지역별 오픈데이터를 활용한 주거지역 추천 시스템

*이현지 **김효정 ***김지호 ****김시은 *****임양미
덕성여자대학교
*han2ul0221@naver.com

Residential Area Recommendation System Using Regional Public Data

Lee, Hyun-Ji Kim, Hyo-Jeong Kim, Ji-Ho Kim, Si-Eun Lim, Yang-Mi
Duksung Women's University

요약

코로나19 이후 사람들이 집에서 보내는 시간이 증가함에 따라 삶을 살아가는 형태가 변화하였고, 집뿐만 아니라 주거 지역의 환경 또한 중요성이 대두되었다. 그러나 주거 지역에 환경과 관련된 정보를 사용자가 보기 쉽게 제공하는 어플리케이션이 거의 존재하지 않는다. 이러한 문제를 해결하기 위해 주거 환경을 결정하는 데 있어서 중요하다고 판단되는 속성에 대해 공공포털에서 제공하는 오픈 데이터를 활용하여 본 연구에서 주거 환경을 결정하는 데 있어서 중요하게 판단되는 속성을 바탕으로 주거 지역 추천 모델을 제안한다. 공공포털에서 제공하는 오픈 데이터를 활용하여 콘텐츠 기반 모델을 통한 지역 추천 시스템을 도입했다. 사용자의 라이프스타일에 대한 설문조사 결과와 온라인 상에서 수집한 뉴스, 통계보고서 등의 데이터 분석 결과를 활용하여 가중치를 도출하고, 이 가중치를 모델에 적용해 사용자 개인 선호도에 맞는 주거 지역을 추천해 주는 추천시스템을 개발하였다.

1. 서론

코로나19는 삶을 살아가는 형태와 사람들의 인식에 새로운 변화를 일으켰다. 온라인 수업의 증가로 인해 집에서 작업하는 시간이 늘어남에 따라 집에서의 생활의 질이 중요해지고 주거 환경과 주거지에 대한 중요성이 주목받게 되었다.[1]

자신이 살고자 하는 집의 지역을 선택할 때 많은 사람은 집값을 고려함과 동시에 주변 편의시설, 환경, 대중교통, 안전 등의 요소를 고려하며 재산이나 부양가족 수에 따른 다양한 목적들을 반영한다. 하지만 기존에 있는 부동산 어플리케이션은 지역과 가격 정보 위주의 정보만을 제공하여 사용자의 선호하는 라이프 스타일에 맞는 지표들을 반영하지 못하는 아쉬운 점이 있다. 본 연구에서는 사용자가 원하는 모든 정보를 오픈 데이터를 활용하여 한눈에 볼 수 있게 함과 동시에 사용자의 나이와 성별, 라이프 스타일에 적합한 지역을 추천해 주는 시스템에 관해 설명한다.[2]

오픈데이터는 다양한 정보에 대한 접근성을 제공하며 데이터의 성질과 의미를 더욱 손쉽게 파악하여 빠른 의사결정을 할 수 있도록 한다. 따라서 지역별로 사용자의 라이프 스타일에 맞는 지표들을 오픈 데이터에서 가져와 지역별로 분석한 후 추천시스템 결과로 반영한다. 추천시스템이 지역별 오픈 데이터와 결합하여 개인화된 추천을 제공하고 지역주민들의 삶의 질을 향상하는데 기여하기 위해 이러한 연구를 하게 되었다.

2. 데이터 처리

2.1 데이터 수집

추천 시스템을 만들기 위해서 지역별 속성으로 병원, 학교, CCTV, 공원 등의 데이터를 활용하였다. 데이터 수집에 있어 크게 2가지 방법을 이용해서 진행했다, 첫 번째로는 공공포털에서 주거 환경을 결정하는 항목을 가지고 오는 것이다. 서울 지역 데이터의 경우 해당 포털과 함께 연동되어 있는 서울열린데이터광장 데이터도 많이 활용했다. 해당 포털의 경우 더 많은 양의 정보를 포함하고 있어서 활용하기에 용이했다. 또한 위치 데이터를 포함하고 있는 특정 데이터에 대해서는 LOCALDATA(지방행정인허가데이터개방) 포털에서 데이터를 수집하였다. 두 번째로 공공포털에서 제공하는 오픈 데이터가 없는 항목에 대해서는 온라인상의 데이터를 크롤링 작업을 통해 수집했다. 예를 들어 학교 데이터 항목의 경우 각 교육청 사이트를 이용해 학교명, 주소 등의 정보를 크롤링해서 데이터를 확보했으며, 다른 항목들에 대해서도 이와 같은 방법으로 데이터를 구축했다.

2.2 데이터 전처리

데이터를 활용하기 위해서는 우선적으로 데이터 상에 오류는 없는지, 누락된 데이터는 없는지, 바로 활용이 가능한 형태인지를 파악해야 한다. 오류 데이터의 경우 검색을 통해 정확한 정보를 파악한 후에 원래 있었던 데이터를 비교하여 틀린 부분을 찾아내고 수정하는 방식으로 진행했다. 누락된 항목을 가진 데이터는 경우에 따라 두 가지 방법으로 전처리를 진행했다. 주어진 데이터를 더 이상 사용할 수 없을 때는 해당 데이터를 제거하였고, 필요한 데이터에 대해서는 누락된 정보에 대해 크롤링 작업을 통해서 정보를 추가하는 작업을 수행하였다.

데이터베이스를 만드는 데 있어 대부분의 데이터 파일은 직접 적용할 수 없는 상태이다. 따라서 해당 데이터를 활용할 수 있도록 형태를 바꾸는 작업이 필수로 수반된다. 주거지역을 추천하기 위해서 자치구를 기준으로 모든 데이터를 구분하기로 하였다. 해당 자치구의 데이터 파악을 위해 주소 데이터를 가지고 있는 항목에 대해서는 주소에서 자치구를 추출하는 방식으로 전처리 작업을 수행하였으며, 특정 항목에 대해서 각 자치구별 데이터의 개수가 필요한 경우에는 자치구를 변수로 설정하여 해당 되는 데이터의 개수를 계산하는 방법을 수행했다. 데이터 정규화 단계에서는 min-max scaling 과정을 거쳐 아이템(지역)의 속성 정보를 item matrix 형태로 정리하였다. 이 과정에서 각 항목 별 모든 데이터에 반복적으로 이뤄지는 작업에 대해서는 Power Automate라는 자동화 프로그램을 활용해서 소요되는 시간을 줄일 수 있었다.

3. 추천 모델

본 연구에서는 주거 지역 추천을 위해 콘텐츠 기반 모델을 사용하였다. 콘텐츠 기반 모델은 아이템의 콘텐츠를 직접 분석하여 아이템과 사용자 선호도를 바탕으로 아이템을 추천하는 방식이다.[3] 주거 지역은 서울시 자치구 단위로 구분하였다.

3.1 가중치 도출

본 연구에서 제안하는 추천 시스템은 사용자의 라이프스타일 설문조사 결과, 연령, 성별, 가족형태 등의 기본정보를 기반으로 주거 지역 선택 시 고려 요소 분석 결과를 활용하여 도출해 낸 가중치를 함께 적용한다.

3.1.1 사용자 대상 설문

사용자 대상으로 사전 설문조사를 실시해 주거 지역 고려 요소별 우선순위와 기본 정보, 취미 등 라이프스타일에 대한 정보를 수집한다. 이는 user preference info, user info로 DB에 저장되며, 가중치 도출 시 활용된다.

표 1. 사용자 설문 결과 예시

사용자	성별	연령	가족 형태	취미 생활	교통	공원	성향
A	여성	20대	4인	테니스	지하철, 버스	매우 중요	활발
B	남성	30대	2인	축구	지하철	보통	활발
C	남성	40대	4인	공연	버스	보통	조용
D	여성	50대	3인	미술관	자차	중요	조용

표 2. 사용자 설문 결과 예시

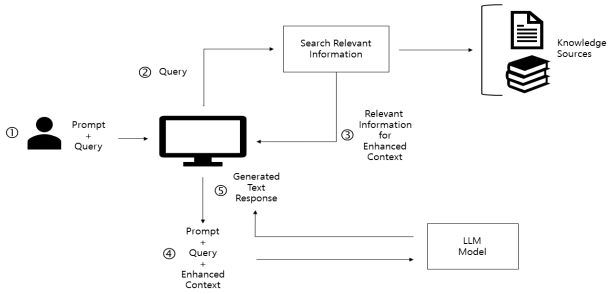
사용자/ 우선순위	안전	생활시설	교육	의료	환경	교통	기타
A	1	3	5	2	6	4	7
B	2	3	6	1	7	4	5
C	4	2	7	1	5	3	6
D	1	4	3	5	2	6	7

3.1.2 문헌 정보 분석

크롤링을 통해 수집한 뉴스 기사, 논문, 분석보고서 등의 자료는 생성형AI를 통해 분석하여 가중치 도출 시 활용한다. 제작 및 배포한 생성형 AI는 RAG(LLM)을 통해 문서를 분석하고 쿼리를 처리한다. 자체 수집한 데이터의 원본 파일에서 문서 데이터를 추출, 문서를 청크 단위로 잘라 벡터 형태로 임베딩한다. 쿼리 발생 시 생성형 AI의 LLM 모델이 질문을 벡터로 임베딩하고, 벡터DB에서 가장 일치하는 청크를 찾아 LLM에 전달하며, 최종적으로 LLM의 응답을 쿼리의 결과로 제공한다. 본 연구에서의 쿼리는 사용자의 성별, 연령 등 기본 정보에 따라 우선적으로 고려하는 주거지 선택 요소가 무엇인지 확인하는 작업에 사용된다. 예를 들어, 사용자가 20대 여성인 경우 거주지 선택 시 가장 중요하게 생각하는 요소 2개가 무엇인지 쿼리한다. 쿼리 결과 교통, 안전 순으로 중요하게 생각한다는 응답을 받는다. 응답 결과에 따라 20대 여성 사용자에게 사용자의 설문조사에 기반한 개인적인 선호도 외에도 문헌 정보 분석 결과를 활용하여 교통, 안전 요소에 추가적인 가중치를 부여하게 된다.

사용자 대상 설문 결과와 문헌 정보 분석을 통해 도출된 가중치의 정보를 바탕으로 user matrix를 제작한다.

그림1. 문헌 정보 분석 과정 도식화



3.2. 주거 지역 추천

데이터 전처리와 사용자 설문, 문헌 정보 분석 과정을 통해 제작한 user matrix와 item matrix를 최종적으로 내적하여 user-item matrix를 생성한다.

표 2. user-item matrix 예시

	강남구	강동구	광진구	노원구	도봉구
A	31.56	40.18	47.81	46.23	42.90
B	30.54	35.87	40.64	41.60	40.90
C	34.29	37.68	46.09	48.60	46.09
D	22.56	27.99	39.86	41.69	37.48

최종적으로 user-item matrix의 값이 가장 큰 3개의 지역을 사용자에게 추천한다. 사용자 A의 경우, 광진구, 노원구, 도봉구를 추천한다.

4. 결론

본 연구에서는 콘텐츠 기반 모델을 사용하여 사용자들의 취향에 부합하는 주거 지역을 추천하는 시스템을 구현하였다. 이를 통해 기존의 한정된 부동산 정보 제공에서 벗어나 사용자들은 자신의 다양한 요구사항을 효율적으로 만족시킬 수 있게 되었다.

그러나 연구 과정에서 콘텐츠 기반 모델의 단독 사용으로 사용자가 본인의 관심사에 맞춰 필터링 된 정보 안에 갇히는 '필터 버블' 현상이 나타났다. 이러한 문제점을 해결하기 위해 사용자 기반 협력 필터링 모델이 필요하다. 따라서 향후 더 정밀하고 만족도 높은 추천 결과를 도출하기 위해 콘텐츠 기반 모델과 사용자 기반 협력 필터링 모델을 함께 사용한 하이브리드 모델을 기반으로 연구를 진행할 예정이다.

참 고 문 헌

[1] 임다혜 and 권영상. (2022). 랜덤 포레스트 모형을 활용한 청년들의 주택 유형별주거환경 만족도 영향 요인 중요도 분석- 주거실태조사(2020) 데이터를 활용하여. 도시설계, 23(6), 103-122

[2] 박선희, 김정호, 유현배(2017), 빅 데이터 가시화 기술을 적용한 공공데이터 콘텐츠 구현 - Map가시화 기법, 한국디지털콘텐츠학회

[3] 김주연, 김희찬, 강우진 and 홍진혁, 2022. 컨텐츠 기반 필터링을 활용한 공예품 추천 시스템 개발. 한국정보과학회 학술발표논문집. 개최지, 개최날짜. 한국정보과학회.