Word2Vec 기반 장르 유사성을 활용한 웹툰 검색

이창민○, 안제정, 강동연, 이현아
금오공과대학교 컴퓨터소프트웨어공학과
eqqt97@naver.com, kwoyoung@naver.com, ehddus158@naver.com, halee@kumoh.ac.kr

Webtoon Search utilizing Genre Similarity with Word2Vec

ChangMin Lee, JeJeong Ahn, DongYeon Kang, Hyunah Lee
Kumoh National Institute of Technology, Dept. of Computer Software Engineering

요 약

본 논문에서는 기존 웹툰 장르 검색 시스템의 단점을 보완하기 위해 키워드 기반 유사 장르 검색 시스템을 제안한다. 기존 웹툰의 장르와 키워드를 분석하여 44개의 장르를 설정하고 해당 장르에 적합한 웹툰을 수집한다. 나무위키와 위키피디아 문서로 학습된 Word2Vec모델에 기반하여 계산한 사용자 입력 키워드와 44개의 장르간 유사도로 사용자 입력에 가장 유사한 장르를 찾는다. 유사 장르에 포함되는 웹툰을 결과로 출력하여 사용자가 선호하는 장르의 웹툰을 제시한다. 실험 결과에서는 나무위키에서 '장르'로 검색하여 얻는 작은 크기의 문서 집합에서 Word2Vec을 학습한 모델에서 가장 높은 검색 성능을 보였다.

주제어: Word2Vec, 기계학습, 장르 유사성, 웹툰

1. 서론

다음과 네이버를 포함한 여러 포털사이트를 통해 다양한 웹툰을 편리하게 접할 수 있게 되면서 웹툰을 생산하는 작가와 독자의 수가 크게 증가하고 있다. 각 독자들의 취향에 따른 다양한 장르가 있지만, 근래 짧은 시간에 대량의 웹툰이 생산되면서 독자는 자신이 선호하는 장르의 웹툰을 찾기가 어려워졌다. 국내 웹툰 사이트 중 가장 대표적인 네이버 웹툰[1]은 13개의 장르별 조회를 제공하지만 검색에서는 웹툰 제목과 작가로만 검색할 수 있다. 예를 들어 장르 분류에는 학교나 학원 관련 장르가 존재하지 않고, 이에 관련된 웹툰을 찾기 위해 '학교'를 검색하면 제목이나 작가명에 반드시 '학교'가 들어간 웹툰만을 제시한다. 다음 웹툰[2]은 다양한 장르와 검색을 지원하지만 사용 편의성의 차이로 네이버보다 사용자 수가 적다. 다음 웹툰은 16개의 장르와 함께 3개의 키워드로 웹툰 특징을 표현할 수 있어 다양성을 보장하지만, '학원물'을 검색하면 제목이나 키워드로 '학원물'을 포함하는 웹툰이 단 하나만 존재하여 사용자가 원하는 충분한 결과를 제시하지 못한다.

본 논문에서는 기존 웹툰 검색의 단점을 보완하기 위해 단어 간 유사도를 사용한 키워드 기반 검색 방법을 제안한다. 이를 위해 네이버와 다음 웹툰, 나무위키의 장르 정보를 분석하여 의미 있는 44개의 장르를 설정하고, 설정된 장르에 해당하는 웹툰 정보를 웹툰과 네이버 블로그를 활용하여 수집하여 저장한다. 사용자가 장르 검색을 위해 '학교'을 입력하면 시스템에서는 '학교'에 유사한 장르를 찾는다. 유사 장르 결정에서는 나무위키와 위키피디아 문서로 학습한 Word2Vec 모델을 사용하며, '학교'에 대해서는 '학원물', '일진' 등이 유사한 장르로 추천된다. 최종적으로 시스템에서는 유사 장르에 해당하는 웹툰을 사용자에게 제시하여 다양한 검색어로도 적합한 장르의 웹툰을 추천한다.

2. 장르 유사성을 활용한 웹툰 검색

본 프로그램에서 제안하는 유사 장르 검색 프로그램의 전체 구조는 그림 1과 같다. 네이버 웹툰과 다음 웹툰에서 크롤링을 하고 네이버 블로그의 정보를 추가로 활용하여 웹툰별 장르를 분류하여 Webtoon DB에 저장한다. 위키피디아 데이터와 나무위키 장르 관련 데이터를 웹 크롤링한 데이터를 전처리 과정을 거쳐 단어를 벡터로 임베딩하는 Word2vec모델을 학습한다. 사용자가 입력한 키워드와 44가지 장르간의 유사도를 구하여 가장 유사한 장르 3개를 찾고 웹툰 데이터베이스에서 해당 장르 3개의 속한 웹툰을 보여준다. 아래에서는 각 단계에 대해서 상세하게 설명한다.

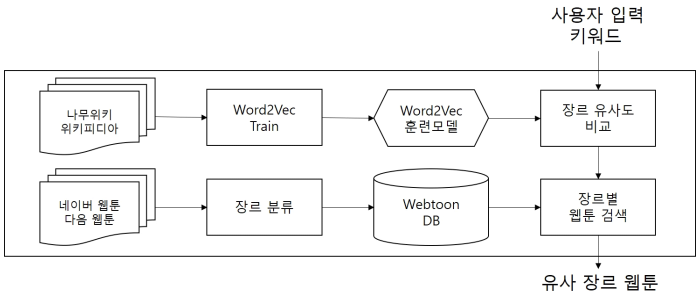

그림 1. 유사 장르 검색 프로그램의 시스템 구조도

2.1 웹툰 장르 설정 및 웹툰 정보 수집

네이버 웹툰은 에피소드, 옴니버스, 스토리, 일상, 개그, 판타지, 액션, 드라마, 순정, 감성, 스릴러, 시대극, 스포츠의 13개의 한정적인 장르만을 제공한다. 이에 반해 다음은 에피소드, 스토리, 공포, 드라마, 무협, 미스터리, 순정, 스릴러, 스포츠, 액션, 일상, 지식, 코믹, 판타지, 학원, 성인의 16개 장르와 함께 최대 3개의 키워드를 설정할 수 있어 다양한 장르를 지원하지만 사용편의성이 떨어지며 검색에서의 정확한 결과를 내기 어

렵다. 이러한 문제를 해결하기 위해 본 연구에서는 네이버 웹툰과 다음 웹툰의 장르와 키워드, 나무위키의 장르 항목[3]의 만화의 장르 정보를 검토하여 웹툰 사용자들이 선호할 수 있는 아래의 44개 장르를 설정한다.

감성, 공포, 괴담, 군대, 권력, 귀신, 귀족, 도깨비, 돈, 동물, 드라마, 로맨틱, 모험, 무협, 바이러스, 병맛, 복수, 생존, 성인, 스릴러, 스토리, 스포츠, 시대극, 신화, 악마, 액션, 에피소드, 옴니버스, 욕망, 음식, 일상, 일진, 재벌, 정치, 좀비, 천사, 초능력, 추리물, 코믹, 토커, 퇴마, 판타지, 학원물, 히어로

네이버와 다음 웹툰의 장르에서 확장된 44개의 장르 분류에 적합한 웹툰을 파악하기 위해 네이버 블로그에서 각 장르별 추천 웹툰을 검색한다. 예를 들어 '액션 웹툰 추천','동물 웹툰 추천'과 같이 '장르명'+'웹툰 추천'으로 네이버 블로그를 검색하고, 검색 상위 블로그를 각 100개씩 크롤링해서 그림 2와 같이 장르별 텍스트 파일로 저장한다. 저장된 문서에서 웹툰 제목의 빈도수를 각 텍스트 별로 파악한다. 웹툰은 네이버와 다음 웹툰에 등록된 웹툰으로 한정한다. 웹툰의 제목이 일정 이상의 빈도수를 나타내면 텍스트 파일의 장르를 해당 웹툰의 장르로 데이터베이스에 저장한다. 얻어진 Webtoon DB에서는 하나의 웹툰이 여러 장르에 소속될 수 있어 한 웹툰이 가지는 다양한 특성을 반영한다.

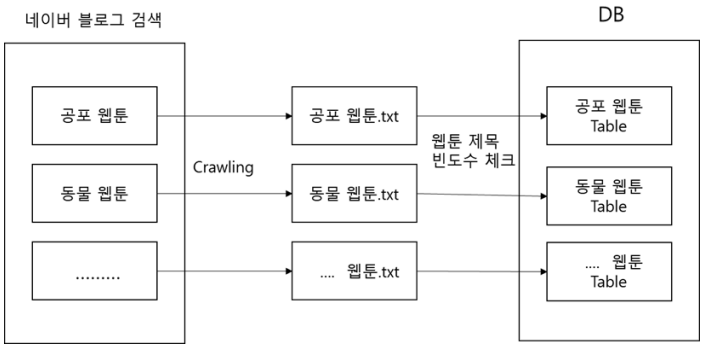

그림 2. 장르 분류를 위한 웹 크롤러 설계

문서처리 과정에서는 데이터 전처리 과정을 통해 특수문자, 중국어, 일본어 같은 불필요한 문자들을 제거한다. 데이터 전처리 과정은 noise, error 등을 포함해 발생하는 문제를 해결하기 위한 필수적인 과정이다[4]. 한글 형태소 분석 및 품사 태깅을 지원하는 KoNLPy의 Okt[5]클래스를 사용하여 형태소 분석을 실시한다. 형태소 분석을 통해 동일한 단어라도 품사를 결합하면 의미를 구분 할 수 있기 때문에 Word2Vec모델의 성능 향상에 기여한다.

2.2 Word2vec 모델 학습

단어를 벡터로 표현하는 방법을 워드임베딩(word embedding)이라고 부른다. 워드임베딩을 위해서는 벡터 공간 모델(Vector Space Models, VSMs)[6]이 주로 사용된다. 벡터공간 모델은 분산 가설(Distributional Hypothesis)[7]에 기반을 둔다. 벡터로 표현된 단어들은 문법적인 부분만 아니라 의미적인 부분까지 반영된다. 워드임베딩을 통해 생성된 벡터는 거리가 가까울수록 서로 비슷한 벡터를 갖는다. 수치화된 벡터들은 벡터 간의 거리를 활용한 벡터 연산이 가능하다.

본 연구에서는 워드 임베딩을 하는 여러 방법 중 Word2Vec을 사용한다. Word2Vec 모델은 Continuous Bag-of-Word(CBOW) 모델과 Skip-Gram 모델[8]로 나뉜다. 통계적으로 CBOW 모델은 많은 수의 분포 상 정보를 바로잡는 효과를 가지며, 대부분 작은 데이터 셋일수록 유용한 것으로 알려져 있다. Skip-Gram 모델은 큰 규모의 데이터 셋을 가질 때 더 잘 동작하는 경향이 있는 것으로 알려져 있다. Skip-Gram 모델을 사용하여 워드임베딩을 진행한다.

본 논문에서 사용하는 Word2Vec 모델을 위한 데이터 셋은 3가지로 나뉜다. 첫 번째 데이터 셋은 유사 장르 검색 시 정확도를 향상시키기 위한 나무위키 장르 관련 텍스트 데이터, 두 번째 데이터 셋은 위키피디아 한국어 전체 데이터, 세 번째 데이터 셋은 나무위키 장르 관련 텍스트와 위키피디아 한국어 데이터를 합친 데이터이다. 위 모델들을 벡터 크기 300으로 하여 학습을 진행한다.

2.3 장르 기반 웹툰 검색

Word2Vec에 기반한 유사 장르 기반 웹툰 검색을 위해 44가지의 장르 벡터와 사용자 입력 키워드 벡터간의 유사도를 비교한다. 유사도 상위 3개의 장르를 선택하고 웹툰 DB에서 해당 장르에 속한 웹툰을 사용자에게 제시한다. 그림 3은 구축된 시스템의 예시화면을 보인다. 시스템에서는 장르에 해당되는 웹툰들을 네이버와 다음의 추천수 내림차순으로 제시한다. 그림의 '학교'에 대한 검색 결과에서 제목에는 '학교'가 포함되지 않지만 학교나 학원에 관련된 적합한 웹툰을 제시함을 확인할 수 있다.

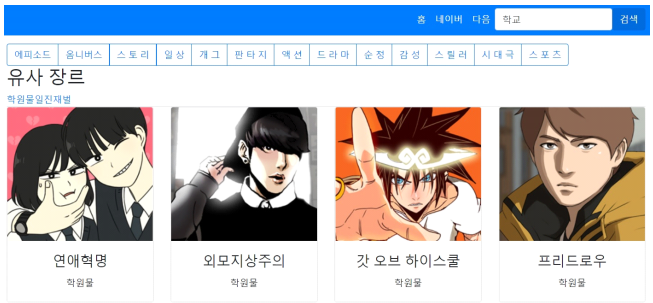

그림 3. 장르 유사성을 활용한 웹툰 검색시스템 UI의 예

3. 실험 및 평가

3.1 실험 데이터

본 연구에서는 Word2Vec 학습을 위해 위키피디아와 나무위키 데이터를 사용한다. 전체 한국어 위키피디아의 638MB 58,685,345어절, 나무위키 장르 관련 글들을 크롤링하여 얻은 3.77MB 383,039어절, 위키피디아와 나무위키 장르 관련 글들을 합친 집합의 총 세 종류의 데이터 셋을 Word2Vec 모델을 사용하여 워드 임베딩을 진행한다. 평가에서는 학교, 괴물, 연애, 좀비, 범죄의 5개의 키워드에 대한 검색 결과를 사용했으며, 웹툰을 자주 보는 대학생 5명의 평가자 중 3명 이상이 정확하다고 평가하면 올바른 것으로 판단하고 정확도를 측정하였다.

3.2 실험 결과 및 평가

우선 Word2Vec 학습에서 사용한 세 종류의 문서에 따른 성능을 비교하였다. 표 1은 각 데이터 모델로부터 얻은 5개의 검색어에 대해 얻어진 3개 유사 장르를 보인다. 평가에서는 가장 작은 크기의 나무 위키에 의한 결과가 가장 적합한 것으로 나타났다. 예를 들어 위키피디아에서는 '학교'에 대해서 '군대'가, '연애'에 대해서 '액션'이 유사한 장르로 추천되는 결과가 나타났다. 나무위키 데이터 모델은 장르와 관련된 글이지만 위키피디아 데이터 모델은 장르와 관련된 글 이외의 글도 포함하며, 위키피디아에 기반한 Word2Vec은 noise가 나무위키 모델보다 증가될 뿐만 아니라 보편적인 문맥에서의 워드 임베딩으로 유사 장르 검색 결과가 적합하지 않은 것으로 분석되었다.

표 1. 각 데이터 모델의 입력 장르와 유사 장르 3개

	나무위키	위키피디아	나무위키+위키피디아
학교	학원물, 일진, 재벌	스포츠, 일진, 군대	군대, 스포츠, 정치
괴물	퇴마, 귀신, 스릴러	귀신, 악마, 좀비	악마, 귀신, 좀비
연애	로맨틱, 스포츠, 학원물	학원물, 로맨틱, 액션	학원물, 로맨틱, 일상
좀비	좀비, 바이러스, 스릴러	좀비, 악마, 귀신	좀비, 귀신, 악마
범죄	추리물, 스릴러, 정치	스릴러, 액션, 로맨틱	스릴러, 액션, 드라마

표 2에서는 유사 장르 검색의 정확률을 보인다. 결과에서 나무위키 모델은 67.1%, 위키피디아 모델 37.5%, 나무위키+위키피디아 모델 41.4%를 보여, 표 1에서 분석한 바와 같이 나무위키를 사용한 유사 장르 추출이 가장 우수한 결과를 나타냈다. 오류 분석에서는 3개의 유사 장르 중 일부가 부적절한 웹툰을 제시한 것으로 파악되었다. 예를 들어 나무위키 모델에서 '좀비'키워드의 유사 장르는 좀비, 바이러스, 스릴러가 나오는데, 좀비, 바이러스로 분류된 웹툰은 대부분이 좀비 웹툰을 제시한다. 스릴러로 분류된 웹툰은 좀비 웹툰이 아닌 웹툰을 제시하는 경우가 많았다.

표 2. 각 데이터 모델에 대한 키워드 입력에 대한 웹툰 출력 결과

	학교	괴물	연예	좀비	범죄	평균
나무위키	62/72 86.1%	26/54 48.1%	135/164 82.3%	21/73 28.8%	48/72 66.7%	292/435 67.1%
위키피디아	21/29 72.4%	9/11 81.8%	101/236 42.8%	6/11 54.5%	60/237 25.3%	197/524 37.5%
나무+위키피디아	7/16 43.8%	9/11 81.8%	123/215 57.2%	6/11 54.5%	82/294 27.9%	227/547 41.4%

표 3은 표 2에서 나무위키 모델에서 검색 결과가 좋지 않은 '괴물'과 '좀비'를 키워드로 네이버 웹툰과 다음 웹툰에서 검색했을 때의 결과를 보인다. 가장 성능이 떨어지는 키워드에 대해서도 제안 방법이 네이버 웹툰의 검색에 비해서는 모든 경우에서 우수한 결과를 보였지만, 다음 웹툰의 결과에 비해서는 낮은 성능을 보였다. 하지만 네이버와 다음 결과를 합집한 경우에 비해 제안 방법의 결과가 더 많은 수의 웹툰을 추천하는 동시에 적합한 웹툰의 수도 더 많아, 다양한 웹툰을 찾기 위하는 사용자에게는 보다 적합한 서비스를 제공할 수 있을 것으로 기대된다.

표 3. 네이버 웹툰과 다음 웹툰, 제안 방법의 검색 결과 비교

	네이버	다음	네이버+다음	제안 방법
괴물	4/25 16.0%	10/18 55.5%	14/43 32.5%	26/54 48.1%
좀비	4/17 23.5%	12/15 80.0%	16/32 50.0%	21/73 28.8%

4. 결론

본 연구에서는 단어간의 유사도를 구하여 키워드기반 웹툰 유사 장르 검색 방식을 제안했다. 사용자가 원하는 장르의 적중률은 높지 않지만 기존 웹툰 검색 시스템보다 효과적인 성능을 보였다.

향후에는 웹툰 분류에 보다 적합한 데이터 셋을를 선별하고 장르들을 레이블링하여 머신러닝을 통해 장르를 추천하는 연구를 진행할 예정이다. 또한 영화, 애니메이션과 같은 다양한 분야에 적용할 수 있는 장르 유사성 기반 검색으로 확장할 예정이다.

참고문헌

[1] 네이버 웹툰, https://comic.naver.com/webtoon/weekday.nhn
[2] 다음 웹툰, http://webtoon.daum.net/
[3] '장르'의 나무 위키 https://namu.wiki/w/장르
[4] 김영수, 이승우, "문서 분류를 위한 신경망 모델에 적합한 텍스트 전처리와 워드 임베딩의 조합", 정보과학회논문지, 2018.
[5] Okt, github.com/open-korean-text
[6] G. Salton, A. Wong, C. S. Yang, "A vector space model for automatic indexing", 1975.
[7] Zellig S. Harris, "Distributional Structure", 1954.
[8] Tomas Mikolov, Kai Chen, Greg Corrando, Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space", 2013.