유의어 사전 기반 환경기술 검색 시스템 설계

PIAO XIANGHUA, YIN HELIN, 구영현, *유성준
세종대학교 컴퓨터공학과
hianghua729@gmail.com, yinhelin0608@gmail.com, yhgu@segjong.ac.kr,

*sjyoo@sejong.ac.kr

Design of environmental technology search system using synonym dictionary

PIAO XIANGHUA, YIN HELIN, Yeong Hyeon Gu, *Seong Joon Yoo
Departments of Computer Engineering, Sejong University

요 약

국가기후기술정보시스템은 국내 환경기술과 국외의 수요기술 정보를 제공하는 검색 시스템이다. 그러나 기존의 시스템은 유사한 뜻을 가진 단일 단어와 복수 단어들을 모두 식별하지 못하기에 유의어를 입력했을 경우 검색 결과가 다르다. 이런 문제점을 해결하기 위해 본 연구에서는 유의어 사전을 기반으로한 환경기술 검색 시스템을 제안한다. 이 시스템은 Word2vec 모델과 HDBSCAN(Hierarchical Density-Based Spatial Clustering of Application with Noise) 알고리즘을 이용해 유의어 사전을 구축한다. Word2vec 모델을 이용해 한국어와 영어 위키백과 코퍼스에 대해 형태소 분석을 진행한 후 단일 단어와 복수 단어를 포함한 단어를 추출하고 벡터화를 진행한다. 그 다음 HDBSCAN 알고리즘을 이용해 벡터화된 단어를 군집화 해주고 유의어를 추출한다. 기존의 Word2vec 모델이 모든 단어 간의 거리를 계산하고 유의어를 추출하는 과정과 대비하면 시간이 단축되는 역할을 한다. 추출한 유의어를 통합해 유의어 사전을 구축한다. 국가기후기술정보시스템에서 제공하는 국내외 기술정보, 기술정보 키워드와 구축한 유의어 사전을 Multi-filter를 제공하는 Elasticsearch에 적용해 최종적으로 유의어를 식별할 수 있는 환경기술 검색 시스템을 제안한다.

Ⅰ. 서론

최근 다양한 환경문제로 인해 환경보호에 대한 인식이 증가함에 따라 각 나라의 환경기술 수요가 늘고 있다. 환경기술은 환경의 자정능력을 향상하고 사람과 자연에 대한 피해를 감소하는 기술이다. 미국, 독일 등 선진국은 체계적으로 환경기술을 개발해 이미 많은 기술을 보유하고 있다. 하지만 개발도상국은 환경기술 개발이 상대적으로 늦기 때문에 선진국의 환경기술을 필요로 한다. 한국의 국가기후기술정보시스템 (Climate Technology Information System, CTIS) [1]은 국내의 환경기술과 국외의 수요기술 정보를 제공하는 검색 시스템이다. 그러나 CTIS의 검색 시스템은 단순 단어를 검색하기 때문에 단일 단어와 복수 단어의 유의어를 모두 식별하지 못해 유사한 뜻을 가진 단어에 대해 검색이 되지 않는 문제점이 있다.

대부분의 정보 검색 시스템은 WordNet과 같은 데이터베이스를 이용해 유의어를 식별한다. WordNet은 단일 단어로 구성되어 있지만 환경기술에 대한 단어는 복수 단어로 구성되는 경우가 많기에 WordNet을 유의어 사전으로 이용하면 적합하지 않다. 예를 들면 복수 단어 “미국 국가”를 단일 단어

“미국”과 “국가”로 나뉠 수 있다. 하지만 ‘국가’는 나라와 노래를 의미하기에 이 단어 하나로만 단어의 의미를 판단하지 못한다. 따라서 본 연구에서는 단일 단어와 복수 단어를 포함한 유의어 사전을 기반으로 하는 환경기술 검색 시스템을 제안한다. 단일 단어와 복수 단어를 포함한 유의어 사전의 구축에는 Word2vec[3] 모델과 HDBSCAN[7] 알고리즘을 이용한다. Word2vec 모델은 한국어와 영어 위키백과 코퍼스를 이용해 단일 단어와 복수 단어가 모두 포함된 단어를 추출하고 벡터화를 진행한다. 그리고 벡터화된 단어를 HDBSCAN 알고리즘에 적용해 군집화 해준다. 위의 과정은 벡터화된 단어들을 Word2vec에 적용해 유의어를 추출하는 과정과 대비하면 시간을 단축할 수 있다. 그 다음 추출한 유의어를 통합해 유의어 사전을 구축한다. 이어서 Elasticsearch 검색엔진에 유의어 사전과 CTIS에서 제공하는 국내외 기술정보와 키워드를 적용해 역색인 목록을 구축한다. 사용자가 기술에 대한 설명을 입력하면 구축한 역색인 목록을 이용해 기술 설명에 포함된 단일 단어와 복수 단어로 구성된 유의어를 포함한 결과가 출력되게 한다.

본 논문의 구성은 다음과 같다. 2장 관련 연구에서는 유의어 사전을 검색 시스템에 적용한 연구, Word2vec 모델을 이용한 연구, Word2vec 모델과 클러스터링 알고리즘을 혼합해 사용한 연구 및 HDBSCAN 알고리즘과 다른 클러스터링 알고리즘의 성능을 비교하는 연구에 대해 소개한다. 3장 시스템 설계에서는 전체 시스템에 대해 간략히 소개하고 세부적으로 데이터 셋, 유의어 사전 구축과 검색엔진 구축에 대해 소개한다. 검색엔진 구축에서는 전처리 과정과 역색인 및 검색에 대해 소개한다. 마지막으로 4장에서는 본 연구의 결론과 향후 계획에 대해 서술한다.

Ⅱ. 관련 연구

기존 연구 [2-8]에서는 다양한 방법을 이용해 문장에서 유의어를 추출했다.

연구 [2]에서는 유의어 사전을 기반으로한 정보검색 시스템을 구축했다. 해당 연구에서는 한-영 사전과 영-한 사전을 이용해 유의어 후보 목록을 얻었다. 후보 목록의 빈도수, 사전의 위치 정보와 입력한 명사 정보를 이용해 유의어를 확정하였고 유의어 사전을 구축했다. 유의어 사전을 이용한 유의어 추출에 대해서는 약 78%의 F-score를 얻었다. 해당 연구의 유의어 사전은 단일 단어 기반이기 때문에 복수의 단어를 식별하지 못하는 문제점이 있다. 예를 들어 '수질 오염'은 단일 단어 사전에서는 '수질'과 '오염'으로 나누어지기 때문에 복수 단어인 '수질 오염'을 식별하지 못한다. 최근에 단일 단어로 구성된 사전보다 복수 단어가 포함된 사전을 더 많이 사용한다.

연구 [3]에서는 법률 문서(law document)를 Word2vec 모델과 Bag-Of-Word(BOW) 모델에 적용하고 코사인 유사도 알고리즘을 이용해 문서 간의 유사도를 분석하는 실험을 진행했다. 실험측정 결과 Word2vec 모델의 정확도는 약 80%이고 이는 BOW 모델의 정확도에 비해 약 20% 정도 향상되었다. 연구 [4]에서는 영어 위키백과 코퍼스의 320,000개의 기사를 Word2vec 모델에 적용해 키워드를 추출하고 코사인 유사도 알고리즘을 이용해 유사도를 계산했다. WordSim-353 데이터 셋과 SimLex-999 데이터 셋을 Window Size가 3, 6, 9이고 Vector Dimension이 50, 150, 300인 모델에 적용해 각각 실험을 진행했다. 실험측정 결과 Window Size가 9이고 Vector Dimension이 300일 때 66.5%와 28.45%의 가장 높은 정확도를 얻었다. 따라서 Window Size와 Vector Dimension은 키워드의 유사도에 영향을 주기에 가장 최적인 값을 찾는 것이 중요하다는 것을 알 수 있다.

연구 [5]에서는 20개 다양한 종류의 뉴스 데이터 셋을 Word2vec 모델과 K-means Clustering 알고리즘에 적용해 각 종류의 뉴스에 대해 분류했다. 이 연구에서는 11,314개의 테스트 데이터를 군집을 나누지 않았을 때 약 75.24%의 F-score를 얻었고 5000초의 시간이 소요되었다. 군집의 개수를 500, 1000, 1500, 2000으로 설정했을 때 각각 75.06%, 75.3%, 76.19%, 77.4%의 F-score를 얻고 300, 400, 600, 800초의 시간이 소요되었다. 연구 [6]에서는 Word2vec 모델을 이용해 영어 위키백과 코퍼스에서 키워드를 추출하고 Spectral Clustering 알고리즘과 K-means Clustering 알고리즘을 각각 이용해 군집화를 진행했다. 유의어 추출 성능 비교 결과 Spectral Clustering은 약 77.5%의 F-score를 얻었고 K-means Clustering은 약 35.1%의 F-score를 얻었다. 연구 [5-6]에서 사용한 Spectral Clustering과 K-means Clustering은 사용자가 군집 개수를 직접 설정해야 하는 단점이 있다. 그리고 Word2vec 모델을 이용해 키워드를 벡터화 하고 클러스터링 알고리즘을 이용해 벡터화된 키워드를 군집화 하고 유의어를 추출하는 것은 Word2vec 모델만 이용해 유의어를 추출하는 것보다 시간이 적게 소요된다는 것을 알 수 있다.

연구 [7]에서는 Doc2vec 알고리즘을 이용해 스웨덴 Bombardier Transportation AB의 온보드 열차 제어 시스템 데이터를 사용해 유사도를 계산했다. 그리고 HDBSCAN 알고리즘과 Fuzzy C-means(FCM) 알고리즘을 각각 사용해 군집화를 진행했다. HDBSCAN 알고리즘의 Min_cluster_size(각 클러스터의 최소 단어 개수)를 2로 정했을 때 제일 높은 80%의 정확도와 75%의 F-score를 얻었고 FCM 알고리즘은 군집의 개수를 6으로 정했을 때 제일 높은 67%의 정확도와 54%의 F-score를 얻었다. 연구 [8]에서는 10,000개의 기사가 포함된 데이터 셋을 이용해 뉴스의 주제를 식별했다. 그리고

VSM(Vector Space Model)을 적용한 후 HDBSCAN 알고리즘, HAC(Hierarchical Agglomerative Clustering) 알고리즘과 K-means Clustering 알고리즘을 각각 이용해 키워드를 추출하고 최종적으로 PMCC(Pearson Product-moment Correlation Coefficient)를 통해 이상치를 탐지했다. 해당 연구에서는 HDBSCAN, HAC와 K-means Clustering에 대해 각각 60.2%, 50.5%, 54.6%의 F-score를 얻었다.

기존 연구에서 구축한 유의어 사전은 단일 단어로 구성되어 복수 단어를 식별하지 못하는 단점이 있다. 최근 유의어를 추출하는 연구에서는 시간을 단축하기 위해 Word2vec 모델과 클러스터링 알고리즘을 혼합해 사용한다. 기존 연구에서 사용한 K-means Clustering, Spectral Clustering 등 알고리즘은 사용자가 군집 개수를 설정해야 하기 때문에 한 개 군집에 몇 개의 유의어가 포함될지 확정하지 못한다. 하지만 HDBSCAN 알고리즘은 Min_cluster_size(각 클러스터의 최소 단어 개수)와 Min_samples(각 트리의 노드에 있는 단어 개수)를 설정할 수 있다. 본 연구에서는 Min_cluster_size를 3, Min_samples를 5로 설정해 15개의 의미있는 단어를 유의어로 추출하게끔 설계한다. 따라서 본 연구에서는 Word2vec 모델과 HDBSCAN 알고리즘을 이용해 단일 단어와 복수 단어로 구성된 유의어 사전을 기반으로 하는 환경기술 검색 시스템을 제안한다.

Ⅲ. 시스템 설계

본 연구에서 제안하는 유의어 사전 기반 환경기술 검색 시스템은 크게 유의어 사전을 구축하는 부분과 사전을 검색엔진에 적용하는 부분으로 구성되고 그 구조도는 그림 1과 같다.

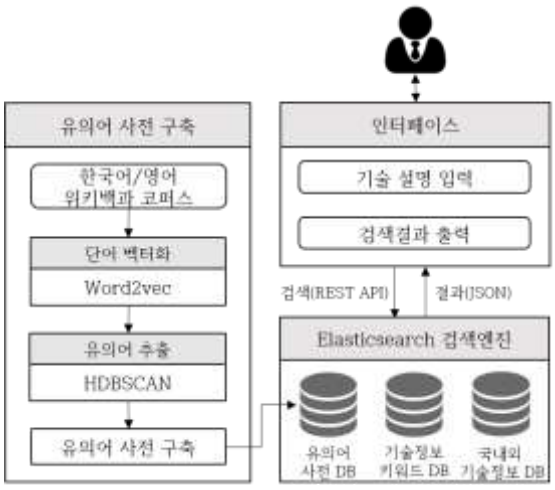

그림 1. 유의어 사전 기반 환경기술 검색 시스템 구조도

그림 1에서 유의어 사전을 구축하는 부분은 Word2vec 모델을 이용해 한국어와 영어 위키백과 코퍼스에서 단어를 추출하고 추출한 단어에 대해 벡터화를 진행한다. 그리고 HDBSCAN 알고리즘을 이용해 유의어를 추출하고 유의어 사전을 구축한다.

구축한 유의어 사전, 국내외 기술정보와 기술정보의 키워드를 Multi-filter 기능을 제공하는 Elasticsearch 검색엔진에 적용시켜 사용자가 단일 단어와 복수 단어가 포함된 기술 설명을 입력하면 단일 단어와 복수 단어의 유의어를 포함한 검색 결과가 출력되게 한다.

1. 데이터 셋

본 연구에서는 유의어 사전을 구축하기 위해 위키백과에서 제공하는 한국어와 영어 코퍼스를 사용한다. 영어 코퍼스의 크기는 17GB이고 한국어 코퍼스의 크기는 701MB이다. 코퍼스에서는 약 50만개의 한국어 기사와 약 1200만개의 영어 기사가 포함되어 있는 Xml 형식의 파일을 제공한다. Xml 파일에는 제목(title), 기사(revision_text) 외 7가지 속성을 제공하지만 본 연구에서는 제목과 기사만 사용한다. 위키백과 코퍼스에서 제공하는 제목은 단일 단어와 복수 단어로 구성되었기에 본 연구에서 목적하는 과와 같이 유의어를 추출하는 데 도움이 된다.

2. 유의어 사전 구축

Word2vec 모델과 HDBSCAN 알고리즘을 사용해 유의어 사전을 구축한다. Word2vec 모델은 CBOW(Continuous Bag-Of-Words)와 Skip-gram을 이용해 단어 임베딩(Word Embedding)을 진행한다. CBOW는 주어진 컨텍스트(Context) 상의 키워드를 예측하는 알고리즘이고 Skip-gram은 키워드로부터 컨텍스트 단어들의 분포를 예측하는 알고리즘이다. 본 연구에서는 주어진 키워드를 이용해 코퍼스에서 유의어를 찾아야 하기 때문에 Skip-gram 알고리즘을 사용한다. Skip-gram 알고리즘은 단어 주변 k개의 단어를 문맥으로 보고 Window 창을 옆으로 이동하는 방식으로 학습 데이터를 생성한다.

Skip-gram 알고리즘은 입력층, 하나의 은닉층과 출력층으로 구성된 신경망(Neural Network)이다. 은닉층은 기사 단어의 개수와 은닉층 뉴런의 개수로 이루어진 가중치 매트릭스이고 가중치 매트릭스의 행이 바로 기사 단어들의 벡터이다. 입력층에 학습 데이터를 입력하고 출력층 대신

Softmax 함수를 이용해 단어를 벡터화 한다. 위의 과정으로부터 얻은 결과물이 벡터화 된 단어이다.

HDBSCAN 알고리즘은 Min_cluster_size와 Min_samples를 기반으로 학습을 통해 단어를 군집화 해준다. 벡터화 한 단어들 사이의 거리를 이용해 MST(Minimum Spanning Tree)를 만든다. 그 다음 구성된 요소들의 클러스터의 계층을 구축하고 설정한 최소 구성요소의 개수를 기반으로 클러스터의 계층을 축약한다. 이어서 축약된 계층에서 설정한 범위내에 있는 클러스터만 선택해 유의어 사전을 구축한다. 유의어 사전의 구조는 표 1과 같이 유의어에 속하는 단어들을 한 Line에 포함되게 한다.

표. 1 유의어 사전 예시

유의어
재정 정책, 경제 정책
액체 수소, 우주로켓의 산화제
고농도 과산화수소, 과산화수소, HTP
대한민국 국가, 애국가, 한국환상곡
석영유리, 유리

3. 검색엔진 구축

입력한 기술 설명에 전처리를 하려면 Multi-Filter 기능이 필요하다. 따라서 본 연구에서는 이러한 기능을 제공하는 Elasticsearch 검색엔진을 선택했다.

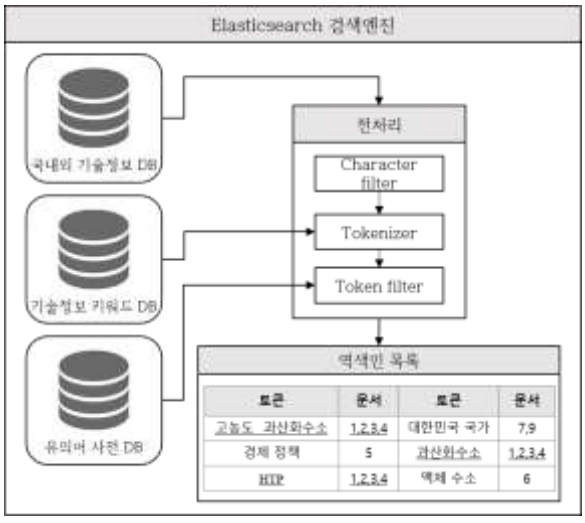

그림 3. Elasticsearch 검색엔진 구조도

Elasticsearch 검색엔진의 구조도는 그림 3과 같이 전처리 단계에서 기술 정보 키워드 DB는 Tokenizer에 적용되고 유의어 사전 DB는 Token filter에 적용된다. 국내외 기술정보 DB를 전처리 과정에 입력해 얻은 토큰과 국내외 기술정보의 문서 번호를 이용해 역색인 목록을 구축한다. 그 다음 기술 설명을 Elasticsearch 검색엔진에 입력하고 기술 설명에 대해 전처리를 진행해 토큰을 얻는다. 그리고 토큰이 역색인 목록에 있는지 확인하고 있으면 해당되는 문서를 결과로 출력한다.

3.1 전처리

Apache Lucene 기반의 분산 검색엔진 Elasticsearch는 Character filter, Tokenizer, Token filter 세 가지의 전처리 기법을 제공한다. Character filter에는 Html 태그를 제거하는 Html strip 방법, 단어를 다른 단어로 바꿔주는 Mapping 방법, 정규식을 이용해 문장 내용을 바꾸는 Pattern replace 방법이 있다. 예를 들면 '&'를 'and'로 바꿔주는 것과 같이 교체하거나 삭제한다. Tokenizer는 분리하는 기준에 따라 토큰을 생성하는 기법이다. 예를 들면 Whitespace tokenizer는 공백을 단위로 나누어 하나하나의 토큰을 얻는다. 그러나 기술 설명에 기술정보 키워드 DB에 존재하는 단일 단어와 복수 단어가 있으면 한 토큰으로 정한다. 예를 들면 '고농도 과산화수소'에 whitespace tokenizer를 적용하면 '고농도'와 '과산화수소'로 나뉘지만 기술정보 키워드 DB에 '고농도 과산화수소'라는 단어가 있을 경우 '고농도 과산화수소'가 한 개 토큰으로 된다. Token filter는 지정하는 규칙에 따라 토큰을 처리하는 기법이다. Synonym filter를 예로 들면 위의 과정을 거쳐 얻은 '고농도 과산화수소'가 유의어 사전에 있는지 확인한다. 유의어 사전에 있으면 "고농도 과산화수소"와 유의어에 속하는 단어(과산화수소, HTP)들을 모두 토큰으로 출력한다. Elasticsearch 검색엔진의 전처리 과정은 그림 4와 같은 Character filter, Tokenizer, Token filter 과정을 통해 토큰화 된 결과물이 출력된다.

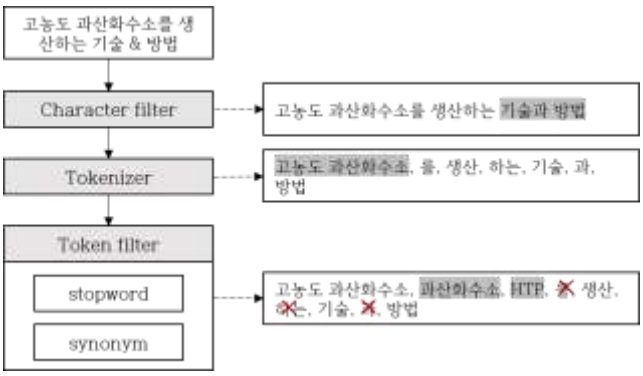

그림 4. Elasicsearch 전처리 과정 예시

3.2 역색인 및 검색

역색인은 토큰을 통해 문서를 찾아내는 목록을 의미한다. 본 연구에서 역색인 목록을 구성하는 방법은 다음과 같다. 국내외 기술정보 DB를 전처리에 입력하여 토큰을 얻는다. 토큰과 대응되는 문서의 번호를 기록해 얻은 목록을 역색인 목록이라 한다.

표 2. 역색인 목록 예시

토큰	문서	토큰	문서
고농도 과산화수소	1, 2, 3, 4	대한민국 국가	7, 9
경제 정책	5	과산화수소	1, 2, 3, 4
HTP	1, 2, 3, 4	액체 수소	6

검색은 전체 문서에서 입력한 쿼리의 조건에 부합되는 문서를 찾아내는 과정이다. 본 연구에서는 기술 설명을 환경기술 검색 시스템에 입력하면 전처리 과정을 통해 여러 개의 토큰으로 나뉜다. 그 다음 토큰이 역색인 목록에 있는지 확인하고 역색인 목록에 있으면 해당되는 문서를 국내외 기술정보 DB에서 찾아 검색 결과로 출력한다.

Ⅳ. 결론 및 향후 계획

본 연구에서는 단일 단어와 복수 단어를 포함한 유의어 사전 기반 환경기술 검색 시스템을 제안한다. 제안하는 시스템은 단일 단어와 복수 단어로 이루어진 유의어를 추출하는 시간을 단축하기 위한 목적으로 Word2vec 모델과 HDBSCAN 알고리즘을 혼합해 유의어를 추출하고 유의어 사전을 구축한다. 이어서 Multi-filter를 제공하는 Elasticsearch 검색엔진에 유의어 사전과 CTIS에서 제공하는 국내외 기술정보와 기술의 키워드를 적용한다. 사용자가 기술 설명을 입력하면 제안하는 시스템은 단일 단어와 복수 단어의 유의어를 포함한 결과를 출력한다. 본 연구에서 제안한 유의어 사전 기반 환경기술 검색 시스템을 CTIS에 적용하면 사용자들이 환경기술을 검색하는데 도움을 줄 수 있을 것으로 기대한다.

향후 연구로는 제안하는 시스템을 구축하고 여러 개의 토큰이 입력되었을 때 토큰의 우선 순위를 정할 수 있게 설계할 예정이다.

감사의 글

이 논문은 2019년도 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임

(2019-0-00136, 스마트시티 산업 생산성 혁신을 위한 AI융합 기술 개발)

참고문헌

[1] Climate Technology Information System: https://www.ctis.re.kr/ko/index.do

[2] Lee, T. W., & Seo, Y. H. (2003). A Synonym Dictionary Construction for Information Retrieval. In *Annual Conference on Human and Language Technology*(pp. 208-213). Human and Language Technology.

[3] Xia, C., He, T., Li, W., Qin, Z., & Zou, Z. (2019, July). Similarity Analysis of Law Documents Based on Word2vec. In *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*(pp. 354-357). IEEE.

[4] Jatnika, D., Bijaksana, M. A., & Suryani, A. A. (2019). Word2Vec Model Analysis for Semantic Similarities in English Words. *Procedia Computer Science*, *157*, 160-167.

[5] Ma, L., & Zhang, Y. (2015, October). Using Word2Vec to process big text data. In *2015 IEEE International Conference on Big Data (Big Data)*(pp. 2895-2897). IEEE.

[6] Zhang, L., Li, J., & Wang, C. (2017, July). Automatic synonym extraction using Word2Vec and spectral clustering. In *2017 36th Chinese Control Conference (CCC)*(pp. 5629-5632). IEEE.

[7] Tahvili, S., Hatvani, L., Felderer, M., Afzal, W., & Bohlin, M. (2019, April). Automated Functional Dependency Detection Between Test Cases Using Doc2Vec and Clustering. In *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*(pp. 19-26). IEEE.

[8] Cao, T. D., Tran, T. H., & Luu, T. T. (2018, December). Hot topic detection on newspaper. In *Proceedings of the Ninth International Symposium on Information and Communication Technology*(pp. 114-121).