상관관계 분석을 통한 소비예측 시 필요 요소 도출 및 LSTM을 이용한 소비예측 모델

이기훈*, 김진아*, 문남미*
*호서대학교 컴퓨터공학과
e-mail:lkh51738@gmail.com

Correlation analysis is needed to predict consumption and consumption prediction model using LSTM

Kihoon Lee*, Jinah Kim*, Nammee Moon*
*Dept of Computer Science and Engineering, Hoseo University

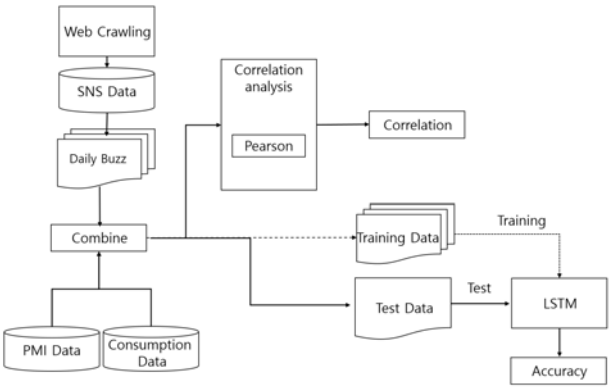
요 약

오프라인 소비자의 의사결정은 크게 라이프스타일, 동기, 개성, 학습 등 개인적인 영향요인과 문화, 기후, 가족 등 기타 상황적 요인을 포함하는 환경적 영향요인에 의해 결정된다. 이러한 요인들을 입력값으로 하는 다양한 딥러닝 모델을 이용한 소비예측 연구들이 진행되고 있다. 딥러닝을 이용한 예측모델을 사용하기 위해서는 먼저 요인들이 의사를 결정하는데 있어 얼마나 상관관계가 있는지 파악하는 작업이 중요하다. 본 논문에서는 이를 위해 다양한 상관관계 분석모델을 이용해 소비 의사결정 요소 중 기후, 문화와 같은 상황적 요인과 소비와의 상관관계를 도출하고, 기후, 문화를 대변하는 미세먼지 데이터와, SNS 버즈량 데이터와 소비데이터를 학습하는 소비예측 LSTM모델을 제안하고자 한다.

1. 서론

인공지능 기술의 발달함에 따라 다양한 빅데이터를 이용해 미래를 예측하려는 시도가 증가하고 있다 [1]. 최근 인공지능의 일환으로, 딥러닝을 이용하여 실생활에 관련한 다양한 예측모델이 연구되고 있다. 예측 모델은 주식, 기상. 번역, 상권 등 다양한 분야에서 사용되고 있다. 그 중 딥러닝을 이용한 경제분야에 대한 예측은 관심도 높은 이슈 중 하나이다[2]. 본 논문에서는 경제예측 중 가계소비에 대한 예측을 하고자 한다. 소비에 대한 예측을 하기 위해서는, 소비자의 의사결정에 대한 예측이 중요하다. 소비자의 의사결정은 크게 라이프스타일, 동기, 개성, 학습 등 개인적인 영향요인과 문화, 기후, 가족 등 기타 상황적 요인을 포함하는 환경적 영향요인으로 나누어진다[3]. 본 논문의 연구는 이러한 요소들 중, 문화를 대표하는 SNS 데이터와, 기후 데이터와 소비와의 상관관계를 분석한다. 또한 상관관계 분석 뿐만 아니라 딥러닝을 이용하여 미래의 소비를 예측할 것이다. Google에서 제공하는 Tensorflow와 Long Short Term Memory networks (LSTM)을 활용하여 소비를 예측한다. 본 논문의 구성은 1장에서 해당 연구 에 대한 간단히 소개 후 본론의 2장에서는 연구를 함에 앞서 어떠한 데이터를 수집하고 정제하였는지, 학습 요소로 왜 기후데이터와 SNS 데이터를 활용했는지 대해서 설명을 한 후 LSTM을 사용한 시스템과 그에 해당하여 본 연구에 적용한 핵심기술에 대해 설명한 후 해당 기술을 적용하여 소비예측에 대한 진행결과에 대해서 설명을 한다. 3장 결론에서는 제안한 연구 주제와 사용한 LSTM의 활용방안, 더 나아가야 할 방향에 대해 언급하고자 한다.

2. 본론



(그림 1) 전체 프로세스

위 그림 1은 본 논문에서 하고자하는 SNS 데이터와 기후데이터, 소비데이터를 이용한 상관관계 분석과 소비예측을 위한 LSTM 예측 모델에대한 프로세스이다. 먼저 크롤링을 통해 SNS 데이터를 수집하고, 수집한 데이터를 파싱을 통해 일별 버즈량을 파악한다. 이렇게 수집된 데이터를 날짜를 키 값으로 기후데이터인 PMI 데이터와 소비데이터를 병합한다. 병합된 데이터를 바탕으로 피어슨상관관계 분석기법을 이용하여 세 데이터 간의 상관관계를 분석한

다. 또한 병합된 데이터를 트레이닝데이터와 테스트 데이터로 나누어 소비예측에 대한 정확도를 파악하는 프로세스이다.

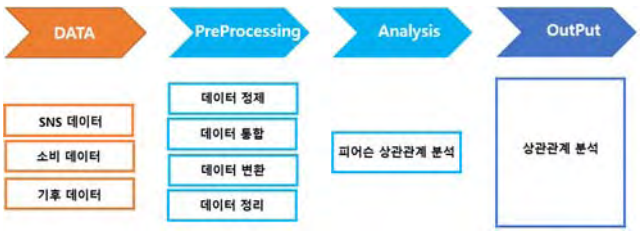
2-1. 데이터 수집

Date	tamp	count	sns	지역	측정소명	측정일시	PM25
20170101	3.12E+08	16209.16	57.17791	충남	천안,성황동	2017010101	43
20170102	2.35E+08	14829.76	48.71165	충남	천안,성황동	2017010102	42
20170103	2.23E+08	14268.32	56.01226	충남	천안,성황동	2017010103	41
20170104	2.4E+08	14302.2	53.68098	충남	천안,성황동	2017010104	10
20170105	2.25E+08	14176.36	59.7546	충남	천안,성황동	2017010105	40
20170106	2.74E+08	16209.16	70.92024	충남	천안,성황동	2017010106	30
20170107	3.48E+08	19253.52	76.07361	충남	천안,성황동	2017010107	43
20170108	3.24E+08	18580.76	67.42331	충남	천안,성황동	2017010108	20
20170109	2.09E+08	13193.84	53.80368	충남	천안,성황동	2017010109	52
20170110	2.5E+08	14641	53.61963	충남	천안,성황동	2017010110	61
20170111	2.38E+08	15028.2	60.49079	충남	천안,성황동	2017010111	41
20170112	2.33E+08	14036	55.15337	충남	천안,성황동	2017010112	42
20170113	2.64E+08	15415.4	60.92024	충남	천안,성황동	2017010113	40
20170114	3.42E+08	17961.24	77.17791	충남	천안,성황동	2017010114	43
20170115	3.14E+08	17273.96	66.07361	충남	천안,성황동	2017010115	39
20170116	2.26E+08	13658.48	50.12269	충남	천안,성황동	2017010116	37
20170117	2.43E+08	14045.68	56.56441	충남	천안,성황동	2017010117	40
20170118	2.24E+08	14157	58.0368	충남	천안,성황동	2017010118	44
20170119	2.26E+08	13798.84	52.94478	충남	천안,성황동	2017010119	52
				충남	천안,성황동	2017010120	54

(그림 2) 수집 데이터 예시

그림 1은 상관관계 예측 및 예측을 위한 딥러닝 모델에 사용될 데이터로 차례로 왼쪽 표는 일별 소득금액, 결제횟수, 소비관련 키워드로 수집한 SNS 버즈량을 나타낸다. 오른쪽 표는 천안지역의 미세먼지농도(PM25)를 일별로 나타내주는 공공데이터이다. 기후데이터 중 미세먼지농도를 선정하여 학습하는 이유는, 최근 국내 미세먼지 농도가 해를 거듭할수록 높아지고, 이에 비례하여 대중의 관심도가 높아짐을 관련 자료와, 각종 언론기사를 통해 알 수 있다[4]. 실제로 이러한 미세먼지 농도가 오프라인 소비에 미치는 영향에 대한 연구가 진행된 바 있다[5]. 이러한 연구에 근거하여 기상데이터 중 미세먼지를 선정하였다.

해당 연구는 미세먼지가 국내 소매판매에 미치는 영향을 분석하였으며, 소매판매는 대형소매점의 판매액으로 측정했다. 연구 결과 미세먼지(PM2.5) 농도가 10μg/m3 증가시 대형 소매점 판매가 약 2%정도 감소한다는 분석 결과를 보였다. 이때 분석방법으로는 상관계수 산출 및 기초회귀분석을 사용 하고자한다.

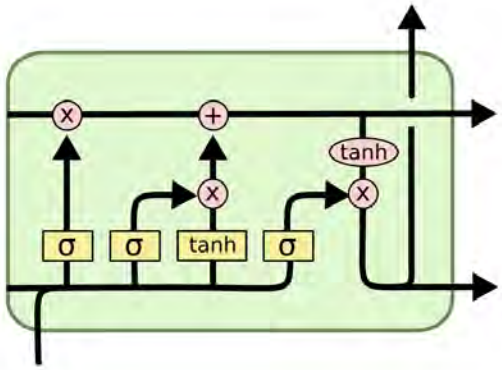


(그림 3) 상관관계 분석 프로세스

위 그림2는 본 연구 중 비정형 데이터인 SNS 데이터와 정형데이터인 소비데이터와 기후데이터 간의 상관관계를 분석하는 프로세스 이다. 수집한 데이터를 분석하기 위해 날짜별로 정리하고 중복되는 날짜의 데이터를 정리하기 위해 정규화를 사용했다. 전처리를 거친 데이터를 이용해서 피어슨 상관관계 분석을 통해 소비와 SNS 및 미세먼지 데이터와의 상관관계를 분석했다.

2-2. LSTM을 사용한 예측모델

본 절에서는 LSTM을 활용한 소비 예측 모델에 관한 연구 내용이다. 본 연구에서는 Google에서 제공하는 Tensorflow를 활용하여 가격 예측 모델을 구현했다. 앞서 예측 모델이 과거 데이터를 바탕으로 학습할 수 있도록 다량의 데이터를 전처리과정을 진행한다.



(그림 4) LSTM 구조

위 그림 2는 RNN을 변형한 모델중 하나인 LSTM으로, RNN의 히든 State에 cell-state를 추가한 구조이다. 이 cell-state로 인해 RNN의 최대 단점은 vanishing gradient problem을 극복한다. 비교적 구조가 간단한 RNN모델을 사용하지 않고 LSTM을 사용한 이유는, 시계열 데이터를 학습하고 데이터의 연속성을 고려하기 위해 본 연구에서는 LSTM을 이용하여 소비를 예측하고자 했다.

3. 결론

본 논문에서는 미세먼지 농도와 SNS 버즈량, 소비데이터 사이의 상관관계를 도출하고, 이를 학습 LSTM 모델로 학습하여 소비를 예측하는 모델을 제시 했다. 실험 데이터는 미세먼지 데이터의 경우 충남 천안시의 관측소에서 2017.01.01.~2018.12.31. 기간 동안 수집했으며, 마찬가지로 소비데이터는 카드사의 충남 천안시 신안동의 2017.01.01.~2018.12.31. 동안 카드소비 데이터를 사용했다. SNS 버즈량의 경우 신안동을 메인 키워드로 소비에 관련된 키워드를 사용하여 인스타그램에서 크롤링한 데이터를 사용했다. 이러한 데이터를 Input 값으로 하는 LSTM을

활용하여 미래의 소비를 예측할 수 있는 모델을 제안했다. 이를 통해 고려해야할 요인이 많아 딥러닝으로 예측이 어려운 소비예측에 있어 지표를 제공하고자 한다.

ACKNOWLEDGEMENT

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2017R1A2B4008886).

참고문헌

[1] 김병운. (2016). 인공지능 동향분석과 국가차원 정책제언. 정보화정책, 23(1), 74-93.
[2] 김유신, 정승렬, 홍성관, 강희주. (2017). 소셜빅데이터를 이용한 온라인 소비자감성지수(e-CCSI) 개발. 인터넷정보학회논문지, 18(4), 121-131.
[3] 황용철(Hwang Yong-Cheol), 김정희(Kim Chung Hee). (2005). 소매 환경에서 지각된 가치의 선행요인과 성과요인에 관한 연구. 소비문화연구, 8(1): 91-114
[5] 유이선, 이재윤 "미세먼지가 국내 소매판매에 미치는 영향" i-KIET 제 30호 2017.7.31.
[4] 조정형(Cho, Jeong-hyung), 차은주(Cha, Eun-joo) , 김영재(Kim, Young-jae). (2018). 여가활동 참여자의 미세먼지 문제인식이 참여행동에 미치는 영향. 여가학연구, 16(03): 1-19