

웹 크롤링을 통한 개인 맞춤형 정보제공 애플리케이션

¹김주현, ¹최정은, ¹신우경, ¹박민준, ²김태국

¹부경대학교 컴퓨터·인공지능공학부 학부생

²부경대학교 컴퓨터·인공지능공학부 교수

¹juju6065@pukyong.ac.kr, ¹cju4103@pukyong.ac.kr, ¹sinariari@pukyong.ac.kr,

¹minjun1997@pukyong.ac.kr, ²king@pknu.ac.kr

Information-providing Application Based on Web Crawling

Ju-Hyeon Kim¹, Jeong-Eun Choi¹, U-Gyeong Shin¹, Min-Jun Piao¹,
Tae-Kook Kim²

School of Computer and Artificial Intelligence Engineering,
Pukyong National University

요 약

본 논문에서는 웹 크롤링을 통한 개인 맞춤형 정보제공 애플리케이션에 관해 연구하였다. 본 서비스는 Java의 Jsoup 라이브러리를 이용해서 웹 크롤링(Web Crawling)한 데이터를 MySQL에 저장한다. 이를 통해 사용자가 지정한 키워드를 필터링하여 사용자에게 정보를 제공한다. 예를 들어 사용자가 지정한 키워드 관련 공지 사항이 업데이트되면 구현한 앱 내에서 확인 가능하며, KakaoTalk 알림톡을 통해서도 업데이트된 정보를 실시간으로 전송받는 서비스를 구현하였다.

1. 서론

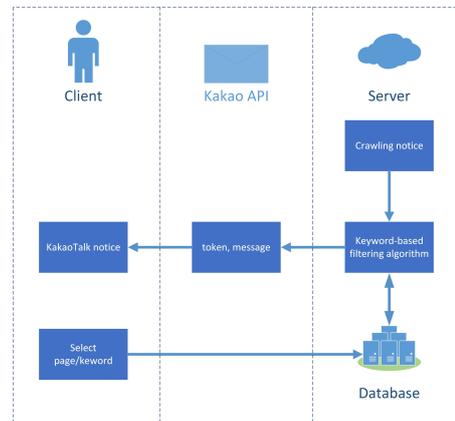
최근 ‘정보 싸움’[1]이라는 말이 있을 정도로 많은 정보가 쏟아져 나오고 있으며, 이러한 정보는 개인별로 필요한 정보가 다르다. 공공기관인 학교가 제공하는 정보에는 학사일정, 장학, 취·창업 등 여러 웹 페이지로 세분화되어 있다. 원하는 정보를 얻기 위해 여러 페이지를 접속해야 하고, 시간도 많이 소요되는 문제가 있다[2]. 이러한 문제를 해결하기 위해 웹 크롤링(Web Crawling)을 통해 사용자가 지정한 정보를 리스트화하여 볼 수 있는 앱을 설계하고 구현하였다.

2. 웹 크롤링 기반의 정보제공 앱 설계 및 구현

‘웹 크롤링 기반의 개인 맞춤형 정보제공 앱 (Dingle)’은 사용자가 원하는 키워드로 정보를 필터링(filtering)하여 실시간으로 알림을 제공한다.

(그림 1)은 제안한 ‘Dingle’ 앱 서비스의 흐름도이다. 사용자는 제공 받고자 하는 정보의 웹 페이지 및 키워드를 선택한다. 서버는 이를 기반으로, 각 홈페이지의 공지사항을 크롤링하고 키워드 기반 필터링 알고리즘을 거쳐 수집한 정보를 데이터베이스에 저장한다. 이 과정에서 사용자가 설정한 페이지 및

키워드 관련 정보가 있으면, 해당 사용자에게 카카오톡 알림톡을 통해 정보를 알려준다.



(그림 1) 제안된 웹 크롤링 기반의 개인 맞춤형 정보제공 앱 서비스의 흐름도

<표 1>은 소프트웨어 생명 주기에 맞춰 사용자에게 요구하는 기능들이 무엇인지 나타내는 요구사항 정의서이다. 앱에서 필수 기능을 중요도 5로 설정하고 우선적으로 구현할 필요가 없는 기능들을 중요도 1로 두어 우선순위가 높은 순서대로 구현을 진행하였다. 홈 화면인 공지 리스트에서는 웹 크롤링을 통

해 얻은 정보 중 제목, 내용, 이미지를 보여주며 앱의 UX를 위해 제목과 내용을 각각 13글자, 17글자로 설정하여 구현하였다.

<표 1> 요구사항 정의서

화면명	요구사항 명	요구사항 내용	중요도
회원 가입	카카오 회원 가입	- 카카오 회원 가입을 진행한다. - 카카오 알림 톡을 이용하여 카카오톡 서비스에 대한 동의화면을 보여준다.	4
로그인	로그인 실행	카카오 로그인 진행한다.	3
	카카오톡 알림	- 로그인한 카카오톡으로 공지사항의 업데이트를 알린다. · 제목, 내용, 이미지, 링크 포함	3
홈 화면	공지사항 리스트	- 선택한 키워드를 기준으로 필터링 된 공지사항을 보여준다. · 제목 · 내용 · 사진 *사진이 없는 공지사항일 경우 기본 사진으로 대체	5
JJIM	찜한 공지사항 리스트화면	- 사용자가 제공 받은 정보 중 관심 있는 정보를 모아 보여 준다.	2
커스텀 초기 설정	커스텀화면	- 알림을 받고자 하는 홈페이지와 카테고리 설정한다. - 홈페이지와 카테고리 모두 중복 선택이 가능하다.	3

앱의 개발환경은 <표 2>와 같다. Java를 이용한 Spring Boot 환경에서 웹 크롤링을 구현하였으며, 결과물을 보여주는 앱은 Dart언어를 사용하는 Flutter를 이용한다. 데이터는 MySQL을 통해 관리되며, AWS 클라우드를 활용하여 서버 인프라를 구축 후 배포한다.

<표 2> 'Dingle' 앱 서비스 개발 환경

구분	상세내용
시스템 개발 환경	통합 개발 환경
	개발 언어

IntelliJ(Spring Boot), Android Studio(Flutter), MySQL, Figma(UI/UX), AWS
Java, Flutter



(그림 2) 'Dingle' 앱 UI

(그림 2)는 Figma로 그린 앱의 주요 화면의 UI이

다. 회원 가입 및 로그인 이후에 홈페이지와 키워드 커스텀 화면이 각각 나온 후 선택한 키워드를 기반으로 필터링된 공지사항이 나타난다.

(그림 3)은 (그림 2)에서 구성한 화면을 기반으로 실제 메인 홈페이지와 학과 홈페이지를 웹 크롤링한 결과를 나타낸 실제 홈 화면이다.



(그림 3) 홈 화면

웹 크롤링은 Java의 Jsoup 라이브러리를 통해 이루어진다. Jsoup의 Elements 클래스의 select() 함수를 통해 CSS 선택자를 인수로 주어 요소를 반환하고 반환된 요소는 text() 함수를 통해 텍스트화하여 MySQL에 저장한다. 이를 통해 홈페이지의 공지사항 목록을 추출 후, 각 항목에 대해 제목, 내용, 링크, 이미지를 가져온다.

4. 결론

본 연구에서는 Jsoup 라이브러리를 통한 웹 크롤링 기술을 이용해 정보 수집 효율을 높이는 앱 'Dingle'을 제안하였다. 또한 '찜' 기능을 통해 관심 정보를 손쉽게 확인하도록 구현하였다. 공공기관의 대상을 부경대학교로 설정하였으나 범위를 특정 지역 또는 계층으로 대상을 넓힌다면 더 다양하고 많은 정보를 얻을 수 있는 효과를 기대한다.

참고문헌

- [1] 동아일보, <https://weekly.donga.com/coverstory/3/01/11/74666>
- [2] 김주현, 최정은, 신우경, 박민준, 김태국, “실시간 웹 크롤링 기반 개인 맞춤형 알림 서비스에 관한 연구,” 2023년 한국멀티미디어학회 춘계학술발표대회, 제26권, 제1호, pp.86-88, 2023.