JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 23, No. 8, pp. 1437-1445, Aug. 2022

Check for updates

휴리스틱 클러스터링을 활용한 다중 문서 요약 시스템 : 반도체 산업 사례

김 지 연[1] · 이 지 은[1] · 이 재 은[1] · 임 예 희[1] · 이 상 민[2] · 조 민 수[2*]
[1]광운대학교 정보융합학부 학사과정
[2*]광운대학교 정보융합학부 조교수

A Multi-Document Summarization System using Heuristic Clustering: The Case of Semiconductor Industry

Ji-Yeon Kim[1] · Ji-Eun Lee[1] · Jae-Eun Lee[1] · Ye-Hee Lim[1] · Sangmin Lee[2] · Minsu Cho[2*]

[1]Bachelor's Course, Information Convergence, Kwangwoon University, Seoul 01897, Korea
[2*]Assistant Professor, Information Convergence, Kwangwoon University, Seoul 01897, Korea

[요 약]

방대한 양의 텍스트 데이터가 축적되고 있는 현대 사회에서 정보를 효율적으로 추출 및 제공하기 위한 문서 요약 연구가 활발히 이루어지고 있다. 하지만 이는 주로 단일 문서 요약에 관한 연구로 다중 문서 요약에 관한 연구는 비교적 적다. 본 논문에서는 휴리스틱 군집화를 통해 다중 문서 요약을 수행하는 방법을 제시한다. 구체적으로, 본 논문은 다양한 주제의 문서 중 유사한 주제를 갖는 최적의 클러스터들을 도출하고 해당 클러스터 내의 대표 문서를 찾고, 그 후 선택된 대표 문서를 요약하는 방법을 제시한다. 본 논문에서는 반도체 산업 관련 기사 데이터를 통한 데이터 분석 결과 및 반도체 산업 다중 문서 요약 시스템을 소개한다. 본 연구는 다중 문서 요약에 관한 새로운 접근 방법을 제시하였다는 것 뿐만 아니라 문서 요약이 필요한 다양한 서비스에 이용 가능하다는 측면에서, 학계와 산업계 모두 기여점을 갖는다.

[Abstract]

In modern society, the plethora of digitalization has encouraged to produce and collect vast amounts of textual data. As such, there has been a keen interest in document summarization, which efficiently extracts and provides information in a couple of sentences. To this end, there have been numerous approaches to single-document summarization in the leading research stream, while relatively few studies on a multi-document summarization. In this paper, we propose a multi-document summarization method with heuristic clustering. More in specific, this paper suggests a method of deriving optimal clusters with similar topics among heterogeneous documents, finding a representative document for each cluster, and then summarizing the selected representative. This paper introduces a data analysis result using the collected news articles in the semiconductor industry and the multi-document summarization system. This research has a contribution both from academia and industry in suggesting a new approach for multi-document summarization and being used for numerous domains requiring document summarization.

색인어 : 문서 요약, 텍스트마이닝, 인공지능, 클러스터링, 반도체 산업
Keyword : Text Summarization, Text Mining, Artificial Intelligence, Clustering, Semiconductor Industry

http://dx.doi.org/10.9728/dcs.2022.23.8.1437

Received 29 June 2022; **Revised** 02 August 2022
Accepted 08 August 2022

***Corresponding Author; Minsu Cho**

Tel: +82-2-940-8454
E-mail: mcho@kw.ac.kr

Ⅰ. 서 론

디지털 콘텐츠의 다양화와 함께 온라인 상 의사소통 및 정보 교류가 활발해지면서 방대한 양의 텍스트 데이터가 축적되고 있다. 그러나, 비정형 텍스트 데이터의 증가 추세와는 반대로 종이책 독서량은 크게 감소하고 있다. 이는 디지털 시대의 많은 현대인들이 길이가 긴 글을 오랜 시간 동안 집중력 있게 읽지 못한다는 것을 의미하며, 독해력의 감소와 더불어 문서 요약 활용 확대로 귀결된다[1]. 과거 문서 요약은 사람이 직접 문서 내 중요한 문장을 추출하거나 재생산하는 형태로 수행되었다. 하지만, 최근 인공지능 기반의 문서 요약 알고리즘이 개발되면서 이를 활용한 서비스 이용이 급속도로 증가하고 있다.

인공지능 기반의 문서 요약은 주로 뉴스 기사에 대한 요약이 주를 이루며 이는 여러 플랫폼에서 제공되고 있다. 네이버의 '요약봇'[2], 연합뉴스의 '기사 3줄 요약 서비스'[3]가 대표적인 예시이며, 이러한 서비스는 사용자에게 중요한 정보를 빠르게 전달하는 점에서 의의가 있다. 하지만 해당 서비스들은 모두 단일 기사에 대한 요약만을 제공한다는 한계점이 있다. 즉, 단일 문서에 대한 요약 서비스는 활발히 제공되고 있는 반면 다중 문서에 대한 요약 서비스는 많지 않다.

다중 문서는 비슷한 주제를 가진 서로 다른 문서들을 말하며, 이를 요약하기 위해서는 단일 문서 요약과는 다르게 여러 문서들 사이에 중요한 핵심을 추출하는 과정이 중요하다. 이에 따라, 여러 문서들을 단일 문서로 취급한 후 요약을 진행하는 사례가 많다. 그러나, 이러한 경우 요약된 내용 간 유사성이 높아서 유의미한 문서 요약이 불가능하다는 한계를 갖는다.

이러한 한계점을 극복하기 위해, 본 연구에서는 휴리스틱 클러스터링 기법을 활용한 다중 문서 요약 방법을 제시한다. 본 연구에서의 다중 문서 요약은 중복되는 정보에 대한 접근을 줄임으로써 정보 검색의 효율성을 높이고자 수행하는 것으로, 중복 제거를 통해 선정된 개별 문서에 대한 요약으로 정의한다. 이를 위해, 단어 임베딩, 클러스터링, 문서 유사성 파악, 문서 요약의 일련의 과정을 수행한다.

본 연구 방법을 검증 및 실험하기 위하여, 반도체 산업을 연구 대상으로 선정하고 일평균 800개의 반도체 산업 관련 뉴스 기사를 수집하였다. 반도체는 4차 산업혁명의 핵심 기반 기술이며 국내 반도체 관련 기업의 성황으로 인하여 연일 많은 뉴스 기사가 생성되고 있다. 이러한 이유로 반도체 산업을 실험 대상으로 선정하였으며, 이와 더불어 해당 데이터 기반 다중 문서 요약 시스템 개발을 수행하였다.

본 논문의 구성으로는 2장에서는 본 연구의 기반이 되는 관련 연구를 제시하며, 3장에서는 데이터 수집부터 요약문의 제공까지 본 연구의 전체적인 흐름을 소개한다. 4장에는 문서의 임베딩, 클러스터링, 요약 기법 성능을 비교한 결과를 정리하며, 5장에서는 실제 반도체 관련 기사를 이용한 다중 문서 요약 서비스에 대한 내용을 포함한다. 마지막으로 6장에서 본 연구의 요약, 기여점 및 한계점을 제시한다.

Ⅱ. 관련 연구

2-1 단일 문서 요약

Kwon[4]의 연구에서는 사전학습 언어모델을 이용하여 한국어 단일 문서 요약을 진행하였다. 해당 연구에서는 인코더-디코더 형태의 사전학습 언어모델에 집중하여, 데이터 미세조정 및 정량적, 정성적 성능 평가를 함께 진행하였다. MASS, MASS+PPLM, mBART, Transformer, KoBART(Korean Bidirectional and Auto-Regressive Transformers) 5개의 언어모델을 학습하였고, 이 중 KoBART 언어모델을 기반으로 한 요약 모델에서 가장 높은 성능을 보였다. 정량적 평가 평균을 제시하여 KoBART가 추출요약 대표 알고리즘인 TextRank와 유사한 요약 결과를 나타냄을 확인하였다.

추출 요약의 일종인 TextRank[5]는 Google의 PageRank 알고리즘을 텍스트에 적용한 알고리즘으로, 문서 내 문장 간 유사도 또는 문장 내 단어 간 유사도를 통해 도출된 값이 높은 상위 n개의 단어 혹은 문장을 추출한다. 주로 문서 요약이나 키워드 추출에 사용되며, 주어진 데이터에서 내용을 추출하기 때문에 여러 항목 간의 연결을 식별하여 전체 문장을 효과적으로 요약할 수 있다. 그러나 문서의 일부를 발췌하는 형식이기 때문에 다양한 요약 표현이 어렵다는 단점이 있다.

KoBART[6]는 SKT에서 개발한 한국어 사전학습 모델로, 입력 텍스트 일부에 잡음을 추가해 원문으로 복구시키는 Auto-Encoder 형태의 BART를 40GB 이상의 한국어 텍스트로 학습시킨 모델이다. 질의응답, 요약, 번역 등 다양한 주요 자연어처리 분야에서 90% 이상의 성능을 보이고 있다. 다만 KoBART의 경우, 기존 연구에서는 원문 하나의 여러 문장을 하나의 문장으로 생성하는 단일 문서에 관한 평가만을 진행하여, 다중 문서 요약에서의 성능을 파악하기 어렵다는 한계가 존재한다. 이에 본 연구에서는 다중 문서에서의 KoBART 활용에 중점을 둔 요약 모델을 구축하고자 한다.

2-2 다중 문서 요약

Cho[7]의 연구에서는 각 문서에서 다루고 있는 소주제를 고려해 요약하기 위한 방법으로 텍스트 랭킹 알고리즘을 제안하였다. 이 연구에서는 그래프 기반의 스펙트럴 군집화로 형성된 유사 문장 군집의 영향력을 고려해 각 문장의 중요도를 계산하였다. 계산된 중요도를 바탕으로 순차적으로 문장을 추출하는 방식으로 요약을 진행하였으며, 기존 다중 문서 요약 연구들이 문장을 단어벡터의 빈도로만 표현하여 문장이 지닌 특징을 반영하지 못하는 문제를 해결하였다.

Song[8]의 연구에서는 동적 연결 그래프를 이용해 자동으로 문서를 요약하는 기법을 구현하였다. 공통으로 포함되는 단어에 대한 문장을 서로 연결해 동적 그래프를 생성할 때,

단어의 포함 여부만을 기준으로 연결 시 모든 문장이 순환 연결되는 문제를 해결하고자 최소 공통 포함 단어 수 개념을 제시하였으며, 이를 문장 길이에 따라 동적으로 제한하는 방법을 제안하였다.

이에 본 연구에서는 군집화를 통해 각 문서의 유사도를 판단하고 군집 내 문서들 간 유사도를 판단하는 과정에 랭킹 알고리즘을 적용하여 군집 별로 대표 문서를 선정하고자 한다. 또한 요약 과정에서는 전체 문장 길이를 고려하여 요약문의 개수만큼 문서를 분할하고, 분할된 각 문단에 대한 요약문을 도출함으로써 문서 전체에서 중요도가 높은 부분이 존재할 때 일부분에 치우치지 않고 전반적인 내용을 포함하도록 요약하는 방법을 제시한다.

Ⅲ. 연구 방법

3-1 연구 개요

본 연구는 그림 1과 같이 진행되며 크게 5단계로 주제 선정, 데이터 수집, 데이터 전처리, 데이터 분석 및 서비스 확장으로 구성되어 있다.

첫 번째 단계는 다중 문서 요약 플랫폼 이용자가 원하는 주제를 선정하는 것이다. 본 연구에서는 주제를 반도체 산업으로 정하고 이와 관련된 문서들을 수집하였지만, 이 외 목적에 맞게 다양한 주제를 선정할 수 있다.

두 번째, 세 번째 단계는 앞서 선정한 주제와 관련된 문서들을 수집 및 전처리한 후 중복을 제거하는 것이다. 사용자는 해당 주제에 맞는 키워드를 하나 또는 다수를 조합하여 문서 수집을 진행할 수 있다.

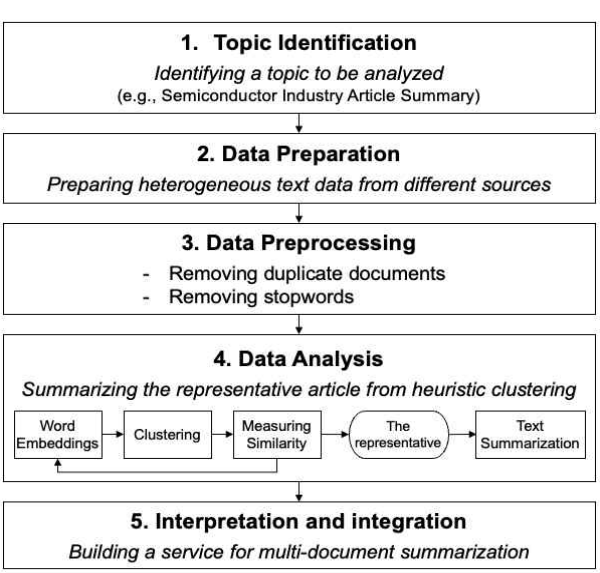

그림 1. 제안된 연구 개요
Fig. 1. An overview for the proposed research

본 연구에서는 특정 포털 사이트 검색란에 반도체 관련 키워드를 검색하여 나오는 기사를 일자별로 수집하였다. 수집 후에는 동일한 문서에 대한 중복 제거, 불용어 제거 등의 기본적인 전처리 과정을 수행한다.

네 번째 단계는 휴리스틱 클러스터링을 통한 대표 문서 추출 및 문서 요약을 목적으로 한다. 이 과정에서 단어 임베딩, 클러스터링, 유사도 측정 등의 과정을 수행한다. 수행 후 도출된 대표 문서에 대하여 요약 알고리즘을 적용한다. 이 때 문장 분리기를 이용하여 문서 전체를 임의의 세 문단으로 나누고, 각 문단마다 중심 문장을 하나씩 추출하여 총 세 문장의 요약문을 제공한다. 또한, 이 과정에서 추출요약인 TextRank, 생성요약인 KoBART 등이 활용될 수 있다. 해당 내용에 관해서는 3.2절에서 자세히 설명한다.

마지막 단계는 도출된 결과를 기반한 문서 요약 시스템 개발을 목적으로 한다. 본 단계에서는 다중 문서에 대한 핵심 요약 문장 도출을 포함한 서비스 확장이 포함된다.

3-2 휴리스틱 클러스터링

휴리스틱 클러스터링은 다수의 문서가 혼재된 상황에서 최적 개수의 클러스터를 도출하는 과정 및 특정 클러스터 내 대표 문서를 찾는 과정을 휴리스틱 방법으로 해결하는 것이다. 이 과정에서 임베딩, 클러스터링, 유사도 측정의 방법이 활용되며, 이에 대한 상세 설명을 아래와 같이 제시한다.

1) 임베딩

텍스트 데이터를 다루기 위해서는 텍스트를 벡터로 나타내는 임베딩 과정이 선행되어야 한다. 본 절에서는 다양한 임베딩 기법 중 TF-IDF(Term Frequency-Inverse Document Frequency)[9]와 Word2Vec[10], BERT(Bidirectional Encoder Representations from Transformers)[11] 총 세 가지 방법을 소개한다.

TF-IDF는 단어 빈도(TF)와 역 문서 빈도(IDF)를 곱을 이용해 전체 문서 내 빈도와 단일 문서 내 빈도 간 특정 단어 등장 비율을 산출한다. DTM(Document Term Matrix) 내의 각 단어가 갖는 중요도를 가중치로 설정하므로 전체 문서에 자주 등장하는 단어일수록 값이 작으며 이를 해당 단어의 중요도가 낮다고 판단한다.

Word2Vec은 단어 간 연관을 학습함으로써 기존 임베딩 기법에서 벡터 또는 행렬의 값이 대부분 0으로 표현되어 단어 벡터 간 유의미한 유사성을 표현하지 못하는 문제를 해결하였다. 비슷한 문맥에 쓰인 단어는 유사한 의미를 가진다는 가정하에, 단어의 의미를 다차원의 공간상에 분산시켜 벡터화한다. 본 알고리즘에는 window size를 기반으로 주변 단어들로부터 중간 단어를 예측하는 CBOW와, 중간 단어들로부터 주변 단어를 예측하는 Skip-Gram 두 가지 학습 방식이 있다.

BERT는 transformer의 encoder를 쌓아올린 형태의 임베딩 기법이다. 기존의 seq2seq처럼 encoder-decoder 구

조를 유지하지만 그 단위가 n개 존재한다는 특징이 있고, 레이블이 없는 대용량의 텍스트 데이터를 사전 훈련한 언어 모델이다. 또한, 출력 임베딩에서 self-attention을 이용하고 subword tokenizer를 사용함으로써 보다 더 문맥을 반영한 임베딩을 얻고 OOV(Out-Of-Vocabulary) 문제를 예방한다.

2) 클러스터링

임베딩 벡터를 이용하여 유사한 내용을 가진 문서를 하나의 군집에 모으기 위해 클러스터링을 진행한다. 대표적으로 K-Means[12], DBSCAN(Density-based spatial clustering of applications with noise)[13], Hierarchical Clustering[14] 까지 총 세 가지 알고리즘을 소개한다.

K-Means Clustering은 주어진 데이터를 k개의 군집으로 묶는 알고리즘으로, 각 군집 간 거리 차의 분산을 최소화하는 방식으로 동작한다. 이는 자율 학습의 일종으로, 레이블이 달려 있지 않은 입력 데이터에 레이블을 달아 주는 역할을 하며, 유사한 데이터 포인트끼리 그룹화하여 패턴을 찾아내는 것이 주된 목적이다. 그러나 이상치가 있어도 이를 확인할 수 없고 중심값에 치명적인 영향을 줄 수 있다는 단점을 갖는다.

DBSCAN은 밀도 기반의 클러스터링 기법으로, 밀도가 높은 한 점을 기준으로 반경 ε(epsilon) 내에 특정 개수 이상의 점이 있으면 해당 부분을 하나의 군집으로 인식하는 방식이다. K-means 군집화와 달리 클러스터의 수를 사전에 설정하지 않아도 되며, 클러스터의 밀도에 따라 서로 연결하기 때문에 기하학적인 모양을 가지는 군집도 잘 찾을 수 있다는 장점이 있다. 노이즈 포인트를 통하여 이상치 검출이 용이한 반면, 부분적으로 비슷한 밀도를 가진 데이터 셋의 경우엔 취약하고 중심점 처리 순서가 매번 달라서 알고리즘을 수행할 때마다 다른 결과가 나온다는 단점이 있다.

Hierarchical Clustering은 계층적 트리 모형을 이용해 개별 개체들을 순차적 및 계층적으로 유사한 개체 혹은 그룹과 통합하여 군집화를 수행하는 알고리즘으로, 최종적으로 하나의 케이스가 될 때까지 군집을 묶는다. 군집과의 거리를 기반으로 진행하면서 DBSCAN과 같이 군집 수를 사전에 정하지 않아도 학습을 수행할 수 있고, 연결이 완성된 후 Dendrogram을 생성하여 이를 시각적으로 확인할 수 있다. 그리고 Dendrogram을 적절한 수준에서 자름으로써 전체 데이터를 여러 군집으로 나눌 수 있다.

3) 문서 유사도 측정

위 과정을 통해 유사한 내용을 가진 문서끼리 군집을 형성한 후에는, 대표 문서 선정을 위한 군집 내 문서 간 유사도를 측정한다. 대표적인 유사도 측정 방법으로는 유클리드 거리 계산법을 이용한 유클리디안 유사도(Euclidean similarity)와 코사인 각도를 이용하는 코사인 유사도(Cosine similarity), 합집합에서의 교집합 비율을 측정하는 자카드 유사도(Jaccard similarity) 등이 있다.

Algorithm 1 Deriving the representative with heuristic clustering

Input : A collection of documents D, *weights* w_{inter}, w_{intra}
Output : A collection of the representative documents RD

Let $clustering_i(D)$ (i = *1* to n) be the functions to derive clusters C from a collection of documents D with different clustering algorithms and parameters. Let $distance(d_i, d_j)$ be the function to measure the similarity between d_i and d_j using the different methods, e.g., Euclidean, Jaccard, and Cosine similarity. Let *getMinimumValue(L)* be the function to identify the minimum value in a list L.

$DOCRD \leftarrow \emptyset$
forall $i \leftarrow 1$ to n **do**
  $RD_i \leftarrow \emptyset$, $C_i \leftarrow \emptyset$
  clusters $C_i \leftarrow clustering_i(D)$
  forall cluster c in C_i **do**
    forall documents D_c in cluster c **do**
    $sim_c \leftarrow inf, rd_c \leftarrow \emptyset$
      forall document $d_{c,i} \in D_c$ **do**
        $dist_{c,i} \leftarrow 0$
        forall document $d_{c,j} \in D_c$ **do**
          $dist_{c,i} \leftarrow dist_{c,i} + distance(d_{c,i}, d_{c,j})$
        if $dist_{c,i} < sim_c$ **then**
          $sim_c \leftarrow dist_{c,i}$
          $rd_c \leftarrow d_{c,i}$
      $RD_i \leftarrow RD_i \cup rd_c$
  $doc_i \leftarrow getDistanceOfClusters(C_i, RD_i, 0.5, 0.5)$
  $DOCRD$.**insert**(doc_i, RD_i)
  $DOC \leftarrow DOC \cup doc_i$
$minDOC \leftarrow getMinimumValue(DOC)$
$RD \leftarrow DOCRD[minDOC]$
return RD

Function $getDistanceOfClusters(C, RD, w_{inter}, w_{intra})$
  $inter \leftarrow 0$, $intra \leftarrow 0$, $DOC \leftarrow 0$
  forall representative documents $rd_i \in RD$ **do**
    forall representative documents $rd_j \in RD$ **do**
      $intra \leftarrow intra + distance(rd_i, rd_j)$
  forall cluster c in C **do**
    forall documents D_c in cluster c **do**
      forall document $d_{c,i} \in D_c$ **do**
        $inter \leftarrow inter + distance(d_{c,i}, rd_c \in RD)$
  $DOC \leftarrow w_{inter} * inter + w_{intra} * intra$
  return DOC

그림 2. 대표 문서 도출 과정
Fig. 2. A pseudocode for deriving the representative

4) 대표 문서 도출 알고리즘

다중 문서 요약 서비스의 핵심 중 하나는 중복되는 문서가 없어야 한다는 점이다. 데이터 수집 이후 중복 문서 제거 과정 없이 바로 요약을 하게 되면 사용자들에게 혼란을 줄 수 있다. 따라서 수집된 데이터 중 내용이 동일한 문서는 제거하고, 내용은 다르나 주제가 동일한 유사 문서에 대해서는 이들을 포괄할 수 있는 각 클러스터별 하나의 대표 문서를 선정할 필요가 있다. 그림 2는 위에서 소개한 기법을 적용하여 대표 문서를 도출하는 과정을 나타낸 의사코드이다.

먼저 이종적인 문서 집합에 대해 클러스터링을 적용한다. 이 때, 앞서 소개한 클러스터링 방법 및 파라미터를 조정하며 클러스터를 추출한다. 이 후, 도출된 군집 내 모든 문서 간 유사도를 계산하여 유사도가 가장 높은, 즉 동일 군집 내 다른 문서와의 거리가 가장 짧은 문서를 도출한다. 이 때, 유사도도 마찬가지로 유클리디안 유사도, 코사인 유사도 등 다양한 방

법이 활용될 수 있다. 이 과정 후, 문서 집합에 대한 군집 도출 및 각 군집의 대표 문서 도출이 적절한지를 판단하기 위한 거리 계산을 수행한다. 그림 2 내 함수 getDistanceOfClusters가 이 과정을 나타내며, 동일 군집 내 대표 문서와 타 문서 간 거리(inter distances) 및 모든 대표 문서 상호 간 거리(intra distances)를 모두 활용한다. 이 때, 가중치를 조정하여 두 값의 중요도를 조절할 수 있다.

이러한 문서 유사도 측정 및 대표 문서 추출 과정을 한 번만 수행하는 것이 아니라 다양한 클러스터링 알고리즘 및 파라미터를 변화시키며 다수 수행하게 된다. 즉, 휴리스틱한 접근 방법을 수행하여 함수 getDistanceOfClusters로부터 도출되는 결과 값이 가장 최소가 되는 군집 각각의 대표 문서를 추출하게 되는 것이다.

Ⅳ. 연구 결과

4-1 데이터 수집 결과

본 연구에서는 반도체 산업을 중심으로 연구를 진행하기 위해 24개의 반도체 관련 기업을 선정하고, 네이버 뉴스에 해당 기업명을 검색한 결과를 수집하였다. 데이터 수집은 오늘을 기준으로 하루 전, 즉 작일 작성된 모든 기사를 수집하였으며, 수집 항목으로는 언론사, 제목, 본문, 본문 URL, 댓글 수, 관심 수 등이 있다. 초기에는 많은 양의 데이터를 필요로 하지 않았기 때문에 테스트용 데이터 수집 시 기업별 하루 100개로 제한을 두었다. 수집한 데이터를 하나로 통합한 후 제목이 동일한 기사는 본문 내용도 동일한 것으로 판단하고 삭제하였으며, 본문의 길이가 세 문장 이하로 짧은 기사는 요약이 불필요하므로 삭제하였다.

전처리 과정에서는 본문 내용과 무관한 문자열(기자명, 이메일 주소, 작성 날짜 등)을 삭제하였으며, 한글, 영어, 숫자, 의미 있는 특수문자(%, ~ 등)만을 남겼다. 또한, 본문 데이터를 모델의 입력값으로 바로 사용할 수 있도록 문장 간 공백 문자나 개행 문자를 삭제하였으며 문서의 의미 파악에 사용되지 않는 의존 명사와 같은 불용어를 제거하였다. 이와 같은 데이터 수집 및 전처리를 거친 결과, 하루 평균 800개의 기사가 수집되었다.

4-2 휴리스틱 클러스터링을 통한 대표 문서 도출 결과

본 연구에서는 중복되는 문서들을 분류해내기 위해, 클러스터링을 통해 한 군집 내에 모인 유사한 내용의 기사 중 해당 군집의 대표 기사를 선정하고 각 군집 별 대표 기사만 활용하기로 하였다. 3장에서 소개한 TF-IDF와 Word2Vec은 문장에 포함된 단어마다 각각 하나의 벡터를 부여하는 단어 수준의 임베딩이며, 문장 단위에서 BERT를 적용한 SBERT(Sentence-BERT)는 문장 자체에 하나의 벡터를 부여한다. 다른 기사와의 유사도를 측정하기 위해서는 기사 하나당 하나의 벡터를 가져야하므로 단어 수준의 임베딩을 적용해 도출한 벡터를 기사 본문 벡터로 변환하였다. 하지만 이 과정에서 각 단어가 가지고 있던 고유한 정보가 손실되는 문제가 발생하여 기사 본문을 하나의 벡터로 임베딩하는 SBERT를 사용하였다.

이어서 데이터 분포에 따라 적합한 클러스터링 기법이 다르므로 K-Means, DBSCAN, Hierarchical clustering 알고리즘을 정성적으로 비교 분석하였다. 2021년 8월 중 하루에 수집된 반도체 관련 기사 936개에 각 알고리즘을 적용한 결과, DBSCAN은 정상적인 군집의 수가 적어 군집화에 실패하거나 하나의 군집 내에 너무 많은 데이터가 몰려있는 현상이 나타났다. K-means와 Hierarchical clustering의 경우 DBSCAN에 비해 준수한 성능을 보였으나, K-means는 문장 내 등장 단어에 치중되어 본문 내용이 유사하지 않은 기사들이 하나의 군집으로 모이는 모습을 보였다. 반면 Hierarchical clustering의 경우 분류 성능이 우수할 뿐 아니라 유사한 기사끼리 분류된 후 남은 기사가 하나의 군집에 모여 남은 기사의 처리가 용이하였다.

표 1. Hierarchical Clustering 결과 예시
Table 1. The result of Hierarchical Clustering

Index	Title	Cluster ID
…	…	…
47	Premium TV shipments of 4 million units in the second quarter… "40% OLED".	2
51	Company C OLED TV, shipments of 940,000 units in the second quarter… greatest ever.	2
53	Company C breaks through 20% of North American TV market share for the first time.	2
60	Market share '33%' Why Company B is wary of '19%' Company C OLED TV.	2
68	"LCD replacement … Mini LED has greater growth potential than OLED".	2
…	…	…
1010	U.S. stock market rises on strong tech stocks… Company D new high·Company E 1%↑. [New York Closing]	4
1012	Company F also 'falling' due to D-RAM warning… Dow·S&P 'new high' for the third day. [New York Closing]	4
1013	The Nasdaq, which towed semiconductors, reached another high… Company G 1.9%↑. [Global briefing before going to work]	4
1043	US stock market rises on strong unemployment indicators and technology stocks… Company H, Company E 2%↑.[New York Closing]	4
1058	[New York Stock Exchange] S&P 500, Nasdaq, all-time high… Semiconductor bullish, Company I, Company J↑.	4
…	…	…

표 2. Hierarchical Clustering 결과 클러스터별 주제
Table 2. The derived cluster topics with Hierarchical Clustering

Cluster ID	Topic
0	Company A's sales and stock price issue
1	Company B's M&A issue
2	Company C's new semiconductor launches and market share
3	Company D's new product launch and related issue
4	Semiconductor industry stock status by company
5	Foreign semiconductor industry related issue
…	…
39	Signed business cooperation with company E and company F

이에 본 연구에서는 가장 분류 결과가 뚜렷하고 본문 데이터에 적합도가 높았던 Hierarchical clustering을 활용하였으며, 이때 군집의 개수는 40으로 설정하여 하루 평균 약 800개의 기사 중 40개의 기사 군집을 생성하였다. 그 후 군집 내 문서 간 유사도를 측정하기 위한 기법 선정을 위해 다양한 기법 비교 및 자료 조사를 진행하였다. Bajusz et al.[15]의 연구에 따르면 Euclidean, Manhattan 등 8개의 유사도 측정법을 비교한 결과 Tanimoto index와 Cosine coefficient가 비교적 높은 유사도 측정 결과를 보였으며, 실제 뉴스 기사 데이터에 적용해 보았을 때 이전 단계에서 선정한 Hierarchical clustering과의 조합 결과가 가장 좋았던 코사인 유사도를 사용하기로 하였다.

표 1과 표 2는 기사 본문을 기준으로 Hierarchical clustering을 적용하였을 때 분류된 결과 예시와 클러스터별 주제를 정리한 표의 일부이다.

4-3 요약 결과

문서의 요약은 생성 요약 알고리즘 KoBART를 이용해 진행하였다. 추출 요약 알고리즘인 TextRank의 경우, 문서에 포함된 문장 중 가장 중요한 정보를 포함한 문장에 우선순위를 부여하고, 상위 n개의 문장을 추출하는 방식으로 요약을 진행한다. 추출 요약 알고리즘은 이와 같이 문장을 단순히 문서로부터 발췌하고 나열하므로 문장이 매끄럽게 연결되지 않는 단점이 존재한다. 따라서, 본 연구에서는 생성 요약 알고리즘을 이용하여 보다 자연스러운 요약문을 생성하고자 하였다.

생성 요약 모델인 KoBART는 많은 문장으로 이루어진 기사와 한 줄로 이루어진 요약문 쌍 데이터를 이용해 사전 학습된 모델로 적은 문장으로 이루어진 기사를 잘 요약하지 못하는 것이 관찰되었다. 이를 해결하기 위해 기사 본문에 포함된 문장의 개수가 10개 이하인 기사는 요약에서 제외하고, 전체 기사 길이의 평균보다 길이가 긴 기사의 경우 요약문에 본문이 포함하고 있는 정보를 충분히 포함시킬 수 있도록 본문을 3등분하여 세 줄의 요약문을 제공하였다.

표 3과 표 4는 이전 단계에서 Hierarchical Clustering을 통해 분류된 5번 클러스터(반도체 산업 전반 관련 기사 군집)의 한 기사에 대해 TextRank와 KoBART를 이용해 요약한 결과 예시이다. 앞에서 언급한 바와 같이 TextRank를 이용한 요약의 경우 기사 본문에 등장한 문장 중 상대적 중요도가 높은 상위 문장을 단순 추출하는 형태이므로 요약 속도가 빠르나 다양한 문장 표현이 어렵고 문맥이 다소 어색하다는 단점이 존재한다. 반면 KoBART를 이용한 요약의 경우 TextRank가 중요 문장으로 추출한 "지난 11일 이하 한국시각 A통신은 러시아의 우크라이나 침공으로 우크라이나 네온가스 공장이 문을 닫아 전세계 반도체대란이 일어날 수 있다고 보도했다. 반도체 제조를 위한 핵심 성분인 네온가스를 생산하는 우크라이나 내 주요 공장이 전쟁으로 가동을 중단했다."라는 문장을 "A 통신은 러시아의 우크라이나 침공으로 인해 우크라이나 네온가스 공장이 문을 닫아 전세계 반도체 대란이 일어날 수 있다고 보도했으며 이에 따라 글로벌 네온 공급량은 절반으로 떨어졌다고 전했다."로 표현함으로써 자연스러운 하나의 문장을 생성하는 모습을 보였다. 본 연구에서는 다중 문서에 대한 요약 알고리즘을 제시할 뿐만 아니라 우수한 성능을 보이는 요약 기법을 선정하여 시스템을 구축하는 것이 목표이므로, 문장 표현의 다양성을 보장하면서도 문맥이 자연스러운 KoBART를 선정하였다.

표 3. TextRank를 이용한 요약 결과
Table 3. The text summarization result using TextRank

#	Sentence
1	News Agency A reported on the 11th local time that the global semiconductor crisis would be worsen due to the suspension of neon gas production plants in Ukraine.
2	At the local time in Korea on the 11th, News Agency A reported that because of the Russia's invasion the Ukrainian neon gas plant was closed, which could lead to a global semiconductor crisis.
3	A major plant in Ukraine that produces neon gas, a key component for semiconductor manufacturing, has been shut down due to the war.

표 4. KoBART를 이용한 요약 결과
Table 4. The text summarization result using KoBART

#	Sentence
1	News Agency A reported that the Ukrainian neon gas plant could be closed due to Russia's invasion of Ukraine, causing a global semiconductor crisis, and global neon supply has halved.
2	If the supply of semiconductors produced at a factory located in Mariupol Kryoin, southern Ukraine, where the semiconductor giant Neon is under attack by Russian troops, cuts off the supply, it is necessary to prepare for a semiconductor crisis.
3	Neon prices in China have skyrocketed since the aftermath of the spread of COVID-19, with semiconductor makers in China pre-securing neon that can withstand at least three months, but small businesses do not, so It is expected to be hit directly.

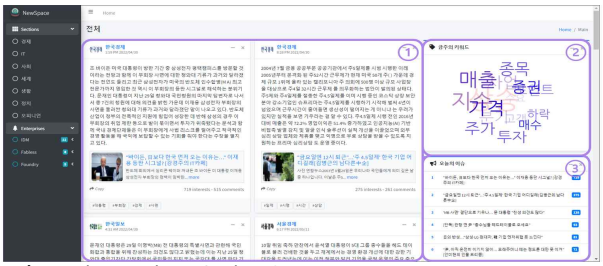

그림 3. 다중 문서 요약 시스템
Fig. 3. A multi-document summarization system

V. 다중 문서 요약 시스템

본 연구의 결과를 사용자에게 제공하기 위한 플랫폼으로 웹을 선택하였다. 해당 웹 서비스는 반도체 산업을 주제로 하고 있으며, 좌측 메뉴에서 뉴스 섹션 분류에 따른 카테고리와 반도체 기업명을 통해 원하는 기사 정보를 얻을 수 있다.

본 시스템은 그림 3에 나타난 것처럼, 카드뉴스, 금주의 키워드, 오늘의 이슈 세 파트로 구성된다. 우선 카드뉴스 내부는 언론사와 원문 기사가 쓰인 날짜, 요약모델을 통해 각 문서의 핵심 내용을 세 줄로 구성한 요약문, 원문 기사, 해당 기사의 해시태그가 포함된다. 금주의 키워드는 각 메뉴에 따라 상위 15개 해시태그로 구성되며, 기사의 관심 수가 높을수록 붉은색에 가깝고 글자의 크기가 크며, 낮을수록 푸른색에 가깝고 글자의 크기를 작게 처리하였다. 오늘의 이슈는 전체 카테고리와 기업을 대상으로 관심 수에 따라 상위 10개 기사를 정렬하였다.

VI. 결 론

오늘날 온라인상에서는 대량의 문서가 끊임없이 생성되고 있다. 다수의 문서를 하나의 문서로 보고 요약하는 연구나 논문은 찾아볼 수 있으나, 수많은 문서 가운데 중복을 제거하고 핵심 내용만을 추출한 요약문과 문서들의 수치 데이터를 분석하여 시각화 한 연구는 찾아보기 어렵다. 이에 본 연구는 반도체 관련 기사를 이용하여 효과적인 다중 문서 요약 방법을 제안하였다. 전체 문서에 클러스터링 기법을 적용하여 유사 기사를 하나의 군집에 모으고, 유사도 측정를 통해 각 군집 내에서 대표 기사를 선정함으로써 중복된 내용을 배제하였다. 대표로 선정된 기사는 각 군집에 포함된 기사 간 유사도 측정 결과 가장 유사도가 높은 기사로, 군집 내 기사에 포함된 정보를 가장 포괄하는 기사라 판단할 수 있다. 대표 기사 선정 후에는 각 기사에 요약 알고리즘을 적용하여 요약문을 생성하였고, 요약문과 함께 관련 키워드를 해시태그로 제공하는 웹 서비스를 성공적으로 구축하였다. 이번 연구에서는 반도체 관련 기사로 실험 범위를 좁혀 진행하였으나, 타 주제와 관련된 새로운 데이터로 동일한 실험을 수행한다면 다른 좋은 결과를 얻을 것으로 기대한다.

6-1 연구 기대효과

본 연구의 기대효과는 다음과 같다. 첫 번째로, 정보 검색의 효율성을 증진시킬 수 있다. 뉴스 기사, SNS, 블로그에는 다른 사람이 작성한 글을 그대로 복사하여 작성된 게시물이 난무하고 있다. 이때 다중 문서 요약 서비스를 이용하면, 중복된 내용이 포함된 여러 문서로부터 각 주제별 하나씩 대표 문서를 선정하기 때문에 보다 다양한 정보에 쉽게 접근할 수 있다.

두 번째는 다중 문서 가운데 대표 문서를 선정하고 이에 대한 요약을 진행하여 사용자에게 신뢰성 높은 요약문을 제공할 수 있다는 점이다. 기존 단일 문서 요약은 다중 문서 요약에 비해 간편하지만 해당 문서에 대해서만 진행되어 편향된 정보가 존재할 수 있다는 단점이 있다. 본 연구에서 진행한 다중 문서 요약은 각 군집을 대표하는 중심 문서를 선정하기 때문에 사용자들은 많은 양의 문서를 직접 읽지 않고도 중요도가 높은 문서의 요약문을 제공받을 수 있다.

세 번째로, 사용자의 시간을 절약할 수 있다. 길이가 긴 문서의 경우 글을 읽는 것과 더불어 핵심 정보를 파악하는 데에도 많은 시간이 소요된다. 하지만 여러 문장으로 이루어진 문서에서 주요 단어 및 요약문을 제공하는 본 서비스를 통해, 더욱 빠른 속도로 정보를 취득할 수 있다.

네 번째로, 데이터 간 연관성을 파악할 수 있다. 본 연구에서는 반도체 관련 기사를 주제로 데이터를 수집하여 주요 기사에 대한 요약문을 제공함과 동시에, 기사마다 키워드를 추출하여 해시태그로 제공하였다. 이를 통해 똑같은 해시태그를 가진 기사들을 작성 기간이나 다른 요인과는 상관없이 키워드 페이지를 통해 한눈에 파악할 수 있다는 장점이 있다.

마지막으로, 사용자가 설정한 특정 주제의 트렌드를 확인할 수 있다. 개인의 취향과 관심사가 마케팅의 주요 척도로 떠오른 요즘, 이러한 웹 플랫폼의 맞춤형 서비스 또한 사용자 관점의 서비스로서 많은 도움과 만족감을 줄 것으로 기대한다. 그리고 주제뿐만 아니라 여러 섹션과 회사 등 조건에 따라 원하는 정보를 확인할 수 있다는 점도 사용자 입장에서는 큰 장점으로 다가올 것이다.

6-2 한계점 및 발전방향

첫 번째, 요약 모델의 성능 향상에 지속적인 노력을 기울여야 한다. 현재는 반도체 관련 기사 데이터에 맞게 요약 모델을 구축하였으나, 추후 이용자가 다른 주제와 목적으로 플랫폼을 활용하고자 한다면 하이퍼파라미터 튜닝과 해당 주제의 수집 데이터를 이용한 요약 모델 업데이트 과정이 필수적이다.

두 번째, 실시간 데이터 수집과 관련하여, 현재는 하드웨어 이슈로 인해 작일 데이터를 크롤링한 후 다음 날 보여주는 형식이다. 이에 이용자가 실시간으로 문서를 제공하는 플랫폼을 원한다면 그에 맞는 하드웨어 사양이 요구되며, 해당 사항과 상관없이 실시간 서비스를 제공하기 위한 방향성 연구가 필

요하다. 이를 보완하기 위해 사용자가 직접 데이터 수집 시간을 설정할 수 있는 부분을 마련해 둔 상태이다.

세 번째, 프로젝트 확장성과 관련하여, 현재는 서비스를 제공자가 설정한 주제에 해당하는 문서만을 보여주게 된다. 즉, 서비스 제공자 외에는 문서 주제를 직접 선정하기 어려운데, 이에 서비스 이용자를 위한 관리 페이지를 개발하여 이용자가 원하는 키워드 입력을 통해 관련 문서를 수집하고 제공하는 방향을 고려하고 있다.

네 번째, 요약 모델 평가 기법에 관한 연구가 필요하다. 모델 평가 지표를 활용한 정량적 평가 방식을 전적으로 신뢰할 수는 없으나, 현재의 정성적 평가 방식을 지속하면서도 정량적 평가를 함께 진행할 수 있는 알고리즘에 대한 계획 및 구축이 필요하다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학 지원사업(2017-0-00096), 한국연구재단의 기초연구사업(NRF-2021R1G1A1094019)의 연구결과로 수행되었음.

참고문헌

[1] Children's Chosun Ilbo. Isn't reading web novels and webtoons also reading? Student reading amount 'tuck'... The use of e-books is increasing [Internet]. Available: http://kid.chosun.com/site/data/html_dir/2022/01/16/2022011600713.html.

[2] Naver. What is a summary bot service? [Internet]. Available: https://help.naver.com/support/contents/contents.help?serviceNo=997&categoryNo=19330.

[3] Yonhap News. Yonhap News launches artificial intelligence article summary service [Internet]. https://www.yna.co.kr/view/AKR20210111095300527.

[4] S. K. Kwon, "The Study on Korean Single Document Summarization Using Pre-trained Language Models," in Proceeding of the Korea Computer Congress, Jeju International Convention Center, pp. 1-3, June 2021.

[5] J. Y. Son and Y. T. Shin, "Music Lyrics Summarization Method using TextRank Algorithm," Journal of Korea Multimedia Society, Vol. 21, No. 1, pp. 45-50, Jan 2018. https://doi.org/10.9717/KMMS.2018.21.1.045

[6] SKT. Korean BART [Internet]. Available: https://github.com/SKT-AI/KoBART.

[7] H. N. Cho, J. H. Lee, "Multi-Document Summarization Based on Graph Clustering", Proceedings of the Korean Information Science Society Conference, Busan, pp. 912-914, 6, 2014.

[8] W. M. Song, Y. J. Kim, E. J. Kim, M. W Kim, "Document Summarization System Using Dynamic Connection Graph", Proceedings of the Korean Information Science Society Conference, Pyeongchang, pp. 112-113, 6, 2008.

[9] Swamy L N and D. J. V. Jordal, "Concept of TF-IDF, Common Bag of Word and Word Embedding for Effective Sentiment Classification," International Journal of Innovative Technology and Exploring Engineering, Vol. 9, No. 4, pp. 2198-2201, April 2020. https://doi.org/10.35940/ijitee.f4582.049620

[10] D. S. Kim, "Text Genre Detection Using Doc2Vec Word-Embedding Language Model," Language and Information, Vol. 23, No. 2, pp. 23–43, July 2019. https://doi.org/10.29403/li.23.2.2

[11] Y. Y. Yoon, Document embedding and classification using BERT, M.E. dissertation, Pusan National University, Pusan, 2022.

[12] R. Khan, Y. Qian, and S. Naeem, "Extractive based Text Summarization Using KMeans and TF-IDF," International Journal of Information Engineering and Electronic Business, Vol. 11, No. 3, pp. 33–44, May 2019. https://doi.org/10.5815/ijieeb.2019.03.05

[13] T. H. F. Khan, N. N. Alleema, N. Yadav, S. Mishra, and A. Shahi, "Text Document Clustering using K-Means and Dbscan by using Machine Learning," International Journal of Engineering and Advanced Technology, Vol. 9, No. 1, pp. 6327-6330, Oct 2019. https://doi.org/10.35940/ijeat.A2040.109119

[14] P. Y. D. Silva, C. N. Fernando, D. D. Wijethunge, and S. D. Fernando, "Recursive Hierarchical Clustering Algorithm," International Journal of Machine Learning and Computing, Vol. 8, No. 1, pp. 1-7, Feb 2018. https://doi.org/10.18178/ijmlc.2018.8.1.654

[15] Dávid Bajusz, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?", Journal of Cheminformatics, May, 2015, 10.1186/s13321-015-0069-3

김지연(Ji-Yeon Kim)

2018년～현 재: 광운대학교 정보융합학부 재학 중(학사과정)
※관심분야 : 데이터 사이언스, 텍스트 마이닝, 프로세스 마이닝 등

이지은(Ji-Eun Lee)

2019년～현　재: 광운대학교 정보융합학부 재학 중(학사과정)
※관심분야 : 데이터 사이언스, 데이터베이스, 인터랙션 디자인 등

이재은(Jae-Eun Lee)

2019년～현　재: 광운대학교 정보융합학부 재학 중(학사과정)
※관심분야 : 비즈니스 데이터분석, 데이터 사이언스, 텍스트 마이닝 등

임예희(Ye-Hee Lim)

2019년～현　재: 광운대학교 정보융합학부 재학 중(학사과정)
※관심분야 : 데이터 사이언스, 텍스트 마이닝 등

이상민(Sangmin Lee)

2004년 : 한국외국어대학교 (공학사)
2007년 : 한국과학기술원 (공학석사)
2019년 : 고려대학교 (공학박사)

2007년～2010년: 티맥스소프트 BP실
2010년～2020년: 삼성전자 DS센터
2020년～현　재: 광운대학교 정보융합학부 조교수
※관심분야 : 제조지능화, 헬스케어AI, 메타휴리스틱스, 기계학습, 인공지능 등

조민수(Minsu Cho)

2013년 : 울산과학기술원 (경영학사)
2018년 : 울산과학기술원 (공학박사)

2018년～2019년: 포항공과대학교 박사 후 연구원
2019년～2020년: 한국생산기술연구원 선임연구원
2021년～현　재: 광운대학교 정보융합학부 조교수
※관심분야 : 프로세스 마이닝, 자연어 처리, 산업 인공지능, 기계학습 등