Doc2vec을 활용한 기억 속의 영화 검색

장진규[1] 박정인[2] 정윤혁[3]
울산과학기술원 융합경영대학원[12], 경영공학과[3]
jingyu12@unist.ac.kr[1] selfcounter@unist.ac.kr[2] yjung@unist.ac.kr[3]

초록

본 연구는 소비자가 묘사하는 영화의 특징으로부터 찾고자 하는 영화를 검색 해주는 모델을 제시한다. 해당 모델은 문서 기반 임베딩 기법인 Doc2vec를 활용하여 구현 하였다. 본 연구에서는 소비자가 묘사한 영화의 특징과 유사한 의미 정보를 가지고 있는 영화들 간의 유사도 비교를 통해 랭킹을 매기고, 랭킹 안에 찾고자 한 영화의 등장 여부에 따라 모델의 정확도를 평가하였다.

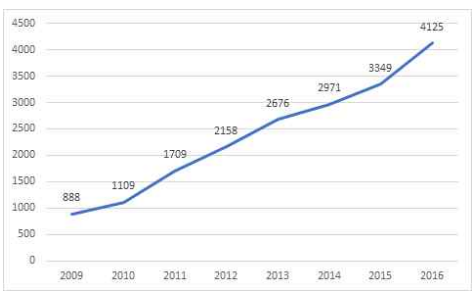
키워드: Doc2vec, 영화 검색, 정보 검색

서론

주문형 비디오(VOD) 서비스는 새롭게 부상하고 있는 영화 소비 채널이다. 국내 VOD 시장은 2016년에 4,125억의 매출을 달성하며 2009년 880억 이래로 5배 성장했고, 매년 평균 25%씩 성장하고 있다.[1] 이런 성장 배경에는 1인 가구의 증가와 경기침체 지속으로 인한 소비 패턴 변화가 있으며, 이러한 성장세는 앞으로도 계속 될 것으로 여겨진다.

전체 디지털 온라인 시장 매출 규모

(단위 : 억원)



출처: 영화진흥연구원

최근 대부분의 VOD 서비스 업체들은 다양한 형식으로 개인화 서비스를 제공하고 있다. 이 서비스는 한 고객의 신상정보, 영화 시청 정보, 그리고 영화 자체에 대한 정보를 가공해 그 고객이 좋아할 것이라고 여겨지는 영화를 추천하는 서비스다. 하지만 아쉽게도 그들은 가장 추가 구매 가능성이 높은 영화를 추천해 주지는 못하고 있다. 그 영화는 소비자의 기억 속에 남아있는 영화이다.

'그 영화 보고 싶은데 제목이 뭐지?' 누구나 한 번씩은 이런 질문을 스스로에게 던진다. 그리고 대부분의 경우에 스스로 답을 찾기란 힘들다. 유년 시절에 봤거나, TV를 시청하거나 SNS를 즐기는 도중에 잠깐 접했던 영화이기에 단서가 충분하지 않다. VOD 서비스 업체들은 지금 당장 보고 싶은 '그 영화'를 찾아주지는 못한다.

이러한 상황에서 일부 사람들은 인터넷에 질문을 올려 영화를 찾고 있다. 그리고 소수의 전문가들은 질문 속에 들어있는 단편적인 단서들을 통해 영화를 찾아주고 있다. 하지만 이러한 교류는 실시간으로 이루어지지 않으며, 전문가들이 반드시 자신의 질문에 답해 주리라는 보장도 없다. 그리하여 본 연구는 인터넷 상에 존재하는 영화 질문-답변 데이터를 통해 기억 속의 영화를 찾아주고자 한다.

관련 연구

1. TF-IDF

Term Frequency Inverse Document Frequency(TF-IDF)는 정보 검색(Information retrieval)이론에서 벡터공간모형을 대표하는 모델이다. Term Frequency(TF)는 특정 단어가 한 문서에서 언급되는 횟수이며, Inverse Document

Frequency(IDF)는 특정 단어가 전체 단어 집합(corpus)에서 언급되는 횟수의 역수이다. TF-IDF는 TF와 IDF의 곱으로, 이 척도가 높은 단어가 그 문서를 대표하게 되며, 그 단어가 질의(query)에 포함될 경우 그 문서를 검색 결과로 반환해준다.[2]

$$FIDF\ t, d) = TF(t, d) \times IDF(t)$$

하지만 TF-IDF는 동일한 의미를 가진 단어들을 다른 단어로 인식하는 단점이 있다. 그렇기에 TF-IDF는 고유 단어들이 일반적으로 쓰이는 전문 분야의 문서나, 정제된 언어를 쓰는 언론 분야의 문서를 분석할 때는 유용하게 쓸 수 있지만, 다양한 어투와 어구가 존재하는 소셜 데이터를 분석하기에는 적합하지 않다.

2. Word2vec

Word2vec는 단어를 벡터화 하는 워드 임베딩(word embedding)의 일종으로 문맥(context)를 통해 단어의 의미를 고려할 수 있는 특성 벡터를 생성한다. 그러므로 다른 단어라도 유사한 문맥에서 자주 사용되면 가까운 벡터 위치를 부여받는다. 벡터간의 거리는 코사인 유사도를 통해 쉽게 구할 수 있다.[3]

Word2vec를 이용해 문서를 검색하기 위해서는 일련의 작업을 더해야 한다. [4]에서는 각 문서마다 단어 벡터들의 중심점(centroid)를 구해 그 문서를 특징하는 지표로 삼았다. 그리하여 질의문서의 중심점과의 거리가 짧은 순서대로 유사한 문서로 판단 한다. [4]

3. Doc2vec

Doc2vec은 Word2vec의 확장판이며, Word2vec과 마찬가지의 방식으로 단어들에게 고유의 벡터를 부여한다. 차이점은 Doc2vec는 각 문서마다 고유 아이디의 역할을 하는 새로운 특징 벡터(feature vector)가 들어가 있다는 점이다. 따라서 단어 벡터를 학습시킬 때 문서 벡터 또한 같이 학습된다. [5]

연구 방법

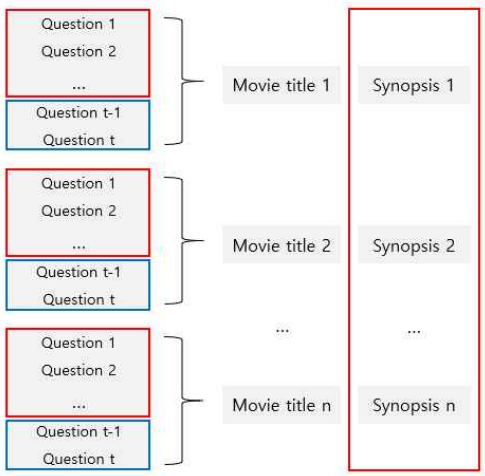
데이터

본 연구에서는 네이버 지식in 서비스의 영화 검색과 관련된 질문·답변을 수집하였다. 그리고 답변된 영화의 시놉시스를 네이버 영화 서비스에서 수집하였다. 수집은 selenium 라이브러리를 활용하였으며, 총 116,377개의 질문과 답변을 수집하였다.

전처리 작업으로 수집된 데이터 중 50글자 미만의 질문, 2건 미만의 질문은 정보가 부족하다고 판단하여 제거하고, 이상치, 중복, 불용어를 제거하였다. 최종적으로 총 38,614개의 질문과 답변, 영화 3866건을 이용하였다.

분석을 위한 POS tagging은 konlpy 라이브러리의 Twitter 모듈을 이용했다. Pos tagging과 함께 스테밍(stemming)과 정규화(normalization)를 하였다. 추출한 형태소 중 명사, 형용사, 동사만 분석에 이용하였다. 의미 파악이 어려운 한 글자 단위의 형태소는 분석에서 제외하였다.

그림 2. word2vec를 활용한 학습 개요



Doc2vec

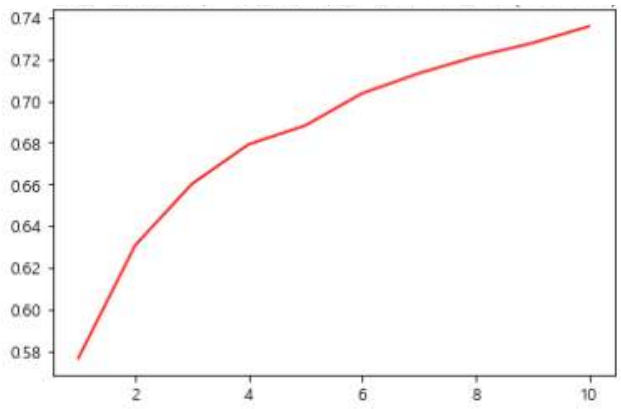
Doc2vec 알고리즘을 학습시키기 위해 질문을 Train set과 Test set 8대 2로 나누고 전체 시놉시스와 Train set을 학습했다. 학습을 위한 데이터는 300차원으로 임베딩 하였다.

Doc2vec 알고리즘은 gensim 라이브러리의 doc2vec 모듈을 이용하였다.

결과

분석 결과, 검색해주는 영화의 개수가 증가할수록 실제로 질문자가 원하는 영화를 찾아줄 확률이 높아졌다. 구체적인 정확도는 영화가 1개 추천될 경우 57.3%, 5개 추천될 경우 70%, 그리고 10개 추천될 경우 73.5%를 보였다.

그림 3. 검색 개수에 따른 정확도 변화



토론

본 연구는 웹 데이터를 활용하여 소비자가 찾고 싶어 하는 영화를 찾는 모델을 만들었다. 시험 세트로 분리된 질문들을 모델에 적용한 결과, 일정 범위 내에서 소비자가 원하는 영화를 찾을 수 있었다.

하지만 영화의 시놉시스를 구할 수 없거나, 질문 자체가 지나치게 짧은 경우 등 모델이 학습할 데이터가 충분하지 않은 경우 영화를 찾기 어려운 한계점이 있다. 그리고 종종 장면을 캡처하여 영화를 찾고자 하는 데이터가 존재하였지만, 이미지 처리 능력의 한계로 분석에서 제외되었다.

그럼에도 본 연구가 VOD 서비스 업체가 개인화 서비스의를 제공하는 새로운 관점을 제공했다는 데에 의의가 있다. 현재 제공되는 개인화 서비스는 '소비자가 인지하지 못한 니즈를 권유'하는 형태의 서비스인데 반해, 본 연구에서 제시된 모델은 '소비자가 요구하는 니즈'를 찾아주는 서비스이기 때문이다. 더불어 소비자가 제시한 질의문을 통해 소비자가 영화의 어떤 부분을 유심히 기억하는가에 관한 정보를 얻을 수 있어, 기존에 니즈를 권유하는 서비스를 더욱 강화할 수 있다.

또한 향후 본 연구에서 사용한 소비자들의 질의 데이터를 분석하면, 영화의 어떤 부분이 소비자의 뇌리에 남는지 분석하여 활용할 수 있을 것으로 기대된다.

참고문헌

[1] 영화진흥연구원 산업정책연구팀. (2016). "한국 영화산업 결산", 영화진흥위원회

[2] Juan Ramos. (2003). "Using TF-IDF to determine word relevance in document queries".

[3] 김우주, 김동희, 장희원. (2016). "Word2vec을 활용한 문서의 의미 확장 검색 방법". *한국콘텐츠학회논문지*, 16(10), 687-692

[4] Georgios-Ioannis Brokos, Prodromos Malakasiotis, Ion Androutsopoulos. (2016). "Using Centroids of Word Embeddings and Word Mover's Distance for Biomedical Document Retrieval in Question Answering.". *Proceedings of the 15th Workshop on Biomedical Natural Language Processings,* 114-118

[5] Quoc Le, Tomas Mikolov. (2014). "Distributed representations of sentences and documents". *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 2014.