 비교 시 이들 사전을 구성하는 용어의 특성에 따라 고유 용어를 정의하고 고유율의 개념을 도입하여 정확도의 차이가 나는 원인을 심층 분석한다.

본 논문의 이후 구성은 다음과 같다. 다음 장인 2장에서는 본 연구와 관련된 선행연구들을 요약하고, 3장에서는 본 연구의 전체적인 개요와 방법론을 제시한다. 4장에서는 제안 방법론을 실제 뉴스 데이터에 적용한 실험 결과를 분석하고, 마지막 장인 5장에서는 본 연구의 기여 및 한계, 그리고 후속 연구 방향을 제시한다.

Ⅱ. 이론적 배경

1. 텍스트 마이닝

텍스트 마이닝은 방대한 양의 텍스트 데이터를 분석하여 의미 있는 정보를 찾아내는 과정으로, 데이터 마이닝, 자연어 처리 등을 포함한 다양한 분야의 기술을 포괄적으로 활용한다(Mooney and Bunescu, 2006; Rijsbergen, 1979; Sebastiani, 2006). 텍스트를 기반으로 작성된 문서는 단순한 수치적(Numerical) 데이터에 비해 상대적으로 더욱 풍부한 정보를 포함하고 있으므로, 이에 대한 분석을 통해 기존의 일반적인 데이터 마이닝에 비해 보다 다양한 지식을 추출할 수 있다. 텍스트 분석은 일반적으로 문서 수집, 파싱(Parsing) 및 필터링(Filtering), 구조화, 빈도 분석 및 유사도 분석의 순서로 수행되며, 텍스트 분석 기술은 비정형 데이터를 정형으로 구조화하는 단계와 구조화된 문서를 분석 및 활용하는 두 단계로 구분하여 살펴볼 수 있다. 특히 국내에서는 문서를 분석하고 활용하는 연구 중 토픽 모델링(Topic Modeling)과 문서 분류(Document Classification) 분야에 관한 연구가 매우 활발하게 이루어지고 있다.

1.1 토픽 모델링(Topic Modeling)

토픽 모델링은 방대한 양의 문서로부터 주요 토픽을 추출하는 과정으로, 문서의 군집화에 기반을 둔 대표적인 방법이다. 즉 문서가 포함하고 있는 용어의 유사도에 따라 문서를 그룹화하고 각 그룹의 주요 용어들로 해당 그룹을 기술한다. 토픽 모델링은 대량의 문서에서 핵심 이슈를 파악하고 시간의 추이에 따른 이슈 변화를 측정하기 위해 주로 사용된다. 또한, 산출된 결과를 가공하여 추가 분석을 하기 위한 방법론이 꾸준히

연구되고 있다.

문서 간 유사도의 측정에는 LSA(Latent Semantic Analysis), PLSA(Probabilistic LSA), LDA(Latent Dirichlet Allocation), 코사인 유사도(Salton and McGill, 1986) 등의 기반 기술이 널리 활용되고 있으며 토픽 모델링 기법을 활용한 연구로는 협업 필터링과 확률론적 토픽 모델링의 결합을 통한 과학 기사 추천 알고리즘을 개발한 연구(Wang and Blei, 2011), 토픽 모델링을 활용하여 투자자 유형을 분리하고 주식시장 내 시장 조치 시 투자자 집단에 따른 투자 시장 참여 유형 변화 추이를 분석한 연구(김정수, 2015), 대선 기간 발생하는 트위터를 실시간 수집하고 분석하여 이용자 네트워크의 특성을 규명한 연구(배정환, 2013), 트위터와 같은 마이크로블로그의 짧은 텍스트 환경에서 토픽 모델을 효과적으로 학습시키는 방법을 제시한 연구(Hong and Davison, 2010) 등이 있다. 또한 호텔 서비스, 정보시스템 분야 등 여러 분야에서 이루어지는 토픽 모델링 활용 연구의 동향을 분석한 시도(박준석, 2016; 박주섭, 2017; 김창식, 2017), 취업 관련 커뮤니티에서 취업 준비생의 관심사를 토픽 분석을 통하여 취업난의 원인을 탐색한 연구(김정수, 2016)도 이루어진 바 있다.

1.2 문서 분류(Document Classification)

문서 분류는 문서 내 용어의 출현 빈도를 분석하여 이에 따라 해당 문서를 특정 분류로 구분하는 과정이다. 주로 지도학습(Supervised Learning) 알고리즘 기반으로 수행되며, 분류된 문서로부터 분류자(Classifier)를 학습하여 미분류 문서에 대한 분석을 수행한다. 문서 분류 연구는 부정 예제로부터 긍정 예제를 분리해낼 수 있는 결정면을 찾아내는 알고리즘인 SVM(Support Vector Machine; 지지 벡터 기계) 기법을 Joachims(1998)가 문서 분류에 적용하면서 활발한 후속 연구가 이루어졌으며, 실시간으로 생성되는 방대한 양의 문서를 체계적으로 관리해야 할 필요성이 증가하면서 최근 중요성이 더욱 높아지고 있다. 문서 분류 성능 개선을 위한 연구로는 다양한 분류 알고리즘의 특질 선택(Feature Selection) 성능을 비교한 연구(Rogati and Yang, 2002), 문서 분류 연구에서 사용된 알고리즘과 정확도를 비교한 연구(Amensisa, 2018) 등이 있으며, 문서 분류의 활용에 관한 연구로는 단일 분류를 가진 문서의 기준을 확장하여 다중 분류를 수행하는 연구(홍진성, 2014) 등을 들 수 있다.

2. 사전구축

텍스트 분석 과정에서 텍스트에 포함된 방대한 용어를 모두 분석에 사용하는 대신, 분석 주제에 집중하고 분석 결과의 품질을 향상시키기 위해 용어 사전 및 불용어 사전을 사용한다. 불용어 사전은 분석 시 배제되는 용어들을 포함하며, 주로 대명사, 관사, 접속사, 전치사 등과 같이 내용 정보가 없는 단어들로 구성된다. 또한 극단적으로 자주 발생하는 단어는 문서 식별력이 없다고 판단하여 불용어로 처리하는 경우도 있다. 이와 반대로 용어 사전은 텍스트 분석 시 사용할 단어들의 집합을 의미하며, 연구에 따라 용어 사전을 형성하는 방법이 다르게 나타난다. 전체 문서에서 명사를 추출하여 엔트로피 기반으로 용어를 선정하는 방법(Hotho et al., 2005), 문서와 용어의 연관성을 한 문서 내에서 특정 단어의 출현 빈도를 나타내는 TF(단순 빈도수)와 해당 용어를 포함하고 있는 문서 수에 대한 전체 문서 수의 비율을 나타내는 IDF(역문서 빈도수)의 곱으로 나타내는 TF-IDF(Term Frequency - Inverse Document Frequency) 가중치를 이용하여 가중치가 크게 나타난 용어들로 분야별 용어 사전을 구성하는 방법(Gupta and Lehal, 2009),

추출한 명사를 빈도순으로 나열하여 상위 단어를 용어 사전으로 사용하는 방법 등이 용어 사전 구축의 대표적인 예이다.

기존에는 범용 용어 사전과 불용어 사전을 활용한 연구가 주로 수행되었으나, 분석에 대한 수요가 세분화되고 다양해짐에 따라 분석 목적에 맞게 특화된 사전들이 정의되고 개발되었다. 그 대표적인 예가 감성 분석으로, 감성 분석에서는 각 용어에 긍정, 부정, 중립의 감성지수를 결합한 감성 사전을 사용한다. 감성 사전 구축에 관한 연구로는 확률론에 기초해 단어의 출현 빈도에 따른 어휘의 연관성을 측정하는 PMI (Pointwise Mutual Information)를 보완한 방법으로 미리 긍정 어휘와 부정 어휘를 정의한 후 도출된 값이 양수일 때는 긍정, 음수일 때는 부정으로 어휘의 극성을 분류하는 방법인 SO-PMI(Semantic Orientation from Point-wise Mutual Information)를 활용하여 분야별 감성 사전을 구축한 연구(이상훈, 2016), 집단 지성을 활용해 단어의 극성을 정의한 연구(안정국, 2015), 범용 감성 사전보다 주제에 특화된 감성 사전을 활용할 때 감성 분석의 정확도가 향상됨을 확인한 연구(송종석, 2011) 등이 있다. 또 다른 예로 개체명 사전의 경우 최근 위키피디아(Wikipedia) 문서에서의 개체명 인식에 대한 연구가 매우 활발하게 진행되고 있다. 구체적으로는 위키피디아 문서의 링크 정보를 이용하여 언어학적 패턴을 생성하고, 생성된 패턴에 일치하는 개체명을 인식하거나 분류체계를 통해 개체명 사전을 구축한다(배상준, 2010; Richman and Schone, 2008).

사전 구축 시 용어 선정은 형태소 분석, 말뭉치(Corpus), 워드넷(WordNet), 단어 빈도수에 기반을 두어 이루어진다. 형태소 분석은 자연어 처리의 가장 기본적인 단계로, 각 문서를 의미를 갖는 가장 작은 단위인 형태소로 분리하고 품사를 명시하는 과정을 나타낸다. 형태소 분석 시 속도와 정확도를 향상시키기 위해 사전 탐색만으로 형태소 분석이 가능한 말뭉치를 이용해 사전을 구축하기도 하며(곽수정, 2013) 형태소로 영화 리뷰 데이터를 추출하여 평점을 예측한 연구(조정태, 2015)도 있다. 말뭉치는 특정한 목적을 갖고 언어의 표본을 추출한 집합으로, 단일 언어 말뭉치, 다중 언어 말뭉치, 구조적 수준의 분석이 가능하도록 구문 분석이 이루어진 분석된 말뭉치 등이 있다. 최근에는 효과적인 언어 연구를 위해 말뭉치에 태그의 형태로 품사 표기를 하는 말뭉치 주석화 과정이 이루어진다. 한편 워드넷(Miller, 1995; Pedersen, 2004)은 어휘 의미망으로 어휘의 중의성 해소나 정보 추출 등과 같은 다양한 자연어 처리에 폭넓게 사용되며 (Fellbaum, 1998), 비슷한 의미를 가진 단어들의 집합인 synset(set of synonym)이 기본 요소이다. 워드넷은 영어의 어휘 데이터 베이스로 약 15만 개의 어휘, 20만 개의 어휘-의미 묶음, 11만 5천 개의 동의어 집합을 제공한다. 워드넷을 활용한 연구로는 워드넷과 어휘 체인(lexical chain)을 결합한 클러스터링 방법을 제안한 연구 (Wei et al., 2015), 워드넷을 기반으로 단어 쌍 사이의 의미적 유사성을 측정하는 새로운 방법을 제시한 연구(Gao et al., 2015), 어휘 간 계층 정보 및 어휘 정보량을 이용하여 문서 클러스터 레이블 선정 방법을 제안한 연구(김태훈, 2017) 등이 있으며, 한국어 워드넷의 구축 및 활용에 관한 연구도 다수 수행된 바 있다(윤애선, 2009; 최석재, 2014; 강상욱, 2015).

이처럼 사전의 유형 및 구축 방식은 더욱 다양화되고 있으며, 사전의 품질이 분석 결과에 미치는 영향 또한 더욱 중요하게 강조되고 있다. 하지만 용어 사전의 품질이 문서 분류 등 텍스트 분석의 결과에 어떠한 형태로 영향을 미치는지 용어 사전의 단어 구성과 그에 따른 품질 비교에 대한 연구는 찾아보기 어렵다. 따라서 본

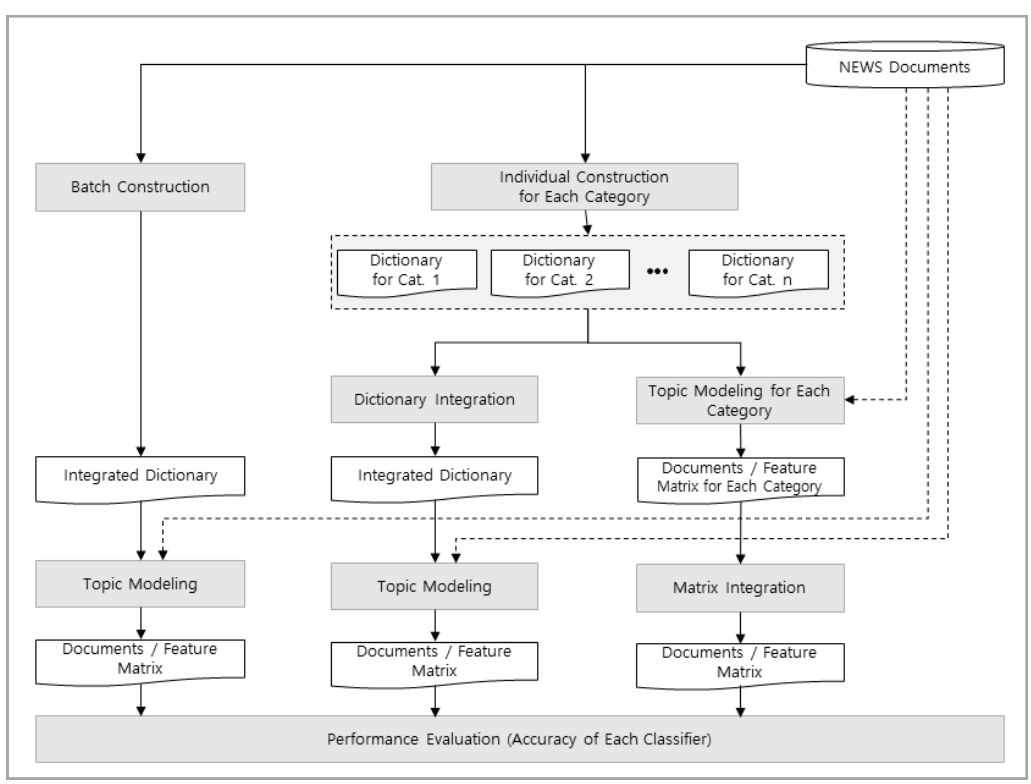
연구는 용어 사전 구축 방법에 따라 각 사전을 구성하는 고유 단어 및 범용 단어의 비율이 어떻게 달라지는지 파악하고, 이러한 구성비가 문서 분류기의 성능에 미치는 영향을 분석한다.

Ⅲ. 제안 방법론

1. 연구모형

본 절에서는 문서 분류 시 사용되는 용어 사전의 다양한 구성 방법을 세분화하여 살펴보고, 각 사전의 성능을 비교하고 분석하여 최적의 사전을 만드는 방안을 제시한다. 구체적으로는 일반적으로 가장 많이 사용되는 방식인 일괄 구축 사전과 앞서 언급한 개별 구축 사전인 용어 통합 사전, 특질 통합 사전을 구성하고, 각각의 용어 사전을 사용했을 때의 문서 분류 예측 정확도를 비교하고 그 차이의 원인을 분석한다. 제안 방법론의 전체적 개요는 <그림 2>와 같다.

우선 <그림 2>에서 NEWS Documents는 용어 사전 구축과 문서 분류 정확도 비교 실험에 사용되는 뉴스 기사의 집합을 나타낸다. 그림에서 좌측의 흐름은 일괄 구축 방식으로, 전체 뉴스 기사에서 파악된 용어의 빈도수에 기반을 두어 사전을 구축한다. 한편 가운데의 흐름은 개별 구축 방식 중 용어 통합 방식으로, 각 카테고리별 뉴스 기사에서 파악된 용어의 빈도수에 기반을 두어 사전을 구성하고 이를 통합한다.



<그림 2> 연구 모형 개요

마지막으로 우측의 흐름은 개별 구축 방식 중 특질 통합 방식으로, 각 카테고리별 사전을 적용하여 특질 분석을 반복적으로 수행하고 그 결과를 통합한다. 이렇게 세 가지 서로 다른 방식으로 구축된 사전의 품질을 간접적으로 평가하기 위해, 이들 사전을 각기 적용한 문서 분류의 정확도를 비교하고 정확도의 차이가 나타나는 원인을 분석한다. 전체 과정을 구성하는 단계별 설명은 이후 절에서 자세히 다루고, 실제 데이터에 대한 실험 결과는 4장에서 소개한다.

2. 용어 사전 구축

본 절에서는 세 가지 용어 사전의 구성 방식을 소개한다. NEWS Documents는 카테고리별로 동일한 수의 기사를 추출하여 구성하며, 각 사전을 500개의 용어로 구성하여 용어의 수를 동일하게 구축하였다. 이후 NEWS Documents로부터 사전을 구축하는 세 가지 방식은 각 부절에서 자세히 다룬다.

2.1 일괄 구축

일괄 구축 방식은 실제 텍스트 분석에서 가장 일반적으로 사용되는 방식으로, 카테고리 구분 없이 전체 문서에서 출현한 용어의 빈도수에 기반을 두어 사전을 구성한다. 구체적으로는 NEWS Documents의 전체 문서에 대해 파싱을 수행하고, 추출된 단어 중 명사만을 추려낸 후, 빈도수가 높은 순서로 정렬하여 정해진 개수의 용어를 선택하여 사전을 구성한다. 요약하면 일괄 구축 방식을 통한 용어 사전은 카테고리에 관계없이 전체 문서에서 빈번하게 나타나는 명사들로 구성된다. 예를 들어 <그림 3>은 전체 카테고리를 망라하여 가장 빈번하게 나타난 명사 500개를 선정한 가상 예를 보이며, 특정 카테고리에 특화된 용어보다는 범용적으로 널리 사용되는 용어가 상위 순위에 나타나는 경향을 보인다.

No	Term	Freq.
W1	기자	21,201
W2	국민	12,487
W3	시장	12,239
W4	서비스	11,704
...	...	...
W500	데이터	1,231

<그림 3> 일괄 구축 사전 구축 가상 예

이후 일괄 구축 사전을 사용하여 NEWS Documents의 기사에 대한 토픽 모델링을 수행한다. 본 연구에서는 용어 사전의 품질을 평가하기 위한 간접적인 척도로 문서 분류의 정확도를 사용하며, 문서 분류를 위해서는 비정형 문서의 구조화가 선행되어야 한다. 본 연구에서는 다양한 방법 중 토픽 모델링을 활용하여 비정형 문서의 구조화를 수행한다. 즉 본 단계에서의 토픽 모델링은 용어 사전의 구축을 위해서가 아니라 구축된 용어 사전의 품질을 문서 분류를 통해 간접적으로 평가하기 위해 수행한다. 구체적으로는 NEWS Documents를 구성하는 기사에 대한

토픽 모델링 과정에서 일괄 구축 용어 사전이 Start List로 적용되게 된다.

토픽 모델링에 관한 내용은 이미 기존 연구에서 상세히 다루고 있으므로(Blei et al., 2003; Deerwester et al., 1990; 김남규, 2017), 본 연구에서는 토픽 모델링 개념에 대한 자세한 소개 대신 주요 활용 과정만을 설명한다. 토픽 모델링의 결과로 토픽을 구성하는 용어(키워드)들과 각 토픽에 대한 문서의 부합 정도를 나타내는 문서/토픽 행렬(Document/Topic Matrix)이 산출된다. 문서/토픽 행렬은 특정 토픽에 대한 문서의 대응 정도를 나타내며 대응되는 값이 클수록 문서와 토픽의 연관성이 높다. <그림 4>는 일괄 구축 방식으로 생성된 용어 사전을 Start List로 적용하여 토픽 모델링을 수행한 가상 결과를 나타내며, 상단 표의 “Topic 1: 기사, 제보, 보도자료, 공감, 언론”과 같이 범용적 용어를 키워드로 갖는 토픽이 다수 도출되는 경향을 보인다. <그림 4>의 하단 표는 각각의 토픽에 분석 대상 문서가 대응되는 정도를 나타내는 가중치로, 용어와 토픽, 용어와 문서의 행렬 곱을 통해 가중치, 즉 문서와 토픽의 연관 정도가 도출된다.

Topic Information	
1	기사, 제보, 보도자료, 공감, 언론
2	스포츠, 미디어, 돈, 서비스, 시장
3	대통령, 청와대, 정부, 정책, 경제
4	통신사, 저작권자, 온라인, 소식, 반응
5	연예, 스타, 사업, 서비스, 선수

	Topic1	Topic2	Topic3	Topic4	Topic5
DOC1	-0.275	0.168	0.262	0.329	0.022
DOC2	0.010	-0.180	-0.050	0.077	0.000
DOC3	0.098	0.378	-0.033	0.000	0.209
DOC4	0.299	0.295	0.000	0.078	0.374
DOC5	0.382	0.002	0.000	0.042	-0.023

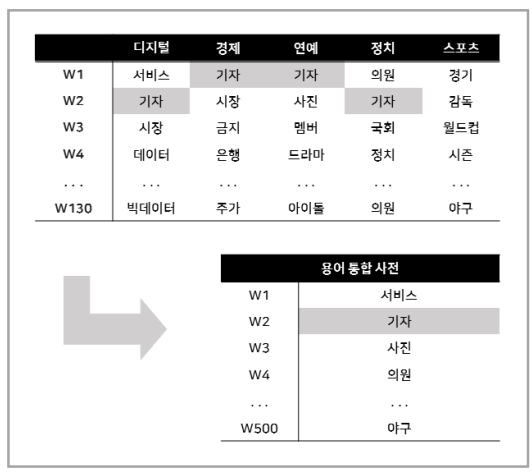
<그림 4> 일괄 구축 방식의 토픽 모델링 결과 예

2.2 개별 구축 방식: 용어 통합

개별 구축 방식은 전체 기사를 각 카테고리별로 나누어 카테고리별 사전을 구축한 뒤 이를 통합하는 방식으로, 통합 방식에 따라 다시 용어 통합과 특질 통합 방식으로 세분화된다. 용어 통합 방식은 카테고리별 사전에 포함된 용어를 단순히 병합하는 방식이며, 특질 통합 방식은 카테고리별 용어 사전을 활용한 카테고리별 토픽 모델링을 통해 특질을 추출한 후 이들 특질을 통합하는 방식을 나타낸다.

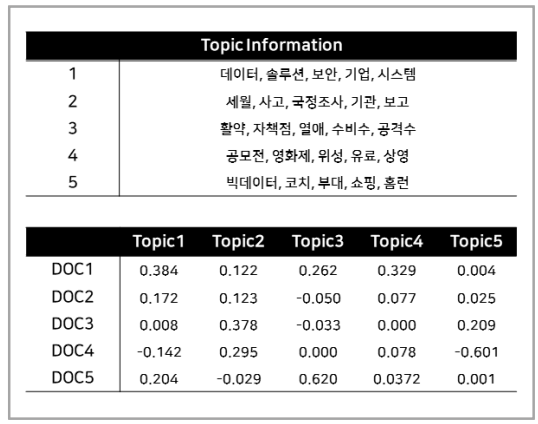
우선 본 부절에서는 개별 구축 방식 중 용어 통합 방식에 대해 소개한다. 우선 NEWS Documents의 전체 기사를 카테고리별로 구분한 후, 카테고리별 파싱 및 빈도 계수를 통해 주요 용어를 빈도순으로 정렬한다. 다음으로 전체 통합 사전의 구성 용어가 특정 카테고리에 치우치는 현상을 방지하기 위해, 각 카테고리마다 고빈도 용어들을 동일 수만큼 추출하여 통합한다. 이

때 둘 이상의 사전에서 중복으로 추출된 용어는 중복을 제거하고 한 번만 사용한다. <그림 5>는 '디지털', '경제', '연예', '정치', '스포츠'의 5개 카테고리 각각에서 고빈도 용어 130개씩을 추출한 후, 중복 제거 및 통합을 통해 용어 500개로 구성된 통합 사전을 구축한 예를 보이고 있다. 이러한 방식으로 도출한 통합 사전은 일괄 구축 방식에 의해 구축된 사전에 비해 범용적 용어보다는 각 카테고리에 특화된 용어를 다수 포함한다는 특징을 갖는다.



<그림 5> 카테고리별 용어 사전의 통합 예

개별 구축 방식 중 용어 통합 방식에 의해 도출된 통합 사전을 활용하여 토픽 모델링을 수행한 결과의 예가 <그림 6>에 나타나 있다. <그림 4>에 비해 <그림 6>에서는 각 토픽을 구성하는 용어가 범용성보다는 카테고리별 특수성을 갖게 되며, "Topic 5: 빅데이터, 코치, 부대, 쇼핑, 홈런"과 같이 둘 이상의 카테고리에서 선정된 용어들로 구성된 융합 토픽도 발견된다는 특징을 갖는다.

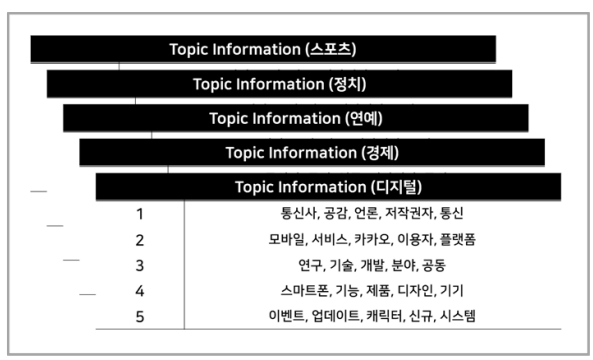


<그림 6> 용어 통합 방식의 토픽 모델링 결과 예

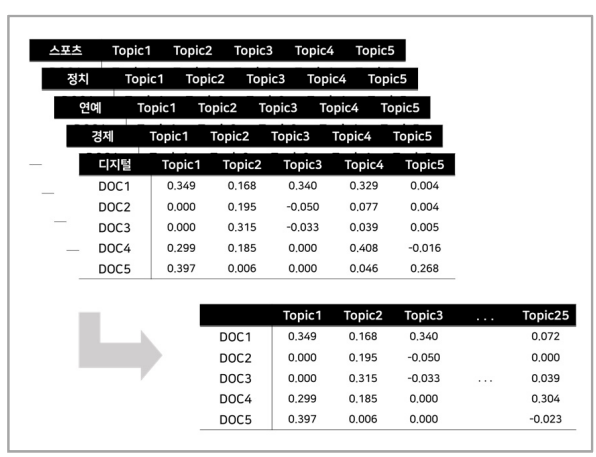
2.3 개별 구축 방식: 특질 통합

본 부절에서는 또 다른 개별 구축 방식인 특질 통합 방식을 소개한다. 우선 카테고리별 파싱 및 빈도 계수를 통해 카테고리별 용어 사전을 도출하는 과정은 3.2.2절과 동일하다. 다만 본 과정에서는 이전과 달리, 각 카테고리별 용어 사전을 통합하는 대신 각 카테고리별 토픽 모델링 분석 결과를 통합한다. 즉 카테고리에 따라 용어 사전을 적용하여 카테고리 수만큼의 토픽 모델링을 수행하고, 그 결과를 통합한다. <그림 7>은 '디지털', '경제', '연예', '정치', '스포츠'의 5개 카테고리에 대해 각각 토픽 모델링을 수행한 결과의 예를 나타낸다. 각 카테고리별 토픽 모델링 과정에서는 해당 카테고리의 용어 사전에 수록된 용어만 의미를 가지므로, <그림 6>과 달리 복수의 카테고리에서 선정된 용어들로 구성된 융합 토픽은 발견되기 어렵다는 특징을 갖는다.

<그림 7>에 나타난 5개 카테고리에 대한 토픽 모델링 결과를 문서 기준으로 병합함으로써, 각 문서가 5개 카테고리 전체의 토픽에 대응되는 정도를 산출할 수 있다. 예를 들어 <그림 8>은 5개 카테고리 각각에서 5개씩의 토픽을 도출한 후, 이를 병합하여 각 문서에 대한 25개 토픽의 가중치를 나타낸 결과를 보인다.



<그림 7> 특질 통합 방식의 토픽 모델링 결과 예



<그림 8> 특질 통합 방식의 특질 통합 예

Ⅳ. 실 험

1. 실험 개요

본 장에서는 실제 뉴스 데이터에 대해 제안 방법론을 적용한 실험 및 분석 결과를 제시한다. 본 실험에서 형태소 분석, 파싱 및 빈도 계수, 토픽 모델링, 인공신경망 분석을 통한 문서 분류는 SAS Enterprise Miner Workstation 14.1을 통해 수행하였으며, 사전 구축 및 문서 분류를 위한 실험 데이터는 인터넷 뉴스 포털인 'D' 사이트에서 수집한 뉴스 기사를 사용하였다.

구체적으로 2014년 6월 넷째 주부터 2014년 7월 첫째 주까지 2주 동안 디지털, 경제, 연예, 정치, 스포츠의 5개 카테고리에 게시된 뉴스 기사 총 39,800건을 사용하였으며, 카테고리별로 7,960건씩 동일한 수의 뉴스 기사를 추출하였다. 이때 하나의 기사는 하나의 카테고리로 분류되어 있다.

2. 방식별 사전 구축 및 토픽 모델링

본 단계에서는 3장에서 제시한 과정에 따라 다양한 방식으로 용어 사전을 구축하고, 각 사전을 Start List로 적용한 문서 분류의 정확도 비교를 통해 사전의 품질을 비교한다.

우선 일괄 구축 방식을 통해 전체 기사 39,800건으로부터 고빈도 용어 500개를 도출하고, 이를 Start List로 적용한 토픽 모델링을 통해 50개의 토픽을 도출하였다(비교 1). 다음으로 개별 통합 방식의 사전 구축을 위해 각 카테고리별 기사 39,800건으로부터 고빈도 용어 130개를 도출하고, 이렇게 도출된 5개 카테고리의 사전의 통합 및 용어의 중복 제거를 통해 용어 500개로 구성된 통합 사전을 구축하였다. 이렇게 구축된 통합 사전을 Start List로 적용한 토픽 모델링을 통해 50개의 토픽을 도출하였다(비교 2).

또한 앞에서 도출한 5개 카테고리별 용어 사전 각각을 적용한 토픽 모델링을 통해 카테고리별로 10개의 토픽을 도출하고, 이를 병합하여 각 문서별로 50개의 토픽에 대응되는 가중치를 산출하였다(비교 3). <그림 9>~<그림 11>은 각각 일괄 구축 방식, 개별 구축 방식 중 용어 통합 방식, 그리고 개별 구축 방식 중 특질 통합 방식에 의해 도출된 토픽 전체의 결과를 보여준다.

TID	Keywords	TID	Keywords
1	연구, 기술, 개발, 교수, 분야	26	정보, 개인, 서비스, 인터넷, 보안
2	조, 조별, 차전, 리그, 승점	27	드라마, 배우, 연기, 역, 촬영
3	공천, 후보, 지역, 전략, 당	28	의원, 국회, 협의, 선거, 세월
4	반응, 누리꾼, 소식, 네티즌, 화제	29	앱, 모바일, 구매, 성공, 주식
5	안타, 루, 홈런, 타자, 선발	30	경기, 프로야구, 한화, 인천, 루
6	이벤트, 업데이트, 캐릭터, 유저, 신규	31	결혼, 사람, 가족, 소식, 사랑
7	병장, 총기, 난사, 사건, 병사	32	데이터, 솔루션, 보안, 기업, 시스템
8	후보자, 장관, 인사청문회, 청문회, 총리	33	스포츠, 미디어, 최고, 모델, 선발
9	스마트폰, 기능, 제품, 기기, 스마트	34	사람, 선수, 모습, 정도, 생각
10	영화, 신, 배우, 전쟁, 제작	35	달러, 규모, 계약, 세계, 해외
11	분기, 주가, 연구원, 실적, 대비	36	회장, 그룹, 지분, 공장, 회사
12	요금, 서비스, 텔레콤, 광대역, 데이터	37	교육, 프로그램, 학생, 학교, 활동
13	대회, 오픈, 여자, 우승, 프로	38	멤버, 인천, 모습, 그룹, 도전
14	주석, 양국, 협력, 관계, 공동	39	위원장, 국회, 위원회, 인사, 인천
15	후반, 슈팅, 전반, 골키퍼, 공격	40	승, 패, 무, 평균, 시즌
16	가구, 아파트, 주택, 층, 시설	41	서비스, 자동차, 차량, 라인, 조사
17	세월, 사고, 기관, 참사, 안전	42	팀, 선수, 시즌, 이슈, 경기
18	축구, 대표팀, 월드컵, 선수, 감독	43	대통령, 청와대, 인사, 총리, 수석
19	앨범, 음악, 팬, 무대, 가수	44	남자, 친구, 여자, 사건, 영화
20	방송, 캡처, 화면, 프로그램, 웃음	45	아이, 부산, 공식, 사이트, 웹
21	타임, 리얼, 연예, 스타, 돈	46	감독, 영화, 제작, 선수, 대표팀
22	문제, 입장, 조사, 관계자, 결과	47	포인트, 외국인, 기관, 상승, 개인
23	제품, 고객, 브랜드, 상품, 소비자	48	게임, 모바일, 시장, 서비스, 이용자
24	금융, 은행, 대출, 상품, 자금	49	사업, 계획, 선수, 지원, 서비스
25	연합, 정치, 민주, 원내, 국회	50	정부, 정책, 경제, 산업, 기업

<그림 9> 일괄 구축 방식을 통한 토픽 도출 결과

TID	Keywords	TID	Keywords
1	분기, 연구원, 주가, 대비, 실적	26	서비스, 지역, 광대역, 고객, 전국
2	조, 조별, 차전, 리그, 승점	27	의원, 국회, 전당대회, 당, 선거
3	공천, 후보, 전략, 지역, 동작	28	드라마, 사랑, 배우, 연기, 역
4	방송, 프로그램, 캡처, 예능, 웃음	29	데이터, 솔루션, 보안, 기술, 기업
5	승, 패, 자책점, 평균, 시즌	30	요금, 텔레콤, 보조금, 단말기, 사업자
6	이벤트, 업데이트, 캐릭터, 유저, 신규	31	회장, 지분, 매각, 그룹, 동부
7	병장, 총기, 난사, 사건, 병사	32	안타, 루, 타자, 홈런, 타석
8	후보자, 장관, 인사청문회, 사퇴, 총리	33	경기, 프로야구, 한화, 잠실, 스포츠
9	스마트폰, 기능, 제품, 기기, 스마트	34	사람, 문제, 결과, 연구, 사실
10	영화, 개봉, 배우, 신, 제작	35	결혼, 연애, 사람, 친구, 감독
11	주석, 양국, 방한, 국가주석, 정상회담	36	대통령, 청와대, 인사, 총리, 수석
12	연구, 기술, 교수, 개발, 과학	37	선수, 팀, 대회, 경기, 시즌
13	골프, 오픈, 라운드, 여자, 프로	38	기업, 경제, 산업, 시장, 장관
14	모바일, 앱, 서비스, 라인, 카카오	39	게임, 시장, 캐릭터, 모바일, 행사
15	세월, 사고, 국정조사, 보고, 특위	40	인천, 도전, 무한, 대회, 멤버
16	제품, 고객, 브랜드, 상품, 행사	41	정보, 개인, 보안, 인터넷, 시스템
17	감독, 축구, 대표팀, 월드컵, 선수	42	멤버, 감독, 카라, 모습, 최자
18	앨범, 곡, 음악, 팬, 가수	43	교육, 프로그램, 지원, 센터, 대상
19	후반, 슈팅, 전반, 골키퍼, 공격	44	화보, 모델, 촬영, 몸매, 라인
20	반응, 누리꾼, 소식, 네티즌, 최자	45	종목, 외국인, 포인트, 지수, 매수
21	달러, 규모, 계약, 지역, 세계	46	자동차, 차량, 연비, 산업, 조사
22	가구, 아파트, 주택, 분양, 단지	47	사업, 시장, 개발, 계획, 서비스
23	금융, 은행, 대출, 상품, 금리	48	정부, 위원회, 위원장, 국회, 행사
24	타임, 리얼, 연예, 돈, 스타	49	아이, 부산, 가족, 딸, 사이트
25	정치, 연합, 민주, 원내, 국회	50	스포츠, 미디어, 최고, 감독, 모델

<그림 10> 용어 통합 방식을 통한 토픽 도출 결과

디지털

TID	Keywords
1	정부, 문제, 국가, 결과, 내용
2	게임, 모바일, 이벤트, 캐릭터, 플레이
3	방송, 화면, 프로그램, 영상, 예정
4	서비스, 스마트폰, 기능, 앱, 제품
5	사람, 관심, 문제, 정보, 개인
6	시장, 투자, 기업, 달러, 제품
7	최고, 결과, 관심, 세계, 영상
8	지역, 전략, 시장, 모바일, 앱
9	사업, 연구, 기술, 기업, 개발
10	고객, 관계자, 예정, 관심, 행사

경제

TID	Keywords
1	서비스, 모바일, 고객, 기존, 상품
2	경기, 결과, 가능성, 가운데, 외국인
3	사람, 정도, 관심, 돈, 가운데
4	정부, 문제, 정책, 결과, 기관
5	시장, 금융, 분기, 투자, 주가
6	지역, 가운데, 관계자, 가능성, 센터
7	사업, 기술, 시장, 개발, 기업
8	제품, 가격, 모델, 소비자, 고객
9	그룹, 예정, 공식, 관심, 회장
10	정보, 고객, 서비스, 센터, 금융

연예

TID	Keywords
1	경기, 번째, 마지막, 가운데, 공식
2	사람, 사건, 가운데, 정도, 사실
3	방송, 모습, 화면, 월드컵, 프로그램
4	영화, 감독, 배우, 제작, 드라마
5	월드컵, 감독, 팀, 마지막, 팬
6	그룹, 활동, 공식, 예정, 아이
7	타임, 리얼, 연예, 제공, 스타
8	팀, 최고, 미디어, 번째, 감독
9	모델, 최고, 예정, 인기, 미디어
10	관계자, 예정, 가운데, 사건, 현장

정치

TID	Keywords
1	시장, 전략, 가능성, 경제, 가운데
2	경기, 결과, 인터뷰, 가능성, 가운데
3	지역, 의원, 정치, 국회, 연합
4	사람, 가족, 사실, 관심, 경제
5	사업, 계획, 협력, 지역, 공동
6	병장, 사건, 총기, 난사, 사고
7	결과, 대상, 기관, 정부, 조사
8	예정, 관심, 계획, 관계자, 내용
9	대통령, 후보자, 인사, 청와대, 총리
10	지역, 협력, 주석, 공동, 양국

스포츠

TID	Keywords	TID	Keywords
1	월드컵, 조, 경기, 조별, 리그	6	감독, 대표팀, 축구, 선수, 처음
2	게임, 오픈, 축구, 시즌, 공격	7	결과, 처음, 평균, 팀, 자리
3	가능성, 기회, 처음, 자리, 경기	8	대회, 여자, 오픈, 팀, 인천
4	경기, 시즌, 선발, 승, 프로야구	9	최고, 미디어, 스포츠, 팬, 처음
5	모습, 팀, 자리, 팬, 마지막	10	가운데, 처음, 자리, 평균, 번째

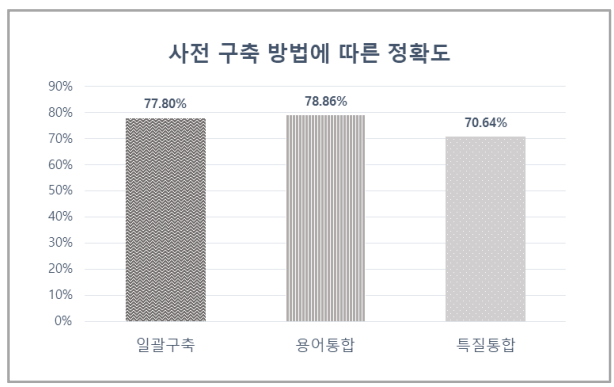
<그림 11> 특질 통합을 통한 토픽 도출 결과

3. 구축 방식에 따른 사전의 품질 비교

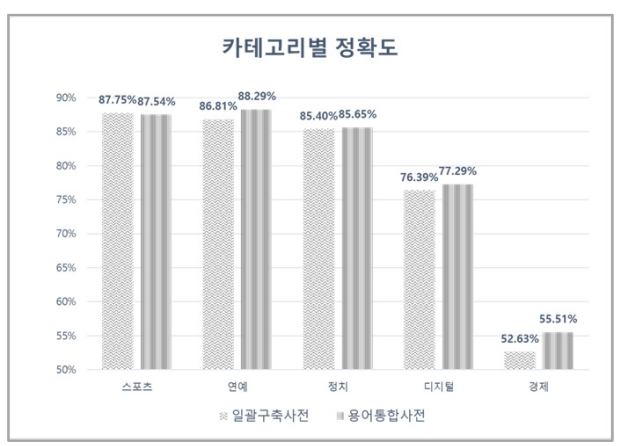
본 연구의 목적은 용어 사전의 구축 방식에 따라 사전의 품질이 어떻게 다르게 나타나는지를 파악하는 것이다. 하지만 사전 품질을 직접적으로 평가할 수 있는 방법은 알려져 있지 않으므로, 본 부절에서는 각 사전을 문서 분류에 적용했을 때 나타나는 분류 정확도를 비교하여 사전의 품질을 간접적으로 평가하고 사전에 따라 문서 분류의 정확도가 상이하게 나타나는 현상을 해석하는 방안을 제시한다.

우선 문서 분류는 SAS Enterprise Miner의 인공신경망을 적용하여 수행하였으며, 인공신경망의 구조는 입력층, 은닉층, 출력층으로 구성된다. 본 실험에서는 39,800건의 뉴스 데이터를 입력층의 데이터로 사용하였으며 은닉 마디 수는 3으로 설정하여 인공신경망 분석을 수행하였다. 또한 학습 데이터의 비중을 70%, 평가 데이터의 비율을 30%로 설정하였다. 일반적으로 학습 데이터와 평가 데이터의 비율을 8:2, 7:3, 6:4로 설정하며 본 연구에서 가장 적합한 모델인 7:3 비율로 설정하였다. 분석 결과 사전에 따른 문서 분류의 정확도는 <그림 12>와 같이 용어 통합 방식이 78.86%로 가장 높게 나타났고, 그 다음으로 일괄 구축 방식(77.80%), 특질 통합 방식(70.64%) 순으로 정확도가 높게 나타났다. 이는

특질 통합 방식의 경우 앞 절에서 살펴본 바와 같이 융합 토픽, 즉 둘 이상의 카테고리에서 선정된 용어들로 구성된 토픽을 형성할 수 없다는 한계로 인해 다른 두 가지 방식에 비해 분류 정확도가 낮게 나타난 것으로 판단한다.



<그림 12> 사전 구축 방법별 문서 분류 정확도



<그림 13> 카테고리별 문서 분류 정확도

다음으로 본 연구에서는 사전별로 분류 정확도가 상이하게 나타나는 원인을 규명하기 위해 추가 실험을 수행하였다. 추가 실험은 <그림 12>에서 정확도가 높게 나타난 일괄 구축 사전과 용어 통합 사전의 두 가지 사전에 대해 수행하였으며, 카테고리별로 문서 분류의 정확도를 살펴보았다. <그림 13>에서 5개 카테고리 전체에 걸쳐 용어 통합 사전의 분류 정확도가 일괄 구축 사전에 비해 높게 나타났으며, 두 사전 모두 스포츠, 연예가 높고, 이후 정치, 디지털, 경제 카테고리 순으로 높은 분류 정확도를 보였다.

이와 같이 카테고리별로 분류 정확도가 상이하게 나타나는 현상은 기존의 유사 연구에서도 여러 차례 발견되었으며, 특히 스포츠와 연예 분

야의 분류 정확도는 타 분야에 비해 일관되게 높게 나타난 바 있다. 또한 이러한 현상의 원인으로는 스포츠와 연예 분야의 기사가 타 분야의 기사에 비해 특수성이 높은 어휘를 다수 포함하고 있음이 주장되어왔다. 따라서 본 연구에서는 사전을 구성하는 용어의 '고유율'이라는 개념을 새롭게 고안하여 사전을 상세 분석하였다. 예를 들어 '월드컵', '투수', '루타' 등의 어휘는 스포츠 이외의 다른 분야에서는 거의 사용되지 않으므로, 해당 어휘를 포함한 문서를 스포츠 카테고리로 쉽게 분류할 수 있다. 하지만 어휘의 특수성과 분류 정확도에 대한 관계는 엄밀하게 다루어지지 않아 고유율이라는 개념을 도입하여 어휘의 특수성과 정확도의 관계를 비교하였다.

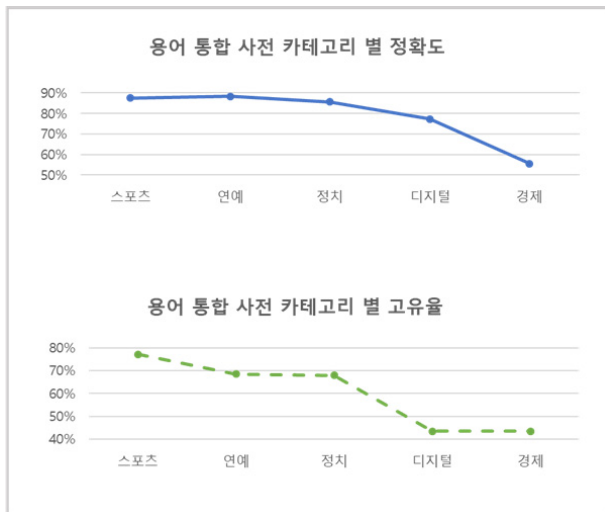
우선 본 실험에서 '고유 용어'는 특정 카테고리의 사전에만 수록되고 다른 카테고리의 사전에 포함되지 않은 용어로 정의된다. 또한 '고유율'은 특정 사전을 구성하는 전체 용어 수 중 고유 용어가 차지하는 비율을 나타낸다. 예를 들어 <그림 14>에서 '연예' 분야 사전의 경우 전체 10개 용어 중 타 분야의 사전에도 중복으로 출현한 용어가 1개, 타 분야에는 나타나지 않은 '연예' 분야의 고유 용어가 9개로, 이 사전의 고유율은 90%로 나타난다.

분야	디지털	경제	연예	정치	스포츠
	시장	기자	기자	의원	경기
	서비스	시장	사진	기자	감독
	기자	은행	멤버	국회	월드컵
	데이터	대출	드라마	정치	시즌
	사용자	아파트	가수	공천	선발
	이용자	서비스	연예	연합	안타
	콘텐츠	주가	생방송	시장	투수
	빅데이터	매매	오디션	트럼프	루타
	모바일	하락	연습생	김정일	기자
	개발자	주식	연애	북한	활약
고유율	70%	70%	90%	80%	90%

<그림 14> 고유 용어와 고유율

본 실험에서 고유율을 정의하고 분석한 이유는 고유율이 분류 정확도에 미치는 영향을 파악하기 위함이며, 카테고리별 고유율과 정확도의 추이를 분석한 결과 <그림 15>와 같이 사전의 고유율이 높은 카테고리에서 분류의 정확도도 높게 나타남을 확인하였다.

이상의 실험에서 사전의 고유율이 높을수록 분류의 정확도도 높게 나타남을 확인하였다. 따라서 본 연구의 <그림 12>에서 확인한 사전 구축 방법에 따른 분류 정확도의 차이를 사전의 고유율 관점에서 추가로 분석하였다. 즉 <그림 16>에서 용어 통합 사전의 고유율은 일괄 구축 사전에 비해 25% 높은 것으로 나타났으며, 이러한 차이가 각 사전을 사용한 문서 분류의 정확도의 차이를 가져온 것으로 판단한다.



<그림 15> 카테고리별 사전의 고유율과 정확도

	전체 용어 수	고유 용어 수	고유율	정확도
용어통합방식	500	393	79%	78.86%

	전체 용어 수	고유 용어 수	고유율	정확도
일괄구축방식	500	269	54%	77.80%

<그림 16> 사전 구축 방법에 따른 고유율과 정확도

본 실험에서는 다양한 방식에 따라 용어 사전을 구축하고, 각 사전의 품질을 해당 사전을 사용한 문서 분류의 정확도 측면에서 살펴보았다. 또한 사전별로 정확도 차이의 원인을 규명하기 위해 사전의 고유율을 정의하고, 각 카테고리별 정확도의 차이를 고유율 관점에서 분석하였다. 분석 결과 사전의 고유율이 높은 카테고리에서 문서 분류의 정확도가 높게 나타났으며, 이와 유사하게 고유율을 높일 수 있는 방식에 따라 사전을 구축한 경우 전체적으로 문서 분류의 정확도가 높게 나타남을 확인하였다. 다만 고유율이 아무리 높아져도 정확도가 무제한 높아지지 않고 어느 정도 선에서 수렴을 하게 될 것이다. 고유율의 증가에 따라 정확도가 같은 비율로 개선되지는 않는 원인은 본 연구에서 사용한 데이터의 정확도가 세츄레이션(saturation)에 가까운 정확도를 보이기 때문에 정확도의 개선 폭이 적게 나타난 것으로 판단한다.

Ⅴ. 결 론

비정형 데이터 분석에서 분석 결과의 정제를 위해 사용되는 용어 사전은 수록된 용어의 관점에서 문서를 바라볼 수 있는 기준을 제공한다는 점에서 더욱 중요하게 여겨지고 있다. 이에 본 연구에서는 용어 사전을 구축하는 다양한 방법을 살펴보고, 각 결과로 도출된 용어 사전을 적용한 문서 분류의 정확도에 따라 용어 사전의 품질을 간접적으로 평가하였다. 또한 카테고리마다 분류 정확도가 상이하게 나타나는 현상에 착안하여 사전의 고유율이라는 개념을 정의하였으며, 용어 사전 구축 방법에 따라 사전의 고유율과 문서 분류의 정확도가 다르게 나타나는 현상을 확인하였다.

본 연구의 의의는 다음과 같다. 우선 본 연구를 통해 카테고리별로 주요 용어를 도출한 후 이를 통합하는 사전 구축 방식이 타 사전 구축 방식에 비해 분류 정확도 측면에서 바람직한 특징을 갖는 것을 확인할 수 있었다. 또한 사전의 고유율을 높임으로써 분류의 정확도를 더욱 향상시킬 수 있는 가능성을 발견하였으며, 이는 본 연구의 실무적 기여로 인정받을 수 있다. 이처럼 본 연구는 고유 용어의 중요성에 주목했다는 점에서 의미를 갖는다. 즉 용어의 단순 빈도수(Term Frequency), 역문서 빈도수(Inverse Document Frequency)와 더불어, 각 용어가 특정 카테고리에서 나타내는 빈도수와 타 카테고리에서 나타내는 빈도수와의 상대적 차이를 고찰했다는 점에서 본 연구의 학술적 기여가 인정될 수 있다.

본 연구는 향후 다음의 측면에서 보완이 필요하다. 우선 본 연구는 문서 분류의 정확도를 비교하는 방법으로 사전의 품질을 간접적으로 평가하였다. 하지만 분류의 정확도는 용어 사전뿐 아니라 다른 여러 변수의 영향도 받으므로, 높은 분류 정확도를 나타내는 사전이 반드시 좋은 품질을 갖는 사전이라고 할 수는 없다. 따라서 향후 사전의 품질을 측정하기 위한 더욱 직접적이고 엄밀한 척도가 개발될 필요가 있다. 또한 본 연구의 수행 과정 중 용어 사전의 생성, 토픽 모델링, 분류 모델 생성, 그리고 정확도의 평가 과정에서 매우 많은 시간과 노력이 소요되었으므로, 향후 다양한 환경에서 더욱 방대한 양의 추가 실험을 수행하기 위해 실험의 많은 부분이 자동화될 필요가 있다. 고유 용어를 정의할 때 카테고리의 증가에 따라 몇 번 이하로 출현하는 용어를 고유 용어의 범주에 넣을 것인지, 사전의 개수에 따른 변화에 대한 부분에 대한 정교화가 필요하며 기본적인 정확도가 다소 낮게 나타나는 데이터에 대해 고유율의 변화에 따라 정확도가 개선되는 정도를 측정해 볼 필요가 있다. 본 연구에서는 2014년의 뉴스 기사를 사용했으나, 최신성을 반영하기 위해 향후 연구에서는 최근 데이터로 실험을 수행할 필요가 있다.

참고문헌

1. 강상욱・김민호・권혁철・전성규・오주현(2015), “세종 전자사전과 한국어 어휘의미망을 이용한 용언의 어의 중의성 해소,” 정보과학회 컴퓨팅의 실제 논문지, 21(7), 500-505.
2. 곽수정・김보겸・이재성(2013), “한국어 형태소 분석을 위한 효율적 기분석 사전의 구성 방법”, *정보처리학회논문지, 소프트웨어 및 데이터 공학*, 2(12), 881-888.
3. 김남규・이동훈・최호창(2017), “텍스트 분석 기술 및 활용 동향”, *한국통신학회논문지*, 42(2), 471-492.

4. 김민철・심규승・한남기・김예은・송민(2013), “트위터 상의 악의적 이용 자동분류”, *한구문헌정보학회지*, 47(1), 269-286
5. 김정수・이석준(2015), “주식시장관리제도와 소셜 미디어의 역할-개인 투자자 집단 유형과 토픽 분석”, *경영과 정보연구*, 34(5), 23-47.
6. 김정수・이석준(2016), “취업준비생 토픽 분석을 통한 취업난 원인의 재탐색”, *경영과 정보연구*, 35(1), 85-116
7. 김창식・최수정・곽기영(2017), “토픽모델링과 시계열회귀분석을 활용한 정보시스템분야 연구동향 분석,” 한국디지털콘텐츠학회 논문지, 18(6), 1143-1150.
8. 김태훈・손미애(2017), “문서 클러스터를 위한 워드넷기반의 대표 레이블 선정 방법”, 인*터넷정보학회논문지*, 18(2), 61-73.
9. 박주섭・홍순구・김종원(2017), “토픽모델링을 활용한 과학기술동향 및 예측에 관한 연구,” 한국산업정보학회논문지, 22(4), 19-28.
10. 박준석・김창식・곽기영(2016), “텍스트마이닝과 소셜네트워크분석 기법을 활용한 호텔분야 연구동향 분석”, *관광레저연구*, 28(9), 209-226.
11. 배상준・고영중(2010), “한국어 위키피디아를 이용한 분류체계 생성과 개체명 사전 자동 구축”, *정보과학회논문지: 컴퓨팅의 실제 및 레터*, 16(4), 492-496.
12. 배정환・손지은・송민(2013), “텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석”, *지능정보연구*, 19(3), 141-156.
13. 송종석・이수원(2011), “상품평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동 구축”, *정보과학회논문지: 소프트웨어 및 응용*, 38(3), 157-168.
14. 안정국・김희웅(2015), “집단지성을 이용한 한글 감성어 사전 구축”, *지능정보연구*, 21(2), 49-67.
15. 윤애선・황순희・이은령・권혁철(2009), “한국어 어휘의미망 [KorLex 1.5]의 구축”, *정보과학회논문지: 소프트웨어 및 응용*, 36(1), 92-108.
16. 이상훈・최정・김종우(2016), “영역별 맞춤형 감성사전 구축을 통한 영화리뷰 감성분석”, *지능정보연구*, 22(2), 97-113.
17. 조정태・최상편(2015), “영화리뷰 감성 분석을 통한 평점 예측 연구”, *경영과 정보연구*, 34(3), 161-177.
18. 최석재・권오병(2014), “빅데이터 분석을 위한 한국어 SentiWordNet 개발 방안 연구,” 한국전자거래학회지, 19(4), 1-19.
19. 최성이・김남규(2014), “토픽 분석을 활용한 웹 카테고리별 방문자 관심 이슈 식별 방안”, *한국데이타베이스*, 21(4), 415-429
20. 홍진성・김남규・이상원(2014), “단일 카테고리 문서의 다중 카테고리 자동확장 방법론”, *지능정보연구*, 20(3), 77-92.
21. Amensisa, A. D., Patil, S. and Agrawal, P.(2018), “A survey on text document categorization using enhanced sentence vector space model and bi-gram text representation model based on novel fusion techniques”, *2018 2nd International Conference on Inventive Systems and Control(ICISC)*, 218-225
22. Blei, D. M., Ng, A. Y. and Jordan, M. I.(2003), “Latent dirichlet allocation”, *Journal of Machine Learning Research*, 3, 993-1022.
23. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R.(1990), “Indexing by latent semantic analysis”, *Journal of the American Society*

for Information Science, 41(6), 391-407.

24. Fellbaum, C.(1998), "A semantic network of english: the mother of all WordNets", *Computers and the Humanities*, 32, 209-220.
25. Gao, J. B., Zhang, B. W. and Chen, X. H.(2015), "A WordNet-based semantic similarity measurement combining edge-counting and information content theory", *Engineering Applications of Artificial Intelligence*, 39, 80-88.
26. Gupta, V. and Lehal, G. S.(2009), "A survey of text mining techniques and applications", *Journal of Emerging Technologies in Web Intelligence*, 1(1), 60-76.
27. Hearst, M. A.(1999), "Untangling text data mining", *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 3-10
28. Hong, L. and Davison, B. D.(2010), "Empirical study of topic modeling in twitter", *In Proceedings of the First Workshop on Social Media Analytics*, 80-88.
29. Hotho, A., Nürnberger, A. and Paaß, G.(2005), "A brief survey of text mining", *In Ldv Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20(1), 19-62.
30. Joachims, T.(1998), "Text categorization with support vector machines: Learning with many relevant features", *In European Conference on Machine Learning*, 137-142.
31. Miller, G. A.(1995), "WordNet: A lexical database for English", *Communications of the ACM*, 38(11), 39-41.
32. Mooney, R. J. and Bunescu, R. C.(2006), "Subsequence kernels for relation extraction", *In Advances in Neural Information Processing Systems*, 171-178.
33. Pedersen, T., Patwardhan, S. and Michelizzi, J.(2004), "WordNet:: Similarity: measuring the relatedness of concepts", *In Proceedings of the 5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 38-41.
34. Richman, A. E. and Schone, P.(2008), "Mining wiki resources for multilingual named entity recognition", *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1-9.
35. Rijsbergen, C. J. V., *Information Retrieval*, 2nd edition, Butterworths, 1979.
36. Rogati, M. and Yang, Y.(2002), "High-performing feature selection for text classification", *In Proceedings of the 11th International Conference on Information and Knowledge Management*, 659-661.
37. Salton, G. and McGill, M. J. Introduction to *Modern Information Retrieval*, McGraw-Hill, 1983.
38. Sebastiani, F.(2006), "Classification of text, automatic", *The Encyclopedia of Language and Linguistics*, 14, 457-462.
39. Vapnik, V. N., *The Nature of Statistical Learning Theory*, Springer, 1995.
40. Wang, C. and Blei, D. M.(2011), "Collaborative topic modeling for recommending scientific articles", *In Proceedings of the 17th ACM SIGKDD International Conference*

on Knowledge Discovery and Data Mining, 448-456.

41. Wei, T., Lu, Y., Chang, H., Zhou, Q. and Bao, X.(2015), "A semantic approach for text clustering using WordNet and lexical chains", *Expert Systems with Applications*, 42(4), 2264-2275.

Abstract

Analyzing the Effect of Characteristics of Dictionary on the Accuracy of Document Classifiers†

Jung, Haegang* · Kim, Namgyu**

As the volume of unstructured data increases through various social media, Internet news articles, and blogs, the importance of text analysis and the studies are increasing. Since text analysis is mostly performed on a specific domain or topic, the importance of constructing and applying a domain-specific dictionary has been increased. The quality of dictionary has a direct impact on the results of the unstructured data analysis and it is much more important since it present a perspective of analysis. In the literature, most studies on text analysis has emphasized the importance of dictionaries to acquire clean and high quality results. However, unfortunately, a rigorous verification of the effects of dictionaries has not been studied, even if it is already known as the most essential factor of text analysis. In this paper, we generate three dictionaries in various ways from 39,800 news articles and analyze and verify the effect each dictionary on the accuracy of document classification by defining the concept of Intrinsic Rate. 1) A batch construction method which is building a dictionary based on the frequency of terms in the entire documents 2) A method of extracting the terms by category and integrating the terms 3) A method of extracting the features according to each category and integrating them. We compared accuracy of three artificial neural network-based document classifiers to evaluate the quality of dictionaries. As a result of the experiment, the accuracy tend to increase when the “Intrinsic Rate” is high and we found the possibility to improve accuracy of document classification by increasing the intrinsic rate of the dictionary.

Key Words: Text Mining, Start List, Document Classification, Topic Modeling, Intrinsic Rate

† This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A5A2A03067632)

* Graduate School of Business IT, Kookmin University, k-kang@kookmin.ac.kr

** Associate Professor, School of Management Information Systems, Kookmin University, ngkim@kookmin.ac.kr