확장 키워드 매칭을 이용한 영화 장르 별 광고 카테고리 추천

서인식○ 황규백
숭실대학교 컴퓨터학부
issuh@ml.ssu.ac.kr, kbhwang@ssu.ac.kr

Advertisement category recommendation for movie genres using expanded keyword matching

Insik Suh○ Kyu-Baek Hwang
School of Computing, Soongsil University

1. 서 론

최근 IPTV가 실용화 단계에 접어들면서 많은 양의 컨텐츠에 효과적으로 광고할 수 있는 방식이 필요해졌다. 기존의 TV 컨텐츠에 광고주가 광고를 입찰하는 방식은 많은 양의 컨텐츠를 가진 IPTV에는 적절치 않으므로 광고주에게 미리 적절한 광고를 추천해주는 방식이 요구된다. 본 논문에서는 IPTV 컨텐츠가 가진 카테고리 중 영화를 선택하여 이 영화 데이터를 장르 별로 묶어 각 장르에 산업 별로 분류된 광고 카테고리를 자동으로 매칭시키는 방법을 제안하고 제안된 방식이 적절한지 평가한다.

2. 관련 연구

[1]은 비디오 자막에서 광고 키워드를 찾는 방법을 제안했다. 구체적으로 이전의 접근 방법이 데이터에서 제공하는 문서에 제한되어 있음을 지적하면서 이런 "문서 내"의 정보 이외에 이미 알고 있는 상황을 활용함으로써 새로운 자질을 찾아내는 시도를 했다. [2]는 온라인 비디오 환경에서 문맥적으로 적절히 광고하는 방법을 제안했다. 문맥적 광고를 선택하기 위해 주로 광고나 비디오의 텍스트 정보를 활용하여 이를 벡터 모델로 접근하여 각 텍스트를 코사인 유사도로 비교했다. 그리고 텍스트를 SVM(support vector machine)을 활용하여 카테고리 별로 분류함으로써 광고 선택의 정확성을 높였다.

3. 광고 추천 시스템

본 논문에서 제안한 시스템은 그림 1과 같은 과정을 거친다. 첫째로, 컨텐츠 메타데이터의 영화 카테고리를 각 장르 별로 묶어 각 영화의 주요 데이터를 추출 및 확장한다. 그리고 확장된 데이터에서 각 장르 별로 핵심 키워드를 추출한다. 둘째로, 광고 데이터를 주요 카테고리 별로 묶어 태그를 추출하고 각 카테고리 별로 핵심 태그만 남긴다. 마지막으로, 이 두 가지 데이터를 매칭한다.

실험에서 사용된 데이터는 두 가지이다. 첫째는 KT에서 제공한 IPTV 컨텐츠 메타데이터로 총 60개의 카테고리로 구성되어 있는데 이 중 장르 별로 구분하기 쉬운 영화 카테고리가 실험에 사용됐다. 그리고 둘째로 애드와플[1]에서 제공된 광고데이터가 사용됐다. 그리고 이 데이터가 가진 24개의 산업별 카테고리를 실험에 활용했다(표 1참조).

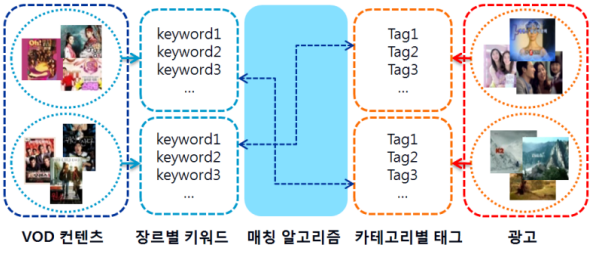

그림 1. 실험 구상도

	컨텐츠 메타데이터	광고 데이터
건 수	2122건	8419건
단어 수	58742개	62205개
분류	20개 장르	24개 카테고리

표 1. 실험에 사용된 컨텐츠 및 광고 데이터

그림 2는 컨텐츠 메타데이터의 처리 과정을 보여준다. 영화 컨텐츠 메타데이터에서 필요한 정보를 추출하여 다음 오픈 API[2]를 통한 영화 데이터의 보완 및 확장을 하고, 이 정보들을 벡터 형식으로 표현하기 위해 줄거리 등의 정

1) http://www.ad.co.kr/
2) http://dna.daum.net/DNALatte/openapi/about/

보를 키워드 형태로 나타낸다. 이 과정에서 한국어 형태소 분석기(Korean Language Technology)[3][3)]가 활용됐다. 그리고 이렇게 남은 각 컨텐츠 메타데이터들을 영화 장르 별로 묶어 각 장르를 대표하는 키워드를 TF-IDF(Term Frequency - Inverse Document Frequency)를 사용하여 추출했다. 각 장르에서 TF-IDF값이 높은 상위 100개를 각 장르의 대표 키워드로 선정했다.

광고 데이터의 처리는 애드와플 홈페이지에서 광고 제목 및 태그 등을 가져와서 이 광고 데이터를 산업별 카테고리 별로 묶는다. 그리고 TF-IDF를 사용하여 각 카테고리를 대표하는 태그 상위 100개를 선정했다(그림 3 참조).

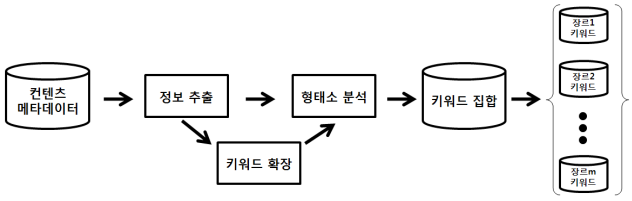

그림 2. 컨텐츠 메타데이터 처리 과정

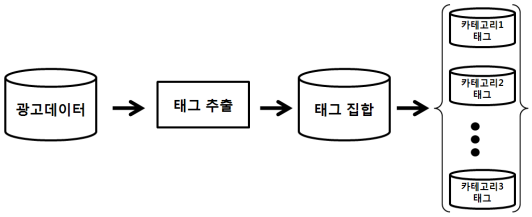

그림 3. 광고 데이터 처리 과정

매칭은 벡터 형식으로 표현되어 있는 각 장르 및 카테고리를 주어진 유사도 측정 방법을 적용하여 가장 높은 값부터 순위를 주었다. 유사도 측정 방법으로 코사인 유사도(Cosine similarity)와 타니모토 계수(Tanimoto coefficient)가 실험에 사용됐다.

4. 실험 결과 및 분석

장르 대 카테고리 정답 테이블은 각 장르에 매칭되는 카테고리에 연구원 3명이 3단계로 적절성을 평가하여 만들어졌다. 정확도는 매우적절과 적절을 정답으로 했고, 정답과 오답(부적절)의 전체 비율은 약 2:1이었다. 실험 결과로 상위 5개가 활용됐고, 평가 측정 방법은 정확도(Precision)와 재현율(Recall) 및 F 척도(F-measure)를 사용했다.

그림 4에서 정확도 그래프를 보면 상위 1개에 대해서는 거의 비슷하다가 급격히 줄어드는데 적절치 못한 키워드가 선정되었기 때문으로 판단된다. 상위 4개 이후로는 크기가 다른 데이터에서 오는 문제를 효과적으로 극복하는 타니모토 계수가 조금 높은 상태를 유지한다. 재현율과 F 척도는 두 실험 모두 단조 증가 형태로 올라가는 모습을 보이고 F 척도가 약간 더 경사가 크다.

표 2는 각 장르에 추천된 광고 카테고리와 이 때 매칭된 키워드를 예로 든 것이다. 장르와 카테고리만 보면 애매할 수도 있지만 매칭된 키워드는 장르, 카테고리 모두에 관련이 깊은 키워드임을 알 수 있다.

코사인 유사도와 타니모토 계수의 정확도 평균은 각 0.7835, 0.7820이고 재현율 평균은 각각 0.1531, 0.1514로 두 가지 방법이 큰 차이를 보이진 않고 유사한 결과를 보여준다.

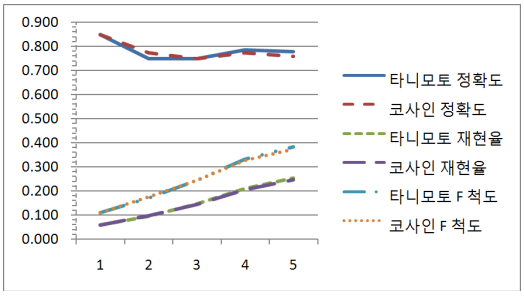

그림 4. 실험 결과 비교

장르	카테고리	매칭된 키워드
액션	수송기기	레이싱,차
스릴러	컴퓨터및정보통신	전화,저녁,게임
공포	화학공업	거울,실내,사진
전쟁	그룹및기업광고	밀림,군인,실화,,
뮤지컬	서비스	뮤지컬,브로드웨이...

표 2. 타니모토 계수를 사용한 추천 결과 예제

5. 결론

본 논문에서 제안한 키워드 확장을 활용한 영화 장르 별 광고 카테고리 추천 방식을 통해 각 영화 장르와 추천된 광고 카테고리의 매칭된 단어가 비교적 적절히 매칭된 것을 볼 수 있었다.

참고 문헌

[1] J. Lee, H. Lee, H. Park, Y. Song, and H. Rim, Finding advertising keywords on video scripts, *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 686-687, 2009.

[2] T. Mei, X.-S. Hua, L. Yang, and S. Li, VideoSense: towards effective online video advertising, *Proceedings of the 15th International Conference on Multimedia*, pp.1075–1084, 2007.

[3] 강승식, "한국어 형태소 분석과 정보 검색", 홍릉과학출판사, 2002.

3) http://nlp.kookmin.ac.kr/