

단문 텍스트의 자연어 처리 기법을 통한 크라우드 펀딩 추천 시스템 개발

이영아*‡, 이선명*‡, 이주연*‡, 이기용**

*숙명여자대학교 컴퓨터과학전공

**숙명여자대학교 소프트웨어학부

1814080@sookmyung.ac.kr, vividlee81@gmail.com, lly012900@gmail.com, kiyonglee@sookmyung.ac.kr

Development of a Recommendation System for Crowdfunding Using NLP in Short Text

Yeong-Ah Lee*‡, Sun-Myung Lee*‡, Ju-Yon Lee*‡, Ki Yong Lee**

*Dept. of Computer Science, Sookmyung Women's University

**Division of Computer Science, Sookmyung Women's University

요 약

최근 자연어 처리에 대한 관심이 증가함에 따라 자연어 처리 기술을 활용한 다양한 추천 시스템이 등장하고 있다. 본 논문에서는 자연어 처리를 이용한 서비스를 개발한다. 본 논문에서 개발한 서비스는 KoNLPy 와 Word2Vec 을 이용하여 크라우드 펀딩 프로젝트 창작자 및 후원자에게 키워드 및 키워드와 유사한 단어가 제목에 포함되는 프로젝트를 추천해준다. 단문 텍스트로서 프로젝트 제목을 사용하여 데이터를 자연어 처리 한 후, 딥러닝 모델에 적용시켜 추출한 데이터를 기반으로 창작자와 후원자에게 추천해주는 방식이다. 따라서 본 서비스는 프로젝트 제목 정보를 통한 추천 시스템의 개발로, 나아가 영화, 도서와 같은 콘텐츠 추천 분야에도 적용할 수 있을 것으로 기대한다.

1. 서론

최근 여러 크라우드 펀딩 사이트에 비슷한 펀딩 프로젝트들이 많다. 여기서 크라우드 펀딩은 대중을 뜻하는 크라우드(Crowd)와 자금 조달을 뜻하는 펀딩(Funding)을 조합한 용어로, 온라인 플랫폼을 이용해 다수의 대중으로부터 자금을 조달하는 방식을 말한다. 이와 같이 창작자가 프로젝트를 개설하면 해당 프로젝트의 상품을 구매하고 싶은 후원자들이 펀딩을 통해 돈을 후원한다. 창작자가 정한 목표 금액에 도달하면 후원자는 상품 구매가 확정되고, 목표 금액에 도달하지 못하면 프로젝트가 무산되어 후원자는 상품을 구매할 수 없게 된다.

한편, 국내 크라우드 펀딩 사이트들과 같이 프로젝트의 정보가 여러 사이트에 분산되어 있는 경우, 사용자들은 자신이 원하는 프로젝트를 찾기 어려워할 가능성이 크다.

따라서 본 논문은 후원자와 창작자 모두에게 도움이 되는 추천 서비스를 개발한다. 이 서비스를 ‘크라우드 펀딩 헬퍼(Crowdfunding helper)’로 칭한다. 크라우드 펀딩 헬퍼의 주 데이터인 프로젝트 제목들은 대개 20 자 내외의 하나의 문장이기 때문에 본 논문

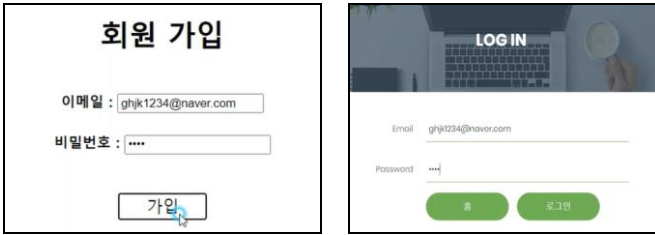
에서는 프로젝트 제목을 단문 텍스트로서 활용하였다. 크라우드 펀딩 사이트에서 프로젝트를 개설하는 유저는 ‘창작자’로 명시하며, 특정 프로젝트에 펀딩하는 유저는 ‘후원자’로 명시한다. 창작자에게는 어떤 펀딩 사이트에 프로젝트를 개설하는 것이 유리한지 추천해주고, 후원자에게는 후원자가 만족할만한 프로젝트를 추천한다.

본 논문의 구성은 다음과 같다. 2 장에서는 본 논문에서 개발한 기능들에 대해 구체적으로 설명한다. 3 장에서는 본 서비스 실제 수행 예를 제시한다. 4 장에서는 결론을 맺는다.

2. 세부기능

‘크라우드 펀딩 헬퍼’의 핵심 기능은 제목을 기반으로 한 유저에게 적합한 프로젝트 추천이다. 특히, 유저가 창작자인지, 후원자인지에 따라 추천 항목이 나뉜다.

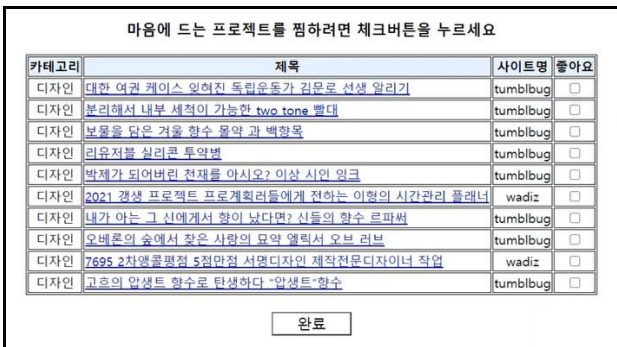
2.1 회원 관리



(그림 1) 회원가입 및 로그인 화면.

그림 1 화면에서 회원가입 및 로그인을 할 수 있으며, 아이디, 비밀번호, 역할 등을 포함한 유저 정보는 자체 데이터베이스에 저장된다.

2.2 관심 프로젝트 관련 기능



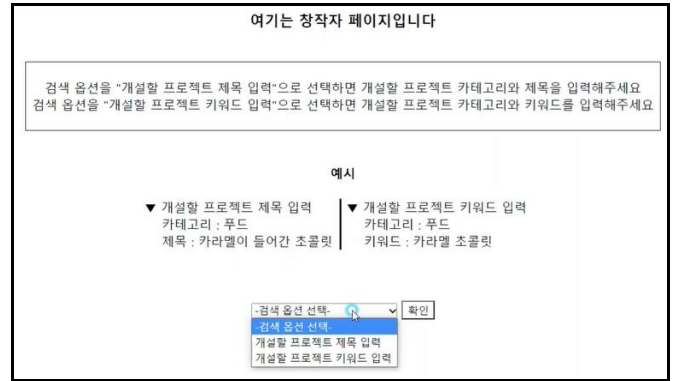
(그림 2) 관심 프로젝트를 찜할 수 있는 목록.



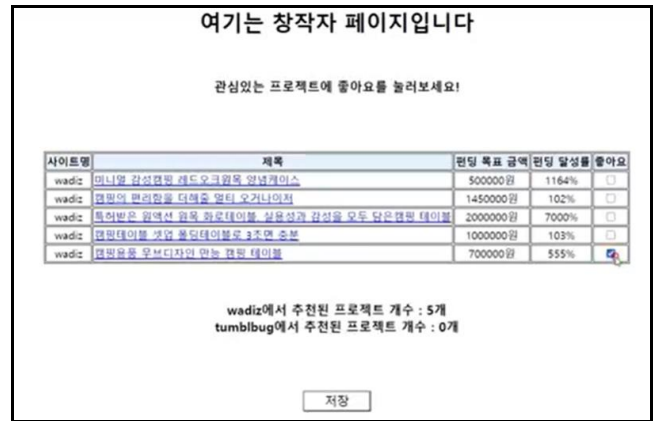
(그림 3) 관심 프로젝트 목록.

관심 프로젝트는 후원자에게는 제목과 설명을 보고 후원하고 싶은 프로젝트를 의미하고, 창작자에게는 프로젝트를 개설할 때 도움이 되는 정보가 있는 프로젝트를 의미한다. 관심 프로젝트에 등록된 것이 없는 경우 유저에게 관심 카테고리를 선택하게 하여, 그림 2 과 같이 해당 카테고리에서 인기있는 프로젝트 중 좋아하는 프로젝트를 선택하게 한다. 창작자 역할에서 등록된 관심 프로젝트 목록과 후원자 역할에서 등록된 관심 프로젝트 목록은 그림 3 과 같다. 또한 이를 각각 또는 통합해서 확인 가능하고, 역할에 따른 추천 결과 데이터를 관심 프로젝트 목록에 등록 및 업데이트가 가능하다.

2.3 프로젝트 추천 관련 기능



(그림 4) 제목 또는 키워드 입력 페이지.



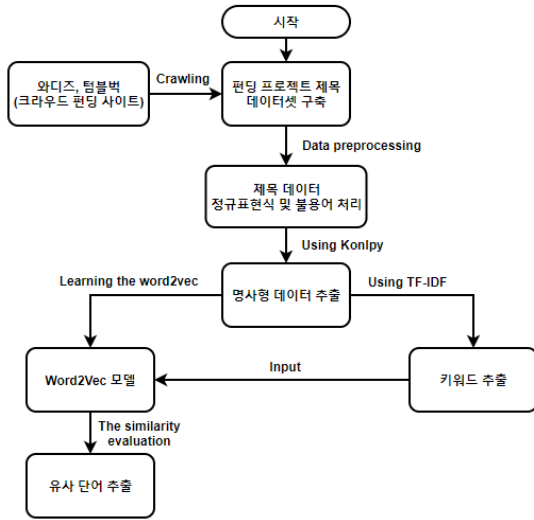
(그림 5) 창작자 역할 프로젝트 추천 결과.

프로젝트 추천 결과를 출력한다. 또한 창작자/후원자 역할에 따라 입력값이 다르다. 두 역할 모두 카테고리를 입력한 후, 그림 4 와 같이 창작자는 자신이 개설하고자하는 프로젝트의 제목이나 키워드를 추가적으로 입력한다. 후원자는 자신이 좋아하는 프로젝트를 관심 프로젝트에 등록한 것이 입력이 된다. 이러한 입력을 통해 그림 5 과 같이 추천 결과를 얻을 수 있다.

3. 추천 기능 구현

본 논문의 서비스는 파이썬을 기반으로 데이터 수집에는 Selenium 을 이용하였다. 데이터 전처리는 KoNLpy[1] 이용하였고, 키워드 추출에는 TF-IDF[2] 기법을 이용하였다. 또한 Gensim 의 Word2Vec 을 활용하여 키워드 유사 단어를 추출하였다. 추출된 키워드가 제목으로 포함된 프로젝트 중, 성공률이 일정 수준 이상이고, 동일 카테고리에 속하는 프로젝트를 찾아 후원자에게는 해당 프로젝트들을, 창작자에게는 해당 프로젝트들이 많이 포함된 펀딩

사이트를 추천한다. 3.1 절에서는 데이터 수집 및 전처리 과정을 설명하고, 3.2 절에서는 키워드 및 유사 단어를 추출하는 과정을 설명한다. 해당 내용의 플로우차트는 그림 6 과 같다.



(그림 6) 3.1 과 3.2 에 대한 플로우차트.

3.1 데이터 수집 및 전처리

펀딩 프로젝트들의 제목, 카테고리, 성공률, 해당 프로젝트에 펀딩한 유저 닉네임 데이터를 수집하기 위해 Selenium 을 이용하여 대중적인 클라우드 펀딩 사이트 2 곳을 크롤링하였다. 프로젝트의 제목들은 대개 20 자 내외의 하나의 문장이기 때문에 본 논문에서는 프로젝트 제목을 단문 텍스트로 취급하였다. 첫 번째로 프로젝트 제목을 전처리하기 위해 KoNLPy 의 Okt 라이브러리를 사용해 명사를 추출하였다. 두 번째로 숫자, 특수문자를 제거하고, 불용어 목록을 구성해 키워드 추출에 불필요한 단어들을 제거하였다. 마지막으로 외래어나 신조어 중 기존 단어들과 뜻이 동일하지만 표기를 다르게 한 단어들을 통일시켰다.

3.2 키워드 및 유사단어 추출

위 전처리 단계에서 제목 토큰화를 마친 후 크롤링을 통한 데이터셋에서 TF-IDF 를 적용하여 해당 프로젝트에 관련된 중요 키워드를 추출하였다. 아래 공식은 TF-IDF 에 대한 것이며 j 는 문서를 뜻하고 i 는 해당 문서에 포함된 용어를 의미한다.

$$tfidf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

추출한 중요 키워드의 유사 단어를 찾아내기 위해 Gensim 의 Word2Vec 을 이용하였다. Word2Vec 에는 Skip-gram[3]와 CBOW 2 가지 모델이 존재하는데, 본 논문에서는 같은 양의 말뭉치로도 업데이트 횟수와

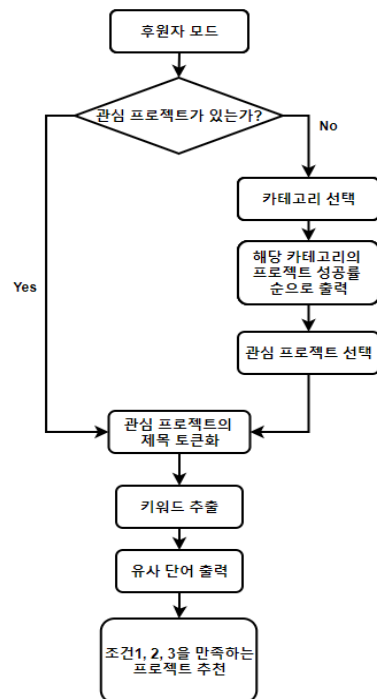
학습량이 많은 Skip-gram 모델을 사용하였다.

3.3 추천시스템

3.2 에서 찾아낸 키워드 또는 키워드의 유사 단어가 제목으로 포함된 프로젝트 중 성공률이 임계값 이상인 것을 찾는다. 그 중 토큰화된 프로젝트의 제목을 가진 프로젝트의 카테고리과 동일한 카테고리를 가진 프로젝트를 추출한다. 3.3 단계는 유저의 역할에 따라 달리 처리된다.

• 후원자

후원자의 경우, 등록된 관심 프로젝트 유무에 따라 처리 방식이 다르며, 구체적 알고리즘은 그림 7 과 같다. 등록된 관심 프로젝트가 있는 경우, 관심 프로젝트에 등록된 프로젝트에 대해서 3.1~3.3 의 단계를 거쳐 추출된 프로젝트를 추천한다. 관심 프로젝트가 없는 경우는 관심 카테고리를 설정한 후, 해당 카테고리에서 인기있는 프로젝트를 보여주어 관심 있는 프로젝트를 선택하게 한다. 여기서 인기있는 프로젝트는 펀딩 달성률(펀딩금액/목표금액(%))이 일정 수치 이상인 프로젝트를 말한다. 이후에는 마찬가지로 3.1~3.3 단계를 거쳐 후원자가 만족할만한 프로젝트들을 추천한다. 이를 통해, 후원자는 이전에 관심있는 프로젝트로 등록한 프로젝트들과 유사한 프로젝트들을 추천받게 된다.

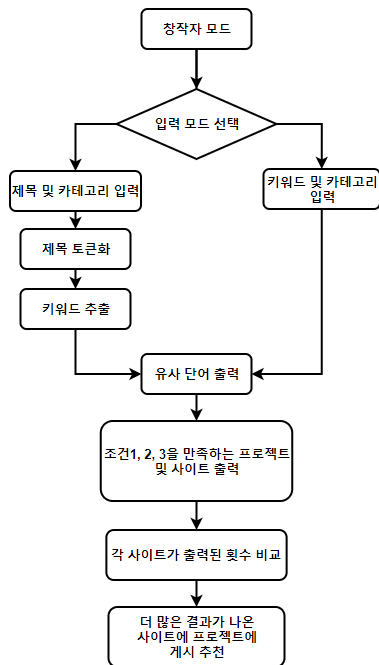


조건1. 키워드 또는 유사 단어가 프로젝트 제목에 포함됨
조건2. 프로젝트 성공률이 임계값 이상임
조건3. 이전에 선택한 카테고리에 속하는 프로젝트임

(그림 7) 후원자 모드에서의 추천 기능 구현 알고리즘.

• **창작자**

구체적 알고리즘은 그림 8 과 같다. 창작자의 경우, 개설할 프로젝트의 제목 또는 키워드를 입력한다. 제목을 입력한 경우에는 입력한 제목에 대해 입력한 키워드에 대해서 3.2~3.3 단계를 거친다. 이때 입력된 키워드는 3.3 단계의 유사 단어 집합에 포함되어 처리된다. 이 과정을 통해 창작자가 입력한 프로젝트 정보와 관련있으며 인기있는 프로젝트와, 해당 프로젝트가 개설된 사이트명을 출력한다. 여기서 출력된 프로젝트들이 많이 개설된 사이트를 창작자에게 추천한다. 이를 통해 창작자는 자신이 개설하고자 하는 프로젝트와 유사한 프로젝트들이 어떤 사이트에 많은지 알 수 있게 되어, 프로젝트를 개설할 사이트 선정에 도움을 받을 수 있다. 또한 키워드 입력을 한 경우 유사 프로젝트들이 어떤 제목을 가지고 있는지 알게 되어 제목 선정에 도움을 받을 수 있다.



조건1. 키워드 또는 유사 단어가 프로젝트 제목에 포함됨
 조건2. 프로젝트 성공률이 임계값 이상임
 조건3. 이전에 선택한 카테고리에 속하는 프로젝트임

(그림 8) 창작자 모드에서의 추천 기능 구현 알고리즘.

4. **추천 기능 구현**

본 시스템을 통한 프로젝트 추천 결과는 표 1 과 같다. 프로젝트 제목으로 ‘오래 사용할 수 있는 강철 루핀 화로대’ 를 입력하고 카테고리 ‘여행·레저’ 를 입력하면 추천 시스템에서 ‘사용’, ‘화로대’ 라는 키워드를 추출한다. 해당 키워드를 통해 ‘필수품’, ‘화로’ 라는 유사단어를 생성할 수 있다. 이렇게 생성된 유사단어 또는 키워드가 포함된 프로젝트를 추천한다. 화로대 관련 프로젝트로부터 화로, 화로 테이블, 화로 봉화로 등 화로와 관련된 프로젝트들이 추천된 것으로 보아, 본 논문에서 제시한 추천 알고리즘의 성능이 준수한 것을 알 수 있다.

5. **결론**

본 논문에서는 자연어 처리와 딥러닝을 이용한 클라우드 펀딩 추천 서비스를 개발하였다. 본 논문에서 개발한 서비스는 제목 데이터를 이용하여 프로젝트 또는 플랫폼을 추천해준다. 또한 본 논문은 프로젝트 제목 정보를 통한 추천 시스템의 개발에 대한 것으로 영화제목과 도서명 같은 짧은 문장의 콘텐츠 추천에도 활용할 수 있을 것으로 기대된다. 추후 연구로는 TF-IDF 대신에 짧은 글에 적합한 키워드 추출 기법을 이용하여 서비스의 성능을 향상시키는 것이다.

6. **참고문헌**

[1] 박은정, 조성준, “KoNLPy : 쉽고 간결한 한국어 정보처리 파이썬 패키지”, 제 26 회 한글 및 한국어 정보처리 학술대회 논문집, 2014.
 [2] PIMPALKAR, A. P.; RETNA RAJ, R. J. Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, v. 9, n. 2, p. 49-68, 18 Jun. 2020.
 [3] Z. . Xiong, Q. . Shen, Y. . Xiong, Y. . Wang and W. . Li, "New generation model of word vector representation based on cbow or skip-gram," *Computers, Materials & Continua*, vol. 60, no.1, pp. 259–273, 2019.

<표 1> 추천 결과

	프로젝트 제목 예시	카테고리	키워드	유사 단어	추천되는 프로젝트 상품 종류
1	오래 사용할 수 있는 강철 루핀 화로대	여행·레저	사용, 화로대	필수품, 화로	- 캠핑 화로 - 원목 캠핑 화로 테이블 - 화로 봉화로
2	스케치업 웹툰 배경 : 중세 다양한 궁전 정원	만화	스케치업, 웹툰, 배경, 중세, 정원	서양식, 판타지	- 서양 웹툰 배경 모음 3종류 - 중세 웹툰 배경 모음
3	잠에만 집중하는 잠의 신, 숙면 안대	홈리빙	숙면, 안대	수면	- 수면 안대 2종류 - 찜질 안대
4	나만의 반지	디자인	반지	루비, 다이아, 실버	- 다이아 반지 2종류 - 실버 반지 3종류