공간 결합과 심층신경망을 활용한 관광지 다중 분류 추천 시스템

안현우, 문남미
호서대학교 컴퓨터공학과
e-mail:nammee.moon@gmail.com

Multiple classification recommendation system using spatial combination and deep learning

An Hyeon Woo*, Moon Nammee**
Hoseo University

요 약

관광지에 대한 관광객의 평가는 날씨, 계절, 관광객의 밀집 정도 등 다양한 환경적 요소에 따라 변화한다. 각 관광지는 객관적인 관점으로 최상의 관광을 경험하게 할 고유한 컨디션이 존재하며 이를 추출하기 위해선 관광에 영향을 주는 여러 환경들에 대한 다중 요인 분석이 가능할 만큼의 정보가 필요하다. 본 논문에서는 심층신경망을 기반으로 한 문장분석기술을 응용하여 관광지 리뷰에 적용, 평점이 포함되지 않은 리뷰에 평점을 추가하여 기상이나 계절, 휴무일 등의 다양한 분류가 가능할 수준의 데이터를 보충하고 축적/보충된 방대한 평점데이터를 토대로 맞춤 추천이 가능하도록 하는 시스템을 설명한다. 이에 본 논문은 학습 환경 구축, 리뷰와 기상 정보의 결합, 최종 추천 방법 등 전반적인 프로세스에 대한 내용을 설명한다.

1. 서론

관광에 대한 만족도는 상품과 달리 복잡한 요소들로 결정된다. 관광 경로의 편의성이나 그 날의 기상, 관광객의 밀집 정도, 문화 관광지의 경우 문화재의 보존 정도 등이 요소의 예인데, 그 중 가장 큰 요소로는 기상 상황이라 볼 수 있을 것이다[1,2]. 그림 1 은 기상이 관광객의 관광계획, 관광에 어떤 영향을 주는지에 대한 그림이다. 때문에 관광지에 대한 관광객의 평가 또한 기상 환경에 따라 천차만별로 다를 수 있다. 정리하자면 비오는 날에 아름다운 관광지, 파도가 높게 칠 때 아름다운 해변 등의 경우가 생길 수 있는 것이다.

(그림 1) 관광 계획과 여행에 기상이 미치는 영향[1]

추천 시스템에 있어서 이러한 환경적 요소가 모두 고려된 추천이 가능하려면 기반데이터는 실로 매우 방대해야 한다. 계절과 휴무일, 비의 유무만 고려하더라도 평가에는 최소한 16개의 기반데이터가 필요한 셈이다. 여기에 온도와 파도의 평균 높이, 강우량, 강설량 등도 함께 고려한다면 구분의 정도 등에 따라 요구사항이 무한히 높아질 수 있는 것이다.

추천 시스템에 사용할 수 있는, 평점이 포함된 유명한 관광 후기 플랫폼들을 조사해 본 결과, 기상과 날씨, 휴무일의 구분에 대한 대략적인 추천이 가능할 만큼의 방대한 데이터를 갖는 관광지는 지역의 랜드마크, 또는 인기관광지로 분류된 관광지들뿐이었으며 비인기 관광지의 경우 10개가 채 되지 않는 관광지도 대다수 발견되었다. 본 논문에서 제안하는 추천시스템은 이러한 데이터의 부족을 극복하기 위해 상대적으로 데이터의 양이 방대한 플랫폼을 대상으로 데이터를 목표치까지 축적하고 평점화시키는 작업을 포함하고 있다. 조사 결과 관광지에 따라 많게는 300배 가량의 데이터를 추가적으로 축적 할 수 있었고 추천을 위한 분류 또한 더욱 다양해질 수 있었다.

제안하는 시스템에서는 계절과 온도, 우기, 휴무일 등으로 분류하고 평가 지표를 구축하였다. 이렇게 다양한 환경으로 평가된 지표를 토대로 평균 평점이 가장 높은 관광지라도 여행 당일 날의 환경에 따라 추천받지 못할 수 있

으며 반대로 평균 평점이 낮은 관광지라도 특정 환경에 대해선 반드시 추천되게 동작하는 것이다.

2. 관련 연구

본 시스템의 구성 요소는 크게 데이터의 수집과 문장 분석 기술, 공간 결합, 추천 메커니즘으로 이루어져 있다. 이 단락에서는 간략하게 기술의 개요와 유사한 추천 시스템을 소개한다.

2-1. 유사한 추천 시스템

문장 분석과 입력 조건을 토대로 한 계절별, 기상별 관광지 추천 기법을 다룬 연구들은 이외에도 많이 있어왔다[3,4]. 하지만 우기나 비가 많이 내릴 경우에 실외 여행이 불필요 하다고 보거나, 단어 사전을 기반으로 한 계절 별 추천을 진행할 뿐 계절이나 특정 기상 상황에 대한 관광객의 관점은 다루지 않는 것이 실정이다.

2-1. 데이터의 수집

본 시스템에서 사용하는 데이터는 관광지의 위치 및 소개 정보를 포함한 관광 데이터와 과거 9년간의 기상청 관측 기록 및 관측소 데이터, 학습을 진행할 평점이 포함된 후기 데이터와 축적에 필요한 수집된 후기 데이터 등이 있다. 관광데이터와 기상 데이터는 공공데이터 포털과 기상청에서 제공하고 있으며 후기데이터는 Python과 Selenium 라이브러리를 이용해 제작한 툴로 수집되었다.

2-2 문장 분석 기술

문장에서 감성을 추출하는 기술은 크게 자질기반 접근 방식과 심층신경망을 이용한 방식으로 나누어진다. 전자는 단어가 갖고 있는 의미적 특징을 이용해 문장 전체가 가리키는 감정을 추론하는 방식이다. 심층신경망 또한 결론적으로 단어의 자질을 토대로 판단한다는 점에서 자질기반 방식과 흡사하나 이를 구성된 신경망 속에서 스스로 학습하고 문맥 수준의 의미 추론이 상대적으로 저렴한 비용으로 가능하며 일반적으로 보다 정확한 결과로 추론 가능하다는 점이 장점으로 볼 수 있다. 하지만 실제 활용에 있어서 필요 데이터의 양이 상대적으로 방대한 편이며 적용을 하기 위해선 반드시 학습 시간이 필요하고 학습된 시스템으로 다른 목적의 추론을 진행하기 힘들다는 점은 단점으로 꼽을 수 있다.

본 제안 시스템에서 문장의 학습과 감성 추론을 진행할 신경망으로는 CNN을 채택하였다. CNN의 컨볼루션 메커니즘을 이어진 단어 사이에 적용시키면 적은 비용으로 N-gram의 효과를 가질 수 있다는 연구 결과가 있다[5].

2-3 공간 결합

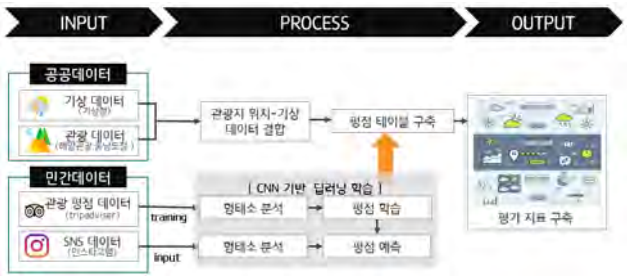
본 시스템은 관광지의 과거 기상정보를 활용하기 때문에 4개의 각기 다른 데이터를 공간적으로 결합하는 과정이 필요하다. 대상 데이터는 관광 데이터, 기상 기록, 관측소 정보, 평점이 포함된 후기 데이터이며, 결합과정은 전체 테이블을 후기 데이터와 결합하여 좌표 정보와 해당 후기 작성 날짜의 가장 가까운 관측소의 기상정보를 붙이는 작업이다.

2-4 추천 메커니즘

추천은 평가지표를 기반으로 진행되며 다양한 추천 메커니즘을 도출할 수 있다. 예를 들어 사용자가 “따듯한 날 가기 좋은 관광지”라는 물음으로 추천을 진행한다면 평가지표에서 따듯한 범주의 온도와 기후조건을 설정하여 평점 순으로 추천한다. 만약 추천을 위한 조건이 따로 없다고 하여도 현재 시점의 날씨 정보를 얻어와 평가지표의 같은 분류의 평가 중 높은 평점 순으로 추천한다면 현재 날씨에 가장 관광하기 좋은 관광지를 추천할 수 있는 것이다.

3. 본론

제안 시스템은 그림 2와 같이 이루어져 있다. 기상청과 공공데이터 포털에서 제공하는 기상정보와 관광지에 대한 정보데이터를 사용하는데, 이는 리뷰 데이터에 포함된 관광지 이름과 게시 일자를 토대로 그 날의 기상과 관광지 정보를 결합하는 데 사용한다. 평점이 포함된 리뷰로서 트립어드바이저 후기를 사용하여 학습하며 학습된 시스템에 평점이 포함되지 않은 인스타그램 리뷰를 입력함으로써 평점을 추가시키는 작업을 할 수 있다. 도출된 평점테이블에는 평점과 관광지 이름, 게시 일자, 리뷰 문장이 담겨있다. 평점테이블의 관광지 이름과 공공데이터인 ‘관광 데이터’ 와 결합하면 관광지의 좌표를 얻을 수 있으며 ‘기상데이터’에서 제공하는 관측소 좌표를 이용하여 가장 가까운 관측소의 기상정보를 참조, 게시일자에 해당하는 기상정보를 붙임으로써 공간 결합을 수행하게 된다.



(그림 2) 추천시스템 전체 프로세스

본 시스템에서 후기를 학습하기 위한 신경망 모델은 CNN(Convolutional Neural Network)을 활용하도록 하였

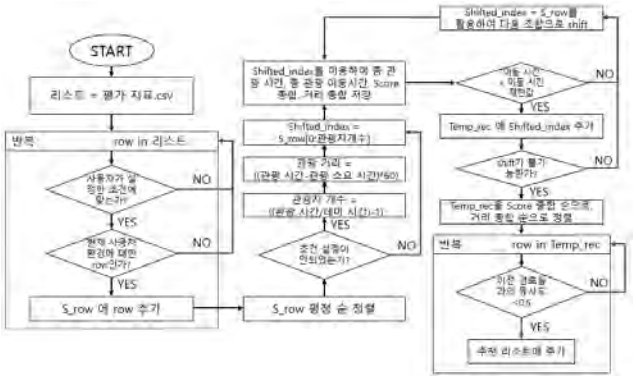
으며 학습과 평가 데이터에 활용한 트립어드바이저의 후기 데이터는 49089개로써 1~5점 평가 형태를 띄우고 있다.

평점이 포함된 테이블이 만들어지면 결합과정과 병합을 거치는데 병합이란 표 1과 같은 정해진 기준을 토대로 결합된 데이터들을 하나의 기준 코드로써 병합하는 과정이다. 최종적으로 결합/병합된 데이터는 그림 2의 평가 지표로써 기상과 평점을 포함한 관광지에 대한 리뷰 정보를 갖고 있다. 결과적으로 이 데이블을 이용하면 “봄의 비 오는 따듯한 휴무날”은 S0C2B1H로 표현 할 수 있으며 이 코드를 활용하면 비슷한 환경을 갖는 날의 각 관광지에 대한 평점을 얻을 수 있을 것이다.

기준	코드	내용
계절	S0~3	3.2 ~ 6.1 봄 6.2~9.1 여름 9.2~12.1 가을 12.2~3.1 겨울
평균 기온	C0~3	6도 미만(추움), 6도~17도 미만(쌀쌀), 17도~28도 미만(선선,따듯), 28도 이상(더움)
일 강수량	B0~3	0mm(없음), 10mm 이하(소), 10mm~50mm(중), 50mm 이상(대)
휴무일	H/W	공휴일/평일

(표 1) 데이터 병합 기준 표

평가지표의 구축이 끝났다면 관광지를 보다 효용적이게 경로 형태로 추천하는 추천 과정이 남아있다. 이는 아래 그림 3과 같은 과정을 통해 요약하여 설명 가능하다.



(그림 3) 추천 과정 순서도

4. 실험 결과

실험은 아래 표 2와 같은 데이터를 기반으로 진행하였다.

데이터	내용	비고
후기데이터(평점 포함)	관광지 명, 후기, 게시 일자, 평점	학습용도로 활용
후기데이터(평점 미포함)	관광지 명, 후기, 게시 일자	평점테이블 구축 용도로 활용
관광 정보	관광지 명, 좌표, 간략한 소개, 주소 등	기상정보 결합을 위해 활용
기상 정보	관측소 식별자, 일자, 기상 정보	
관측소 정보	관측소 식별자, 좌표	

(표 2) 사용 데이터 개요표

후기 데이터를 학습함에 있어서 사전 전처리 작업은 문서를 학습에 적합한 크기로 자르는 Cutting 과정과 Konlpy의 Okt(과거 Twitter)라이브러리를 활용한 형태소 분류, 단어와 평점 정보를 배열에 매핑하는 과정으로 이루어져있다.

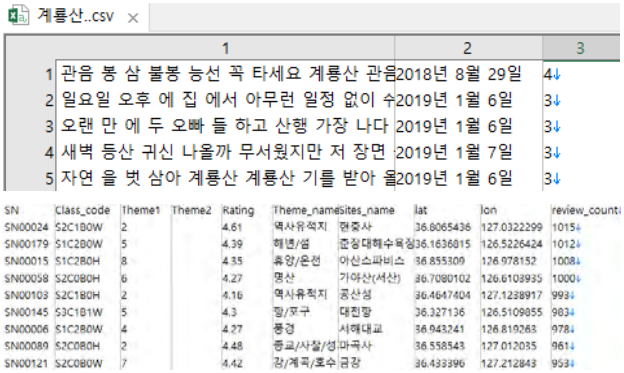
학습의 파라미터 세트는 표 3과 같이 이루어져 있으며 이는 경험적으로 얻어진 결과로써 더 좋은 세트가 존재할 수 있다.

파라미터	값	내용
embedding_dim	32	임베딩 단어 벡터의 차원
filter_sizes	(3,4,5)	필터의 크기, 이미지 분석 시의 커널과와 같은 역할을 한다.
num_filters	128	컨볼루션 채널의 수
dropout_keep_prob	0.2	학습 시 학습될 뉴런의 비중을 다룸. 과적합을 예방할 수 있다.
l2_reg_lambda	0.2	l2 정규화의 람다 값, 정규화 정도를 조정할 수 있다.

(표 3) 사용 데이터 개요표

64개의 배치사이즈를 갖고 6000번의 학습을 진행하였을 때 학습의 결과는, 트레이닝 세트의 경우 64%의 정확도와 평균 1.00의 오차를 가졌고, 테스트 세트의 경우 53%의 정확도와 평균 1.07의 오차를 가졌다.

결합 및 합병의 과정은 본론에 상기한 순서대로 진행하였으며 아래 그림 4는 이러한 평점 테이블과 결합의 결과물을 의미한다. 평가지표를 구축하는 관광지 후기의 게시글은 충남에서 제공하는 “관광명소” 데이터를 활용하여 약 200개의 관광지를 활용하였으며 전체 822,547개의 후기를 평점화하고 결합하였다.


계룡산..csv

	1	2	3
1	관음 봉 삼 불봉 능선 꼭 타세요 계룡산 관음	2018년 8월 29일	4
2	일요일 오후 에 집 에서 아무런 일정 없이 수	2019년 1월 6일	3
3	오랜 만 에 두 오빠 들 하고 산행 가장 나다	2019년 1월 6일	3
4	새벽 등산 귀신 나올까 무서웠지만 저 장면	2019년 1월 7일	3
5	자연 을 벗 삼아 계룡산 계룡산 기를 받아 올	2019년 1월 6일	3

SN	Class_code	Theme1	Theme2	Rating	Theme_name	Sites_name	lat	lon	review_count
SN00024	S2C1B0W	2		4.61	역사유적지	갑사	36.8065436	127.0322299	1015
SN00179	S1C2B0W	5		4.39	해변/섬	춘장대해수욕장	36.1636015	126.5226424	1012
SN00015	S1C2B0H	8		4.35	휴양/온천	아산스파비스	36.855309	126.978152	1008
SN00058	S2C0B0H	6		4.27	명산	가야산(서산)	36.7080102	126.6103935	1000
SN00103	S2C1B0H	2		4.16	역사유적지	공산성	36.4647404	127.1288917	993
SN00145	S3C1B1W	5		4.3	항/포구	대진항	36.327136	126.5109855	983
SN00006	S1C2B0W	4		4.27	풍경	서해대교	36.943241	126.810263	978
SN00089	S2C0B0H	2		4.48	종교/사찰/성	마곡사	36.558543	127.012035	961
SN00121	S2C0B0W	7		4.42	강/계곡/호수	금강	36.433396	127.212843	953

(그림 4) 평점 테이블 및 병합 과정 결과물

5. 결론

본 논문에서는 다양한 환경적 요인을 고려한 관광지 추천 시스템을 소개하였다. 기상/계절/인구밀집도 등이 관광지에 미치는 영향은 무시할 수 없는 정도이나 조사한 바에 따르면 밀집 정도를 유추하고 제시하는 시스템은 있었으나, 계절의 변화에 따른 추천이나 강수량을 판단하고 실

내/실외의 관광지를 추천하는 이원법적인 추천이 주류를 이었다.

제안하는 시스템의 구조는 학습 시스템의 성능이 높아지면 전체 신뢰도가 자연스럽게 올라가며, 데이터가 확충되면 확충되는 만큼의 추천 범위 확장을 의미하는 구조를 갖고 있어 개선 및 확장에 용이하다고 볼 수 있다.

추후 발전 사항으로는, 클래스 코드를 기준화 하고 분류 있는 방법에서 클래스 코드를 없애고 비선형 그래프화 시켜 보다 세밀한 추천이 가능하도록 하는 것을 계획하고 있으며 학습 시스템 또한 관광지 문장 분석에 특화되는 구조를 설계하는 것을 목표로 하고 있다.

사사

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

참고문헌

[1] Scott, D., and Chr Lemieux. "Weather and climate information for tourism." Procedia Environmental Sciences 1 (2010): 146-183.

[2] Becken, Susanne, and Jude Wilson. "The impacts of weather on tourist travel." Tourism Geographies 15.4 (2013): 620-639.

[3] 홍유식, et al. "인터넷 기반 스마트여행 추천시스템." 대한전자공학회 학술대회 (2016): 1903-1904.

[4] 최진우, et al. "SNS 태그 분석 기반의 계절별 여행지 추천 기법." 한국통신학회 학술대회논문집 (2015): 498-499.

[5] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).