Journal of The Korea Society of Computer and Information
Vol. 23 No. 10, pp. 103-109, October 2018
www.ksci.re.kr
https://doi.org/10.9708/jksci.2018.23.10.103

Dynamic Text Categorizing Method using Text Mining and Association Rule

Young-Wook Kim*, Ki-Hyun Kim*, Hong-Chul Lee*

Abstract

In this paper, we propose a dynamic document classification method which breaks away from existing document classification method with artificial categorization rules focusing on suppliers and has changing categorization rules according to users' needs or social trends. The core of this dynamic document classification method lies in the fact that it creates classification criteria real-time by using topic modeling techniques without standardized category rules, which does not force users to use unnecessary frames. In addition, it can also search the details through the relevance analysis by calculating the relationship between the words that is difficult to grasp by word frequency alone. Rather than for logical and systematic documents, this method proposed can be used more effectively for situation analysis and retrieving information of unstructured data which do not fit the category of existing classification such as VOC (Voice Of Customer), SNS and customer reviews of Internet shopping malls and it can react to users' needs flexibly. In addition, it has no process of selecting the classification rules by the suppliers and in case there is a misclassification, it requires no manual work, which reduces unnecessary workload.

▸Keyword: Document Categorizing, Topic Modeling, Association Rule, Hierarchical Clustering

I. Introduction

정보통신기술이 발달함에 따라 대용량 데이터의 저장 및 관리가 원활하게 되었고, 개인용 컴퓨터와 스마트폰의 보급으로 사용자는 장소와 시간에 구애받지 않고 인터넷을 사용할 수 있게 되었다.

이에 따라 과거엔 사회구성원 간에 지식공유와 소통에 폐쇄적이었던 반면, 트위터, 페이스북 등 소셜네트워크서비스(Social Network Services : SNS)와 카카오톡, 라인, 위챗, 스카이프 등 메신저(Messenger)를 이용해 정보를 양방향으로 공유할 수 있는 매개역할을 하고 있다[1].

오프라인으로만 이루어졌던 산업서비스도 스마트 기기를 이용해 유통, 금융, 제조, 교통, 물류 등 산업 전반에 걸쳐 소비자와 소통할 수 있는 채널이 다양화되고 온라인으로 제품을 구매하는 비대면 채널이 확장되고 있다[2].

인터넷을 활용한 비대면 서비스의 다양화는 텍스트데이터의 증가로 이어지고, 처리해야할 정보와 문서의 양이 점점 방대해지고 복잡해질 뿐만 아니라 데이터 처리와 저장에 많은 비용을 소비하고 있다[3].

최근 비정형화된 텍스트데이터의 비율의 증가로 텍스트 마이닝의 중요성이 커지고 있으며, 텍스트 마이닝 기술은 크게 정보검색(information retrieval), 기계학습(machine learning), 시맨틱 웹(semantic web), 자연어처리(natural language processing)가 있다[4].

텍스트 마이닝은 문서를 자동 분류하는데 사용되며, 사전의 사용 유무에 따라 문서 클러스터링(Document Clustering) 방식과 문서 분류(Document Classification) 방식으로 나뉜다[5]. 문서 클러스터링은 분류 기준 없이 문서 간 거리를 계산하여 연관된 문서들을 반복적으로 군집화하는 방법이고, 문서 분

• First Author: Young-Wook Kim, Corresponding Author : Hong-Chul Lee
*Young-Wook Kim (kimyw2@korea.ac.kr), School of Industrial and Management Engineering, Korea University
*Ki-Hyun Kim (kingyou@korea.ac.kr), School of Industrial and Management Engineering, Korea University
*Hong-Chul Lee (hclee@korea.ac.kr), School of Industrial and Management Engineering, Korea University
• Received: 2017. 11. 20, Revised: 2017. 12. 02, Accepted: 2017. 12. 02.
• This paper is the result of the research carried out by the Ministry of SMEs and Startups(MMS) Business Administration's (2016) industry-university cooperation technology development project(the first step project)(No : C0397738)

류는 분류 기준이나 잣대가 되는 문서를 기반으로 범주화하는 기법이다[3].

이 중 토픽모델링(Topic Modeling)은 정형화되지 않은 구조의 문서들을 클러스터링 방식을 이용하여 주제를 도출하는데 사용되었다[3]. 이런 특징을 이용하여 특허문서를 분석해 최근 특허출원 트렌드를 파악하거나, 특정 분야의 논문을 단서로 최근 연구동향을 파악하는 등 다양한 분야의 분석 도구로 사용되고 있다.

LDA(Latent Dirichlet Allocation)는 토픽모델링 기법 중 가장 많이 활용되는 연구방법으로 단어와 개념 사이의 관계를 수치화하는 LSI(Latent Semantic Index)와 단어와 문서 간의 관계를 확률모형으로 계산한 pLSA(probabilistic Latent Semantic Analysis)에서 발전된 기법으로 확률모델을 이용하여 주요 토픽을 파악, 토픽 간의 관계를 나타내는 모형이다[4].

LDA는 SNA(Social Network Analysis)를 이용해 도출된 키워드 간의 관계를 분석하거나[6], 상승하는 키워드(Hot topic)와 하강하는 키워드(Cold topic)를 찾아내어 문헌의 동향을 파악한다[7].

이처럼 LDA와 같은 토픽 모델링은 문헌의 주제를 나타내기엔 좋은 모델이지만 토픽 간의 연관성은 알 수 없기 때문에 도출된 주제 키워드의 관계를 설명하고 시각화 할 수 있는 SNA 기법을 함께 활용한다.

본 논문에서는 기존의 공급자 중심의 문서분류를 위한 인위적인 범주규칙을 탈피하고, 사용자 니즈 혹은 트랜드에 따라 변하는 동적 문서 분류방법을 제안해 텍스트 문서의 처리 및 저장 비용을 줄이고 SNS나 메신저, 웹사이트의 댓글에서 발생되는 비격식 텍스트데이터의 분석에도 활용하고자 한다.

논문의 구성은 다음과 같다. 2장에서는 본 논문과 관련된 문서분류의 개요와 실제 연구에 사용한 LDA와 연관성분석에 대해 소개하고 3장에서는 앞서 설명된 두 가지 기법을 이용한 실험 설계 내용에 대해 기술한다. 4장에서는 3장의 설계 내용을 바탕으로 실험을 수행하여 결과를 분석하고 5장에서는 본 연구의 결론과 향후 연구방향을 제시한다.

II. Related Works

1. Document Classification

문서분류는 문서를 대표하는 특징으로 구성된 키워드와 유사한 문서들을 같은 그룹으로 분류하는 기법으로 대용량 텍스트 문서를 관리하는 방안으로 연구되어 왔다[3]. 서론에서 기술한 것처럼 사전에 정의된 분류기준에 따라 특정한 범주에 유사한 문서를 분류하는 것을 문서 분류(Document Classification), 문서 간 유사도 거리에 따라 연관된 문서를 반복적으로 군집화하는 과정을 문서 클러스터링(Document Clustering)이라 한다.

사전 정의된 분류 기준의 경우도 각 문서에 나타나는 단어의 빈도에 대한 정보를 추출해 분류에 반영하는 통계적인 방법과 기업의 정보관리 담당자 같은 전문가가 문서의 내용을 바탕으로 기준을 정하는 지식기반 분류 방법이 있다. 통계적인 방법은 주로 베이지안 확률(Bayesian Probability)모델을 이용하여 각 범주에 문서가 속할 확률을 계산하는 방법이 활용되고[7] 신경망, 의사결정나무, SVM(Support Vector Machine)과 같은 기계학습 알고리즘도 사용된다[8]. 지식기반 분류 방법은 몇 개의 문서 샘플을 분석한 후 분류 규칙을 만드는 방법과 문서 내용 외에 전문가의 의견이 반영된 방법이 있다.

2. Topic Modeling

토픽 모델링은 비구조화 된 대용량의 문헌에서 주제를 도출하기 위한 방법으로 유사한 의미를 가진 단어들을 군집화하는 방식으로 문헌의 주제를 파악하는 알고리즘이다[9,10].

토픽 모델링 알고리즘 중 하나인 LDA(Latent Dirichlet Allocation)은 pLSI(Probabilistic Latent Semantic Index)가 발전된 문헌 모델로 이전 모델과는 다르게 각 주제들의 분포로 문헌을 계산해 토픽을 생성하는 사후 추론 방법으로[4] 문서 내 용어들의 존재 여부가 중요하다고 가정하여 방대한 양의 문서에서 Dirichelt 확률 분포를 사용해 유사한 키워드를 군집하는 방법으로 전체 문헌의 주제를 도출하는 모델이다[9].

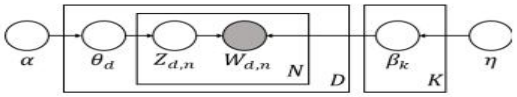

Fig. 1. LDA Architecture

위 그림을 보면 α, β는 코퍼스 단위로 정해지고 N과 Θ는 문서 단위로 값이 정해진다. β는 각 주제별로 특정 단어가 생성될 확률이 담긴 테이블 이며, N은 문서의 길이, Θ는 해당 문서에서 각 주제의 가중치를 나타낸다. Z는 문서의 I번째 단어에 대한 주제벡터이다. 이 모델에서는 주제 개수가 K로 고정되어 있으며, 따라서 Θ와 Z는 길이가 K인 벡터이다[4].

Table 1. Description of LDA model

Sgn	Presentation
K	Topic
α	Dirichlet parameter
η	Topic hyperparameter
θ_d	Per-document topic proportions
β_k	Per-corpus topic distributions
$Z_{d,n}$	Per-word topic assignment
$W_{d,n}$	Observed word

문서분류에 관련된 연구는 일반인이 이해하기 어려운 법원 판례문서에 LDA를 이용하여 주제어를 도출하고 코사인 유사도로 유사 판례 범주를 분류하는 시스템을 제안하였고[11], 논문,

특허, 뉴스기사 등 정제된 글이 아닌 자유롭게 기술되는 비격식 텍스트 데이터의 분류를 위해 LDA의 단어 분포를 활용해 상위 랭크된 단어를 기준으로 다른 단어를 분해, 병합하는 방법으로 교정해 모델의 성능을 개선하였다[12].

본 논문에서는 LDA를 이용해 각 범주의 상위 주제가 되는 후보 키워드를 생성하고, 반복적인 실험을 통해 주제를 선정하였다.

3. Association Rule

연관성 분석은 자료에 존재하는 항목 간에 if-then 형식의 연관규칙을 찾는 비지도 학습법으로 웹사이트에서 손님의 장바구니를 분석하거나 기업의 데이터베이스에서 유효한 규칙을 찾아 감사(Audit)시스템에서의 부정거래 탐지(FDS : Fraud Detection System), 신용카드사에서 대출판매를 위한 마케팅 모델링, 쇼핑몰에서 고객의 장바구니를 분석한 추천시스템 등 여러 가지 문제에 적용되고 있다[13].

연관규칙은 지지도(support)와 신뢰도(confidence)를 사용하여 적절한 X와 Y를 선택하게 된다. 지지도는 빈발하게 나타나는 항목을 고려하기 위한 척도로 X, Y 동시에 포함하는 수의 비율을 말하며 아래와 같이 표현된다[3].

$$\mathrm{Supp}(R) = P(X \cap Y) = \frac{n(X \cap Y)}{N}$$

Fig. 2. Support of association rule

신뢰도는 발견된 규칙의 영향력을 나타내는 척도로 X가 발생할 때 Y도 동시에 발생할 확률을 나타낸다. 즉, 신뢰도가 높은 규칙일수록 의미가 크다[3].

$$\mathrm{Conf}(R) = P(Y \mid X) = \frac{P(X \cap Y)}{P(X)}$$

Fig. 3. Confidence of association rule

문서분류에서의 기존 연구는 뉴스기사를 이용해 범주 내의 문서들 간에 연관성 있는 키워드들의 집합을 추출하고 각 범주별로 의미적으로 대표성을 지닌 키워드를 계층적으로 분류했고[3], 연관규칙과 유전 알고리즘을 이용해 연관단어 지식베이스를 만들어 문서 분류의 정확도를 높였다[13].

본 논문에서는 토픽모델링을 이용해 추출된 키워드를 계층적으로 분류하기 위해 연관규칙을 찾는 대표적인 방법인 Apriori 알고리즘을 사용하였다.

III. Dynamic Text Categorizing Method

본 논문은 LDA와 연관성 분석을 적용한 동적 문서분류 방법을 제안한다. 먼저, LDA 분석을 통해 후보 키워드를 선정하고 Association rule을 통해 단어의 관계를 설정하는 방법으로 진행했다.

이 장에서는 첫 번째로 본 연구에서 제안한 연구의 프로세스를 살펴보고, 두 번째로 실험환경에 대해 설명 후 마지막으로 동적 문서분류의 결과를 분석한다.

1. Process

그림 4는 동적 문서 분류 연구를 위한 프로세스이다. 먼저 연구에 사용된 데이터는 2016년 3월부터 2017년 3월까지 “4차 산업혁명”으로 검색한 뉴스기사를 수집한 데이터이며 총 9169개의 뉴스 데이터이다. IT_과학 2093개, 경제 3489개, 국제 263개, 문화 691개, 사회 516개, 정치 1393개, 지역 724개로 구분되어 있다. 수집된 데이터 중 기사원문을 제외한 기사작성날짜(또는 수정날짜), 언론사정보, 기자정보 등은 삭제하였고 언론사마다 대·중·소분류 방식이 차이가 남으로 대분류를 제외한 중·소분류는 제거한 상태에서 실험하였다.

두 번째로 Java개발 환경에서 KOMORAN 2.0을 이용해 형태소 분석을 시행하였다. 기존 형태소 분석기와는 달리 여러 어절을 하나의 품사로 분석 가능함으로써 공백이 포함된 고유명사를 더 정확하게 분석할 수 있다. 아울러 불용어와 특수문자, 숫자 등을 제거하였다.

세 번째로 대표 키워드를 선정하기 위해 python의 gensim package를 이용해 파라미터 값을 달리하여 실험을 진행하였다.

네 번째로 연관규칙을 이용해 도출된 단어 간의 관계를 계산하고 계층적으로 분류하였다.

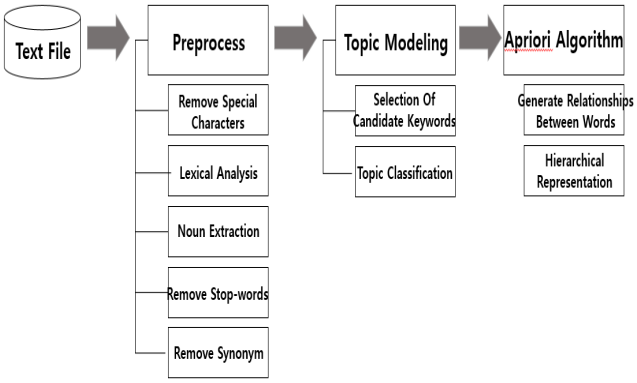

Fig. 4. Dynamic Text Categorizing Process

2. Environment

연구에 사용된 데이터와 테스트 환경은 1절과 같으며, 후보 키워드 생성을 위해 Python의 gensim package를 이용하여 LDA를 실행하였고 R의 apriori 함수를 이용해 단어 간의 관계를 계층적으로 구분하였다.

3. Result

3.1 Generate Keywords

해당 뉴스 데이터에서는 뉴스 기사 원문만을 사용하였으며 뉴스기사의 단어를 추출하기 위해 형태소 분석기 KOMORAN을 이용하여 명사형태의 형태소만을 추출하였다. 형태소 분석

기와 전처리 과정을 통해 총 101,336개의 단어를 추출하였으며 이를 Document Term Matrix 형태로 구성하였다. 단어를 추출하는 과정에서 검색어로 사용되었던 "4차 산업혁명"은 모든 문서에서 관찰되었기 때문에 관련된 단어들을 제거하였다.

후보키워드를 선정하는 과정에서 사용한 LDA모델은 python의 gensim package를 이용하여 실험하였으며 기사 셋에 나타나는 주제의 개수와 하나의 주제가 포함하는 단어의 개수는 반복적인 실험을 통해 15개의 주제로 재분류 하고 10개의 주요단어를 선정 하였다. 주요단어로 선정된 150개의 단어에서 특정시기를 언급한 단어 및 중복된 단어, 인물, 화폐단위 등을 제거하고 142개의 단어를 주요단어로 선정하였다.[Table 1] 선정된 주요단어를 후보키워드 사전으로 등록한 이후에 Document Term Matrix를 재구성 하였다. 이를 통하여 각 기사에 키워드를 설정했다.

Table 2. Using LDA, we classify them into 15 topics, extracting 10 key words for each topic, and then preprocessing them.

Topic	Keyword
0	국민의당,새누리당,더불어민주당,일자리,바른정당
1	산업부,지식재산,광주,ict,산업통상자원부,특허청,친환경, 경쟁력
2	충북,세종,세종시,하이닉스,청주,반도체,한전,성균관대
3	대구,정보통신,인공지능,신산업,시상식,산업용 로봇
4	미국,중국,한국,일본,불확실성,기업,경제,경쟁력,투자자
5	3d,상상력,창원,원내대표,프린터,한국,중국,프린팅
6	부총리,울산시,울산,한국은행,저출산,일자리,기획재정부,개인정보,근로자
7	20대국회,제조업,스마트공장,중소기업,경쟁력,서비스업,독일,팀장,제조업 혁신
8	창조경제,청와대,국회의원,국무총리,기득권
9	인공지능,알파고,클라우스,다보스포럼,다보스,세계경제포럼,슈밥,스위스,일자리
10	미국,구글,자동차,인공지능,한국,증강현실,실리콘밸리,스타트업,가상현실,블록체인
11	부산,미래창조과학부,미래부,전문가,기업인,연구개발,서울시
12	미래,산업,기업,경제,시대,주제,교육,변화,혁신
13	학생,교육부,교육과정,학부모,아이들,인재 양성,산학협력,협의회,대학원,지역사회
14	자율주행차,인공지능,빅데이터,사물인터넷,클라우드

3.2 Select Keyword and Create keyword Relation

후보키워드 사전을 이용하여 재구성된 Document Term Matrix를 이용하여 단어의 관계를 설정하기 위하여 R의 aruls library를 이용하여 Apriori 알고리즘을 사용했다. 본 연구에서는 Apriori알고리즘을 제한적으로 사용한다. Apriori 알고리즘의 Left-hand side을 제한함으로 써 특정 단어가 다른 단어와의 관계를 보기 위해서다. Left-hand side를 제한하기 위해서 대표키워드를 선정해야 한다. 본 연구에서는 Document Frequency를 이용하여 DF값이 0.1에 근접한 5개의 후보키워드를 대표 키워드로 선정 하였다[그림 5]. DF값은 단어 t가 포함된 문서의 수를 전체 문서의 개수 D로 나누는 값이다. 선정하는 과정에서 국가명은 제외하였다.[Table 2] 대표키워드는 단어 관계에서 1차 구조가 된다. 즉, "4차 산업혁명"과 대표키워드의 관계로 볼 수 있다.

Table 3. Words selected as a representative keyword using Document Frequency.

Keyword	Document Frequency	Value
인공지능	2218	0.241902
일자리	1321	0.144072
경쟁력	1024	0.111681
빅데이터	910	0.099247
전문가	897	0.09783

2차 구조를 만들기 위해서 Apriori 알고리즘의 Left-hand side을 대표키워드인 "인공지능"으로 제한한 결과 값을 사용한다. [Table 4]에서 Right-hand side가 2차 구조의 후보키워드가 되며 support 값이 높은 값이 2차 대표키워드가 된다. 단, confidence값이 높으면 복합 명사일 가능성이 있기 때문에 복합명사는 제외 할 수 있도록 한다. 3차 구조를 만들기 위해서는 Left-hand side를 2개의 키워드로 고정한 이후에 2차 구조를 만드는 방법을 반복한다. count는 문서의 개수를 의미 한다.[Table 5]

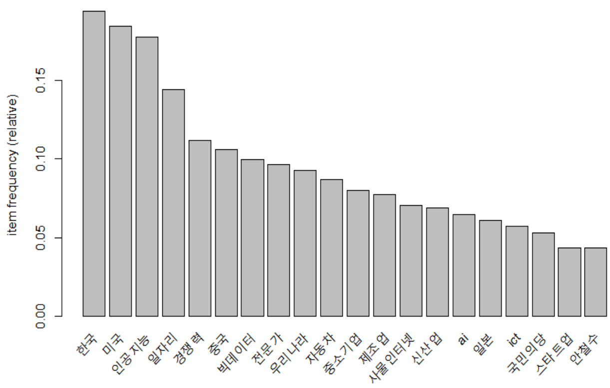

Fig. 5. Keyword comparisons related to the Fourth Industrial Revolution.

Table 4. The result of limiting the left-hand side of the Apriori algorithm to "Artificial Intelligence"

```
{인공지능}
lhs rhs support confidence
lift count
[1] {인공지능} => {미국} 0.041443996 0.2337023
1.269441 380
[2] {인공지능} => {빅데이터} 0.039480859 0.2226322
2.243203 362
[3] {인공지능} => {사물인터넷} 0.037408660 0.2109471
2.998719 343
```

Table 5. The result of limiting the left-hand side of the Apriori algorithm to {US, Artificial Intelligence}, {Big Data, Artificial Intelligence}, {Internet of things, Artificial Intelligence}

```
{미국,인공지능}
lhs            rhs            support            confidence
lift       count
[1] {미국,인공지능} => {한국} 0.011124441 0.2684211
 1.386565 102
[2] {미국,인공지능} => {중국} 0.009379431 0.2263158
 2.132672  86
[3]    {미국,인공지능}     =>     {빅데이터}0.008725052
0.2105263    2.121226   80
[4]    {미국,인공지능}    =>{사물인터넷}    0.008397862
0.2026316    2.880510   77

{빅데이터,인공지능}
lhs            rhs            support            confidence
lift       count
[1] {빅데이터,인공지능} => {사물인터넷} 0.013305704
0.3370166    4.790860 122
[2] {빅데이터,인공지능}  => {미국}         0.008725052
0.2209945    1.200414   80

{사물인터넷,인공지능}
lhs            rhs            support            confidence
lift       count
[1] {사물인터넷,인공지능} => {빅데이터}    0.013305704
0.3556851    3.583821 122
[2]{사물인터넷,인공지능} => {미국}         0.008397862
0.2244898    1.219400   77
```

그림 6을 보면 인공지능과 빅데이터 두 개의 키워드로 연관성 분석을 실시한 결과이다. 그림 7을 보면 인공지능과 미국을 주제로 연관성 분석을 실시하였고, 그림 8은 인공지능, 사물인터넷을 주제로 연관성 분석을 실시한 그래프이다. 각기 선택되는 키워드에 따라 그룹핑이 다르게 되며, 범주화 개수도 변경됨을 알 수 있다.

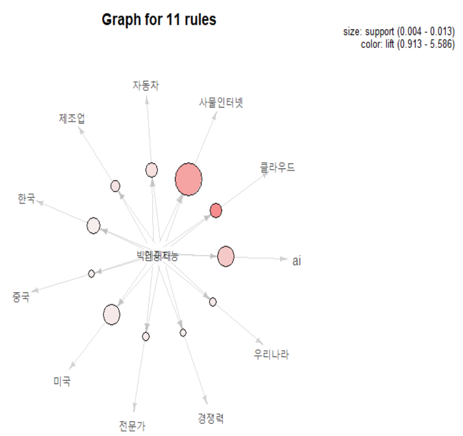

Fig. 6. "Artificial Intelligence", "Big Data" Keyword association rule graph

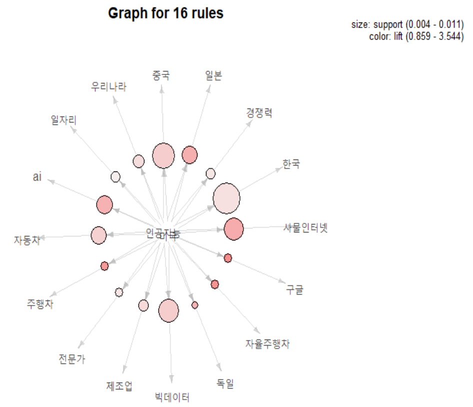

Fig. 7. "Artificial Intelligence", "US" Keyword association rule graph

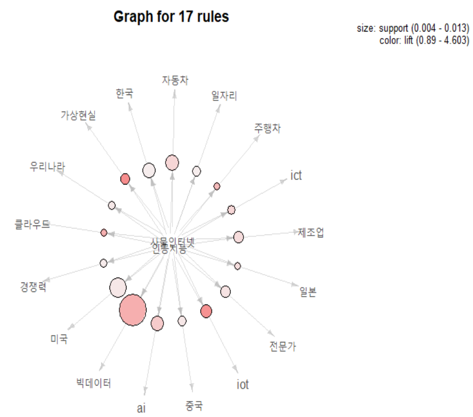

Fig. 8. "Artificial Intelligence", "Internet of things" Keyword association rule graph

Ⅳ. Conclusions

본 논문은 토픽모델링 기법 중 하나인 LDA와 데이터마이닝 기법 중 통계적인 방법이 아닌 연관성 분석을 이용해 분류규칙을 생성하였다. 또한 기존의 정형화된 범주에서 벗어나 사용자의 니즈에 맞춰 범주화 할 수 있도록 고안하였다.

연구한 알고리즘을 통해 토픽 결과를 네트워크 구조로 만들어 특허를 출원하였다. LDA 분석 결과인 6개의 토픽과 각 토픽의 키워드를 통해 DW(Data Warehouse)를 구성한다. DW는 네트워크 구조로 이루어지며 키워드가 노드, 키워드 간의 거리가 아크가 된다. 노드의 크기는 해당 키워드의 TF*DF 값으로 결정하고, 노드 간의 거리를 의미하는 아크의 길이는 전체 문서 중 두 키워드가 동일 문서에 언급되는 비율을 의미하는 Support Value의 역수로 계산한다. 특정 문서에서만 언급되는 단어를 중요단어로 판단하는 기존 TF*IDF 방식과 달리, 본 알고리즘에서는 개별 문서에서도 많이 언급되며 많은 문서에서도 자주 언급되는 단어를 보다 중요한 단어로 판단한다. 노드의 크기와 아크의 길이가 결정되면 가장 큰 노드를 중심 노드로 삼고 중심 노드와의 거리가 가깝고 노드의

크기가 큰 노드를 후보 노드로 선정한다. 다음으로는 중심 노드와 선정한 후보 노드를 통합하여 새로운 하나의 중심 노드로 삼고 다시 타 노드와의 거리를 계산한다. 이 과정을 키워드의 수만큼 반복한 뒤 종료한다. 확장을 종료하면 Hierarchy 구조의 최종 DW가 생성된다.

이 알고리즘은 비대면 채널이 많은 콜센터나 쇼핑몰에서 텍스트 데이터로 고객 동향이나 이슈사항을 분석할 때나 포털사이트에서 연관검색어의 도출방식을 새롭게 고안할 때 사용될 수 있을 것이다.

연구의 한계점은 각 키워드마다 연관성 분석을 실시하다보니 연산과정에서 시간과 자원이 많이 소요된다. 단어와 단어 간의 관계를 계산하는 방법에서 다른 알고리즘의 적용도 고려해봐야 할 것이다.

연구를 하며 보완해야할 점은 LDA를 이용해 범주화될 후보 키워드를 추출할 때 파라미터의 문제이다. Python의 Gensim의 경우 파라미터 값을 Auto로 설정 후 실험하였지만, 자동분류를 위해선 최적화된 파라미터를 선택하는 방안이 제시되어야 한다. 두 번째로 범주화에서 복합명사가 대표 범주키워드로 들어가 분류가 이질적일 수 있다. 이 역시 최적화된 기준을 찾는 방법이 제시되어야 할 것이다.

향후 연구방안은 타겟이 되는 문서에 유사한 문서를 찾아내는 문제에 대해 제안된 알고리즘과 HLDA(Hierarchical LDA)의 성능비교를 진행할 것이다. 이는 방대한 양의 문서를 검색할 때 얼마나 유사도 높은 문헌을 추천해 줄 수 있는지 정확도에 대한 비교로 문서 분류 작업을 더욱 정교하게 만들 수 있을 것이다.

REFERENCES

[1] Joo. J, “Mediating Effects of Swift Trust on Konwledge Sharing in Social Network Services” Korean Academic Society of Business Administration, Vol.43 No.3, pp589-612, Jun. 2014.
[2] NIA, IT Future Strategy report No.9, pp1-25, Dec. 2014.
[3] Joo. K., Shin. E., Lee. J., Lee. W., "Hierarchical Automatic Classification of News Articles based on Association Rules" Journal of Korea Multimedia Society Vol.14 No.6, pp. 730-741, Jun. 2011.
[4] Song. M, ”Text Mining”, Aug. 2017.
[5] Hwang. H, Lee. J, ”A study of a Knowledge Inference Algorithm using an Association Mining Method based on Ontologies” Journal of Korea Multimedia Society Vol.11 No.11, pp. 1566-1574, Nov. 2008.
[6] Kim. H, Rhee. H, ”Trend Analysis of Data Mining Research Using Topic Network Analysis” Journal of The Korea Society of Computer and Information Vol. 21 No.5, pp.141-148, May. 2016.
[7] Ko. S., Lee. J., "Weighted Bayesian Automatic Document Categorization Based on Association Word Knowledge Base by Apriori Algorithm" Journal of Korea Multimedia Society Vol.4 No.2, pp. 171-181, Apr. 2001.
[8] Y. Yang, J. O. Pederson, "A Comparative study on feature selection in text categorization" In Proceedings of the 14th International Conference on Machine Learning Jul. 1997.
[9] Blei, D. M., "Probabilistic topic models" Communications of the ACM, pp. 77-84, Apr. 2012.
[10] Lee. H, Yang. S, Ko. Y, "Feature Expansion based on LDA Word Distribution for Performance Inprovement of Informal Document Classification" Journal of KIISE, Vol.43 No.9, PP. 1008-1014, Sept. 2017.
[11] Sim. J, Kim. H, "A Searching Method for Legal Case Using LDA Topic Modeling" Journal of The Institute of Eletronics and Information Engineers Vol.54 No.9, Sept. 2017.
[12] Park. J, Song. M, ”A Study on the Research Trends in Library & Information Science in Korea using Topic Modeling” Journal of Korea Society for Information Management, Vol.30 No.1, pp. 7-32, Mar. 2013.
[13] Park. C, Kim. Y, Kim. J, Song. J, Choi. H, “Data Mining using R”, Jul. 2011.

Authors

Young-wook Kim was born in Seoul, Korea in 1981. He received the M.S degrees in School of Industrial and Management Engineering at Korea University, Korea in 2018. Currently He works as data analyst at Penta systems in Seoul. His research interests are machine learning and text mining.

Ki-hyun Kim was born in Seoul, Korea in 1984. He finished the coursework in School of Industrial and Management Engineering at Korea University, Korea in 2015.He is currently working as a researcher in School of Industrial and Management Engineering at Korea University, since 2011. His research interests are Management Information System(MIS) and Big Data analysis and data minig.

Hong-chul Lee was born in Seoul, Korea in 1960. He obtained the Ph.D. in School of Industrial Engineering at Texas A&M University, USA in 1993He is currently working as a professor in School of Industrial and Management Engineering at Korea University, Korea since 1996. His research interests are production information system and logistics information system and Supply Chain Management (SCM)