일반논문 (Regular Paper)
방송공학회논문지 제24권 제3호, 2019년 5월 (JBE Vol. 24, No. 3, May 2019)
https://doi.org/10.5909/JBE.2019.24.3.515
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

심층신경망 기반 데이터 보충과 영향요소 결합을 통한 하이브리드 추천시스템

안 현 우[a)], 문 남 미[a)‡]

Influential Factor Based Hybrid Recommendation System with Deep Neural Network-Based Data Supplement

Hyeon-woo An[a)]and Nammee Moon[a)‡]

요 약

특정 상품에 대한 사용자의 선호도는 상품의 질 외에도 많은 요소들에 의해 결정된다. 추천시스템에 있어 이러한 외적 요소들의 반영은 데이터의 부족을 포함한 여러 가지 근본적인 문제가 존재하여 지난한 일이었다. 그러나 공공데이터의 개방과 다양하고 방대한 양의 데이터를 가진 평가 플랫폼의 등장 등 기반 환경이 갖춰짐에 따라 외적 요소들의 접근이 용이해 졌다. 이러한 변화에 따라 본 논문은 상품의 품질 외에 사용자의 선호도에 영향을 주는 요소들을 반영할 수 있는 추천시스템 구조를 제안하고 사례를 적용하여 이러한 요소가 실제 선호도에 미치는 영향을 관찰하고자 한다. 제안하는 시스템의 구조는 크게 영향요소를 선정하고 추출하는 과정과 문장 분석을 활용하여 부족한 데이터를 보충하는 과정, 평가데이터와 영향요소를 결합하고 병합하는 과정으로 나눌 수 있으며 제안시스템의 결과 그룹과 실제 사용자 선호도 그룹 간 비교를 통해 구조 변수 설정의 적절성 등을 판단할 수 있는 검증 과정 또한 함께 제안한다.

Abstract

In the real world, the user's preference for a particular product is determined by many factors besides the quality of the product. The reflection of these external factors was very difficult because of various fundamental problems including lack of data. However, access to external factors has become easier as the infrastructure for public data is opened and the availability of evaluation platforms with diverse and vast amounts of data. In accordance with these changes, this paper proposes a recommendation system structure that can reflect the collectable factors that affect user's preference, and we try to observe the influence of actual influencing factors on preference by applying case. The structure of the proposed system can be divided into a process of selecting and extracting influencing factors, a process of supplementing insufficient data using sentence analysis, and finally a process of combining and merging user's evaluation data and influencing factors. We also propose a validation process that can determine the appropriateness of the setting of the structural variables such as the selection of the influence factors through comparison between the result group of the proposed system and the actual user preference group.

Keyword : Hybrid Recommendation, influencing Factor, Recommendation System

Ⅰ. 서 론

상품에 대한 평가를 SNS나 후기 플랫폼에 공유하여 다른 유저들과의 공감대를 형성하거나 정보를 교환하는 것은 인터넷에서 하나의 문화처럼 자리 잡았다. 이로 인해 상품을 사용한 사용자들의 의견은 급속도로 축적되고 있고, 이 방대한 데이터는 상품 공급자 입장에서 마케팅이나 사용자 관리, 신제품 전략 등의 영역에서 매우 유용하게 사용 될 수 있다. 이 흐름이 가져올 수 있는 긍정적 영향 중 하나는 바로 추천 시스템의 성능 증대이다. 기본적으로 잘 알려진 추천시스템의 핵심 기술인 협업필터링을 포함한 많은 기술들이 반드시 데이터의 축적이 선행되어야 추천을 진행할 수 있으며 데이터의 양과 질에 따라 추천 성능이 크게 좌우된다.

의견 데이터가 이처럼 방대해진다면 의견 그룹화를 통한 활용이 더욱 수월해진다. 예를 들면 월별로 나누어진 상품의 평균 평점 그래프라던가 나이나 성별, 사는 지역에 따른 평점 등이 있다. 협업 필터링의 군집화가 이러한 가능성을 활용한 예시이다. 비슷한 성향의 사용자들을 그룹으로 군집화하고 군집내의 평가를 구성원의 평가와 비슷하다고 판단, 추천에 응용하는 것이다.

방대한 데이터는 내용기반 추천 시스템에도 큰 질적 향상을 가져온다. 사용자의 선호 내용을 판단할 근거가 풍족할수록 복잡한 사용자의 니즈를 적합하게 수용할 수 있기 때문이다.

사람의 감성은 환경에 지대한 영향을 받는다. 선호도 또한 사용자의 당시 감성에 크게 영향을 받는데, 이렇게 독립적인 요소로써 사용자의 감성에 영향을 주는 요소들은 명백히 존재하나 적절히 활용되지는 못하고 있다. 이것에는 여러 이유가 있는데 가장 큰 이유로는 영향요소의 수집이 매우 힘들다는 점이다. 예를 들어 음악 평가와 관련된 영향요소들을 본다면 크게는 사용자의 현재 감성, 운동 형태, 주변 날씨 등이 있을 것이고 작게는 사용자의 프로파일, 인기순위 등을 들 수 있을 것인데 평가 데이터에서 직접적으로 수집하기 힘든 것들이다. 또 다른 이유로는 데이터의 부족이다. 만약 적절한 수집 과정을 만들어 주변 날씨를 수집하여 영향요소를 활용한 분석이 가능해졌고, 나아가 온도별로 4개의 분류(추움, 선선, 따듯, 더움)와 강수별로 5개의 분류(없음, 약우, 강우, 약설, 강설)로 평가를 나누어 분석한다면 최소한의 활용을 위해선 적어도 20개의 독립적인 영향요소를 갖는 평가가 존재해야 할 것이다. 물론, 각기의 분류를 독립적인 영향요소로 분리하여 다룰 수 있겠지만 이 경우 “따듯한 날의 비 오는” 환경과 “추운 날의 비 오는” 환경의 특징을 명확히 구분하지 못할 것이므로 제외한다.

추천시스템은 추천 성능을 향상시킬 수 있다면 가능한 많은 정보를 이용하여야 한다[1]. 명백히 사용자의 감상에 영향을 주는 요소가 존재하며 이를 적절하게 활용할 수 있다면 추천시스템의 성능 또한 향상될 것임이 자명하다. 본 논문은 이러한 영향요소를 간접적인 접근을 통해 수집하는 방법을 보여주고 심층신경망을 활용한 문장 분석으로 부족한 데이터를 보충하여 두 가지 문제를 해결하고 영향요소를 활용하는 방법을 제안하는 구조에 관광지 추천이라는 사례를 대입시켜 보여준다.

논문은 각각의 추천 기법과 유사 기법을 소개하는 관련 연구 장과, 제안하는 영향요소 기반 하이브리드 추천시스템(IFBHR : Influence Factor Based Hybrid Recommendation system)을 상세히 설명하는 장, 관광이라는 영역에 IFBHR 구조를 적용하고 테스트 결과를 설명하는 사례적용의 장, 본 연구의 발전방향과 한계점에 대해 논의하는 결론 장으로 구성된다.

Ⅱ. 관련연구

1. 협업필터링과 내용기반 추천

추천 시스템의 알고리즘은 크게 협업필터링과 내용기반 추천 방식을 들 수 있다. 협업 필터링이란 사용자의 구매이

a) 호서대학교 컴퓨터정보공학부(Department of Computer Engineering, Hoseo University)
‡ Corresponding Author : 문남미(Nammee Moon)
E-mail: nammee.moon@gmail.com
Tel: +82-2-2059-2310
ORCID: http://orcid.org/0000-0003-2229-4217
※ This work has supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2017R1A2B4008886).
※ 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2017R1A2B4008886).
· Manuscript received April 30, 2019; Revised May 16, 2019; Accepted May 16, 2019 .

력이나 과거 의견 데이터들을 토대로 평가 이력을 쌓고 여러 사용자들 간 비슷한 패턴을 보이는 사용자들을 군집화하여 평가되지 않은 아이템들을 군집내에서 추천하는 방식이다[2]. 과거 내역을 기반으로 추천을 진행하는 협업 필터링의 기본 원리 때문에 만약 분석할 과거 내역이 존재하지 않는다면 추천리스트에 포함되지 않는 초기 평가자 문제(first rater problem)라는 문제가 존재한다[3]. 파생되는 알고리즘으로 사용자 기반, 아이템 기반, 모델 기반 협업필터링이 존재한다.

내용 기반 추천이란 아이템의 콘텐츠를 분석하여 얻어낸 아이템과 아이템, 또는 아이템과 사용자 선호도 간 유사성을 활용한 추천 기법이다[4]. 사용자가 작성한 프로파일이나 과거 평가를 기반으로 아이템들의 속성들 간 유사성을 찾고 이를 이용하여 추천하는 방식이다. 때문에 협업필터링이 갖는 초기 평가자(first rater)문제는 존재하지 않지만 유사성이 깊은 아이템만을 추천하는 과도한 특수화(Over Specialization) 문제를 갖고 있다.

많은 추천 시스템이 두 가지 알고리즘을 기반하여 연구되었으며 동시에 두 추천 기법이 갖는 문제점을 해결하고자 많은 연구가 진행되었다.

2. 하이브리드 추천시스템

내용 기반 추천은 평가가 존재하지 않아도 추천할 수 있으나 추천 결과가 과도하게 특수화 된다는 단점이 존재하고 협업 필터링의 경우 반대로 다양한 아이템을 추천할 수 있지만 과거 평가 내역이 없으면 추천을 진행하지 못하는 초기 평가자(first rater)라는 문제점이 존재한다. 하이브리드 추천 시스템은 이러한 두 가지의 장・단점을 적절히 융합한 방식이다.

하이브리드 시스템의 모델은 형태에 따라 크게 네 가지로 분류 될 수 있다[5]. 첫 번째는 여러 추천 기법의 결과를 다양한 형태로 조합하는 것이다. 이것은 여러 추천 기법의 결과의 가중평균합을 구해 하나의 결과로 도출하는 방식이 될 수도 있고, 상황에 맞는 추천 기법 하나만을 선택하여 활용하거나 각 추천 기법의 변수를 다양하게 혼합하여 활용하는 방식도 될 수 있다. 두 번째는 내용기반 추천에 사용되는 정보를 협업 필터링에 활용하는 형태의 모델로 과거 평가 내역이 아닌 아이템의 속성을 활용한 사용자 프로파일 구축이 그 예이다. 세 번째는 협업 필터링의 특징을 내용기반에 입히는 방법이다. 토픽 모델을 활용하여 내용 기반의 프로파일을 축소하는 기법이 대표적인 예이다[6]. 마지막 모델의 형태는 협업 필터링과 내용기반 추천 방식을 동시에 고려하는 단일 모델이다.

본 논문에서 제시하는 구조는 영향요소라는 평가데이터나 상품과 독립적인 요소를 기반으로 그룹화 되어 각각의 상품 프로파일에 등록되고 추천을 진행한다는 점에서 세 번째 형태의 하이브리드 추천 시스템이라고 말할 수 있다.

3. 유사한 추천 시스템

본 논문에서 소개하는 IFBHR과 비슷하게 사용자의 환경요인을 고려한 추천 시스템은 이전에도 많은 연구가 이뤄졌었다. 그 중 하나는 사용자의 위치에 따라 선호하는 뉴스의 주제가 변한다는 가정을 전제로 진행한 연구로써 LDA (Latent dirichlet Allocation) 토픽 모델링을 통해 집, 음식점, 직장 등에서 접근하는 기사들을 조사하여 유의미한 결과를 얻어내었다[7].

컴퓨팅 성능이 증대되고 수집 가능한 정보들이 다양해짐에 따라 추천시스템에 사용할 수 있는 정보들도 같이 늘어나고 있다. 본 논문에서 다루는 영향요소도 이러한 정보들 중 하나인데, 비슷한 연구로 아이템 프로파일과 사용자 프로파일이 포함된 정보들과 상황 프로파일이라는 날씨, 계절, 시간 등의 감상 환경을 정의하는 프로파일을 구축하여 추천 알고리즘을 구축하여 좋은 결과를 얻었다[8]. 해당 연구는 이러한 다양한 상황 프로파일이 음악 감상에 영향을 준다는 가정을 전제로 두고 진행하였으며 각각의 프로파일을 다차원의 가중치로 설정하여 다중회귀분석을 통해 영향력을 분석하였다. IFBHR과의 차이점은 상황 프로파일을 가중치로 다루기 때문에 개별 상품 혹은 장르에 대한 영향력은 다루지 못하고 전체 상품에 대한 관계 및 영향력만 저장되기 때문에 아이템 전체에 대한 관계성은 반영할 수 있으나 각 영향요소가 각 장르에 미치는 영향력을 관찰하기는 힘들다.

딥러닝을 기반으로 영향요소를 고려한 연구 또한 존재한다. 딥 오토인코더를 활용한 추천시스템이 그 예인데 본 논

표 1. 유사 시스템 비교 표(●: 가능,존재 ◐: 부분 가능,부분 존재, X: 불가능, 존재하지 않음)
Table 1. Similar system comparison chart(●: possible, existence ◐: partial possible, partial existence, X: impossible, not present))

	LDA-based article recommendation[7]	Multi-profile-based music recommendation[8]	Deep auto-encoder based personalized recommendation[9]	IFBHR
Source of influence factor	External	Individual / External	Individual	External
Recommendation type	Hybrid	Hybrid	CF(Collaborative Filtering)	Hybrid
Utilization Learning Model	STPM (Spatial Topical Preference Model)	Multiple regression analysis	SDAE (Stacked Denoising AutoEncoder)	Appropriate model selection can be applied according to supplementary data
Recommended results	article	music genre	rating score	preference
Personalized recommendations	●	◐	●	◐
Evaluation data supplement	X	X	X	●
No user evaluation history required	X	●	X	●
Solving Over Specialization Problems	●	◐	●	◐
application expandable	X	X	●	●
Relative operation speed (Required operation)	normal (location estimation, topic calculation)	fast (Profile acquisition)	Slow (User review extraction)	fast (Extraction of influence factor)

문이 사용자가 속한 환경요인을 고려했다면 해당 시스템은 사용자 고유의 성질을 고려하는 것을 목적으로 진행된다. 요약하자면 평점과 사용자가 남긴 리뷰, 오피니언 데이터를 토대로 평점간의 상관성을 Stacked Denoising Auto- Encoder (SDAE)를 활용하여 학습하고 적용시킨 연구이다[9].

표 1은 이러한 유사 시스템들과 IFBHR 간의 특징을 정리한 표이다.

Ⅲ. IFBHR(Influence Factor Based Hybrid Recommendation-system)

IFBHR(Influence Factor Based Hybrid Recommendation-system)은 다음과 같은 세 가지의 조건을 전제 조건으로 갖는다.

첫째, 앞서 설명했듯이 수집 가능한 영향요소가 존재해야 한다. 이는 직접적으로나 간접적으로나 접근 가능한 데이터를 의미하며 반드시 사용자의 감상에 영향을 주는 요소로 설명된다. 둘째, 사용자의 선호는 객관적이고 일반적인 선호로 판단한다. 본 추천 시스템은 영향요소가 감상에 미치는 영향을 이용하여 추천을 진행하는 시스템이다. 셋째, 선호를 표현할 단위가 존재해야 한다(이하 선호도라 표기). 사용자의 감상을 대체할 수단으로써 영향요소가 감상에 미치는 영향을 수치화하고 직관적으로 변화를 관찰할 수 있게 한다.

1. 시스템 개요

시스템은 그림 1과 같이 동작한다.

- phase 1 : 영향요소와 연결 데이터의 원천데이터를 수집하는 과정과 원천데이터에서 영향요소를 추출하는 과정이다. 데이터의 수집은 데이터의 테이블을 얻어오는 과정이 될 수 있고 제공을 받기 위한 경로의 확보(API를 통한 제공, 수집 기법을 통한 즉각적인 수집 방법 구축 등)가 될 수도 있다. 일반적인 경우 영향요소는 평가데이터와 독립적인 위치에 존재하므로 간접적인 접근 방식에 의해 수집된다. 때문에 개별 평가데이터와의 연결점을 찾고 필요로 하는 영향요소만을 가려내는 일련의 정제과정이 필요하다. 이 과정은 산포되어있는 영향요소와 평가데이터의 연결을 위한 필요데이터를 정확하게 잘 찾아내는 것이 중요하다. 이를 연결데이터라 칭한다. 일반적으로 원천데이터는 필요한 영향요소 외의 수많은 부가 정보들이 포함되어있다. 이후 과정의 부하를 줄이기 위해 데이터를 정제하는 과정이 필요하다. 수치화된 영향요소를 시스템

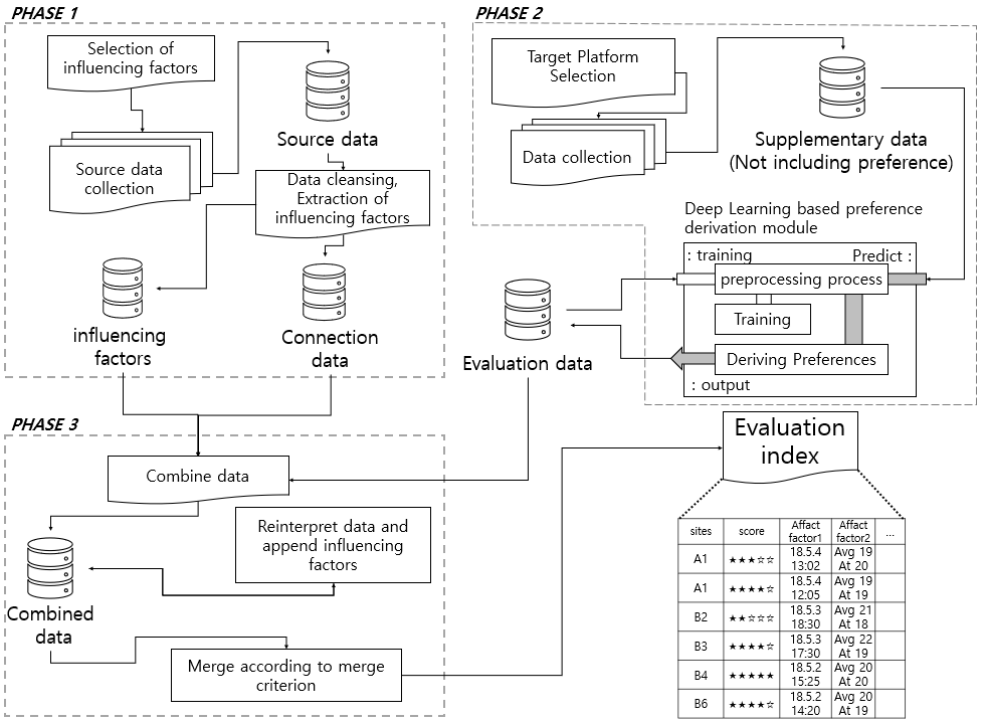

그림 1. IFBHR 전체 흐름도
Fig. 1. The overall flow diagram of the IFBHR

에 맞게 범주화 한다거나, 필요 없는 데이터를 제거하여 크기를 줄이는 것들이 정제과정에 포함된다.

- phase 2 : 평가 데이터의 보충을 진행하는 역할을 한다. 기존의 평가 데이터가 충분하다면 이 과정은 생략될 수 있다. 보충되는 데이터는 기존 선호도가 표현된 평가 데이터와 마찬가지로 영향요소와의 결합이 가능한 속성을 지니고 있어야 한다. 이때 선호도를 추출하기 위해 딥러닝 기반의 문장 분석 기법이 들어가며 학습의 오차는 최종 추천 성능에 큰 영향을 미친다. 만약 보충되는 데이터에 선호도를 대체할 속성이 존재한다면 문장 분석을 통한 선호도 추출은 생략될 수 있다.
- phase 3 : 영향요소와 평가 데이터를 결합하고 병합되는 과정이다. 경우에 따라 이 과정에서 새로운 영향요소가 결합될 수 있으며 이러한 상황은 평가데이터에 내재되어 있는 속성을 2차적으로 해석하는 과정에 의해 발생된다. 주된 예로 날짜를 기반으로 한 계절 분류, 위치를 기반으로 한 지역 분류 등을 들 수 있다. 병합 과정은 평가지표를 만들어내기 위한 마지막 과정으로 평가 지표란 임의의 병합 알고리즘을 통해 동일한 범주의 영향요소에 따라 선호도가 병합되고 표현하는 지표로 설명된다.

2. 영향요소

영향요소란 사용자의 감상에 영향을 미치는 요소이다. 일반적으로 영향요소와 선호도 사이의 관계가 선형을 이루고 있을 때, 선형 회귀 분석을 통해 감상에 미치는 영향력을 관찰할 수 있으며 하나가 아닌 여러 개의 영향요소를 검증하고자 할 때는 아래와 같은 다중 회귀 분석 식을 활용할 수 있다.

• Yi : 선호도, k 개의 영향요소 X1i, X2i .. Xki

: $Yi = \alpha + \beta 1\ X1i + \beta 2\ X2i + \cdots + \beta k\ Xki + \varepsilon$

하지만 일반적인 경우 선호도와 영향요소 사이의 관계는 비선형으로 이루어져 있으며 이를 검증하고자 할 때는 훨씬 복잡한 계산이 필요하다. 근본적으로 Y항에는 자연현상

의 결과가 아닌 사람의 평가가 들어가기 때문에 수학적으로 명백히 검증될 수 없으며 데이터의 보충이 필요한 사례의 경우 선호도와 영향요소 간의 상관관계를 검증하기 위한 최소조건(i>=k)에도 미달하는 경우가 생길 수 있기 때문에 수식을 통한 검증은 매우 제한적이다.

이러한 이유 때문에 결과적으로 영향요소의 선정은 실험적, 직관적, 경험적인 관점에 기반하여 선정된다. 예를 들어 관광지의 경우 사용자의 감상에 미치는 영향요소로 관광객 밀집 정도, 계절, 날씨 등을 꼽을 수 있다. 실제로 관련된 몇몇 연구에 따르면 기상 상황이 여행에 미치는 영향력은 검증되어 있다[10,11]. 음악의 경우 앞선 예시와 비교했을 때 상대적으로 감상자의 감정 상태에 더 크게 영향을 받는다고 볼 수 있다. 따라서 감정상태의 변화를 볼 수 있는 시간[12]을 포함하여 날씨나 인기 정도, 버즈량 등을 영향요소로 들 수 있다.

영향요소를 얻기 위한 과정은 원천데이터의 수집에서 시작한다. 여기서 원천데이터란 영향요소를 포함하거나 간접적으로 영향요소를 표현하거나, 평가데이터와 영향요소와의 결합을 위한 데이터 등 결합을 통해 영향요소를 도출하는데 도움을 주는 모든 데이터를 의미한다. 때문에 원천 데이터의 선정은 시스템적인 관점에서 접근하여야 한다. 만약 영향요소를 추출하기 위해 한 개의 원천데이터만 필요하다 하자. 원천데이터를 S라 하고, 영향요소를 I, 결합을 위한 원천데이터와 평가데이터의 속성을 각각 S.c와 E.c 라 한다면 영향요소 I는 다음과 같이 추출할 수 있을 것이다.

- 원천데이터 S와 평가데이터 E 사이에서 I 추출 과정
  : $I = (S) \bowtie (S.c = Ei.c) (E)$

하지만 일반적으로 원천데이터에서 제공되는 정보의 한계로 인해 위와 같은 직접적인 연결은 찾기 힘들고 간접적인 연결 즉, 연결을 위한 또 다른 원천데이터가 필요할 경우가 있다.

예를 들면 영화 평가를 대상으로 지역을 유추할 속성이 포함되어 있다고 가정하고 해당 일자의 지역별 관광객 수를 통해 밀집 정도를 유추하는 경우가 그 예이다. 영향요소는 관람한 지역의 해당 영화에 대한 관람객 밀집 정도이며 이는 관람객 수로써 간접적으로 표현된다. 관람객 수를 얻어오기 위한 데이터는 영화진흥위원회에서 공공 API로 제공된다. 이때, 필요한 지역에 대한 관람객 수를 얻기 위해서 지역에 대한 코드가 필요하므로 이를 연결하기 위한 또 다른 원천데이터가 필요하다. 이는 평가데이터에서 존재하는 위치데이터가 어떤 형식으로 이루어져 있는가에 따라 다른데 가장 일반적인 포멧인 주소 정보로 이루어져 있다고 할 때, 같은 제공처인 영화진흥위원회에서 제공하는 지역 코드 테이블에서 해당 위치데이터를 대조하고 획득하여 사용하여야 한다. 결과적으로 두 개의 원천데이터(관람 정보 API, 지역 코드 테이블)와 세 개의 연결 속성(영화 제목, 관람 일자, 위치데이터)을 필요로 하게 된다. 이 경우에는 API를 통한 데이터의 수집이 이루어져 결합 과정에서 영향요소를 추출하는 것이 비용적인 면이나 구현의 용이성 면에서 합리적이므로 정제와 추출이 결합과정에서 이루어진다고 볼 수 있다.

영향요소를 선정하고 추출하기까지의 전체적인 흐름은 아래 그림 2와 같다.

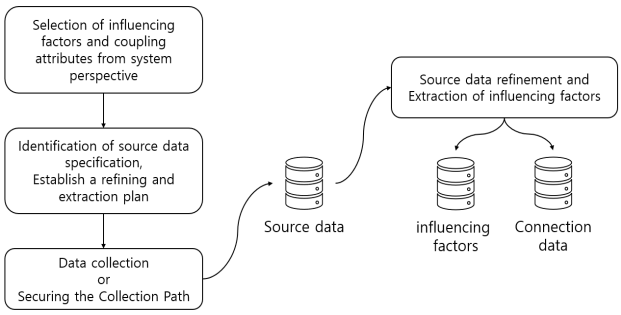

그림 2. 영향요소 추출 과정
Fig. 2. Factor extraction process

3. 데이터 보충

추천리스트에 포함될 아이템들에 대한 각각의 평가데이터가 모든 영향요소와의 결합이 가능할 만큼 충분히 존재한다면 이 과정은 필요치 않다. 데이터가 모자라거나 적정한 수준에 이르지 못했을 때 보충을 진행하는데, 보충을 위한 수집 플랫폼을 선정하는데 있어 아래와 같은 조건을 갖는다.

첫째, 기존 선호도가 포함된 평가 플랫폼과 평가의 기준이 크게 다르지 않아야 한다. 평가 플랫폼은 저마다의 성격을 지니고 있다. 만약 한쪽은 엄격한 기준의 성격을 갖고

있고 다른 쪽이 상대적으로 느슨한 기준의 성격을 갖고 있을 경우 둘 사이의 불균형이 최종 추천 성능을 저해하게 된다.

둘째, 데이터의 형식은 달라도 영향요소와의 결합에 필요한 속성은 반드시 지니고 있어야 한다.

셋째, 문장 분석이 가능한 형태를 지니고 있어야 한다.

모든 조건을 만족하고 보충할 데이터를 수집하였다면 딥러닝을 통한 선호도 도출을 진행한다. IFBHR의 사례적용에서는 CNN을 통해 문장 분석을 진행하였으나 이는 변경이 가능한 부분이며 분석할 데이터의 특성에 맞춰 신경망을 재구성하거나 새로운 구조를 통해 진행하는 등 도출되는 선호도의 오차를 최소로 줄이는 것을 목적으로 한다.

만약 수집 대상 데이터에 선호도를 대체할 수단이 있다면 셋째 조건은 무시할 수 있다. 하지만 첫째 조건에 합치하기 위해 기준이 크게 다르지 않은 플랫폼을 선정하거나 선호도를 조정하여 기존 데이터와 보충 데이터의 불균형을 맞추어야 할 것이다.

보충 과정의 전체 흐름은 아래 그림 3과 같다.

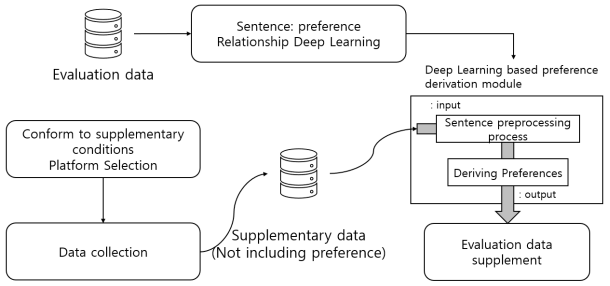

그림 3. 데이터 보충 과정
Fig. 3. Data Supplementation Process

4. 결합 및 병합

최종적으로 도출된 영향요소와 평가데이터를 결합/병합하는 과정이다. 평가데이터와 영향요소와의 결합 속성은 적용 사례에 따라 달라질 수 있다. 본 논문의 사례를 예로 들면 관광지에 대한 영향요소로 평가 관광지에 대한 평가 작성 시기의 날씨를 선정하였고 결합 속성은 평가 일자이다.

이처럼 영향요소의 수집과 결합이 마무리 되었다면 병합 과정이 남아있다. 병합 과정이란, 영향요소를 의미적으로 분류하여 구분의 정도를 낮추는 작업으로써 수치만으로 이루어진 영향요소들을 의미 있는 범주로 그룹화하는 작업으로 볼 수 있다.

병합 과정은 상황에 따라 두 가지로 나뉜다. 첫 번째는 단순히 병합되는 선호도들을 합산하고 평균치를 내어 병합하는 방법이다. 두 번째로는 데이터를 보충하는 작업을 진행하였고 선호도를 도출하는 기법의 결과가 확률분포로 이루어져 있다면 최종 스코어의 확률을 아래 가중평균합 수식을 통해 좀 더 안정적으로 병합하는 방법이 존재한다.

• S: 선호도, P: 스코어 예측 확률, m: 범주에 포함된 스코어 개수

$$: \ AVG(S) = \frac{\sum_{i=1}^{m} P_i * S_i}{\sum_{i=1}^{m} P_i}$$

평가데이터와 영향요소의 결합 및 병합에 관한 전체 흐름은 아래 그림 4와 같다.

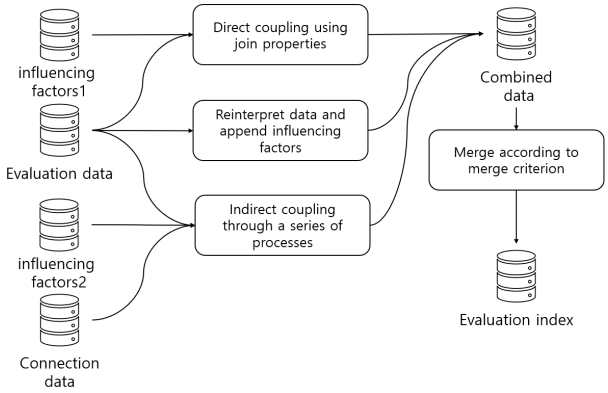

그림 4. 가중평균합 의사코드
Fig. 4. Weighted average sum pseudo code

최종적으로 병합 시 평가지표가 출력되며 이를 활용해 피추천자가 속한 다양한 영향요소에 따라 알맞은 추천이 가능해진다. 이 때 영향요소와의 결합 신뢰도를 판단하는 근거로써 병합된 데이터의 수를 포함시키는데 이는 결합/병합된 데이터가 다양한 영향요소에 따른 실제 사용자 선호도를 얼마나 잘 반영하였는지 확인하는 검증과정에서도

활용 가능하다.

5. 영향요소 검증 과정

데이터의 보충을 진행했을 때 영향요소가 결합되고 병합된 데이터가 어느 정도의 신뢰도를 갖는지 판단하는 과정이다. 이 과정을 통해 영향요소와의 결합이 적절히 이루어졌는지 판단할 수 있으며 병합 과정의 개선이나 보충된 양의 적절성, 병합 데이터 수에 기반한 선호도 신뢰 가중치 등을 판단할 수 있다.

과정은 실제 사용자들의 선호도가 포함된 평가데이터에 영향요소를 결합하고 병합하는 작업과 실제 병합된 선호도의 차를 MAE(Mean Absolute Error) 값으로 구하는 작업으로 이루어진다. 결합 및 병합 과정은 상기 서술한 과정을 그대로 진행하나 잡음을 줄이기 위해 20 이상으로 병합 개수가 넘지 않으면 대조 리스트에서 제거한다. 검증 과정의 흐름은 그림 5와 같다.

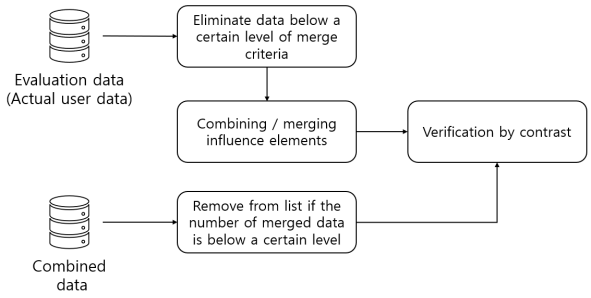

그림 5. 검증과정
Fig. 5. Verification process

Ⅳ. 사례적용

1. 사례 개요

영향요소를 보다 뚜렷이 관찰할 수 있는 좋은 사례로 관광지 추천을 들 수 있다. 이미 관련한 연구를 통해 기상이 관광에 미치는 영향력이 검증되었으며 계절과 관광인구의 밀집을 또 다른 영향요소로 다루었다. 대상 관광지는 충남에 소재한 197개 관광지로 정하였다.

2. 활용 데이터 선정

관광지에 대한 영향요소로는 크게 기상과 밀집도, 계절을 들 수 있다. 기상과 계절은 직접적인 수집이 가능하며 밀집도는 휴무일을 통한 대략적인 유추를 통해 간접적으로 획득하여 사용할 수 있다. 영향요소를 얻기 위한 원천 데이터는 기상청 과거 기록 데이터, 관측소 정보 데이터, 관광지 정보 데이터이다. 이를 이용하여 결합과정에서 관광지와 가장 가까운 관측소의 게시일자 기준 날씨 정보를 얻어오며 밀집을 간접적으로 대신할 휴무일과 계절에 대한 영향요소는 평가데이터에 포함된 게시 일자를 토대로 구하였다. 제공되는 원천 데이터는 이후 작업의 부하를 줄이기 위해 사용하는 정보 외의 데이터를 지우게 되는데 기상청의 과거 기록 데이터에서는 평균 강우량과 평균 온도를 제외한 모든 열을 지우고 관광지 정보 데이터에서는 좌표와 관광지 이름을 제외한 모든 열을, 관측소 데이터에서는 관측소 번호와 좌표를 제외한 모든 열을 지우게 된다.

수집 모듈의 개발은 python selenium을 통해 개발하였으며 평가데이터 수집 플랫폼으로 유명 관광 후기 플랫폼인 TripAdvisor로 선정하였으나 충남 소재의 관광지 평가 데이터의 수가 너무 적어 영향요소로의 결합이 불가능하였다. 따라서 데이터 보충 과정을 진행하기 위하여 지역 구분 없이 가능한 많은 데이터를 목적으로 49089개의 후기 수집을 진행하였다. 수집된 데이터는 학습용도로 활용되며 평가데이터의 역할은 보충된 데이터가 담당하도록 설계하였다.

보충데이터의 수집 플랫폼은 인스타그램으로 선정하였다. 선정 이유는 다음과 같다. 학습에 적당한 문장 길이를 갖고 있으며, 영향요소와의 결합을 위한 게시 일자를 포함하고 있고, 비슷한 기준의 평가와 영향요소를 토대로 한 다양한 분리가 가능할 정도의 방대한 데이터를 포함하고 있다. 일례로, ‘대천해수욕장’과 관련된 후기가 170개 가량이 TripAdvisor에 존재했고 보충데이터 플랫폼인 인스타그램에는 같은 키워드로 111,652개가 존재함을 확인하였다.

3. 추천시스템 구축

TripAdvisor의 후기 데이터를 학습함에 앞서 사전 전처

리 작업으로 문서를 학습에 적합한 크기로 자르는 Cutting 과정과 Konlpy의 Okt(과거 Twitter)라이브러리를 활용한 형태소 분류, 단어와 평점 정보를 배열에 매핑하는 과정이 존재한다. 마찬가지로 보충 데이터의 선호도 도출 작업 또한 사전에 같은 문장 전처리 과정을 진행하게 된다.

학습하기 위한 신경망 모델은 문장분석에도 활용 가능한 CNN(Convolutional Neural Network)[13]을 활용하도록 하였으며 학습에 활용한 후기 데이터는 1~5점 평가 형태를 포함하고 있다. 충분히 학습된 모델을 통해 데이터 보충을 끝마쳤다면, 결합 과정을 진행한다.

결합은 총 4개의 서로 다른 데이터(평가데이터, 기상청 과거 기록 데이터, 관측소 정보 데이터, 관광지 정보 데이터)를 활용한 과정으로 간략하게 요약하자면 관광지의 좌표를 얻고, 가장 가까운 관측소를 얻고, 해당 관측소의 과거 기록에서 평가데이터의 게시 일자에 해당하는 날씨 정보를 얻어오는 과정이다. 결합 이후에 각각의 후기 데이터는 게시일자에 맞는 기상정보를 갖고 된다. 아래는 이를 적용할 의사코드이다.

• 평가 관광지에 대한 좌표정보 lat,lon을 얻어오기 위한 과정(E:i개로 이루어진 평가데이터, S:관광지 공공데이터)
: lati, loni = π(S.lat,S.lon)(S)⋈(S.name=Ei.name)(Ei)

```
get_coord(Ei, S){
  for(s_item in S){
    if (s_item.name == Ei.name){ |
      coord = [s_item.lat, s_item.lon]
      break
    }
  }
  // Return to sightseeing spot coordinates
  return coord
}
```

그림 6. 관광지 좌표 획득 의사코드
Fig. 6. Pseudo code to acquire tourist spot coordinates

• 관광지 좌표와 가장 가까운 거리의 관측소 넘버(O.id)를 얻기 위한 과정(O:관측소 데이터, dist(lat1, lon1, lat2, lon2):두 좌표 사이의 거리)
: Closest_obsi = π(O.id) (DNO O.id MIN(id) KEEP (DENSE_RANK LAST ORDER BY dist(O.lat,O.lon,lati,loni) DESC)O)

```
get_closest_observation(O,coord){
  min_val = 99999
  for(o_item in O){
    // dist(coord1, coord2) : Distance-returning function
    d = dist(o_item.coord, coord)
    if(min_val > d){
      min_val = d
      // Save Observatory Code
      closest_obs_code = o_item.code
    }
  }
  // Return Observation Code at Minimum Distance
  return closest_obs_code
}
```

그림 7. 가장 가까운 관측소 획득 의사코드
Fig. 7. Obtain the nearest station Pseudo code

• 해당 관측소의 과거기록 중 게시 일자(E.date)에 해당하는 기상기록을 얻기위한 과정(W:기상 과거기록 데이터)
: weather=(Ei)⋈W.date=Ei.date(σ(W.obs_id=Closest_obsi)(W))

```
get_weather(closest_obs_code,Ei,W){
  for(w_item in W){
    // Filter records from the closest observatory station
    if(w_item.obs == closest_obs_code){
      // Weather records that match the evaluation date
      if(Ei.date == w_item.date){
        weather_condition = w_item
        break
      }
    }
  }
  // Return the weather record
  return weather_condition
}
```

그림 8. 영향요소 획득 의사코드
Fig. 8. Influence factors acquisition pseudo code

병합은 아래와 같은 표 2를 기준으로 진행된다.

표 2. 병합 기준 표
Table 2. Merge criteria table

standard	code	Contents
season	S0~3	3.2 ~ 6.1:Spring 6.2 ~ 9.1:Summer 9.2 ~ 12.1:Autumn 12.2 ~ 3.1:Winter
Average temperature Daily	C0~3	Less than 6C(cold) 6C~17C(cool) 17C~28C(worm) 28C or more(hot)
precipitation	B0~3	0mm(none) Less than 10mm(small) 10mm~50mm(medium) 50mm or more(huge)
Closed days	H/W	Holyday/Week day

결과적으로 이 테이블을 이용하여 병합한다면 각각의 영향요소를 코드로써 분류할 수 있을 것이며 보다 직관적인

판단이 가능해지고 영향요소로의 분류가 가능해진다. 그림 9는 결합 및 병합을 마친 평가지표의 부분이다. 속성은 순서대로 [관광지의 식별번호, 분류코드, 테마1, 테마2, 평점, 테마이름, 관광지 명, 좌표[lat, lon], 병합된 개수] 으로 이루어져있다.

SN	Class_code	Theme1	Theme2	Rating	Theme_name	Sites_name	lat	lon	review_count
SN00024	S2C1B0W	2		4.61	역사유적지	현충사	36.8065436	127.0322299	1015
SN00179	S1C2B0W	5		4.39	해변/섬	춘장대해수욕장	36.1636815	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	127.1238917	993
SN00145	S3C1B1W	5		4.3	항/포구	대천항	36.327136	126.5109855	983
SN00006	S1C2B0W	4		4.27	풍경	서해대교	36.943241	126.819263	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 9. 평점 테이블 및 병합 과정 결과물
Fig. 9. Rating table and merge process output

4. 실험결과

실험은 2절과 같은 구조에서 진행하였으며 데이터 보충을 위한 CNN 학습 파라미터는 아래 표 3과 같다.

표 3. 실험 사용 학습 파라미터
Table 3. Experimental use learning parameter

parameter	value	Contents
embedding_dim	32	Dimension of Embedding Word Vector
filter_sizes	(3,4,5)	Size of filter. It acts like the kernel in image analysis.
num_filters	128	Number of convolution channels
dropout_keep_prob	0.2	It deals with the weight of neurons to be learned during learning. It is possible to prevent over-fitting.
l2_reg_lambda	0.2	The lambda value of the l2 normalization. The degree of normalization can be adjusted.

64개의 배치사이즈를 갖고 6000번의 학습을 진행하였을 때 학습의 결과는 트레이닝 세트의 경우 53%의 정확도를 가졌고, 테스트 세트의 경우 54%의 정확도를 가졌다. 보충되는 데이터는 같은 형태의 822,547개 후기 데이터이다.

결합과 병합은 3장 4절에서 상기한 방식과 표 1과 같은 기준으로 진행하였다. 병합 된 데이터가 갖는 선호도의 정확도를 검증하기 위하여 실제 데이터의 선호도 차를 MAE (Mean Absolute Error) 방식을 사용하여 구하였다. 검증 신뢰도의 향상을 위해 적어도 5개 이상의 병합 데이터 수를 갖고 있는 데이터를 대상으로 검증하였을 때 총 0.406 값이 도출되었고 10개 이상의 데이터를 대상으로 하였을 때는 0.3011, 20건으로 하였을 때는 0.2154가 도출되었다. 이는 보충된 양과 학습 오차를 고려하여 볼 때, 결합된 영향요소가 실제 사용자가 적용한 선호도에 많은 영향을 끼치며 보다 객관적인 평가를 대변한다고 해석될 수 있다.

가장 많은 데이터를 가진 관광지 중 하나인 '대천해수욕장'의 경우 병합된 데이터에서 휴무일을 제외한 64개 범주 중 병합 데이터 수 기준 상위 10개 범주를 추출하였을 때 총 95,207개의 개수를 가졌고 비율 분포는 그림 10, 11과 같았다.

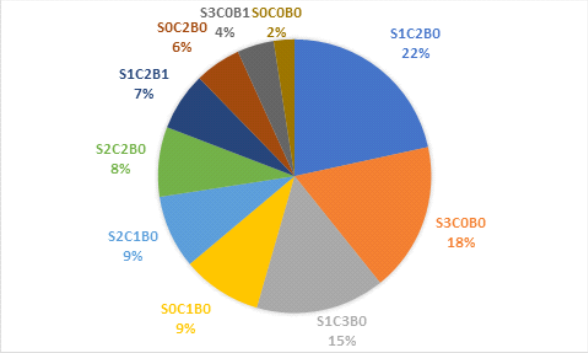

그림 10. 상위 10개 범주의 데이터 분포 비율
Fig. 10. Data Distribution Ratio for the Top 10 Categories

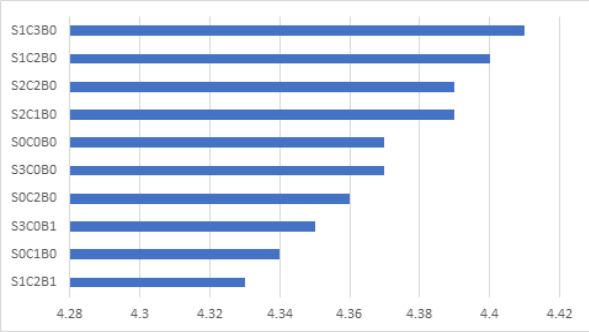

그림 11. 상위 10개 범주의 선호도 차트
Fig. 11. Affinity charts for the top 10 categories

결과를 해석해 볼 때 바다 관광지의 특성 상 여름에 해당하는 'S1'의 범주가 도합 44%이고 겨울('S3',22%), 봄('S0',17%), 가을('S2',17%)가 뒤를 이었다. 또 여름에 해당

하는 'S1'의 평점이 높은 것으로 확인되고 가장 낮은 랭크를 기록한 여름의 선호도는 강우량이 'B1'인 약간의 비가 내리는 여름 바다를 의미하므로 직관적인 관점에서 서해와 습한 바다의 조화가 그리 좋지 못하다는 것을 알 수 있다. 비슷한 예로 8위인 'S3C0B1'(추운 겨울의 약간의 비) 범주가 있다.

V. 결 론

본 논문에서는 실제 사용자의 선호도에 영향을 끼치는 영향요소를 고려한 추천 시스템 구조를 소개하였다. 관광지 추천 사례를 적용한 실험 결과로 영향요소의 결합이 실제 사용자 선호도에 미치는 영향을 뚜렷이 볼 수 있었고 해수욕장의 예시에서 직관적인 관찰도 가능하였다.

IFBHR의 구조는 적절한 영향요소의 선택, 보충 데이터를 담당하는 학습 시스템의 성능 향상 등 부분적 개선이 이루어지면 시스템의 전반적인 신뢰도가 올라가며, 데이터가 확충되면 확충되는 만큼 추천 범위의 확장을 의미하는 구조를 갖고 있어 개선 및 확장에 용이하다고 볼 수 있다.

참 고 문 헌 (References)

[1] J. Son, S. Kim, H. Kim and S. Cho. "Review and Analysis of Recommender Systems" Journal of the Korean Institute of Industrial Engineers, Vol. 41, No. 2, pp. 185-208, April 2015, https://doi.org/10.7232/JKIIE.2015.41.2.185 (accessed April. 15, 2015).

[2] Goldberg, David, et al. "Using collaborative filtering to weave an information Tapestry." Communications of the ACM, Vol. 35, No. 12, pp. 61-71, Dec 1992. (https://go.galegroup.com/ps/anonymous?id=GALE%7CA13039895&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=00010782&p=AONE&sw=w)

[3] Su, Xiaoyuan, and Taghi M. Khoshgoftaar. "A survey of collaborative filtering techniques." Advances in artificial intelligence, Vol. 2009, Article ID 421425, 19 pages, 2009, https://doi.org/10.1155/2009/421425 (accessed Aug. 3, 2009).

[4] Wu, Yi-Hung, and Arbee LP Chen. "Index structures of user profiles for efficient web page filtering services." Proceedings 20th IEEE International Conference on Distributed Computing Systems. IEEE, April 2000. (DOI. 10.1109/ICDCS.2000.840981)

[5] Adomavicius, Gediminas, and Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." IEEE Transactions on Knowledge & Data Engineering 6. vol. 17, pp. 734-749, June 2005. (DOI. 10.1109/TKDE.2005.99)

[6] Soboroff, Ian, and Charles Nicholas. "Combining content and collaboration in text filtering." Proceedings of the IJCAI. Vol. 99. pp. 86-91, sn, 1999. (https://www.csee.umbc.edu/csee/research/cadip/1999Symposium/mlif.pdf)

[7] Noh, Yunseok, Yong-Hwan Oh, and Seong-Bae Park. "A location-based personalized news recommendation." 2014 International Conference on Big Data and Smart Computing (BIGCOMP). IEEE, 2014. (DOI. 10.1109/BIGCOMP.2014.6741416)

[8] Park, Kyong-Su, and Nam-Me Moon. "Multidimensional Optimization Model of Music Recommender Systems." The KIPS Transactions: PartB. Vol. 19, No. 3, pp. 155-164, June 2012, https://doi.org/10.3745/KIPSTB.2012.19B.3.155 (accessed Feb. 31, 2012)

[9] Hyunwoo Je, Junwoo Kim, Mun Y. Yi. "Deep AutoEncoder based Personalized Recommendation System : Considering user's intrinsic characteristics." KOREA INFORMATION SCIENCE SOCIETY. 773-775. June 2017. (http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07207377&language=ko_KR)

[10] Scott, D., and Chr Lemieux. "Weather and climate information for tourism." Procedia Environmental Sciences. Vol 1, pp. 146-183, 2010, https://doi.org/10.1016/j.proenv.2010.09.011 (accessed Nov. 18, 2010)

[11] Becken, Susanne, and Jude Wilson. "The impacts of weather on tourist travel." Tourism Geographies. Vol. 15, No. 4, pp. 620-639, Feb 2013, https://doi.org/10.1080/14616688.2012.762541 (accessed Feb. 12, 2013)

[12] Dzogang, Fabon, Stafford Lightman, and Nello Cristianini. "Diurnal variations of psychometric indicators in Twitter content." PloS one. Vol. 13, No. 6, e0197002, June 2018 (https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0197002&type=printable)

[13] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882, Aug 2014. (https://arxiv.org/abs/1408.5882)

저 자 소 개

안 현 우

- 2018년 : 호서대학교 컴퓨터소프트웨어전공 공학사
- 2018년 ~현재 : 호서대학교 대학원 컴퓨터공학과 석사과정
- ORCID : https://orcid.org/0000-0003-2880-5639
- 주관심분야 : 빅데이터 처리 및 분석, 추천시스템, 인공지능(AI)

문 남 미

- 1985년 : 이화여자대학교 컴퓨터학과 공학사
- 1987년 : 이화여자대학교 공학석사
- 1998년 : 이화여자대학교 공학박사
- 1999년 ~ 2003년 : 이화여자대학교 조교수
- 2003년 ~ 2008년 : 서울벤처정보대학원대학교 디지털미디어학과 교수
- 2008년 ~ 현재 : 호서대학교 컴퓨터소프트웨어전공 교수
- ORCID : http://orcid.org/0000-0003-2229-4217
- 주관심분야 : Social Learning, 빅데이터 처리 및 분석, HCI, 메타데이터, User Centric data analysis