「정보시스템연구」 제28권 제3호
한국정보시스템학회
2019년 9월, pp. 105~121

http://dx.doi.org/10.5859/KAIS.2019.28.3.105

TF-IDF를 활용한 한글 자연어 처리 연구

이종화* · 이문봉** · 김종원***

〈목 차〉

Ⅰ. 서 론

2016년 27,300 백만 달러에 이른 빅데이터 세계시장 규모가 매년 22% 이상의 성장세를 이어가면서 2019년은 49,000백만 달러의 전망을 내놓고 있다(wikibon, statista 2017). <표 1>은 국내 빅데이터 시장은 세계 시장보다 더 팽창하고 있는 것을 확인할 수 있다. 국내 빅데이터 시장규모는 KISTI의 마켓리포트에 의하면 매년 고성장을 이룰 것으로 전망되며 국내 빅데이터 시장 규모는 2016년 약 3,300억 원을 시작으로 2021년까지 1조 원대 규모로 성장할 것으로 보고 있다(한국과학기술정보연구원, 2017).

IDC 보고에 따르면 세계 데이터양은 2016년 16ZB에서 2025년 163ZB로 10배 증가할 전망이라고 한다(IDC, 2017). IT분야의 시장조사 및 컨설팅 전문기관인 가트너(gartner)에서 발표한 자료에 따르면 지난 10여 년간 컴퓨터 파워와 모바일 디바이스의 발달로 사회 각 영역에 빠른 확산과 시멘틱 웹(semantic web) 환경의 보급으로 데이터의 생성은 폭발적으로 늘고

<표 1> 빅데이터 시장 규모 및 전망

구분	2016	2017	2018	2019	2020	2021	CAGR
세계시장	27,300	33,500	40,800	49,000	57,300	62,000	22.10%
국내시장	333	423	539	692	894	1,134	26.90%

* 동의대학교 e비즈니스학과, jhlee6050@deu.ac.kr(주저자)
** 동의대학교 경영학과, mblee@deu.ac.kr
*** 동의대학교 경영정보학과, jokim@deu.ac.kr(교신저자)

그 양도 급속히 증가하고 있다(gartner.com, 2018). 특히 데이터의 형태면에서 레거시(legacy) 시스템에서 많이 발생하는 정형데이터보다는 SNS의 확산으로 비정형데이터 증가 폭이 매우 크다. 비정형 데이터는 텍스트 형태의 오가는 대화뿐 아니라 일상 속에서의 사진이나 동영상 등과 같이 대용량이면서 복잡한 형태이며 비구조적인 데이터라 할 수 있다. 가트너는 기업 내 빠르게 늘어나는 데이터 가운데 정형적 데이터보다는 비구조적이며 비정형인 데이터의 비중이 80%이상으로 보고 있다. 비정형데이터 중 고객의 취향, 감성, 선호도 등 고객의 니즈 분석에는 텍스트 데이터의 중요성도 함께 높아지고 있다.

수많은 텍스트 기반 비정형 데이터분석을 통해 패턴을 찾고 통찰력(insight)을 얻기 위해서는 복잡한 프로세스가 필요하다. 먼저, 고객의 니즈를 통찰하기 위해 그들의 마음을 읽어 들이기 위해서는 자료 수집(crawling)이 선행되어야 한다(김은우・금득규, 2014). 블로그, 소셜미디어, 웹 사이트 콘텐츠 등 복잡한 웹 페이지 구조를 분석하여 필요한 데이터를 수집 및 저장하는 과정이다. 두 번째는 수집된 데이터를 분석(analysis)하는 과정이 필요하다. 텍스트 분석은 표준어를 상대적으로 많이 사용하는 인터넷뉴스 이외의 인터넷어, 채팅어 등과 같이 비표준어가 상대적으로 많은 자연어 처리를 텍스트마이닝(text mining) 처리를 통하여 이슈를 발견하는 과정이다(서새남, 2017). 텍스트마이닝은 분석의 목적과 관점에 따라 키워드 정제 작업과 필요한 데이터 추출을 통해서 시각화된 결과를 도출한다. 마지막으로 데이터 분석 목적에 따른 데이터 분석(user-driven analysis)이 필요하다. 문제의 키워드 선정과 그에 따른 시각화 데이터 추출을 활용하여 의사결정에 활용하게 된다(Amado et al., 2018; 양낙영 등, 2018; Lee, 2013; 유은지 등, 2012).

특히, 한글 자연어 처리(KoNLP)는 표준어로 등록되지 않은 비속어 및 채팅어 등의 비표준어 전처리 과정이 필요하다. 비표준어를 태킹으로 자음과 모음을 분리하여 비표준어 사전과 비교하여 유사한 키워드를 구분기도 한다(An and Kim, 2015; 이종화・이현규, 2016).

우리말 한글을 이용하여 텍스트마이닝을 연구하는 연구자들은 대부분 빈도 분석을 기초하여 정보 추출(extraction), 문서 분류(classification), 문서 군집(clustering) 등 분석 방법을 활용하고 있다(Lee and Lee, 2017; 남민지 외, 2015).

TF-IDF, BOW(bag of words), N-gram, NMF(non-negative matrix factorization), Word2Vec 등의 마이닝 기법을 활용하여 의미분석 연구가 진행되고 있다(Christian et al., 2016; Zhang et al., 2010; Hoyer, 2004). 대부분 영문 기반의 연구로 진행되며 우리말 한글 기반 연구의 미진함이 보인다. 물론, 한글 기반은 영문 기반보다 감정이나 형태 표현이 자유롭게 네티즌의 축약형 표현으로 지속적으로 새로운 단어들이 만들어지고 있다. 이러한 한계점으로 한글 기반 자연어 처리의 연구자 더욱 의미 있다고 본다.

본 연구는 한글 기반 문장을 이용한 단순 빈도 위주의 분석인 TF(term frequency)기법과 문장 내 이슈 단어를 추출하여 분석하는 TF-IDF(term frequency-inverse document frequency)기법의 결과를 군집분석을 통해 비

교하고자 한다. 본 논문의 2장은 TF분석과 TF-IDF분석 그리고, 군집분석에 관하여 선행 연구를 살펴본다. 3장에서는 한글기반 SNS 연구 데이터를 TF와 TF-IDF 기법으로 결과값을 추출하기 위한 프레임워크와 분석에 필요한 스크립트 소스를 제시한다. 4장은 현장 연구 결과를 웹으로 공유하며 특정 키워드를 통한 분석 결과를 제시하였다. 마지막 장에는 연구의 결과와 향후 연구를 제시하고자 한다.

Ⅱ. 선행 연구

2.1 TF(term frequency)

한 문장이 주어졌을 때 문장을 구성하는 것은 단어이며 각 단어 별로 문서의 연관성을 수치로 확인하고자 할 때 단어의 빈도를 활용한다. “문서에서 단어의 연관성은 무엇인가?”라는 질문에 문장에서 단어들이 얼마만큼의 정보를 가지고 있는지를 나타낸다고 볼 수 있다.

TF기법은 문서가 주어졌을 때 이 단어가 몇 번 출현했는지를 나타내는 수치라 볼 수 있다. TF기법의 예증은 “문서가 있을 때 단어가 여러 번 출현되었다면 그 여러 번 출현한 만큼 연관성이 높을 것이다.” 라는 가설로 TF값을 사용한다(Ye et al., 2016; Christian et al., 2016; Xia et al., 2016).

$$TF = tf(t, d) \qquad \text{식1}$$

식1의 tf(t, d)는 연구 문서(문장) d에서 단어 t가 몇 번에 걸쳐서 나타났는지 빈도를 구한 것으로 단어의 빈도라 볼 수 있다.

"a new car, used car, car review" --- 예문A

예문1의 문장 구성 단어는 <표1>과 같다.

<표 2> 예문A - TF Values

word	TF
a	1/7
new	1/7
car	3/7
used	1/7
review	1/7

<표 2>은 일곱 개의 단어로 구성된 예문A로 TF값을 정리하면 다음과 같다. ‘a’는 한 번 출현으로 ‘a’의 TF값은 1/7이 되며 ‘new’, ‘used’, ‘review’도 출현 빈도가 동일하여 TF값이 동일하다. 하지만 ‘car’의 출현 빈도는 3회로 <예문A>의 문장에서 가장 높은 빈도이며 높은 TF 값을 갖게 된다. 즉, 단어 ‘car’는 예문A 문장에서 가장 중요한 단어이다.

하지만 TF값의 치명적 오류를 아래의 예문을 통하여 확인해 볼 수 있다.

"a friend in need is a friend indeed" --- 예문B

예문B의 문장은 8개의 단어로 구성되어 있으며 단어 빈도 때 따른 TF값을 정리한 것은 <표 3>과 같다.

<표 3> 예문B - TF Values

word	TF
a	2/8
friend	2/8
in	1/8
need	1/8
is	1/8
indeed	1/8

<표 3>의 TF값을 살펴보면 ‘a’와 ‘friend’가 빈도 2회 출현으로 같은 TF값을 갖는다. 하지만 예문2에서 주어진 문장에서 중요한 단어는 'friend'이다. 자주 출현된 단어가 TF값이 높게 적용된다는 가정에서 본다면 예문2에서 부정관사 'a'처럼 연관성 없는 단어가 발생된다. 즉, TF값만으로 한 문장내의 단어와의 연관성을 나타내기 힘든 결과를 얻게 된다. 단순 단어 빈도가 높다고 문장의 연관성을 높게만 판단하기엔 오류가 발생할 수 있다. 특정 단어가 문서나 문장의 전체에서 얼마나 공통적으로 나타나는지를 확인하여 문장 내 자주 등장하는 단어를 연관성 없는 단어들에 제한할 필요가 있다.

2.2 TF-IDF(term frequency-inverse document frequency)

이러한 TF값의 치명적 오류를 바로 잡기 위하여 IDF를 활용하고 있다. 어떤 단어는 문장의 연관성이 낮음에도 불구하고 자주 출현하는 경우가 발생한다. 이런 연관성 없는 단어들에 제한을 주기 위한 기법으로 TF-IDF기법을 활용한다(Salton and Buckley, 1988; Christian et al., 2016; Xia et al., 2016).

TF-IDF는 검색엔진에서 사용하는 텍스트 데이터 처리 알고리즘이다. 연구 문서나 문장이 있을 때, 특정 단어가 연구 문서 내에서 어느 정도 중요한 의미를 갖는지 통계적 수치를 나타낸다. 또한 문서들 사이에 비슷한 정도인 유사도(similarity)나 검색 결과의 순위를 결정하는 검색엔진에 활용하기도 한다. 먼저 IDF (inverse document frequency)를 살펴보면 식2와 같다.

$$\text{IDF}= \log\frac{D}{1+df(t)} \qquad \text{식2}$$

식2를 살펴보면 역문서의 빈도를 표현하기 위한 식이다. 총 문장의 개수를 단어가 출현한 문장의 개수로 나누어 주면 된다. 즉, 한 단어가 문서 전체에서 얼마나 공통적으로 나타나는지를 나타내는 값이다. 전체 문서의 수를 해당 단어를 포함한 문서의 수로 나눈 뒤, 로그를 취해 얻을 수 있다(Salton and Buckley, 1988).

"a new car, used car, car review" --- 예문A
"a friend in need is a friend indeed" --- 예문B

앞 TF절을 설명할 때 사용하였던 예문A, B를 이용하여 역문서 빈도(IDF) 값을 확인할 수 있으며 <표 4>와 같다.

<표 4> 예문A, B - IDF Values

word	IDF
a	Log(2/2)=0
new	Log(2/1)=0.3
car	Log(2/1)=0.3
used	Log(2/1)=0.3
review	Log(2/1)=0.3
friend	Log(2/1)=0.3
in	Log(2/1)=0.3
need	Log(2/1)=0.3
is	Log(2/1)=0.3
indeed	Log(2/1)=0.3

<표 4>는 예문 두 개의 문장을 분자에는 '2'가 삽입되며 각각의 단어는 몇 개의 문장에 출현되었는지를 분모에 삽입한다. ‘a’인 경우는 두 문장에서 사용되었으므로 IDF 값이 ‘0’으로 나타난다. ‘new’경우는 전체 두 문장 중 한 문장에서 출현되었으므로 Log(2/1)의 대입으로 0.3의 IDF 값을 갖는다. ‘a’를 제외한 단어는 두 문장 중 단일을 빈도를 보이고 있으며 ‘a’는 두 문장 모두 나타난 단어로 제한할 필요가 있다.

$$\text{TF-IDF} = tf(t,d) * \log\frac{D}{1+df(t)} \qquad \text{식3}$$

식 3을 살펴보면 TF-IDF의 수학적 정의는 단어 빈도와 역문서 빈도의 곱이라 할 수 있고 여러 문서가 있을 때, 특정 문서 안에서 특정 단어가 어느 정도 중요한 의미를 갖는지 수치를 보여준다(Salton and Buckley, 1988).

<표 5>는 예문 A와 B를 사용하여 TF-IDF 값을 정리하였다. 예문 A의 TF_IDF값을 살펴보면 0.13의 값이 가장 높으며 ‘car’의 단어가 문장에서 가장 중요한 단어이다. 예문 B를 살펴보면 TF값만을 보았을 때는 ‘a’와 ‘friend’가 같은 값을 갖고 있다. 하지만 TF-IDF값을 살펴보면 ‘friend’가 0.08로 가장 높은 점수로 나타났다. ‘a’는 IDF 값이 '0'이기 때문에 TF-IDF값은 ‘0’이다. 문장 B에서 가장 많은 정보를 함유한 단어 즉, 가장 연관성 높은 단어는 ‘friend’이며 0.08로 가장 높다.

이와 같이 TF-IDF는 어떤 문장이 주어졌을 때 각 단어별로 문장의 연관성을 수치로 나타낸 값이다. 대부분의 연구자는 영문 기반 TF-IDF 분석에 많은 결과를 표현하고 있다 (Salton and Buckley, 1988; Christian et al., 2016; Xia et al., 2016). 한글 기반 연구의 부재는 국민의 니즈 분석이 늦어진다는 뜻으로 해석된다. 본 연구는 연구 대상 문서를 문장 단위로 구분하여 TF 기법의 특정 키워드를 포함하는 문장과 TF-IDF 기법을 사용하여 문장의 주요 단어가 특징 키워드인 문장을 각각 추출 및 구분하여 두 기법의 차이를 살펴보고자 한다.

<표 5> 예문A, B - TF-IDF Values

word	TF		IDF	TF * IDF	
	A	B		A	B
a	1/7	2/8	Log(2/2)=0	0	0
new	1/7	0	Log(2/2)=0.3	0.04	0
car	3/7	0	Log(2/2)=0.3	**0.13**	0
used	1/7	0	Log(2/2)=0.3	0.04	0
review	1/7	0	Log(2/2)=0.3	0.04	0
friend	0	2/8	Log(2/2)=0.3	0	**0.08**
in	0	1/8	Log(2/2)=0.3	0	0.04
need	0	1/8	Log(2/2)=0.3	0	0.04
is	0	1/8	Log(2/2)=0.3	0	0.04
indeed	0	1/8	Log(2/2)=0.3	0	0.04

2.3 군집분석(cluster analysis)

군집분석은 객체(object)들의 집합을 군집 즉, 여러 그룹으로 묶어주는 것이다. 같은 그룹에 있는 객체들은 다른 그룹에 있는 객체들에 비해 더 유사한 특성을 가진다고 볼 수 있다(Lee et al., 2018; Nowak and Tibshirani, 2007).

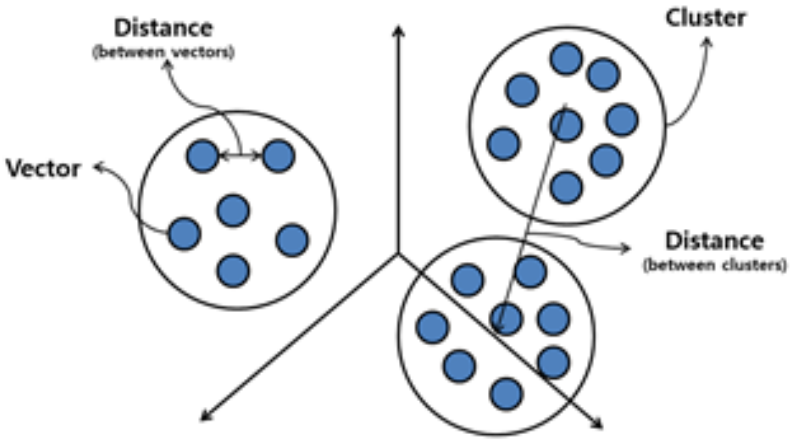
가령 통신사에서 고객에 대한 정보는 성별, 나이, 지역 등의 개인 인적 정보 외에도 사용 요금제, 결제 방법, 가입 년 수, 통화량 등 통신 서비스의 다양한 정보가 자사에서 관리되고 있을 것이다. 이렇게 다른 특성으로 표현되는 수 많은 가지 수를 단 몇 개의 그룹 즉, 군집으로 분류할 수 있다면, 각각의 군집에 속한 고객들에게 군집에 맞는 적당한 서비스를 맞춤형 마케팅으로 진행할 계획을 기획할 수 있을 것이다. 고객을 여러 집단으로 나누거나 적절한 목적으로 특성 및 차이를 분석하기 위해 군집분석이 사용될 것이다.

<그림 1>는 군집 분석을 위한 거리 개념도로 벡터 거리는 가깝고 군집 거리는 먼 경우를 좋은 군집화 품질로 볼 수 있다. 같은 클러스터로 묶인 객체들 간에는 단어의 유사한 속성을 가지며 다른 군집으로 분류된 클러스터 간에는 다른 속성을 가질 때를 말하는 것이다. 군집 분석에서 거리 측정 기준은 최단연결법, 최장연결법, 와드연결법, 평균연결법, K-means으로 구분한다(Hartigan, 1975).

최단연결법(single linkage method)은 두 군집 사이의 거리를 각 군집에 속하는 임의의 두 개체들 사이의 거리 중에서 최단 거리로 정의하여 가장 유사성이 큰 군집을 묶어 나가는 방법이다(Topchy et al., 2004; Wu et al., 2018).

최장연결법(complete linkage method)은 두 군집 사이의 거리를 각 군집에 속하는 임의의 두 개체들 사이의 거리 중에서 최장거리로 정의하여 가장 유사성이 큰 군집을 묶어 나가는 방법이다(Ferreira and Zhao, 2016; Rong et al., 2018).

와드연결법(ward's linkage method)은 두 군집을 묶을 때 생기는 새로운 군집이 속해 있는 객체들의 오차 제곱 합을 측정하며 군집으로



<그림 1> 군집 분석

인하여 파생되는 오차 제곱합의 증가량을 두 군집 사이의 거리로 정의하여 가장 유사성이 높은 군집으로 만들어가는 방법이다(김성현 · 김동재, 2018).

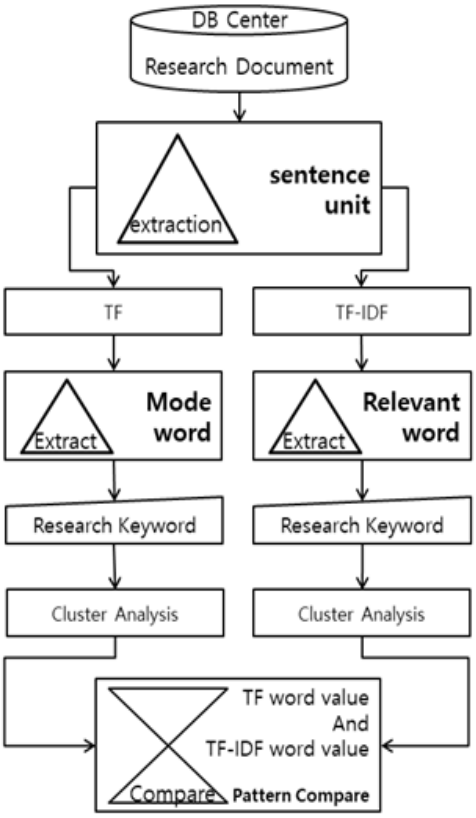
평균연결법(average linkage method)은 크기가 각각인 두 클러스터 사이의 거리를 각 군집에서 개체를 하나씩 선택해 연결한 모든 가능한 경우의 거리 평균을 계산하여 가장 유사 가중치가 큰 군집을 묶어 나가는 방법이다(Jia et al., 2011; Eder, 2017).

K-means는 군집을 형성하지 않고 관찰값을 몇 개의 군집으로 구분시키는 형태로 전체 데이터를 몇 개의 집단으로 그룹화하여 각 집단의 성격을 파악함으로써 전체의 구조에 대한 이해를 돕고자 하는 분석 방법이다(Javadi, 2017).

군집 분석을 이용하여 벡터 사이 간격과 군집 사이 거리를 이용하여 마이닝 처리에 적용하고자 한다. 많은 연구자들은 영문 기반 마이닝 처리를 이용하고 있다. 본 연구는 한글 기반 TF와 TF_IDF 기법의 결과를 군집분석을 이용하여 비교하고자 한다. 특히, 현장 연구 개발로 결과을 웹 페이지를 통하여 실험이 가능하며 실시간 분석을 확인 할 수 있으며 빈도 차이와 군집 간 거리를 시각화 결과로 살펴보고자 한다.

Ⅲ. 연구방법과 프레임워크

본 연구는 텍스트마이닝 연구에서 키워드 중심 빈도를 기준의 TF 기법과 문장 내 중심 단어를 추출 후 빈도 기준의 TF-IDF 기법을 군집분석으로 비교하고자 한다. 군집 분석은 두 군집 사이 거리를 각 군집에서 하나씩 개체를 선택해 연결한 모든 가능한 경우의 거리 평균을 계산하여 가장 유사성이 큰 군집을 묶어 나가는 방법인 와드연결법을 활용하며 가중치를 함께 표현하고자 한다.



<그림 2> 본 연구의 프레임워크

<그림 2>은 본 연구의 프레임워크이며 연구문서 DB에서 문장단위로 구분하여 연구문장을 추출한다. TF 기법은 각 문장의 최빈수를 찾아내어 해당 단어로 문장과 연결하며 TF-IDF 기법을 활용한 방법은 연구 문장을 설명할 수 있

는 정보를 갖는 단어를 찾아내어 문장과 연결한다. 문장을 대표하는 키워드를 도출한 후 동일한 키워드를 입력하여 각각의 군집분석을 진행한다. 군집분석은 키워드 간 평균값을 이용한 군집 분석 방법으로 진행하여 서로 두 기법의 차이를 확인하고자 한다.

<그림 3>은 본 연구의 프레임워크에서 'TF' 'TF-IDF'의 가중치를 찾아내기 위한 R 스크립트의 일부이다.

<그림 3> 예시 알고리즘은 두 문장을 백터로 받아들여 말뭉치로 변환하여 추출 방법을 "weightTF"와 "weightTfldf"로 각각을 추출하여 두 문장 내 단어의 가중치를 확인하고자 한다. 다음 장의 실험은 한글 기반 인터넷 뉴스를 이용하여 문장 내 단순 빈도 기준 군집분석과 잦은 문장 내 출현 단어를 제한 후 빈도 기준 군집 분석 결과를 표현하고 있다. 군집분석을 통하여 두 기법간의 벡터 거리와 군집거리의 변화를 통해 의미 있는 군집을 확인하고자 한다. 또한, 현장 연구를 통하여 분석의 전 과정을 웹페이지로 구현하였다.

Ⅳ. 연구 알고리즘 실험과 결과

본 연구는 인터넷 뉴스 기사를 이용하여 문장 내 단어의 단순 빈도를 이용한 군집분석 결과와 문장마다 자주 등장하는 단어를 제한하여 문장 내 주요 단어 빈도를 이용한 군집 분석 결과를 비교하고자 한다. 영문 기반 TF-IDF 기법을 한글 기반 연구에 적용한 사례로 TF와 TF-IDF 각각의 결과를 군집하여 키워드 간 벡터거리, 군집 거리를 확인 할 수 있을 것이다. 데이터 분석에 앞서 연구 재료가 되는 뉴스 기

```
   # 패키지로드
> library ( tm )
   # 백터형으로 데이터 저장
> txts <- c("a new car used car car review", "a friend in need is a friend indeed")
   # 백터형 분석을 위한 말뭉치로 변환
> corp <- Corpus(VectorSource(txts))
   # 문서들에 단어를 추출하여 행은 문서를 열은 단어를 나열
   # 'weightTf' 속성을 이용하여 'TF' 가중치 함께 도출
> tf <- DocumentTermMatrix(corp, control=list(weighting=weightTf))
   # 'weightTfIdf' 속성을 이용하여 'TF-IDF' 가중치 함께 도출
> tfidf <- DocumentTermMatrix(corp, control=list(weighting=weightTfIdf))
   # 일회 이상 출현된 단어 출력
> findFreqTerms(tf, lowfreq = 1)
   # 문서와 단어의 빈도를 출력
> as.matrix(tf)
   # 0.33이상의 가중치 출력
> findFreqTerms(tfidf, lowfreq = 0.33)
   # 문서와 단어의 가중치 출력
> as.matrix(tfidf)
```

<그림 3> TF, TF-IDF Value 추출 알고리즘

사는 본 연구자의 선행연구로 수집 시스템이 구축되어 있다. “Python을 이용한 SNS 크롤링 시스템 구축”을 이용하여 매일 새롭게 등장되는 뉴스 콘텐츠를 명사, 형용사 등 네티즌들의 감성만을 전처리 과정을 거쳐 DB에 저장되어지고 있다(이종화, 2018). 이종화(2018) 연구는 피이썬 웹드라이버의 가상 웹프라우저를 이용하여 인스타그램, 트위터, 유튜브 등 소셜 데이터를 실시간 수집 가능한 시스템을 연구하였다. 뉴스 기사가 인터넷을 통하여 업로드가 완료되면 1차 전처리 과정 과정을 거쳐 DB로 추출되어 저장된다. 이러한 시스템을 이용하여 웹 페이지 구축을 통하여 현장 연구를 진행하였다. <그림 4>는 연구를 위한 실제 웹페이지(http://14.7.122.142/dashboard/ python_crawling.php)이다. 뉴스와 댓글 중 희망하는 콘텐츠를 선택하고 이슈 범위인 날짜와 이슈 키워드를 입력하면 해당 조건의 뉴스 기사를 분석할 수 있는 웹페이지이다.

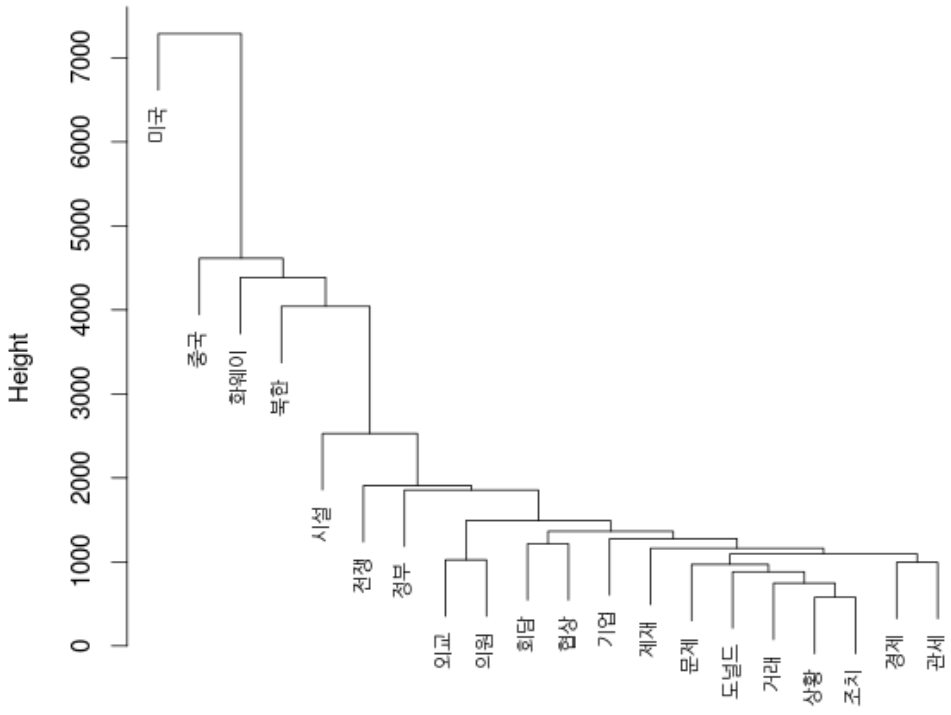
연구의 실험은 인터넷 뉴스 기사의 업로드 일을 기준으로 2019년 5월 20일 ~ 26일(1주일)이며 키워드는 “트럼프” 이슈를 살펴보는 실험을 진행하였다. 실험 기간 내 790건의 기사가 확인되었으며 댓글은 무려 21,079개의 네티즌 의견을 확인할 수 있었다.

<표 6>은 <그림 4>의 웹페이지에서 조건에 따른 뉴스 기사 제목의 리스트이다. 지면상 크롤링 일부 데이터를 기재하였으며 관련 뉴스의 내용은 별도의 콘텐츠 항목에 저장되어 있다. 실험 조건을 “트럼프”로 하여 북한, 일본, 이란, 중국 등의 외교적 뉴스가 대부분을 차지하고 있다.

<그림 4> 분석 웹 페이지

<표 6> 2019-05-20(7일간) “트럼프” 뉴스 DB

기사 제목
트럼프 “김정은, 핵시설 5곳 중 1~2곳만 없애려 했다” 첫 언급
구글·인텔·퀄컴 등 미 기업들, 줄줄이 화웨이와 거래 중단
로이터 “구글, 중국 화웨이와 거래 중단키로”
6월 한·일서 잇단 ‘빅 이벤트’.. 북핵·세계경제 판도 변화 주목
이란과 전쟁하긴 싫고.. ‘겁은 주고 싶은’ 트럼프
트럼프 “김정은, 北핵시설 5곳 중 1~2곳만 폐쇄 원해..하노이회담 결렬 이유”
트럼프 “전쟁 원치 않지만, 이란 핵보유 용인 못해”
주말휴가 보내고 백악관 돌아온 트럼프
“美정부, 中제품에 추가관세 매기면 美기업 수익 6%↓” 골드만삭스
구글, 화웨이의 안드로이드 OS 및 구글앱 접근 차단(종합)
미-중 갈등에 車 관세까지..세계 경제 살얼음판에 수출 시름
화웨이 제재에 국내 부품사 우려 확산..“영향 제한적”
美·日, 트럼프 방일 앞서 고위급 무역협상
中외교부 “트럼프, 무역합의 파기 책임 中에 떠넘겨”
무역전쟁 장기화되면 중국이 유리한 이유 3가지
...

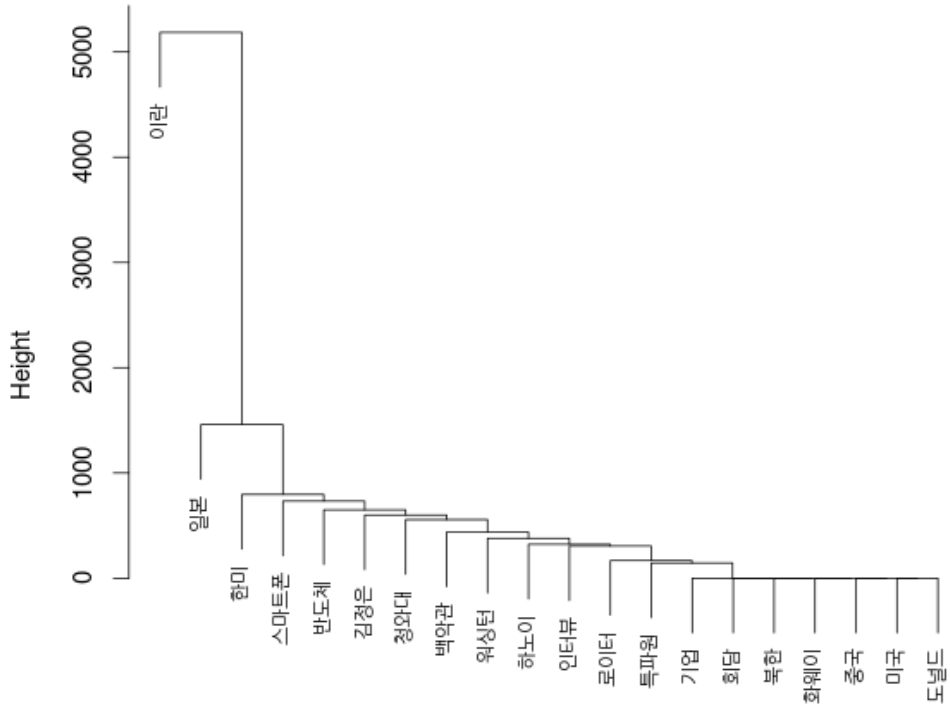


<그림 5> 2019-05-20(7일간) “트럼프” TF 군집분석

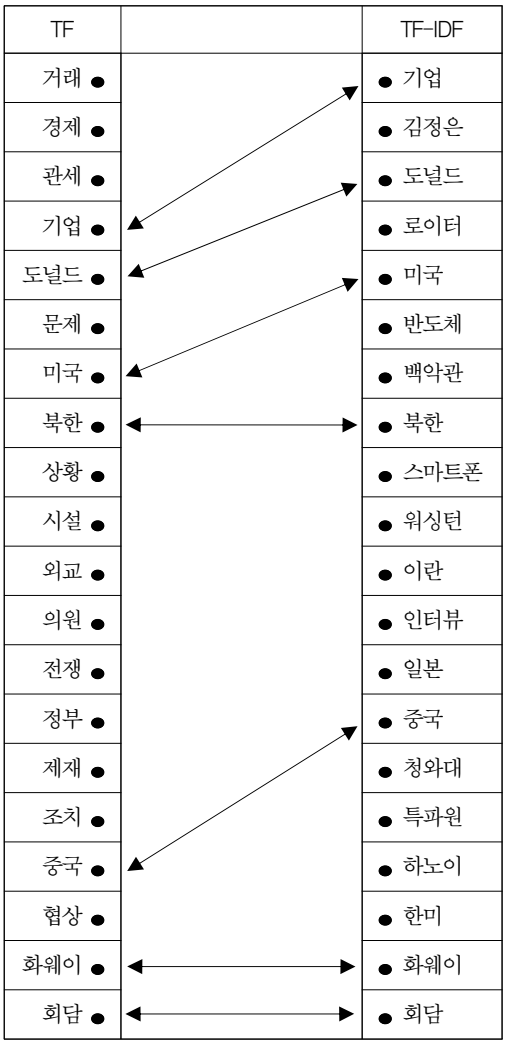
<그림 5>는 5월20일부터 7일간 인터넷 뉴스에 “트럼프” 이슈를 TF 기법을 이용한 군집분석 결과이다. {미국}과 {중국,{화웨이}}이 같은 군집으로 묶여있는 것은 실험 기간 동안에 이슈인 미국과 중국의 무역 장벽으로 인하여 미국 내 화웨이 제품의 제한에 관한 기사들에 공통적으로 나타난 키워드라 볼 수 있다. {북한}과 {시설,{전쟁}}의 묶음은 북한 단거리 미사일 발사 실험에 관한 이슈에 대하여 “트럼프”의 발언 기사들에 공통적으로 나타났다고 볼 수 있다. 그밖에 {회교, 의원}, {상황, 조치}, {경제 관세}등 “도널드 트럼프”정부 이슈들에 관한 키워드라 볼 수 있다.

<그림 6>은 5월 20일부터 7일간 “트럼프” 이슈를 대상으로 TF-IDF 기법을 이용한 군집분석 결과이다. TF 기법과 같은 실험 데이터를 이용하여 TF-IDF기법을 적용한 결과이다. {이란}과 {일본, {한미}}가 같은 군집으로 묶였으며 {기업, 회담, 북한, 화웨이, 중국, 미국, 도널드}가 같은 묶음임을 알 수 있다.

<그림 7>은 TF기법과 TF-IDF 기법에 나타난 키워드를 비교한 결과이며 공통적으로 나타난 데이터는 “기업”, “도널드”, “미국”, “북한”, “중국”, “화웨이”, “회담”이 두 기법에서 공통으로 나타난 결과이다. 미국과 중국의 무역 장벽 공방 관련기사와 북미 2차 회담인 하노이 회담에서의 책임을 물어 북한의 고위급 관료 처형설 관련 기사로 인하여 나타난 공통된 키워드라 볼 수 있다. 또한 TF 기법에만 존재하는 “거래”, “경제”, “관세”, “문제”, “상황”, “시



<그림 6> 2019-05-20(7일간) “트럼프” TF-IDF 군집분석



<그림 7> 2019-05-20(7일간) “트럼프” TF, TF-IDF 비교

설”, “외교”, “의원”, “전쟁”, “정부”, “제재”, “조치”, “협상”은 7개의 공통으로 나타난 데이터와 함께 실험 데이터 600여개 뉴스 기사에서 상위 빈도를 나타난 키워드라 볼 수 있다. “김정은”, “로이터”, “반도체”, “백악관”, “스마트폰”, “워싱턴”, “이란”, “인터뷰”, “일본”, “청와대”, “특파원”, “하노이”, “한미” 키워드는 TF-IDF 기법에서만 등장한 키워드이다. TF기법은 문장에 있는 그대로의 단어를 이용하여 군집 추출된 키워드이며 실제 분석에 의미 없는 단어 “상황”, “시설”, “외교”, “조치” 등이 상위 순위에 있는 단어로 실험 데이터 전체적으로 빈번이 나타난 키워드라 볼 수 있으며 TF-IDF 분석에서 제한될 단어이기도 하다. TF-IDF 기법은 “김정은”, “반도체”, “백악관”, “워싱턴”, “이란”, “일본”, “청와대”, “하노이”, “한미” 등의 지명이나 인면등이 등장하여 관련 이슈에 관련 있는 단어가 등장한 것을 확인 할 수 있다. 즉, TF 기법의 의미 없는 단어의 상위 순위를 확인 활 수 있는 설험이다. 문장 단위 자주 등장되는 키워드에 제한을 설정하여 문장 주요단어 추출을 군집에 이용한 것이다.

Ⅴ. 결론 및 향후 연구과제

전 세계적으로 데이터의 급팽창과 특히, 비정형 데이터의 압도적인 증가로 데이터 가치가 4차 산업혁명의 새로운 원료로 등장하고 있다. 기업에서 정보시스템 영역이 확장되는 이유 중 하나는 늘어나는 데이터 분석의 효율성이 높기 때문이다. 특히, 급속히 증가하는 데이터 형태에는 영상, 음성, 이미지, 소셜네트워크에서 오고 가는 대화 등과 같이 복잡하고 비정형적 데이터가 양적인 증가를 주도하고 있다. 이러한 웹 환경에 있는 고객의 소리 즉, 텍스트 데이터들의 분석이 고객의 니즈 분석으로 이어진다.

텍스트 문장에서 특정 단어 빈도가 높아짐에 따라 주요단어의 기준이 자칫 잘못되면 다른 의미로 분석되어지는 경우가 발생한다. 즉, 단

어의 빈도를 기준으로 분석이 이루어지다 보니 문장의 단어 빈도가 그 문장을 해석하는 단어로 추출되며 연구 결과에 영향을 주고 있다.

본 연구는 대부분의 연구자들이 선택하는 연구 대상 재료를 영문기반에서 한글 기반 미디어 데이터로 분석을 진행하였으며 특정 문장에서 빈도만으로 문장을 대표하는 단어 추출이 아닌 실제 문장에서 중요한 키워드를 추출하는 TF-IDF 기법을 이용하여 문장 분석 연구를 진행하여 결과를 정리한다.

먼저, TF기법을 적용한 연구들은 단어 빈도 중심으로 분석하면 문장의 단어 빈도가 높다면 의미 없는 단어도 그 문장을 해석하는 단어로 추출되며 연구 결과에서 영향을 주고 있다. 본 연구는 문장에서 단순 빈도만으로 문장을 대표하는 단어 추출 기법인 TF기법이 아닌 실제 문장에서 중요한 키워드를 추출하는 TF-IDF 기법을 비교하여 한글 연구에 적용하고자 한다. 두 기법을 군집분석으로 시각화하여 그 차이를 기술하고자 했다.

TF기법과 TF-IDF기법을 구현하기 위한 알고리즘을 한글 자연어 처리(KoNLP)에 응용하여 연구 실험을 진행하며 빈도 분석 기법에서 의미 분석으로 지속적 연구의 가치가 있다는 점을 확인하였다. 영어 표현보다 한글 표현이 다양하다보니 한글 자연어 처리 분석이 더 어렵고 연구 설계에 많은 시간적 투자가 가중될 수밖에 없다. TF-IDF기법을 한글 처리에 활용하여 단순 빈도(TF) 분석과 비교하여 진행한 만큼 기존 텍스트마이닝 연구자에게 제공되는 연구 기법의 변화가 기대된다.

실제 실험은 2019-05-20에서 일주일 뉴스 중 “트럼프” 관련 기사를 기준으로 분석하였다. TF 기법에만 사용된 단어들은 “거래”, “경제”, “관세”, “문제”, “상황”, “시설”, “외교”, “의원”, “전쟁”, “정부”, “제재”, “조치”, “협상”등으로 실제 기사 문장에서 자주 등장되는 단어다보니 의미 분석 방법론에서 어떤 의미를 부여하기 힘든 키워드들이다. 반면, TF-IDF기법을 통해서만 등장한 단어는 “김정은”, “로이터”, “반도체”, “백악관”, “워싱턴”, “이란”, “인터뷰”, “일본”, “청와대”, “특파원”, “하노이”, “한미” 등이 있었으며 <그림 6>의 군집분석을 통해서 단어 간 빈도 차이나 군집 간 거리를 확인할 수 있다. 실험에서 TF기법에서 의미 없는 단어의 등장으로 상대적으로 빈도가 낮게 측정된 의미 있는 단어의 부재가 발생되었으며 TF-IDF 기법에서는 이슈와 관련 있는 의미 있는 단어들을 볼 수 있었다.

TF 기법만의 한계를 실험을 통하여 증명하였고 두 기법의 결과를 시각적으로 구분하기 위하여 군집분석을 병행하였다. 연구 과정에서 개발된 알고리즘을 함께 공유하며 시제품 서비스 웹 페이지를 통하여 독자들에게 연구 개발된 웹을 제공하였다. 연구 결과는 웹 페이지로 공개 및 공유되며 2019년 1월부터 6개월간 인터넷 뉴스 기사 전체와 기사별 네티즌들의 댓글을 함께 분석 대상에서 선택할 수 있으며 본 논문 게재일로부터 6개월 연구 결과 페이지를 공유하고자 한다(http://14.7.122.142/dashboard/python_crawling.php).

TF기법과 TF-IDF기법을 구현하기 위한 알고리즘을 한글 기반 데이터에 적용하여 연구 설계를 진행하므로 빈도 분석 기법에서 의미 분석으로 지속적 연구의 가치를 확인하였다. 또한, 향후 BOW, N-gram, NMF Word2Vec 기법

을 한글 기반 실시간 의미 분석에 적용하여 지속적 연구가 필요하다.

참고문헌

김성현, 김동재, “군집화 및 특성도를 이용한 결측치 대체 방법,” 응용통계연구, 제31권, 제1호, 2018, pp. 29-40.

김은우, 금득규, “특집명: 빅데이터 분석: Social BigDate 서비스 분석플랫폼 구축기술 소개,” 정보처리학회지, 제21권, 제3호, 2014, pp. 35-42.

남민지, 이은지, 신주현, “인스타그램 해시태그를 이용한 사용자 감정 분류 방법,” 멀티미디어학회논문지, 제18권, 제11호, 2015, pp. 1391-1399.

서새남, “4차 산업혁명 주요기술에 대한 법적 고찰-한국 및 중국을 중심으로,” 문화・미디어・엔터테인먼트법, 제11권, 제1호, 2017, pp. 141-172.

양낙영, 김성근, 강주영, “텍스트 마이닝 방법론과 메신저 UI 를 활용한 융합연구 촉진을 위한 연구자 및 연구 분야 추천 시스템의 제안,” 정보시스템연구, 제27권, 제4호, 2018, pp. 71-96.

유은지, 김정철, 이춘열, 김남규, “시맨틱 텍스트 마이닝을 위한 온톨로지 활용 방안,” 정보시스템연구, 제21권, 제3호, 2012, pp. 137-161.

이종화, “Python을 이용한 SNS 크롤링 시스템 구축,” 한국산업정보학회논문지, 제23권, 제5호, 2018, pp. 61-76.

이종화, 이현규, “오픈소스 소프트웨어를 활용한 자연어 처리 패키지 제작에 관한 연구,” 정보시스템연구, 제25권, 제4호, 2016, pp. 121-139.

Amado, A., Cortez, P., Rita, P., and Moro, S., “Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis,” *European Research on Management and Business Economics*, Vol. 24, No. 1, 2017, pp. 1-7.

An, J., and Kim, H. W., “Building a Korean sentiment lexicon using collective intelligence,” *Journal of Intelligence and Information Systems*, Vol. 21, No. 2, 2015, pp. 49-67.

Christian, H., Agus, M. P., and Suhartono, D., “Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF),” *ComTech: Computer, Mathematics and Engineering Applications*, Vol. 7, No. 4, 2016, pp. 285-294.

Eder, M., “Visualization in stylometry: Cluster analysis using networks,” Digital Scholarship in the Humanities, Vol. 32, No. 1, 2017, pp. 50-64.

Ferreira, L. N., and Zhao, L., “Time series clustering via community detection in networks,” *Information Sciences*, Vol. 326, 2016, pp. 227-242.

Hartigan, J., A., *“Clustering Algorithms,”* New York: Wiley, 1975.

Hoyer, P. O., "Non-negative matrix factorization with sparseness constraints," *Journal of machine learning research*, Vol. 5, No. Nov, 2004, pp. 1457-1469.

Javadi, S., Hashemy, S. M., Mohammadi, K., Howard, K. W. F., and Neshat, A., "Classification of aquifer vulnerability using K-means cluster analysis," *Journal of hydrology*, Vol. 549, 2017, pp. 27-37.

Jia J, Xiao X, Liu B, and Jiao L., "Bagging-based spectral clustering ensemble selection," *Pattern Recognit Lett*, Vol. 32, No. 10, 2011, pp. 1456-1467.

Lee, J. H., "Big data, data mining and temporary reproduction," *The Journal of Intellectual Property*, Vol. 8, No. 4, 2013, pp. 93-125.

Lee, J. H., and, Lee, H. K., "A Research on Real-time Analysis of Association Rules Using Hash Tags," The Journal of Internet Electronic Commerce Research, Vol. 17, No. 4, 2017, pp. 105-117.

Lee, Y. S., Lee, J., and Gil, J. M., "A Study on Research Paper Classification Using Keyword Clustering," *KIPS Transactions on Software and Data Engineering*, Vol. 7, No. 12, 2018, pp. 477-484.

Nowak, G., and Tibshirani, R., "Complementary hierarchical clustering," *Biostatistics*, Vol. 9, No. 3, 2007, pp. 467-483.

Rong H, Ma TH, Tang ML, Cao J., "A novel subgraph k+-isomorphism method in social network based on graph similarity detection," *Soft Comput*, 2018, Vol. 22, No. 8, pp. 2853–2601.

Salton, G., and Buckley, C., "Term-weighting approaches in automatic text retrieval," *Information processing & management*, Vol. 24, No. 5, 1988, pp. 513-523.

Topchy A, Jain AK, and Punch W., *"A mixture model of clustering ensembles,"* Proc SIAM Intl Conf Data Mining 2004, pp. 379-390.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J., *"Data Mining: Practical machine learning tools and techniques,"* Morgan Kaufmann, 2016.

Wu, X., Ma, T., Cao, J., Tian, Y., and Alabdulkarim, A., "A comparative study of clustering ensemble algorithms, *Computers & Electrical Engineering,"* No. 68, 2018, pp. 603-615.

Zhang, Y., Jin, R., & Zhou, Z. H., "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, Vol. 1, 2010, pp. 43-52.

이 종 화 (Lee, Jong-Hwa)

부경대학교 경영학 박사학위를 취득하고, 현재 동의대학교 상경학부 e비즈니스학전공 교수로 재직 중이다. 주요 관심분야는 BigData, Mining, Content Analysis 등이다.

이 문 봉 (Lee, MoonBong)

연세대학교 경영학사, 석사와 박사학위를 취득하였다. 현재 동의대학교 경영학과 교수로 재직하고 있으며, 주요 관심분야는 정보시스템 성과, ERP, SNS 등이다.

김 종 원 (Kim, Jong-Weon)

인하대학교 경영학과를 졸업하고, University of Nebraska-Lincoln에서 MBA와 경영학박사를 취득하였다. 관심분야로는 기업성과, CSV, CSR, IT diffusion, SCM 등이다.

<Abstract>

A study on Korean language processing using TF-IDF

Lee, Jong-Hwa · Lee, MoonBong · Kim, Jong-Weon

Purpose

One of the reasons for the expansion of information systems in the enterprise is the increased efficiency of data analysis. In particular, the rapidly increasing data types which are complex and unstructured such as video, voice, images, and conversations in and out of social networks. The purpose of this study is the customer needs analysis from customer voices, ie, text data, in the web environment..

Design/methodology/approach

As previous study results, the word frequency of the sentence is extracted as a word that interprets the sentence has better affects than frequency analysis. In this study, we applied the TF-IDF method, which extracts important keywords in real sentences, not the TF method, which is a word extraction technique that expresses sentences with simple frequency only, in Korean language research. We visualized the two techniques by cluster analysis and describe the difference.

Findings

TF technique and TF-IDF technique are applied for Korean natural language processing, the research showed the value from frequency analysis technique to semantic analysis and it is expected to change the technique by Korean language processing researcher.

Keyword: Big Data, TF-IDF, Text Mining, Cluster Analysis, KoNLP

* 이 논문은 2019년 6월 17일 접수, 2019년 6월 25일 1차 심사, 2019년 8월 8일 게재 확정되었습니다.