TF-IDF를 활용한 k-means 기반의 효율적인 대용량 기사 처리 및 요약 알고리즘

장민서*, 오수진**, 김응모***
*성균관대학교 문과대학
**성균관대학교 정보통신대학
***성균관대학교 소프트웨어대학
e-mail : minseo461@gmail.com, bgbanana4@gmail.com, ukim@skku.edu

Article Analytic and Summarizing Algorithm by facilitating TF-IDF based on k-means

Minseo Jang*, Sujin OH**, Ung-Mo Kim***
*College of Humanities, Sungkyunkwan University
**College of Information and Communication Engineering, Sungkyunkwan University
***College of Software, Sungkyunkwan University

요 약

본 논문에서는 뉴스기사 데이터를 활용하여 대규모 뉴스기사를 소주제로 분류하는 군집 분석 방법을 제안한다. 또한, 분류된 뉴스기사를 사용자가 빠르게 이해하고 접할 수 있도록 핵심 문장을 추출하여 제공하는 방법을 제안한다. 분석 데이터는 포털 사이트 점유율 1위인 네이버의 경제 분야 뉴스기사를 크롤링하여 수집한다. 뉴스기사의 분석을 위해 전 처리를 통해 특수문자, 조사, 어미, 구두점 등의 불용어 처리를 수행한다. 또한, k-means 알고리즘을 이용하여 대용량의 뉴스기사를 주제 별로 분류하는 것을 진행하며 그것을 토대로 핵심 문장을 추출한다. 추출된 핵심 문장은 분류된 뉴스기사의 주제를 나타내며 사용자에게 빠르게 정보를 전달하기 위해 활용한다. 본 논문의 연구 내용이 여러 언론사 사이트에 반영되면 사이트 품질과 사용자 만족도 향상에 기여할 수 있을 것으로 보인다.

1. 서론

1.1 연구배경

방송매체의 급격한 발전과 인터넷 및 온라인 정보서비스의 기하급수적인 증가가 정보의 폭발적인 증가를 불러왔다. 특히, 인터넷 뉴스의 특성 상 다양한 뉴스가 한꺼번에 제공되기 때문에 사용자가 원하는 뉴스에 접근하는 것이 힘든 실정이다. 또한, 기사를 제공하는 포털 사이트 뿐만 아니라 정보를 제공하는 여러 사이트에서도 동일한 문제가 발생한다. 따라서 사용자에게 원하는 정보를 효과적으로 제공하기 위해서는, 주제 별로 분류하여, 중복되는 내용을 제거하고 핵심 문장으로 요약하여 이를 제공하는 것이 필요하다.

1.2 연구목표

인터넷 상의 정보들을 분석하여 사용자에게 양질의 정보들을 제공하여 사용자의 검색을 최소화하고 원하는 정보에 빠르게 접근함으로써 사용자 만족도 향상에 목적을 둔다.

이에 본 논문에서는 관련된 정보를 탐색하여 제공하기 위해 뉴스기사 데이터를 활용하여, 대규모 뉴스기사를 소주제로 분류하는 군집 분석을 진행한다. 또한, 분류된 뉴스기사를 사용자가 빠르게 이해하고 접할 수 있도록 핵심 문장을 추출한다. 추출된 핵심 문장은 분류된 뉴스기사의 주제를 나타내며 이는 사용자에게 빠르게 정보를 전달하기 위해 활용될 것이다.

1.3 overview

본 연구에선 방대한 양의 뉴스기사를 소주제로 분류하는 군집 분석 방법을 제안하며 또한, 분류된 뉴스기사를 사용자가 빠르고 쉽게 이해하고 접할 수 있도록 핵심 문장을 추출하여 제공하는 방법을 제안한다. 1장에선 연구배경과 목표에 대하여 설명하고 2장에선 본 연구와 관련 있는 선행 연구들에 대한 조사를 진행한다. 3장에선 데이터의 수집 및 처리 작업을 진행하며 4장에선 구현과 데이터의 분석 결과를 소개한다. 마지막 5장에선 본 연구의 결론과 향후 연구에 대하여 언급한다.

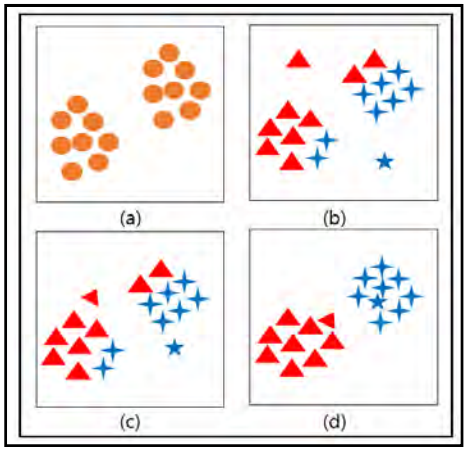
2. 관련 연구

2.1 데이터마이닝

데이터마이닝(Data Mining)은 자동화되고 지능을 갖춘 데이터베이스 분석기법이다. 데이터마이닝은 일반적으로 데이터 선택, 데이터 정제, 데이터 변환, 데이터마이닝, 패턴평가, 지식 표현의 6단계로 되어있다. 데이터 패턴을 추출하기 위해서 실제 데이터마이닝 알고리즘이 적용되는 단계는 데이터마이닝 전 과정 중 핵심 단계로서 일반적으로 데이터마이닝이라고 한다.

2.2 k-means 알고리즘

k-means 알고리즘은 가장 널리 알려진 클러스터링 알고리즘이다. 비계층적 클러스터링 기법으로 문서와 군집의 중심 값을 나타내는 센트로이드와의 유사도를 측정하여 문서를 적합한 군집에 재배치하는 기법이다. 여기에서 클러스터 센트로이드(중심)는 클러스터에 속하는 문서들의 평균 백터 값을 이용한다. k-Means 알고리즘의 기본 개념은 주어진 데이터를 k개의 클로스터로 묶는 것이다. 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다. k-means의 장점은 모든 형태의 데이터에 적용이 가능하며 대용량에 대한 처리가 가능하다는 점이다. 먼저 군집에 속한 데이터들의 평균값인 중점을 잡은 후, 평균값을 기준으로 데이터를 묶는다. 그 다음 중점을 계속 변경시키며 데이터를 묶는 과정을 반복하며 더 이상 중점이 변하지 않으면 수행을 중지하고 결과를 반환한다. 자세한 과정은 (그림 1)과 같다[1, 2].



(그림 1) k-means 알고리즘 과정

2.3 TF-IDF

TF-IDF는 주로 정보 검색과 텍스트마이닝에서 이용하는 가중치 지표로 어떤 단어가 특정 문서내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다. 문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정하거나, 문서들 사이의 비슷한 정도를 구하는 등의 용도로 사용된다. TF (단어빈도, term frequency)는 특정한 단어가 문서 내에 등장 빈도를 나타내며 이 값이 클수록 문서에서 자주 사용하는 단어이다. 하지만 문서 내에서 자주 사용된다고 하여서, 중요한 의미를 가진다고는 할 수 없다. '오늘', '그리고' 와 같은 다수의 문서에서 자주 사용되지만, 그 의미는 중요하지 않는 단어를 처리하기 위해서 IDF (역문서빈도, inverse document frequency)를 사용한다. IDF 값은 문서의 성격에 따라 결정된다. 특정 문서 내에서 단어 빈도가 높을수록, 그리고 전체 문서들 중 그 단어를 포함한 문서가 적을수록 TF-IDF값이 높아진다. 본 논문에서 사용하는 문서 가중치 값(TF-IDF)은 아래와 같은 수식으로 표현된다.

$$TF-IDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (1)$$

이를 통해, 문서 내에서 사용 빈도가 높으며, 의미를 가지는 단어를 얻을 수 있다[3].

3. 데이터 수집 및 처리

3.1 데이터 수집

본 논문은 온라인 뉴스에 초점을 맞춘다. 따라서 수집된 데이터 역시 포털사이트에서 제공하는 온라인 뉴스이다. 여러 포털사이트 중 점유율 1위[1]를 차지하고 있는 네이버의 온라인 뉴스 기사를 연구 데이터로 선정하였으며 이 중에서도 경제 분야의 기사를 대상으로 한다. 기사 데이터는 크롤링 작업을 통해 수집한다. 크롤링을 위한 코드는 R 프로그래밍 언어로 작성되었으며, 네이버 뉴스 기사에 관한 naverNewsParser[2] 라이브러리를 활용하여 수집되었다. 수집된 데이터는 2018년 1월 11일자 네이버 경제 분야의 기사이며, 총 2,379건의 기사가 수집되었다. 수집된 데이터의 속성은 기사의 카테고리, 제목, 작성자, 작성일, 내용으로 구성된다[4].

3.2 데이터 처리

수집된 데이터는 자연어 처리(Natural Language Processing)를 통한 정제되었다 자연어 처리란 인간의 언어를 기계적으로 분석하여 컴퓨터가 이해할 수 있는 형태로 가공하는 것을 말하며, 이를 위해 뉴스기사 내의 특수문자, 불 용어와 같이 자체적으로 큰 의미를 가지지 않는 성분과 정보를 제거하는 작업을 수행한다. 불 용어에 포함되는 품사는 조사, 어미, 구두점 등이 있다. 이 작업엔 R 프로그래밍에서 제공하는 한국어 자연어 처리(KoNLP) 패키지를 사용한다. KoNLP에서 지원하는 NIA 사전의 98만여 개의 단어를 활용하여 수집한 데이터에서 명사만을 추

1) internettrend
2) https://github.com/mrchypark/naverNewsParser

출한다. 그 후, 구두점과 숫자, 영단어를 처리한다. 이는 <표 1>과 같은 방식으로 이루어진다. 이후, 각각의 뉴스 기사에서 관련성이 적은 문장을 제거하여 키워드 추출 시 의미 없는 키워드가 선택되지 않도록 한다. 문장 제거는 문장 간의 코사인 유사도를 계산하여 관련성이 낮은 문장을 제거한다. 주요 키워드 추출을 위한 방안으론 TF-IDF를 활용한다[5].

<표 1> 원본 문장과 단어 처리 후 문장

원본 문장	단어 처리 후 문장
수집한 데이터의 속성은 기사의 카테고리, 제목, 작성자, 작성일, 내용이며 핵심 문장 요약과 세부 분류는 기사 내용을 분석하여 수행하였다.	수집 데이터 속성 기사 카테고리 제목 작성 자 작성 일 핵심 문장 요약 세부 분류 기사 내용 분석 수행

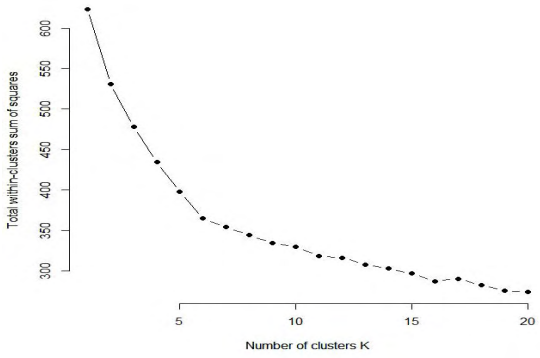
4. 구현 및 결과 분석

4.1 제안 시스템

본 논문에서 제안하는 시스템은 R 프로그래밍을 사용하여 구현되었으며, k-means 알고리즘과 TF-IDF 가중치 모델을 이용하여 데이터 분석을 진행한다. 제안 시스템은 크게 두 단계로 구성되어 있다. 우선 k-means 알고리즘을 적용하여 대용량 기사를 k개의 군집으로 분류한다. 그리고 TF-IDF 가중치 모델을 적용하여 불필요한 문장을 제거하고 핵심 문장을 추출한다. 본 연구에선 텍스트마이닝을 진행하기 위한 data.table 라이브러리와 한글 형태소 분석 라이브러리인 KoNLP, rJava 라이브러리를 사용한다. 이 후, 데이터 분석을 위해 arules, tm, proxy 라이브러리를 추가적으로 이용한다.

4.1.1 k-means 군집 분석

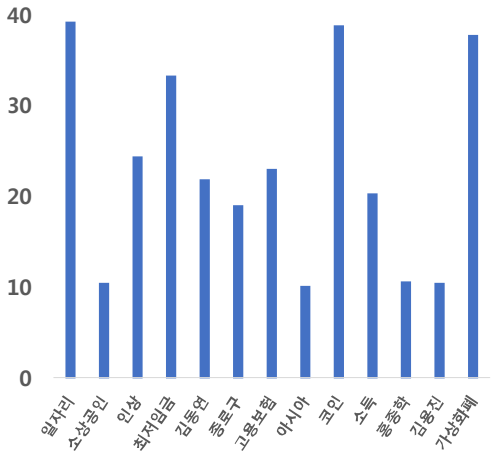
본 단계에서는 k-means 알고리즘을 활용하여 대용량 기사를 처리하여 소주제의 군집으로 분류하는 군집 분석을 진행한다. k-means 알고리즘의 효율은 군집의 개수 k에 따라 결정되며, 가장 적은 오차 값을 가지는 k 값을 선정해야 한다. 본 논문에서는 최적의 k값을 정하기 위해 elbow방법을 이용했다. 문서 군집에서 가장 적합하다고 여겨지는 거리 코사인 유사도를 사용하며, 초기값으로 20을 지정한다. (그림 2)에서 군집 내 분산이 급격하게 줄어들다 점점 완만해지는 경향을 띔을 알 수 있다. 기울기가 완만해지는 지점이 가장 적은 오차 값에 가지는 지점으로 볼 수 있기 때문에, 본 논문에서 최적의 k값으로 15를 선정하였다. 따라서 수집된 데이터들은 총 15개의 군집으로 나눠졌으며, 각 군집에 속하는 기사들을 군집별로 리스트화되어 저장된다.



(그림 2) 군집의 개수에 따른 데이터 군집

4.1.2 TF-IDF 가중치 모델

다음으로 TF-IDF를 활용하여 뉴스기사 내에서 불필요한 문장들을 제거하고 그와 동시에 중요한 문장들을 추출한다. TF-IDF을 이용하여, 각 문장에 속해있는 단어들을 비교하고 문장의 유사도를 측정한다. 비슷한 공간에 매핑 될 때를 유사성이 높다고 판단하며, 전체 문서 집합에서 여러 문장에 걸쳐서 함께 등장하는 단어 간에는 관련성이 있다고 판단한다. 이와 같은 방식으로 얻어진 문장 유사도와 단어 간 관련성의 평균을 이용하여 불필요한 문장들을 제거한다. 문장 간의 유사도를 측정하기 전, TF-IDF를 이용해 추출한 단어의 예로는 일자리, 가상화폐, 소득, 고용보험 등이 있으며 이의 분포도는 (그림 3)과 같다[6].



(그림 3) TF-IDF 단어 분포

그 다음으로, 앞서 구한 빈출 빈도가 높은 단어와 문장의 유사도를 기반으로 각 군집별로 한 개의 핵심문장을 추출한다. 다른 문장들과의 유사도 합이 높을수록 문서 내에서 중요한 문장으로 구분된다. 문장 간의 유사도 측정을 통해, 기사 내에서 해당 문장이 가지는 비중을 구할 수 있다. 어떤 한 문장이 기사 내에서 큰 비중을 가진다면, 이는 중요한 문장이라 할 수 있으며, 이 중 가장 큰 비중을 차지하는 문장을 군집의 핵심 문장이라고 판단한다.

4.2 제안시스템 분석

본 논문에서 제안하는 시스템은 k-means 알고리즘과 TF-IDF 가중치 모델을 활용하여 대용량의 기사를 소주제별로 군집화하고 빈출 단어와 문장의 유사도를 기반으로 핵심 문장을 추출한다. 본 논문에서 사용한 데이터에서는 최적 k값을 15로 정하였기 때문에, 기사들은 총 15개의 군집으로 분류되었으며, 각 군집마다 1개씩의 핵심문장을 추출하여 <표 2>와 같이 핵심 문장이 추출되었다.

<표 2> 핵심 문장 추출 데이터

	핵심 문장
1	김동연 경제부총리 겸 기획재정부 장관이 11일 오전 …
2	박 장관은 “정부는 (가상화폐 거래가) 매우 위험한 …
3	11일 이마트 서울 용산점에서 모델들이 '오이스터 …
4	"한재수 삼성전자 메모리사업부 전략마케팅팀 부사…
5	"고형권 기획재정부 1차관은 11일 정부서울청사에 …
6	오는 18일 개항하는 인천국제공항 제2여객터미널은 …
7	세계 최대 가전·IT 박람회인 'CES(Consumer Ele …
8	산업부는 준회원국 가입을 통해 우리의 10대 수출국 …
9	정부, 시장 활성화 방안유명무실한 펀드 …
15	홈플러스 노사는 임금체계 개편 없이 직원들의 실질 …

핵심 문장 추출에 이용되는 문장 간의 유사도는 (그림 4)와 같이 행렬 형태로 나타낼 수 있다. 가로 축은 기사에 존재하는 문장의 개수를 의미하며 행렬 값은 유사도로 볼 수 있다. 같은 문장이 존재한다면 두 문장 사이의 유사도는 1이며, 문서 내의 다른 문장과의 유사도 합이 높을수록 중요한 문장으로 구분된다. 문장의 중요도 순서는 좌측 열의 순서이며 맨 우측 열인 Total을 기준으로 정렬된 결과이다. 각 문장에 대한 유사도 합이 높을수록 문서 내에서 중요한 문장이라 판단되어 핵심 문장 후보 문장이 된다.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
0	1.000	0.138	0.000	0.173	0.128	0.130	0.177	0.259	0.000	0.000	0.000	0.000	0.044	0.211	0.043	0.080	2.383
7	0.259	0.202	0.000	0.137	0.000	0.190	0.171	1.000	0.000	0.088	0.000	0.000	0.065	0.205	0.000	0.000	2.317
3	0.173	0.103	0.000	1.000	0.263	0.148	0.088	0.137	0.000	0.112	0.000	0.000	0.000	0.158	0.128	0.000	2.310
13	0.211	0.155	0.000	0.158	0.085	0.146	0.078	0.205	0.000	0.000	0.000	0.000	0.000	1.000	0.049	0.000	2.087
6	0.177	0.000	0.000	0.088	0.143	0.076	1.000	0.171	0.000	0.067	0.000	0.088	0.072	0.078	0.082	0.000	2.042
1	0.138	1.000	0.000	0.103	0.000	0.143	0.000	0.202	0.000	0.000	0.000	0.000	0.098	0.155	0.050	0.044	1.933
5	0.130	0.143	0.000	0.148	0.000	1.000	0.076	0.190	0.000	0.000	0.082	0.000	0.000	0.146	0.000	0.000	1.915
4	0.128	0.000	0.000	0.263	1.000	0.000	0.143	0.000	0.000	0.000	0.191	0.000	0.000	0.085	0.097	0.000	1.907
12	0.044	0.098	0.047	0.000	0.000	0.000	0.072	0.065	0.082	0.000	0.000	0.077	1.000	0.000	0.020	0.000	1.505
14	0.043	0.050	0.000	0.128	0.097	0.000	0.082	0.000	0.000	0.000	0.000	0.025	0.020	0.049	1.000	0.000	1.494
11	0.000	0.000	0.057	0.000	0.000	0.000	0.088	0.000	0.040	0.061	0.065	1.000	0.077	0.000	0.025	0.000	1.413
10	0.000	0.000	0.000	0.000	0.191	0.082	0.000	0.000	0.000	0.000	1.000	0.065	0.000	0.000	0.000	0.000	1.338
9	0.000	0.000	0.000	0.112	0.000	0.000	0.067	0.088	0.000	1.000	0.000	0.061	0.000	0.000	0.000	0.000	1.328
8	0.000	0.000	0.078	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.040	0.082	0.000	0.000	0.000	1.200
2	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.078	0.000	0.000	0.057	0.047	0.000	0.000	0.000	1.182
15	0.080	0.044	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.124

(그림 4) 문서 내 문장 유사도

5. 결론

본 연구에서는 기술의 발달로 스마트 기기와 컴퓨터를 통해 무수히 쏟아지는 무분별한 정보 속에서 사용자에게 양질의 정보를 제공하는 방법을 제안한다. 대규모의 뉴스 기사를 k-means알고리즘을 활용하여 다시 소주제로 군집화한다. 그리고 분류된 기사를 TF-IDF 가중치 모델을 이용하여 각 군집마다 하나의 핵심 문장을 추출한다. 추출된 핵심 문장은 각 군집내의 문서에서 가장 큰 비중을 차지하는 문장이다.

본 논문은 인터넷 뉴스기사 중 경제 분야 뉴스기사에 대해서만 연구를 수행하였기 때문에 타 웹문서나 분야에 대한 다른 뉴스기사에 대한 검증도 필요하다. 하지만 본 연구의 결과는 사용자가 짧은 시간을 투자하여 핵심 문장만을 읽음으로써 용을 한 눈에 알아볼 수 있다는 점에서 사용자 만족도 향상에 도움을 줄 것으로 예상되기 때문에, 향후 연구 가치를 가진다.

참고문헌

[1] Hongtao Liu, Chen Fang, Yu Wu, Ke Xu, Tian Dai. (2015). Improved K-means Algorithm with the Pretreatment of PCA Dimension Reduction. International Journal of Hybrid Information Technology, 8(6), 195-204.

[2] 윤태식, 심규석. (2012). 고차원 대규모 데이터 처리를 위한 K-means 클러스터링. 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 18(1), 55-59

[3] 유은순, 최건희, 김승훈. (2015). TF-IDF와 소설 텍스트의 구조를 이용한 주제어 추출 연구. 한국컴퓨터정보학회논문지 , 20(2), 121-129.

[4] 최승주, 김종배. (2017). Examine the Relationships Between Portal Article of Naver and Real Time Search Word Using Web Crawling. Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology. Vol.7, No.11, November (2017), pp. 787-794

[5] 남길임, 이수진, 최준. (2017). 대규모 웹크롤링 말뭉치를 활용한 신어 사용 추이 조사의 현황과 쟁점. 한국사전학, (29), 72-106.

[6] 박대서. (2018). 효과적인 뉴스기사 검색을 위한 키워드 추출 방법 연구. 한국교육학술정보원