정형 데이터와 비정형 데이터와 상호학습을 통한 영화 흥행 예측 정확도 향상 방안

저자
(Authors)

고은정, 김남규

출처
(Source)

한국지능정보시스템학회 학술대회논문집 , 2017.8, 11-12 (2 pages)

발행처
(Publisher)

한국지능정보시스템학회
Korea Intelligent Information Systems Society

URL

http://www.dbpia.co.kr/Article/NODE07282098

APA Style

고은정, 김남규 (2017). 정형 데이터와 비정형 데이터와 상호학습을 통한 영화 흥행 예측 정확도 향상 방안. 한국지능정보시스템학회 학술대회논문집, 11-12.

정형 데이터와 비정형 데이터의 상호학습을 통한 영화 흥행 예측 정확도 향상 방안

고은정
국민대학교 비즈니스 IT 전문대학원
sbtm3459@kookmin.ac.kr

김남규
국민대학교 경영정보학부
ngkim@kookmin.ac.kr

Abstract *– 영화는 최신 트렌드에 매우 민감하며, 영화 흥행 전략은 흥행 정도에 따라 실시간으로 변경된다. 또한 전체 개봉 영화 중 흥행작이 차지하는 비율은 매우 적기 때문에, 효율적 투자를 위해 영화의 흥행을 사전에 예측하는 것은 매우 중요하다. 이러한 이유로 최신 영화의 개봉 시점 데이터를 분석하여 총 누적 관객수를 예측하는 연구가 다수 수행되었으며, 이들 대부분은 목적 변수(Target data)의 값이 사전에 알려져 있는 기분류(Labeled) 데이터를 활용한 지도학습 기반으로 이루어졌다. 지도학습의 예측 정확도 향상을 위해서는 많은 수의 기분류 데이터에 대한 학습이 이루어져야 하는데, 최신 영화의 개봉 시점 데이터를 충분히 확보하는 것은 현실적으로 매우 어렵다. 따라서 본 연구에서는 미분류(Unlabeled) 데이터의 목적 변수 값을 식별하고, 이를 다시 학습에 활용하는 준지도학습(Semi-supervised learning) 기반의 영화 흥행 예측 방안을 제시한다. 특히 영화 데이터의 구조적 특성과 텍스트 특성을 구분하여, 구분된 속성들 간 상호학습(Co-Training)을 통해 영화 흥행 예측의 정확도를 향상시키는 방안을 제시하고자 한다.*

Key Terms *– 빅데이터, 텍스트 마이닝, 토픽 모델링, 준지도학습, 상호학습*

I. 서론

2016 년 전체 한국 영화산업 매출은 2 조 2,730 억 원으로 2015 년 대비 7.6% 증가했으며, 2014 년 이후 2 조원 대 매출을 유지하고 있다. 2016 년 극장 입장권 매출액은 1 조 7,432 억 원으로 2015 년 대비 소폭 상승(1.6%)한 반면, 관객 수는 2 억 1,702 만 명으로 0.1% 감소했다. 연간 평균 관람 횟수는 4.20 회로 세계 최고 수준이다(영화진흥위원회 산업정책연구팀, 2017). 이러한 산업의 규모에 비해 실제 한국에서 개봉하는 영화 중 흥행에 성공하는 작품의 수는 많지 않기 때문에 투자를 통해 수익을 창출하는 경우는 적다. 따라서 효율적인 투자를 위해 영화의 흥행을 예측하는 모델을 세우는 것은 매우 중요하다.

영화는 최신 트렌드에 민감하기 때문에 예측 모델을 학습할 때 최신 영화 데이터를 사용해야 하며, 흥행 전략은 실시간으로 수립 및 수정되어야 하기 때문에 예측된 내용을 즉각적으로 활용할 수 있도록 개봉 전 혹은 개봉 직후 시점에 총 누적 관객수 예측이 이루어져야 한다. 이 두 가지 사항을 반영한 경우 기분류 데이터가 현저하게 줄어들게 된다. 기분류 데이터를 활용한 지도학습을 기반으로 영화 흥행 예측 모델을 제안한 다양한 연구가 존재하지만, 학습용 데이터가 충분히 확보되지 못한다는 한계가 존재한다. 따라서 이러한 한계를 극복하기 위한 대안으로 본 연구는 준지도학습을 활용하고자 한다.

특히 본 연구에서는 영화에 관련된 데이터가 영화 관객수 등 구조적 특징과 영화 리뷰 등 텍스트 특징으로 명확하게 나뉜다는 점을 활용하여, 준지도학습의 모형 중 상호학습을 기반으로 한 연구 방안을 제안하고자 한다. 우선 데이터의 특징에 따라 원본 데이터를 분할하고, 그 중 텍스트 데이터는 토픽 모델링을 통해 구조화한다. 각 특징 데이터는 훈련용(Training), 검증용(Validation), 스코어링용(Scoring) 데이터 역할에 맞게 구성한다. 훈련용과 검증용 데이터를 학습 시키고, 각 특징 데이터에 따라 분류 모델을 만들고, 해당 목적 변수 값 중 일정 기준을 만족하는 경우 다른 특징의 훈련용 데이터에 적용한다. 그리고 확장된 학습용 데이터를 통해 다시 분류 모델을 만드는 과정을 반복하는 상호학습을 기반으로 한 제안 방법론의 정확도를 파악하고자 한다.

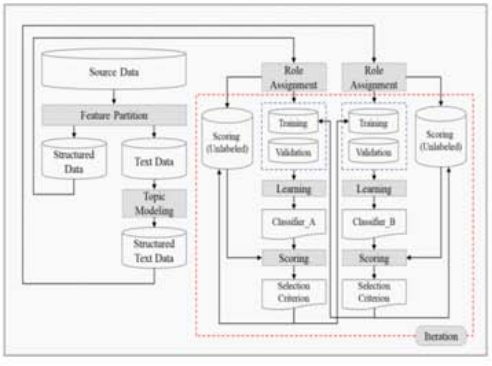
II. 관련 연구

준지도학습(Liu et al., 2003; Lu et al., 2013; Zhu and Goldberg, 2009)은 훈련용 데이터를 통해 학습된 모델을 사용하여 미분류 데이터 중 일부를 예측 값을 통해 기분류 데이터로 분류하고, 해당 데이터들을 훈련용 데이터에 반영하여 다시 모델을 학습시키는 반복적인 패턴을 사용한다. 준지도학습의 대표적인 모형으로 자기학습(Self-Training)(Yarowsky, 1995)과 상호학습(Blum and Mitchell, 1998)이 널리 사용되고 있으며, 준지도학습 기반 분류기의 성능은 기분류 문서와 미분류 문서 비율(Triguero et al., 2014), 학습 데이터 및 목적 데이터의 분포(Lee et al., 2015; Nigam et al., 2006; Silva and Ribeiro, 2004) 등의 영향을 받는 것으로 알려져 있다.

상호학습은 독립적인 특징에 따라 데이터를 나누어 각 특징에 따라 분류 모델을 만들고, 해당 분류 모델을 통해 나온 예측 값을 통해 미분류 데이터를 기분류 데이터로 바꾸어 다른 특징 데이터의 훈련용 데이터에 반영한 후 다시 모델을 학습시키는 것을 반복하는 방법이다.

III. 제안 방법론

본 연구의 제안 방법론은 <그림 1>과 같다.



<그림 1> 제안 방법론의 연구 모형

우선 원본 데이터를 각 특징에 따라 구조적 데이터와 텍스트 데이터로 구분한다. 텍스트 데이터는 원본 데이터 그대로는 예측 모델에 활용할 수 없으므로, 토픽 모델링(Topic modeling)을 통해 구조적 데이터처럼 구조화된 텍스트 데이터로 만들어 분석을 진행한다. 구조적 데이터와 구조화된 텍스트 데이터를 각각 훈련용, 검증용, 스코어링용 데이터 역할에 맞게 분할한다. 이 때 훈련용과 검증용 데이터는 기분류 데이터이고, 스코어링용 데이터는 미분류 데이터이다. 해당 역할에 맞게 데이터를 설정하고, 훈련용과 검증용 데이터를 학습에 사용하여 각 분류 모델를 만들고, 해당 분류 모델을 통해 각 스코어링 데이터를 스코어링한다. 스코어링을 마친 데이터는 일정 기준을 넘기면 다른 특징의 훈련용 데이터에 반영하고, 기준에 미치지 못하면 스코어링된 내용을 삭제한 후 다시 스코어링 데이터에 남아있도록 한다. 이 과정을 반복(Iteration)하여 훈련용 데이터는 지속적으로 확장되고, 스코어링 데이터는 지속적으로 축소된다.

IV. 참고문헌

영화진흥위원회 산업정책연구팀, 2016 년 한국 영화산업 결산, 2017.

Blum, A. and Mitchell, T., "Combining Labeled and Unlabeled Data with Co-Training," COLT: Proceedings of the Workshop on Computational Learning Theory, 1998.

Lee, S., Kim, J., and Myaeng, S. H., "An Extension of Topic Models for Text Classification: A Term Weighting Approach," Proceedings of the 2015 International Conference on Big Data and Smart Computing(BigComp), (2015), 217-224.

Liu, B., Dai, Y., Li, X., Lee, W. S., and Yu, P. S., "Building Text Classifiers Using Positive and Unlabeled Examples," Proceedings of the 3rd IEEE International Conference on Data Mining, (2003), 179-188.

Lu, Y., Okada, S., and Nitta, K., "Semi-supervised Latent Dirichlet Allocation for Multi-label Text Classification," Proceedings of 26th IEA/AIE, (2013), 351-360.

Nigam, K., McCallum, A., and Mitchell, T., "Semi-Supervised Text Classification Using EM," Supervised Learning, MIT Press, 2006.

Silva, C. and Ribeiro, B., "Labeled and Unlabeled Data in Text Categorization," Proceedings of the IEEE International Joint Conference on Neural Networks, (2006), 2971-2976.

Triguero, I., Saez, J. A., Luengo, J., Garcia, S., and Herrera, F., "On the Characterization of Noise Filters for Self-training Semi-supervised in Nearest Neighbor Classification," Neurocomputing, Vol.132, (2014), 30-41.

Yarowsky, D., "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, (1995), 189-196.

Zhu, X. and Goldberg, A. B., Introduction to Semi-Supervised Learning, Morgan & Claypoll Publishers, 2009.