성균관대학교
소프트웨어대학

연구작품

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

장민서
성균관대학교 소프트웨어대학

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

Jang Min Seo
Sungkyunkwan University

지도 교수: 김응모 교수님

연구실명: 데이터베이스 연구실

개 요

최근 개인에게 제공되는 통신 기술들이 발전하고, 개인 PC와 Mobile 기기가 보편화되어 많은 사람들이 인터넷에 있는 정보들에 접근하기가 쉬워졌다. 이런 기술의 발달로 사용자는 과거와는 비교할 수 없을 만큼 편리하게 정보에 접근 할 수 있게 되었지만 양질의 정보를 찾는데 아직 한계가 존재한다.

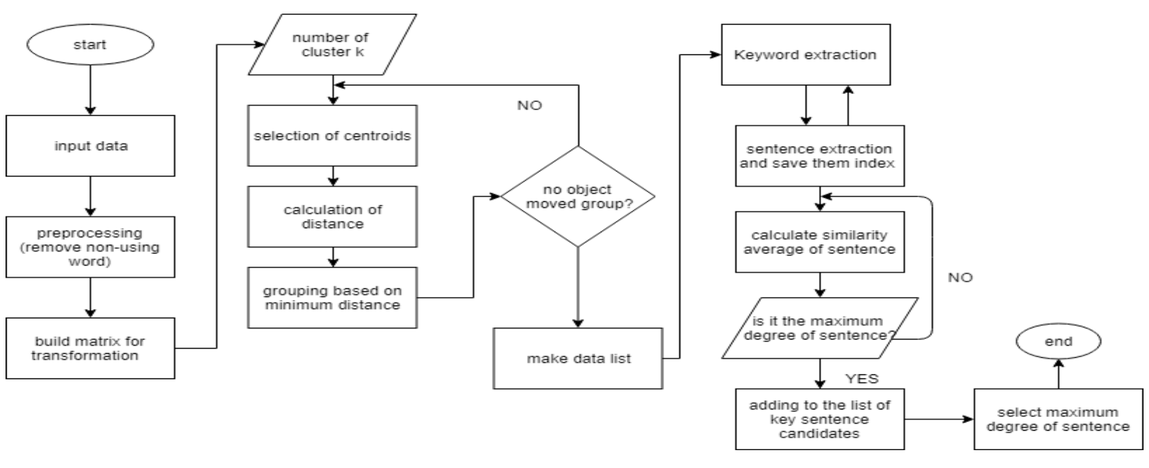
이에 본 논문에서는 관련된 정보를 탐색하여 제공하기 위해 뉴스기사 데이터를 활용하여, 대규모 뉴스기사를 소주제로 분류하는 군집 분석을 진행하였고 또한, 분류된 뉴스기사를 사용자가 빠르게 이해하고 접할 수 있도록 핵심 문장을 추출하였다.

본 논문의 연구 내용이 여러 언론사 사이트에 반영되면 사이트 품질과 사용자 만족도 향상에 기여할 수 있을 것으로 본다.

시스템 구성

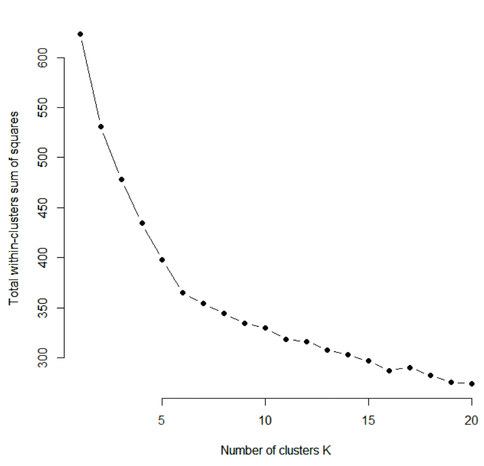
뉴스기사 수집 및 전처리 → 군집분석 -K-means 알고리즘 적용 → TF-IDF 가중치 적용 → 핵심문장 추출

데이터 수집	- **네이버 뉴스 경제분야 약 3,000개** - **오픈소스 이용 크롤링**
데이터 정제	- **R 라이브러리 KoNLP의 백과사전 사용** - **불용어 제거 및 단어 추출**
데이터 변환	- **단어와 문서 매트릭스 생성**
데이터 분석	- **K-means 알고리즘을 통한 군집분석** - **TF-IDF 가중치를 이용한 키워드 분석 및 핵심 문장 추출**



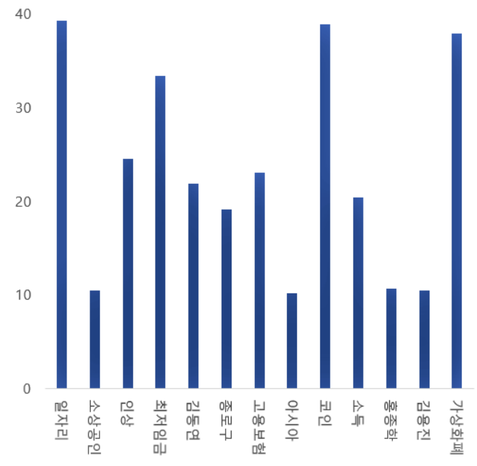
본 논문에서는 R 프로그래밍을 분석 도구로 사용하였다. 텍스트마이닝을 진행하기 위한 data.table과 KoNLP, rJava 또한 데이터 분석과 표현을 위해 arules, tm과 proxy 라이브러리 패키지를 사용하였다.
시스템은 크게 두 단계로 구성되어 있다. 먼저 K-means 알고리즘을 적용하여 기사를 k개의 군집으로 분류한다. 그 후 TF-IDF 가중치 모델을 적용하여 불필요한 문장을 제거하고 핵심 문장을 추출한다.

시스템 분석



(a) K-means 적용한 군집분석

- 대용량 기사를 소주제의 군집으로 분류
- 이 과정에서 적합한 군집 개수를 정하기 위해 elbow기법을 적용
- 수집된 데이터는 15개로 군집분류



(b) TF-IDF 적용한 키워드 분석

- 문서 내 단어에 가중치를 부여
- 주어진 가중치의 합으로 문서 내 핵심 문장 추출
- 여러 개의 핵심 문장을 후보로 1개의 핵심 문장 추출

결 과

	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 …

- K-means 알고리즘과 TF-IDF 모델을 활용하여 대용량의 기사를 소주제로 군집화 하였다.
- 빈출 단어와 문장의 유사도를 기반으로 핵심 문장을 추출하였다.
- 기사들은 총 15개의 군집으로 분류되었으며, 총 15개의 핵심 문장이 추출되었다.

	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

- 문장 간의 유사도는 행렬 형태로 나타낼 수 있다. 가로 축은 기사에 존재하는 문장의 개수이며 좌측 열은 문장의 중요도 순서이다. 맨 우측 열인 Total이 높을 수록 문서 내에서 중요한 문장이라 판단되어 핵심 문장이 된다.

결론

본 연구에서는 기술의 발달로 스마트 기기와 컴퓨터를 통해 무수히 쏟아지는 무분별한 정보 속에서 핵심 문장을 추출하여 사용자에게 양질의 정보를 제공하는 방법을 제안하였다.

하지만 인터넷 뉴스기사 중 경제 분야 뉴스기사에 대해서만 연구를 수행하였기에 타 웹문서 혹은 타 분야에 대한 검증도 필요하다.

본 연구의 결과는 사용자가 짧은 시간을 투자하여 핵심 문장만을 읽음으로써 내용을 한눈에 알아볼 수 있다는 점에서 사용자 만족도 향상에 도움을 줄 것으로 예상되기 때문에, 향후 연구 가치를 가진다.