http://www.jsebs.org
pISSN 2288-3908
eISSN 2765-3846

The Journal of Society for e-Business Studies
Vol.27, No.4, November 2022, pp.77-90
https://doi.org/10.7838/jsebs.2022.27.4.077

국내외 기술 이전 활성화를 위한 랭킹 정보 제공 기법

A Ranking Information Provision Technique for Activating Domestic and Foreign Technology Transfer

고효진(Hyojin Ko)*, 김민지(Minji Kim)**, 정옥란(Okran Jeong)***

초 록

정보기술이 발달함에 따라 시공간의 규제를 넘어 많은 정보를 이용할 수 있게 되었지만 이를 활용하지 못하는 경우 정보의 불균형이 발생한다. 정보 불균형은 기술이전 시에도 발생하는데, 이에 따라 국내 기술 이전을 희망하는 국내/해외 사업자들의 원활한 기술 이전이 어려운 상황이다. 본 연구에서는 국내 기술 정보를 통합 및 분석하여, 기술 이전에 유용한 정보를 제공하는 방안을 제안한다. 특정 도메인의 구조에 따라 크롤링을 진행하기 때문에 다중 도메인을 대상으로 크롤링을 수행하기엔 비용적으로 효율적이지 않고 구조가 복잡한 기존 동작 기반 웹 크롤러가 아닌 새로운 크롤링 기법이 필요하다. 또한 단어와 문서의 빈도를 기반으로 한 단순 요약 기법은 새로운 정보 응용과 참조가 빠른 기술 이전 시장에 적용하는 데 한계가 있기 때문에 페이지에 가중치를 부여하여 중요도에 따라 요약하는 추출적 요약 방식을 적용하고자 하였다. 본 논문에서는 다양한 도메인에서 정보를 수집하기 위한 전역적인 크롤러 기법과 수집한 정보를 기반으로 텍스트랭크(TextRank)를 이용한 요약 모델을 생성한다. 다중 도메인에서 한 번에 정보를 수집하고, 수집된 페이지에 가중치를 부여해 중요도가 높은 정보를 추출함으로써 최신 기술 정보의 동향을 파악할 수 있으며, 영문의 기술이전 웹사이트의 프로토타입을 구축하여 국내외 기업의 기술 이전의 활성화를 촉진하고자 하였다. 우리는 실험을 통해 크롤링 기법과 요약 모델의 성능을 검증한다.

ABSTRACT

Continuous development of information technology has allowed the use of vast information without time and space constraints. However, imbalance of information is a common problem arising from improper or lacking utilization of readily accessible information. Such phenomenon can pose a risk in trading technology, complicating smooth transaction for both domestic and foreign enterprises. This paper aims to propose a novel crawling technique of technology transfer, based on integration and analysis of domestic technology information. This novel crawling technique is more cost effective than an action-based web crawler technique, as

본 논문은 2022년 과학기술정보통신부의 재원으로 한국연구재단의 기초연구사업의 연구결과로 수행되었음. (No. 2022R1H1A2092567111298207687010 1).

* First Author, Undergraduate, School of Computing, Gachon University(2rhgywls@gachon.ac.kr)
** Co-Author, Undergraduate, School of Computing, Gachon University(rang5000@gachon.ac.kr)
*** Corresponding Author, Ph.D., Professor, School of Computing, Gachon University(orjeong@gachon.ac.kr)
Received: 2022-08-19, Review completed: 2022-11-17, Accepted: 2022-11-23

the crawling is performed on the structure of a specific domain. An extractive summarization method is introduced to weigh and summarize pages according to their importance. This combats the limitations of simple summarization techniques in the application of new information and references in fast technology trading markets. A global crawler is developed to collect information from various domains, which then allowed development of a summarization model using TextRank. Information collection from various domains, summarization according to importance and establishing a prototype of an English technology transfer website not only enables the user to capture the trend of the latest technical information, but also promotes the activation of technology transactions of domestic and foreign companies.

키워드 : 웹 크롤러, 요약 모델, 텍스트 랭크, 키워드 추출, 기술 이전
Web Crawler, Summarization Model, TextRank, Keyword Extraction, Technology Transfer

1. 서 론

인터넷의 도입 이후 전 세계는 하나의 망으로 연결되어 있다. 자료 계산적 측면이 강했던 기존의 정보 기술과 달리 현재의 정보 기술은 정보 통신의 역할이 주가 되며 물리적 제약이 거의 없이 다양한 정보를 습득하고 사용할 수 있게 되었다. 필요한 정보를 얻을 수 있다는 편의성이 있지만 넘쳐나는 '정보의 과잉' 시대 속에서 필요한 정보만을 효율적으로 얻어오지 못해 정보 격차가 생긴다면, 이로부터 오는 피해는 정보화 시대 이전보다 더 막대할 것이다. 특히 기술 이전 시 발생하는 정보 불균형은 국내외 원활한 기술 교류를 방해하여 단순 정보 불균형 문제를 넘어 더 큰 문제가 발생할 가능성이 있다.

한국산업기술진흥원에서 집계한 기술 이전 사업화 실태 조사보고서에 따르면, 국내 연도별 기술 이전 계약 체결 건수가 매년 지속적인 성장 추세를 보인다. 이러한 상황 속에서 국내 기술 소상공인들의 사업 성장 및 보유 기술에 대한 경쟁력 제고를 위해 하나의 사이트에서 모든 정보를 확인하고 교류할 수 있는 정보 제공 웹사이트가 절실히 필요하다.

이를 위해 본 연구에서는 다양한 국내 공공기관 사이트에 게시되어 있는 기술들을 분석하여 최신 트렌드와 연구 동향을 하나의 포털 사이트로 제공하는 방법을 제시한다. 한 개의 특정 사이트가 아닌, 여러 사이트를 포괄적으로 탐색하며 기술 이전 시장의 전반적인 정보를 빠르고 쉽게 얻기 위한 전역적인 크롤링 방식이 요구된다. 하지만 다양한 기술 정보를 제공하는 여러 기술 사이트들은 도메인별로 다른 구조로 되어 있어 원하는 정보를 한 번에 수집하는 것에 제약이 있다. 이러한 도메인별 특이성에 구애받지 않는 크롤링을 위해서는 각 사이트가 공통으로 제공하는 정보를 이용해야 한다. 기술 사이트들은 도메인별로 다른 구조로 되어 있으나 대부분의 사이트가 기술 정보를 제목이나 페이지 요약문에 기재하였다. 이러한 특성을 가진 다중 도메인에서 공통된 형식의 정보를 수집하기 위해서는 도메인 특징적인 기존의 크롤러가 아닌 전역적인 크롤러가 필요하다. 웹 크롤러를 통해 얻은 데이

터로 하이퍼링크 구조를 가지는 문서에 상대적 중요도에 따라 가중치를 부여하는 방법인 페이지랭크(PageRank)를 corpus에 적용했다. 페이지랭크는 더 많은 링크를 받는 사이트가 더 중요한 사이트라는 가정에 기초하여 동작하는 알고리즘이기 때문에 이를 적용해 다른 사이트들로부터 참조를 많이 받는 기술이 중요한 기술이란 판단을 내릴 수 있다. 따라서 트렌드에 민감한 기술 이전 분야에 페이지랭크 기술을 도입하는 것이 적합하다. 페이지랭크를 기반으로 한국어에 적용할 수 있는 랭킹 알고리즘과 요약 모델을 구성하였고, 그 결과를 딥러닝 번역 모델을 통해 해외에서도 정보를 이용할 수 있게 하는 프로토타입을 구현하였다.

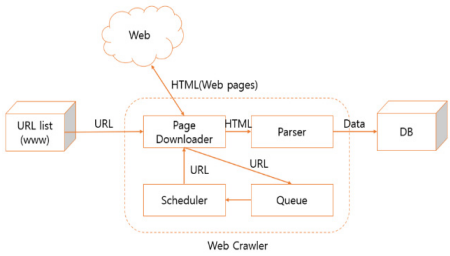
본 연구에서는 특정 도메인의 구조에 구애받지 않는 크롤링 기법과 순위화 기법을 통한 키워드, 핵심 문장 추출로 국내외의 최신 기술 동향을 효율적이고 빠르게 파악할 수 있게 하였고, 이를 다국어로 번역하여 제공하는 프로토타입을 구현한다. 여러 기술 사이트에서 얻어온 정보를 통합하여 핵심 내용을 언어적 한계에 구애받지 않고 제공하는 플랫폼을 제안함으로써 흐름이 빠르게 변화하는 기술 이전 시장에서 정보의 불균형을 완화할 수 있을 것으로 기대한다.

본 논문의 구성은 다음과 같다. 제1장은 서론이며 제2장은 프로토타입 구현과 관련된 연구를 소개한다. 제3장은 전체적인 모델과 제안 모델 내 세부 기법들의 기능을 설명하고 제4장에서 본 모델의 성능은 실험을 통해 평가한다. 제5장에서는 제안된 프로토타입의 전체적인 흐름을 소개하고 마지막 제6장에서는 결론을 다루며 마무리한다.

2. 관련 연구

2.1 다중 도메인 크롤러

빅데이터 분석을 통한 예측이 여러 분야에 걸쳐 사용되면서 빅데이터의 가치가 더욱 커지고 있다. 빅데이터 수집을 위해 여러 정보가 저장된 웹페이지의 HTML 구조를 파싱 하며 데이터를 모으는 작업을 크롤링이라고 한다. 웹 크롤러의 기본 구조는 <Figure 1>과 같다.



〈Figure 1〉 Basic Architecture of Web Crawler

Na and On[9] 기존 웹 크롤러는 URL에서 얻어온 페이지들로부터 데이터를 수집해 DB에 저장하는 방식으로 동작한다. 도메인별 html의 구조에 맞춰 크롤링을 수행하므로 이에 파생되어 웹 크롤링의 목적과 구조에 따른 다양한 크롤링 알고리즘이 연구되고 있다. 하지만 본 연구에서는 다중 도메인을 대상으로 도메인별 구조에 구애받지 않고 정보를 한 번에 수집하는 전역적인 크롤러가 필요하다.

Han and Lee[2] 최근 연구에서는 버튼 클릭, 스크롤 등과 같은 동작을 통해 로딩된 데이터에서 HTML을 파싱 하는 동작 기반 웹크롤러에 대해 다루고 있다. 하지만, 여러 기술 사이트는 도메인별로 다른 구조로 되어 있으나 대부분

공통으로 기술 정보가 제목이나 페이지 요약문에 기재 되어 있다. 따라서 해당 논문에서는 다중 도메인을 위한 정적 페이지에서 파싱 하는 크롤러를 제안한다. 주로 HTML을 파싱하기 위한 도구로는 BeautifulSoup[13] 파이썬 라이브러리를 사용한다. BeautifulSoup는 HTML이나 XML 파일에서 데이터를 파싱하는 파이썬 라이브러리이다. Requests와 urllib를 사용해 서버를 요청하고 html, xml 데이터를 로컬로 저장한 후, 데이터 파서를 통해 저장한 데이터를 분석한다. requests 모듈을 사용하여 http 페이지에 요청을 보내고 get(url) 메소드를 통해 URL에서 정보를 가져와 Html.parser로 받아온 결과의 html tag 구조를 분석하는 과정이 진행된다. Beautiful soup는 동적인 크롤링이 아니기 때문에 웹페이지의 모든 데이터를 한 번에 가져온 뒤 필요한 정보를 추출하는데, 정보 추출 시 find, find_all 함수 등을 통해서 조건에 맞는 tag의 데이터를 가져오거나 텍스트와 링크 등을 가져온다[20].

본 논문에서는 두 개의 스레드를 생성하여 도메인별 하이퍼링크와 인덱스를 추출한다. BeautifulSoup로 Depth에 따라 링크를 크롤링하고 이를 모든 도메인에 공통으로 적용하기 위해 'a' 태그, 'href' 속성으로 하이퍼링크를 얻는다. 그 이후 의미 있는 정보만을 가져오기 위해 웹페이지 내부의 Internal/External link를 구분해 크롤링하고 이 중 Internal link만을 별도로 저장하였다. 이렇게 크롤링 된 내부 링크에서 필요한 기술 정보에 관한 내용만이 추후 랭킹 알고리즘에서 사용된다.

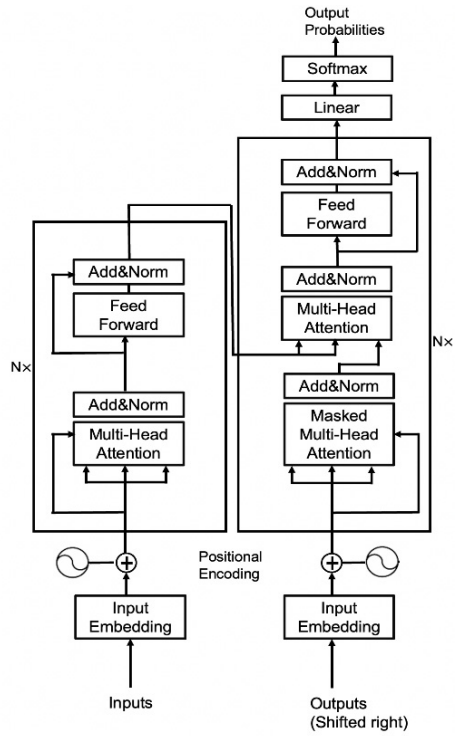
2.2 키워드 추출과 요약 기법

최근 인공지능 기술의 빠른 발전으로 이젠 인공지능이 사람의 능력을 따라잡을 정도의 수준으로 개발되는 추세이다. 특히 시청각 분야의 발전으로 인해, 사람과 못지않은 정확도로 사람의 언어를 이해할 수 있게 되었다. 이때까지 쌓아온 광범위한 양의 빅데이터와 함께 기계가 이를 스스로 학습할 수 있게 되면서 사람의 언어를 이해, 추출, 분류하는 것을 넘어 직접 텍스트를 생성하는 NLP 기술이 빠르게 발전했다[4]. 이에 따라 이에 관련한 여러 연구가 진행되고 있다. 본 연구에서는 추출된 정보를 이용해 최신 기술 동향을 파악하는 요약문을 생성하기 위해 코버트(KoBERT), 텍스트랭크(TextRank) 등 몇 가지 요약 모델을 적용했다. 텍스트 요약은 크게 추출적 요약과 추상적 요약으로 분류된다. 추상적 요약은 원문에 없는 문장과 단어 구도 사용해 핵심 문맥을 반영한 새로운 문장을 생성한 요약 방법이다. 사람이 요약하는 것과 같은 방식으로 볼 수 있다. 이는 인공신경망을 사용하는 방법이기 때문에 추출적 요약보다 난이도가 높고 지도 학습을 위해 실제 요약문이라는 Label data가 필요하기 때문에 데이터를 구성하기 까다롭다. 대표적인 모델로는 Seq2Seq[16]가 있다.

Seq2Seq 모델 중 가장 활발한 연구가 진행되고 있는 트랜스포머(Transformer)[14] 모델은 시퀀스 데이터 분석을 통해 맥락과 의미를 학습하는 방식이다. 이는 시퀀스 데이터를 기반으로 하는 기계 번역, 추상 요약, 음성 인식 처리에 적합하다. 최근 자연어 처리 분야에서 가장 많이 채택하는 기법이며 트랜스포머 모델이 제안된 이후 추상 요약 관련 연구의 성능이 크게 향상되었다. 트랜스포머의 주요 요소는 크게 Self-Supervision과 Self-Attention[17]으로 나뉜다. 이는 NLP에서도 Self-Supervised Learning을 통해 거대한 데이터 세트에서 generalizable

representation을 배울 수 있게 하였고, Self-Attention[18]은 스스로 Attention 연산을 수행하게 하여 CNN, RNN을 사용한 모델들과 다르게 최소한의 inductive bias를 가정한다. 최종적으로 주어진 시퀀스의 관계를 학습하며 넓은 범위의 문맥을 고려할 수 있게 되었다.

<Figure 2>에서 볼 수 있듯이, 제안된 트랜스포머의 구조는 기존의 언어 모델과 다르게 RNN 없이 언어 처리가 가능하다. 하지만 여전히 시퀀스 내 거리가 먼 데이터들 사이의 관계를 학습하는 데 있어 한계가 존재한다. 기존 모델의 한계를 새로운 형태의 위치 인코딩 방법으로 해결하였으며, key와 value 벡터에 pseudo-recurrent connection[1] 형태를 도입하였다. 이 외에도 Decoder만을 포함한 트랜스포머 모델인 OpenAI의 GPT 시리즈[11], Bi-directional 구조를 도입한 BERT[19] 등 기존 트랜스포머를 변형한 연구들이 진행되고 있다. 특히 BERT를 과학 기술 분야 데이터에 적용한 SciBERT[5]이 제안되었다. SciBERT는 여러 도메인으로부터 얻은 데이터에 BERT를 활용하여 훈련하고, 각 도메인 데이터별 파인튜닝을 수행하는 사전학습 언어모델이다.



〈Figure 2〉 Model Architecture of Transformer

본 연구에서 역시 BERT를 한국어 처리에 적합하게 변형한 모델인 KoBERT와 masked language modeling task로 pre-training을 하는 기법을 적용하는 ELECTRA[7] 역시 시도해 보았으나, 지도 학습이기 때문에 라벨 데이터가 필요하였고, 적합한 라벨 데이터를 생성하지 못하여 추상적 요약 방식은 적합하지 않다고 판단하였다.

추출적 요약은 중요도가 높은 핵심 문장이나 구절을 원문에서 추출해, 이를 이용한 요약문을 만드는 방법이다. 따라서 추출적 요약을 통해 생성된 요약문의 문장과 단어들은 모두 원문에서 찾아볼 수 있다. 추출적 요약으로 대표되는 알고리즘은 텍스트랭크(TextRank)[8]가 있다.

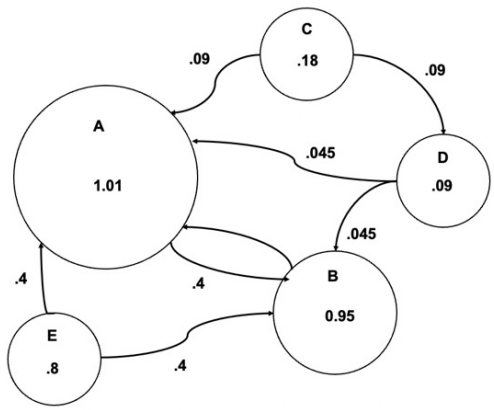
텍스트랭크는 Google의 페이지랭크(PageRank)[10]를 활용한 텍스트 그래프 기반 순위 모델이며, 키워드 추출과 문장 추출 방법을 제공한다. 페이지랭크는 문서 간 인용과 참조로 연결된 웹 문서들에 상대적 중요도에 따른 가중치를 부여하는 방법이다. 외부 사이트가 참조를 많이 한 페이지는 높은 페이지랭크를 가진다. <Figure 3>에서 볼 수 있듯이 참조를 가장 많이 당한 Page A는 가장 높은 페이지랭크를 가지며 이 페이지랭크는 Page A를 참조한 Page C, D, B, E 각각의 PR을 해당 page의 전체 링크(L)로 나눈 값의

합이다. 텍스트랭크는 이 알고리즘에 착안하여 문서 내의 문장과 단어를 이용해 문장의 랭킹을 계산하는 알고리즘이다. 텍스트랭크의 식은 (1)과 같다.

$$TR(V_i) = (1-d) + d^* \sum_{V_j \in \ln(V_i)} \frac{W_{ji}}{\sum_{V_k \in Out(V_j)} W_{jk}} TR(V_j) \quad (1)$$

식 (1)에서 w_{ji}는 문장 또는 단어 V_i에 대한 텍스트랭크 값을 의미하며, 는 문장 또는 단어 i와 j사이의 가중치이다. d는 페이지랭크에서 웹 서핑을 하는 사람이 다른 페이지로 이동하는 확률을 뜻한다. 텍스트랭크는 모든 문장 또는 단어 V_i에 대해 $TR(V_i)$를 계산한 뒤 이를 높은 순으로 정렬하여 텍스트 요약에 활용한다. 텍스트랭크는 페이지랭크와는 다르게 가중치 그래프를 기반으로 한다. 텍스트랭크 키워드 추출은 두 단어 사이의 동시 발생 관계를 사용하고 있으며 문장 추출 시에는 두 문장에 공통으로 포함되는 단어의 개수를 각 문장의 단어 개수의 로그 값의 합으로 나눈 값을 사용한다. 또 다른 최신 연구의 경우 PDF 형태의 논문을 대상으로 질의를 던지고, 질의에 부합하는 문서를 대상으로 추출 요약을 수행하는 IBM Science Summarizer가 제안되었다. 논문에서는 크롤링 된 데이터에 텍스트랭크 모델을 적용해 키워드와 요약문을 추출해 최신 기술 트렌드 분석에 사용한다.

또한, TF-IDF[12], Kr-WordRank[3]를 사용해 랭킹 알고리즘을 생성하고 이를 통해 키워드를 추출하였지만, 이를 이용해 요약문을 생성하는 데에는 한계가 있고 비효율적인 모델 구조가 필요하기 때문에 랭킹 알고리즘과 텍스트 요약을 한 번에 진행할 수 있는 모델인 텍스트랭크가 가장 적합하다고 판단하였다. 랭킹 알고리즘을 이용한 키워드 추출과 요약문 생성을 따로 진행하지 않고 동시에 수행할 수 있는 텍스트랭크를 제안 모델에 적용해 키워드와 요약문을 추출하였다.

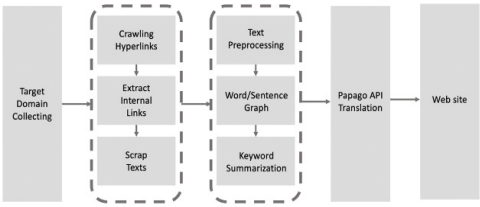


〈Figure 3〉 Example of PageRank

3. 제안 모델

본 연구에서는 도메인의 HTML 구조 특이성에 구애받지 않고 모든 사이트에서 사용할 수 있게 구현한 전역적인 웹 크롤러 모델에 단어 그래프와 문장 그래프를 구축한 뒤 그래프 랭킹 알고리즘인 페이지랭크를 이용해 랭킹 알고리즘과 요약 모델을 구현하였다. 또한 그렇게 생성된 한국어 결과를 번역 모델을 통해 다국어로 번역하여 해외에서도 이용할 수 있는 사이트로 정보 제공이 가능한 모델을 제안하였다. 제안 모델은 크게 세 단계로 이루어져 있다. 첫 번째는 크롤링 과정이다. 이는 DFS(Depth-First Search)를 사용해 웹페이지 내부의 하이퍼링크를 탐색한다.

두 개의 스레드를 생성하여 도메인별 하이퍼 링크들과 인덱스를 추출하여 저장하고, 추출된 정보 속에서 기술 관련 정보만을 가져오는 과정이다. 두 번째는 첫 번째 과정에서 추출된 정보들을 정제하고 이를 이용하여 그래프를 구축하고, 랭킹 알고리즘을 통해 키워드와 요약문을 생성하는 과정이다. 텍스트랭크를 이용해 단어와 문장별 그래프를 구축하고 그래프의 랭킹 스코어를 계산하여 최신 기술 동향을 파악할 수 있게 한다. 다음은 이 결과를 가지고 해외 기업이나 단체가 언어에 구애받지 않게 하기 위해 번역 모델을 통해 다국어로 번역하고, 이를 웹페이지에서 시각화하여 정보를 이용할 수 있게 한다.

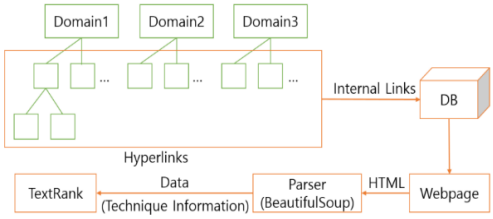


〈Figure 4〉 Overall Model Architecture

<Figure 4>는 전체적인 모델의 구조를 보여준다. 모델의 구조는 크게 3가지로 되어 있다. 타깃 도메인에서 정보를 가져오는 웹 크롤러, 수집한 정보를 요약을 위한 텍스트랭크 요약 모델, 요약된 결과를 영어로 변환한 한영 번역 모듈이다.

3.1 웹 크롤러

크롤러는 깊이 우선 탐색을 기반으로 한 크롤러 구조를 설계한다. 깊이 기반 크롤링은 Root URL에서부터 Child URL을 통과하는 깊이 우선 탐색으로 정보를 수집한다. 각 웹페이지 내에 하이퍼링크가 존재하면 깊이를 증가시키고, 부모-자식 관계에 우선순위를 두어 그 내부의 하이퍼링크까지 탐색한다. 해당 연구에서는 링크 내 자식 노드인 하이퍼링크를 계속 추가하고, 이후 링크를 내부 링크와 외부 링크를 구분해 정보를 가져온다. 이 결괏값을 랭킹 알고리즘에도 적용한다. 링크 구분 과정을 다시 살펴보자면 저장된 URL들을 외부, 내부 링크로 나누어 팝업, 광고 사이트와 같은 외부 링크는 제외하고 유효성이 있는 내부 링크만 별도로 저장한다. 내부 링크의 HTML에서 BeautifulSoup 라이브러리를 통해 기술 정보를 가져와 이를 추후에 키워드 추출과 요약문 생성에 사용한다.



〈Figure 5〉 Proposed Architecture of Crawler

<Figure 5>의 크롤러 모델은 링크 내 하이퍼링크 중 내부 링크만을 따로 저장하고, 내부 링크를 대상으로 웹페이지에서 HTML 콘텐츠를 가져와 데이터를 수집하는 과정을 보여준다.

3.2 키워드 추출과 요약문 생성

문서 요약을 위해 키워드와 핵심 문장을 선택하는 추출적 요약 기법으로 텍스트랭크를 사용하였다. 크롤링 된 텍스트를 받아와 문장과 단어 단위로 분절 후, 단어와 문장에 대해 각각의 TF-IDF 그래프를 구축하고 이 그래프를 이

용해 텍스트랭크 알고리즘을 적용한다. 랭킹값이 높은 순으로 정렬한 뒤 높은 순서의 키워드와 요약문을 반환하는 구조이다. 우선 키워드를 추출하기 위해 단어 그래프를 생성하였다. 주어진 문서 집합을 벡터 공간으로 표현하고 문서의 특징을 단어 가중치의 집합으로 표현하고 문서 내에 나타난 단어의 가중치를 TF-IDF를 사용해 계산하였다. 그 뒤 만들어진 단어 그래프에 페이지랭크를 학습시킨다. 단어 간의 빈도수와 관계성을 기반으로 단어들을 학습시킨 후 텍스트랭크를 이용해 핵심 문장을 추출한다. 이를 위해 우선 문장 그래프를 만들어야 한다. 각 문장을 노드로 두고 문장 간 유사도를 edge weight로 지정한다. 이때 문장 간 유사도를 구하기 위해 코사인 유사도와 텍스트랭크의 유사도를 모두 구현한다. 제안된 텍스트랭크의 문장 간 유사도 척도는 식 (2)와 같다.

$$sim(s1, s2) = \frac{|\{wk|wk \in S_1 \& 1 \& wk \in S_2|\}}{log|S_1| + log|S_2|} \quad (2)$$

위 식에서 볼 수 있듯이, 두 문장 i, j에 공통으로 등장한 단어 w_k의 개수를 각 문장에 공통으로 등장한 단어의 개수 S_i, S_j의 값의 합으로 나누어 두 문장 간 유사도를 구할 수 있다.

이렇게 문장과 문장 간 가중치로 생성된 그래프를 기반으로, 핵심 문장 추출을 진행한다. 추출된 핵심 문장은 곧 요약문이기 때문에 이를 최종 결과물로 결정한다.

4. 실 험

4.1 실험 설정

본 연구에서는 제안 모델과 다른 요약 모델을 비교하는 지표로 Rouge[6] 점수를 사용하였다. ROUGE(Recall-Oriented understudy for Gisting Evaluation) 점수는 사람이 만든 요약본과 자동으로 만든 요약본을 비교하는 지표로, n-gram이나 word sequence 등의 개수를 비교

〈Table 1〉 List of Websites for Crawling

Company	Link address
SM TECH	https://www.tipa.or.kr/
NTB	https://www.ntb.kr/main/mainPortal.do;jsessionid=AE473AA8E405FEEAD88E9EAA6E3E737B
ETRI	https://www.etri.re.kr/intro.html
IP Market	https://www.ipmarket.or.kr/ko/
KIOM	https://www.kiom.re.kr/
KARI	https://www.kari.re.kr/kor.do
ETechS	https://rnd.compa.re.kr/web/irndMain.do;jsessionid=jDayhx9bGxy1lTsNbaqKnx0sLGiwyt4lPWm5gPa7D35tXtZcbxYTqagxyH7w5VWs.VM-INT-01_servlet_irnd
KETEP	https://www.ketep.re.kr/
KIMST	https://www.kimst.re.kr/
KIOST	https://www.kiost.ac.kr/kor.do
KAERI	https://www.kaeri.re.kr/
KORAD	https://www.korad.or.kr/korad/index.do
NFRI	https://blog.nfri.re.kr/kor/index
KRISS	https://www.kriss.re.kr/

하여 측정한다. uni-gram, bi-gram으로 평가하는 Rouge-1, Rouge-2, 가장 긴 문자열을 매칭하여 평가하는 Rouge-L을 사용하였다. 데이터세트는 국내 기술 정보를 모아놓은 다양한 사이트를 대상으로 하였다. <Table 1>은 사이트 목록 중 일부를 나타낸다.

4.2 비교 모델

우리는 요약문 성능을 비교하는 모델로 텍스트랭크와 파이썬 summarization 라이브러리[15]를 사용하였다. 텍스트랭크란 그래프 랭킹 알고리즘의 일종으로, 단어나 문장을 노드로 표현하여 동시 출현 관계의 단어를 엣지로 나타낸다. 문장 단어 간 유사도를 통해 각 노드 간의 랭킹을 계산하여 핵심 문장과 단어를 추출한다. 파이썬 Summarize library는 자연어 처리 패키지인 NLTK를 기반으로 하여 만들어진 간단한 다국어 요약 패키지이다.

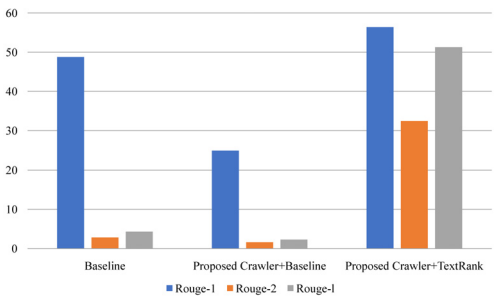
4.3 실험 결과

키워드, 핵심 문장 추출을 통한 주요 기술 요약이 제대로 진행되고 있는지 확인하기 위해, 타 요약문과의 결과 비교를 수행하였다.

<Table 2>와 <Figure 6>은 본 논문에서 직접 구현한 모델이 생성한 요약문, 자체적으로 직접 생성한 요약문, 타 모델이 생성한 요약문과의 결과 비교이다. 직접 생성한 요약문의 경우 타당성을 위해 각 사이트에서 제공하는 핵심 키워드들을 조합하여 생성하였다. 대조군의 노이즈를 최소화하기 위해 타겟 사이트를 “국가 지식재산거래 플랫폼(https://www.ipmarket.or.kr/ko/)”과 “ETRI기술이전홈페이지(https://itec.etri.re.kr/itec/sub06/sub06_01.do)” 두 사이트를 대상으로 실험을 진행하였다. 대상 사이트는 기술 이전 사이트 목록 중 가장 다양한 분야의 기술을 제공하고 있고, 해당 사이트의 이용자가 가장 많기 때문에 대표성을 가진다고 판단하여 위 두 사이트를 선정하였다.

〈Table 2〉 Performance Result

	Rouge-1	Rouge-2	Rouge-L
Baseline	48.76	2.8	4.3
Proposed Crawler +Baseline	24.88	1.6	2.25
Proposed Crawler +TextRank	56.41	32.43	51.28



〈Figure 6〉 Model Comparison

Baseline은 도메인 특징적인 기존의 크롤링 기법을 사용해 각 사이트의 정보를 수집한 후, 얻어진 데이터를 합쳐 하나의 새로운 데이터세트를 생성하였고, 해당 데이터세트에 파이썬의 자동 요약 라이브러리인 “summarization”을 적용하였다. 또 다른 대조군으로는 제안된 전역적인 크롤러를 사용해 두 사이트의 정보를 한 번에 크롤링하고, summarization 라이브러리를 적용해 요약을 수행하였다. 마지막으로

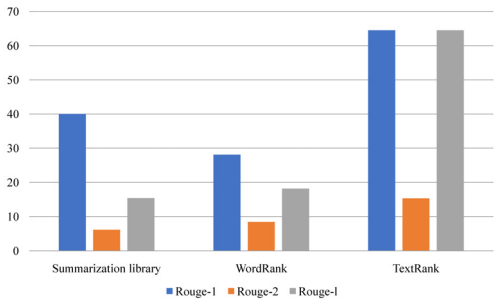
제안 모델의 성능을 파악하기 위해 전역적인 크롤러를 사용해 생성된 데이터세트에 텍스트랭크를 적용해 생성된 요약문을 실험군으로 설정하였다. 요약문의 정확도를 평가하기 위한 연구자 본인이 직접 생성한 요약문을 참조 요약본으로 사용하였다. 또한 추가로 텍스트랭크의 유효성을 판단하기 위해 같은 타겟 도메인에 특징적인 크롤링을 사용해 얻은 문서에 텍스트랭크, Summarization, WordRank으로 요약문을 생성하는 Ablation 연구를 진행하였다.

제안 모델로 생성된 요약문의 경우, Rouge-1 56.41%, Rouge-2 32.43%, Rouge-l 51.28%로 보다 높은 성능을 보였다. 제안 요약 모델의 Rouge-1, Rouge-2는 문장 간 중복되는 유니그램과 바이그램의 수가 가장 높다는 것을 나타내며, 문서 전체를 통틀어 문자열 내 최장 길이의 매칭 수를 나타내는 Rouge-l 역시, 보다 우수한 결과가 나왔다. Rouge 점수 분석을 통해 타 모델들로부터 생성된 요약문보다 높은 품질의 요약문을 생성하는 것으로 판단된다. 하지만 제안된 크롤러와 타 요약 라이브러리를 사용하였을 때 baseline보다 좋은 성능을 보이지 못했는데, 이는 전역적인 크롤러를 사용하였을 때보다 각 도메인의 구조에 맞춰 얻어온 정보가 보다 사용자가 필요로 하는 정보를 가져올 수 있기 때문이라고 보인다. 본 논문에서는 제안 방안의 불완전성과 노이즈 최소화로 인해 대표성을 가진 기술 사이트 2개를 선정하여 실험을 진행하였다. 하지만 시장의 전반적인 동향 파악에는 한계가 있기 때문에 차후 대상 사이트를 확장한 실험을 진행하고, 직접 모델에 적용하며 더욱더 유의미한 기술 동향 파악을 위한 확장이 필요하다. 현재까지 진행된 연구를 기반으로 노이즈 최소화를 위한 방안과 한국어에 적합한 새로운 요약 기법을 적용한다면 더 좋은 성능으로 기술 이전 시장에 크게 기여할 것으로 기대된다.

Ablation 연구 결과, 역시 텍스트랭크를 적용하였을 때 Rouge-1 64.5로 가장 높은 성능이 나온 것을 확인할 수 있으며, 이는 WordRank의 키워드를 많이 포함한 문장을 핵심 문장으로 판단하는 방식보다 텍스트랭크의 토크나이징이 된 문장 간 유사도를 이용하는 방식이 더 적합한 요약 기법이라고 판단된다.

〈Table 3〉 Result of Ablation Study

Model	Rouge-1	Rouge-2	Rouge-L
Summarization library	39.99	6.15	15.38
WordRank	28.1	8.42	18.18
TextRank	**64.5**	**15.32**	**64.5**

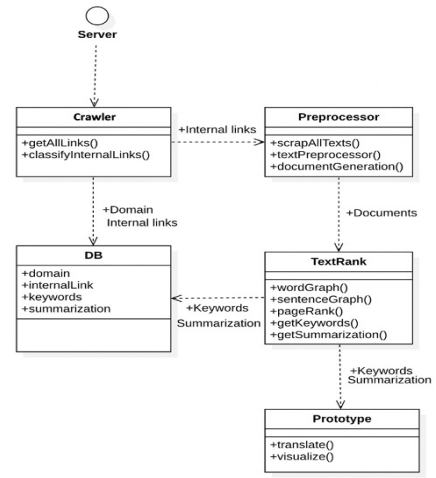


〈Figure 7〉 Model Comparison of Ablation Study

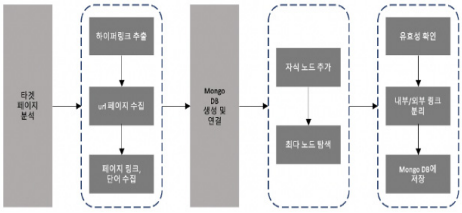
본 연구의 목적은 도메인 구조에 구애받지 않는 전역적인 크롤링 기법이 필요하고, 실제 그렇게 얻어온 정보에 필요한 정보를 얻을 수 있게 전처리를 진행하고 한국어 요약에 더욱 효과적인 제안 요약모델을 적용하였을 때, 좋은 성능을 보였다.

5. 구현 결과

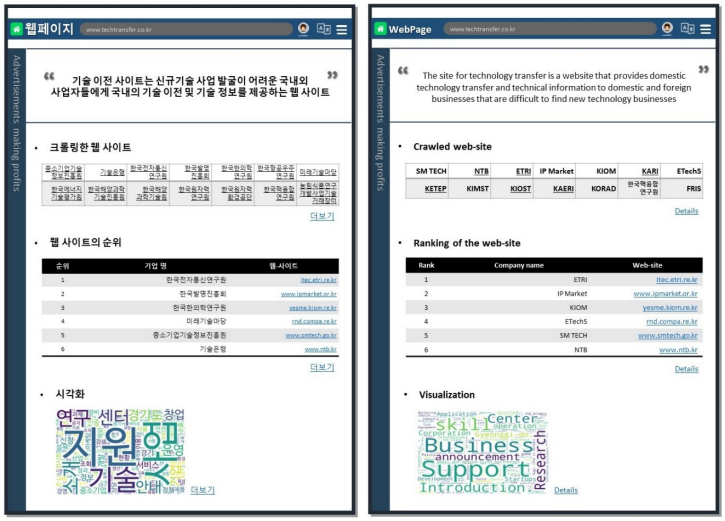
전체적인 구조는 <Figure 8>과 같다. 서버의 요청을 받은 크롤러는 주어진 도메인 내의 모든 링크를 수집하고, 이를 외부 링크와 내부 링크로 분류해 얻어진 내부 링크를 전처리기로 넘겨준다. 이때 파이썬의 BeautifulSoup로 <Figure 9>와 같은 진행 과정을 통해 HTML 태그를 파싱하고, 필요한 텍스트 데이터만을 추출하여 사용한다. 추출된 텍스트를 문서화시켜 텍스트랭크를 적용한다. 크롤링한 텍스트를 받아와 문장과 단어 단위로 분절 후, 단어와 문장에 대해 각각의 TF-IDF 그래프를 구축한다. 여기에 텍스트랭크 알고리즘을 적용해 랭킹값이 높은 순서로 정렬한 뒤, 설정한 개수대로 키워드와 요약문을 반환한다. 반환된 결과를 딥러닝 번역 모듈을 사용해 텍스트를 번역한 결과를 반환한다. 최종 결과물을 확인할 수 있는 웹사이트의 구조는 <Figure 10>과 같다.



〈Figure 8〉 Model Architecture



〈Figure 9〉 Process of Web Crawling



〈Figure 10〉 Website Structure

대상 사이트들과 중요도에 따른 웹사이트의 순위를 볼 수 있다. 또한 트렌드를 파악할 수 있게 중요 키워드를 국문과 영문의 워드 클라우드 형태로 사용자들에게 가시적으로 제공한다. 이러한 웹 사이트를 통해 국내/외의 개인, 사업가들이 최신 동향을 파악하고, 이에 따라 기술 이전을 쉽고 효과적으로 할 수 있다.

6. 결 론

여러 사이트에 분포되어 있는 정보를 한 번에 확인할 수 있는 모델을 설계하여 필요한 정보들을 시공간, 언어의 제약을 넘어 이용할 수 있게 하였다. 기술 사이트가 가진 특성을 활용하여 다양한 구조의 웹사이트에서 정보를 동시에 추출하였으며, 텍스트랭크 기반의 랭킹 알고리즘과 요약기법으로 기술 동향을 파악할 수 있게 하였다. 또한, 인공지능 기계번역 API를 통해 국내외 사용자들의 정보 이용이 가능하다. 하지만 현재는 실제 요약문이 존재하지 않기 때문에 객관적인 성능 평가가 쉽지 않고 추출적 요약 방식으로는 모델 자체의 언어 표현 능력 제한으로 인해 생성된 문장들이 자연스럽지 않다는 한계점이 있다. 차후 충분한 학습이 이루어지고 라벨 데이터로 활용할 수 있는 실제 요약문이 생성된다면 추출적 요약 방식이 아닌 Attention을 이용한 추상적 요약 방식을 도입할 수 있게 된다. 추가로, 실험 대상 사이트의 확장도 가능하다. 따라서 NLG(Natural Language Generation) 영역의 요약을 도입해 위 과정을 진행한다면 더 자연스러운 요약문 생성이 가능하고, 현재보다 설득력 있는 실험을 진행할 수 있다. 결론적으로 기술 동향을 쉽게 파악할 수 있을 것이라 기대된다.

References

[1] French, R. M., “Using pseudo-recurrent connectionist networks to solve the problem of sequential learning,” Proceedings of the 19th Annual Cognitive Science Society Conference. Vol. 16, 1997.

[2] Han, D.H. and Lee, Y.-K., “Design of action-based web crawler structural configuration for multi-website management,” KIISE Transactions on Computing Practices, Vol. 27, No. 2, pp. 98-103, 2021.

[3] Kim, H., Cho, S., and Kang, P., “KR-WordRank an unsupervised korean word extraction method based on wordrank,” Journal of the Korean Institute of Industrial Engineers Vol. 40, No. 1, pp. 18-3, 2014.

[4] Lee, D. and Kim, K., “Web site keyword selection method by considering semantic similarity based on word2vec,” The Journal of Society for e-Business Studies, Vol. 23, No. 2, pp. 83-96, 2018.

[5] Lee, Y.-J. and Choi, H.J., “Joint learning-based KoBERT for emotion recognition in Korean,” Korea Advanced Institute of Science and Technology, pp. 568-570, 2020.

[6] Lin, C.-Y., “Rouge: A package for automatic evaluation of summaries,” Text

summarization branches out," Text Summarization Branches Out, pp. 74-81, 2004.

[7] Luong, M.-T., Le, Q. V., and Manning C.D., "Electra: Pre-training text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, 2020.

[8] Mihalcea, R. and Tarau, P., "Textrank: Bringing order into text," Conference on empirical methods in natural language processing, pp. 404-411, 2004.

[9] Na, C.-W. and On, A.-W., "A proposal on a proactive crawling approach with analysis of state-of-the-art web crawling algorithms," Internet Comput. Serv., Vol. 20, No. 3, pp. 43-59, 2019.

[10] Page, L., Brin, S., Motwani, R., "The PageRank Citation Ranking: Bringing Order to the Web., Technical Report. Stanford InfoLab, 1999.

[11] Radford, A., Narasimhan, K., Salimans, T., and Sutskevet, I., "Improving language understanding by generative pre-training," OpenAI, 2018

[12] Ramos, J., "Using tf-idf to determine word relevance in document queries," Proceedings of the first instructional conference on machine learning, Vol. 242, No. 1. 2003.

[13] Richardson, L., "Beautiful soup documentation," Dosegljivo: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Dostopano: 7. 7. 2018], 2007.

[14] Song, E.-S., Kim, M., Lee, Y.R., and Ahn H., "Transformer-based text summarization using pre-trained language model," Korean Society of Computer Information, Vol. 29, No. 2, pp. 395-398, 2021.

[15] Srinivasa-Desikan, B., "Natural language processing and computational linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras," Packt Publishing Ltd, 2018.

[16] Szűcs, G. and Huszti, D., "Seq2seq deep learning method for summary generation by LSTM with two-way encoder and beam search decoder," IEEE 17th International Symposium on Intelligent Systems and Informatics (SISY), IEEE, 2019.

[17] Tay, Y., Bahri, D., Metzler, D., and Juan, D. C., "Synthesizer: Rethinking self-attention for transformer models," International conference on machine learning, pp. 10183-10192, PMLR, 2021.

[18] Vaswani, A., Shazeer, N., and Parmar, N., "Attention is all you need," Advances in neural information processing systems, Vol.30, 2017.

[19] XBathija, R., Agarwal, P., Samanna, R., and Pallavi, G.B., "Guided interactive learning through chatbot using bi-directional encoder representations from transformers (bert)," 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), IEEE, pp. 82-87, 2020.

[20] Zheng, C., He, G., and Peng, Z., "A Study of Web Information Extraction Technology Based on Beautiful Soup," J. Comput., Vol. 10, No. 6, pp. 381-387, 2015.

저 자 소 개

고효진 (E-mail: 2rhgywls@gachon.ac.kr)
2019년~현재 가천대학교 AI소프트웨어학부 소프트웨어학과 (학사)
관심분야 빅데이터, 머신러닝, 딥러닝, Conversational AI

김민지 (E-mail: rang5000@gachon.ac.kr)
2018년~현재 가천대학교 AI소프트웨어학부 소프트웨어학과 (학사)
관심분야 빅데이터, 머신러닝, 딥러닝, Conversational AI

정옥란 (E-mail: orjeong@gachon.ac.kr)
2005년 이화여자대학교 컴퓨터공학과 (공학박사)
2005년~2006년 서울대학교 컴퓨터공학부 박사후 연구원
2007년 Univ. of Illinois of Urban Campaign
2008년~2009년 성균관대학교 정보통신공학부 연구교수
2017년 Univ. of California, Irvine 방문교수
2009년~현재 가천대학교 AI소프트웨어학부 교수
관심분야 빅데이터, 머신러닝, 딥러닝, Conversational AI