웹 크롤링을 활용한 셰어하우스 데이터 수집 시스템 개발

(Development of Sharehouse Data Collection System using Web Crawling)

요 약

최근 셰어하우스 시장의 성장으로 인하여 많은 셰어하우스 매물들이 생겨나고 있다. 이러한 상황에 상응하여 셰어하우스 사업을 시작하려는 사람들의 수가 증가하고 있다. 하지만 사업을 시작하기 전에 셰어하우스 시장 현황 및 입지에 대한 정보는 필수적일 수밖에 없다. 그러나 일반인이 일일이 웹 페이지에 방문하여 이를 수집하기에는 시간적, 경제적 어려움이 따른다. 따라서 문제점을 해결하기 위하여 이 논문에서는 크롤러를 활용하여 셰어하우스 플랫폼으로부터 데이터를 수집하는 시스템을 개발하였다.

ABSTRACT

Due to the recent growth of the share house market, many share house listings have been created. In response to this situation, the number of people who want to start a share house business is increasing. However, before starting a business, information on the current status and location of the share house market is essential. Nevertheless, it takes time and economic difficulties for the general public to visit the web page and collect it. Therefore, in order to solve the problem, in this paper, I developed a system that collects data using web crawling from a share house platform.

키워드 : 크롤링, 빅데이터, 자동화

Keywords : Crawling, Big Data, Automation

Ⅰ. 서 론

통계청에 따르면 1995년 12.7%에 그쳤던 1인 가구는 2000년 15.5%, 2005년 20.0%, 2010년 23.9%로 지속적인 증가를 보이고 있다. 2015년에는 27.2%인 520만 3000가구로 우리나라에서 가장 흔한 가구 형태가 되었다.[1] 청년 1인 가구의 증가와 함께 이러한 양적·물적 문제들을 해결하기 위한 주거대안 중 한 가지가 바로 셰어하우스이다. '셰어하우스'란 다수가 한 집에 살면서 개인공간은 확보하되 거실, 주방, 화장실 등과 같은 공간이나 설비는 공유함으로써 주거공간을 보다 효율적으로 사용하도록 한 주택유형이다.[2] 이러한 상황에 상응하여 셰어하우스 사업을 시작하려는 사람들의 수가 증가하고 있는 추세이다. 하지만 사업을 시작하기에 앞서서 셰어하우스 시장 및 입지에 대한 정보 파악은 필수적일 수밖에 없다. 그러나 일반인이 일일이 이를 수집하기에는 시간적, 경제적 어려움이 따른다. 따라서 문제점을 해결하기 위하여 크롤러를 활용하여 셰어하우스 플랫폼으로부터 데이터를 수집하는 시스템을 개발하였다.

본 논문에서는 셰어하우스 데이터를 수집 및 처리 과정에 대해 상세히 설명한다.

Ⅱ. 본 론

1. 크롤링 시스템 설계

아래 그림 1은 이 논문에서 소개하는 데이터 수집 시스템의 구조이다.[3]

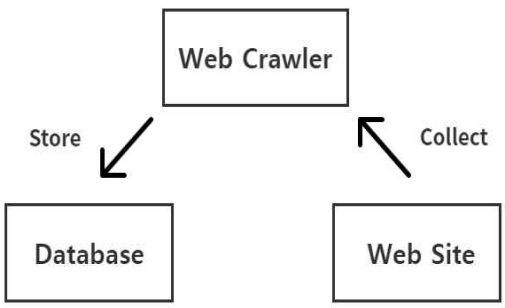

Fig. 1. System Architecture

먼저 웹사이트로부터 데이터를 수집한 후, 이를 데이터베이스에 저장하는 과정을 거친다.

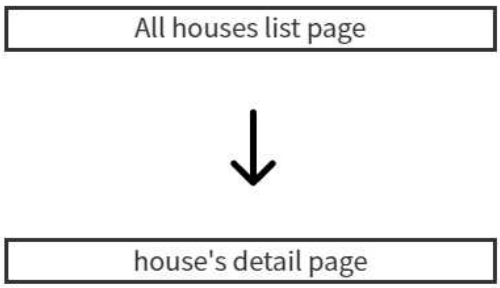

Fig. 2. Data Collection Path

전체 셰어하우스 매물 리스트가 나오는 링크를 첫 번째 루트로 잡아 위의 그림 2처럼 이동하면서 텍스트 데이터들을 리스트에 key-value 구조로 저장해두었다가 해당 작업이 모두 완료되면 마지막에 데이터베이스 테이블에 저장하는 방식으로 데이터 수집을 진행한다.

먼저, 첫 번째 루트로 잡은 전체 셰어하우스 매물 리스트 페이지에 접속해서 각각의 셰어하우스 매물의 하이퍼링크 태그를 찾아내어 리스트에 저장한다. 해당 작업이 완료되면 리스트에 있는 셰어하우스 매물 페이지들을 순차적으로 방문한다.[4]

두 번째 루트인 셰어하우스 매물 페이지로 넘어오면 텍스트 데이터를 추출한다. 텍스트 데이터는 매물명, 성별전용, 보증금, 월세, 면적, 인실 등으로 이루어져있다. 각각의 houses, rooms, beds 테이블에 key-value 형태로 저장한다.

2. 데이터베이스 설계

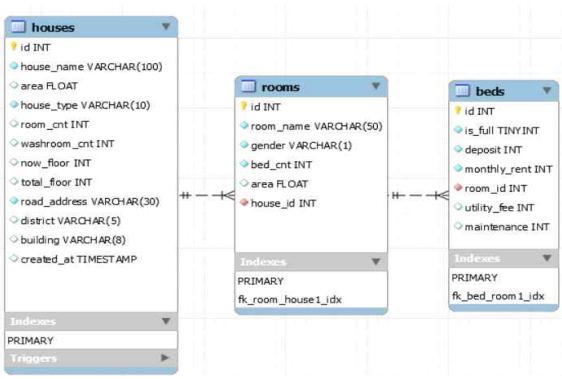

Fig. 3. Database Schema

데이터베이스 테이블은 houses, rooms, beds 총 3개의 테이블로 구성되어있다. 위의 그림 3과 같이 텍스트 데이터를 수집해서 저장한다. 하나의 house에 n개의 room(1:N)이 하나의 room에 n개의 bed(1:N)가 매핑되도록 설계하였다.

Ⅲ. 실 험

크롤러를 실행하면서 발생했던 에러를 케이스별로 아래 표 1과 같이 나타내었다.

Table 1. System Test

Case	Result
No Element	NoSuchElementException: Message: no such element: Unable to locate element
room_name length is short	"Data too long for column 'room_name' at row X"

본 논문에서 실행한 코드는 파이썬으로 작성하였으며 자동화를 위해 selenium을, 데이터 파싱을 위해서 beautifulsoup을 사용하였다.[5][6]

먼저, 첫 번째 케이스의 경우 Xpath 경로를 이용해서 DOM 구조에서 특정 div 안의 텍스트 데이터를 가져오도록 설계하였는데, 각 매물의 페이지마다 div의 구조가 다르기 때문에 설정해 놓은 div를 찾을 수 없다고 에러가 발생했었다. 따라서, Xpath와 tag_name을 적절히 이용해서 문제점을 해결하였다.

두 번째 케이스의 경우 Database Schema에서 room_name의 length를 30으로 설정했었는데, 소수의 room_name이 이를 넘어서 계속해서 rooms와 beds의 데이터가 수집되지 않는 문제가 발생했었다. 따라서, room_name의 길이를 더 크게 설정함으로서 문제점을 해결하였다.

Ⅳ. 결 론

본 논문의 데이터 수집 시스템을 이용하여 여러 셰어하우스 플랫폼의 데이터를 분석할 수 있다. 또한, 자치구별 매물 수 순위와 공실률 등 빅데이터 분석을 통해 셰어하우스 현황 파악을 할 수 있다. 보다 더 나은 데이터 수집을 위해 월마다 업데이트 가능하고 더 많은 셰어하우스 플랫폼으로부터 데이터를 수집해서 추가할 수 있도록 데이터를 수집하고 데이터베이스에 저장하는 기능을 분리시켰다.

하지만 크롤링을 보안상 막아놓은 사이트로부터는 데이터를 얻을 수 없는 제약사항이 존재한다. 또한, 수집하는 데이터의 정확성과 신뢰성이 우선시되어야 데이터 수집 시스템을 사용하는 것이 의미가 있다.

*본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음(2016-0-00017)

References

[1] Population and Housing census, (2010, 2015), Statistics Korea

[2] Oh, Jung, Choi Jung-Min, “A study on the perceptions and demand characteristics of share house in Korea”, Journal of the Korean Housing Association 28(2), 73-78, Nov. 2013.

[3] Chaeeun Lee, Jinwook Jang, “Development of Social Data Collection System using Web Crawling”, Journal of the Korea Information Science Society, 1787-1789, Jun. 2016.

[4] Jae-Ho Shin, Tae-Woo Kim, Chul-Yun Kim, “XPath based crawling method for specific online shopping mall”, Journal of the Korea Information Science Society, 240-242, Dec. 2014.

[5] Jong-Hwa Lee, “Building an SNS Crawling System Using Python”, Journal of the Korea Industrial Information Systems Research 23(5), 61-76, Oct. 2018.

[6] Li Seung, Sujin Yun, Young Woon Woo, “Crawling Methods for Web Data of Various Formats Using Python”, Journal of the Korea Institute of Information and Communication Engineering 23(2), 343-346, Oct. 2019.