자동으로 웹크롤링된 데이터를 이용한 위치정보 시스템

임승환, 조형석, 이성욱
한국교통대학교 컴퓨터정보공학과
e-mail:leesw@ut.ac.kr

A Location Information System using Automatically Web-Crawled Data

Seung Hwan Lim, hyeong Seok Jo, Songwook Lee
Dept of Computer Science & Information Engineering,
Korea National University of Transportation

요 약

우리는 웹상에 존재하는 여러 포털사이트의 위치정보와 그에 대한 데이터들을 수집하고 수집된 데이터들로부터 각각의 정보를 분류하여 웹상에 표시하는 페이지를 만들었다. 그리고 사람들이 자유롭게 위치정보에 대한 내용을 추가하고 공유할 수 있도록 하였고, 검색기능을 활용하여 해당위치를 빠르게 찾을 수 있도록 하였다. 그리고 각각의 데이터들에 대해 자유롭게 댓글을 달 수 있게 하여 사람들이 해당 장소에 대해 자유롭게 의견을 나눌 수 있도록 하였다.

1. 서론

스마트폰의 보급 및 SNS의 발달과 더불어 포털사이트의 블로그 및 카페가 활발히 운영되면서 사용자들은 많은 양의 정보를 빠른 속도로 찾을 수 있고 얻게 되었다. 뿐만 아니라 사용자들이 원하는 장소에 대한 정보를 쉽게 얻을 수 있다. 하지만 그에 비한 단점도 있다. 정보가 무수히 생산되는 반면, 찾고자 하는 정보의 존재 유무에 대한 문제가 나타날 수 있다. 그리고 검색을 통해 수집된 정보가 믿고 사용할 수 있는 신뢰성에 대한 문제도 포함한다.

본 논문에서는 사용자에게 여러 사이트에서 수집된 정보를 가공하고, 또한 여러 사용자의 참여를 통해 시스템에서 제공되는 정보에 신뢰성을 높여주고자 시스템을 설계 및 구현하였다. 본 논문에서는 사용자들이 검색에 따른 시간 투자를 감소시키고, 사용자들간의 정보공유를 통한 정보의 신뢰성을 향상시키기 위한 방법에 대해 제안한다.

본 논문은 2장에서는 시스템의 설계에 대해 설명한다. 3장에서는 설계에 따른 시스템의 구현에 대해 설명한다. 끝으로 4장에서는 본 시스템의 결론을 맺는다.

2. 시스템 설계

본 시스템은 아래 그림 1과 같은 구조를 가진다.

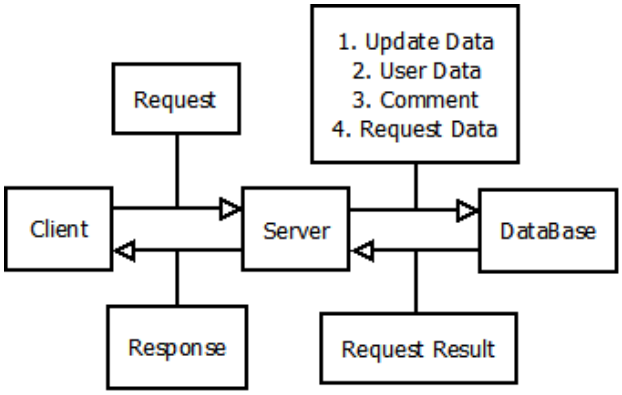

그림 1. 시스템 구성도.

본 시스템은 클라이언트-서버 구조로 이루어졌다. 서버에서는 클라이언트로부터 요청받은 정보를 처리하여 제공해주고, 일정시각에 데이터를 수집하여 처리, 가공해주는 역할을 한다. 클라이언트는 사용자가 접속한 후, 제공되는 모듈들을 이용하여, 서버에 요청을 전송하게 되면, 서버에서 처리된 정보들을 전송받아 제공해주는 역할을 한다.

본 시스템에서 사용된 개발환경은 아래 표에 나타나있다.

O/S	Ubuntu(16.04.1)
Language	PHP, Python, Java
Tools	eclipse, Spyder
API	Daum API(Address, Local) Naver API(기계번역, Search) 도로명주소 개발자센터API
DataBase	MySQL(5.7)
WebServer	Apache(2.4)

표 1. 서버 개발환경

Language	HTML, CSS
FrameWork	JQuery, BootStrap
Tools	DreamWeaver
API	Google API(Map)

표 2. 클라이언트 개발환경

3. 웹크롤링 및 데이터베이스 구축

본 시스템에서는 포털사이트를 통해 수집된 위치정보를 데이터베이스에 구축하였다. 초기 데이터를 구축하기 위해서 부동산개발안내도[1]의 개발에 사용되었던 도로명주소를 참조하였다. 그림 2는 사용된 도로명 주소 데이터의 형태를 보여준다.

그림 2. 축적된 주소 데이터

수집된 데이터에서 주소정보만을 추출한 후, 추출된 주소를 이용하여 다음 로컬API[2]를 이용하여 해당주소에 맞는 장소정보를 JSON형태로 추출하였다. 그림 3은 이를 나타낸 것이다.

```
{
  "channel": {
    "item": [
      {
        "related_place_count": 0,
        "zipcode": "137855",
        "related_place": "",
        "distance": "",
        "direction": "",
        "placeUrl": "http://place.map.daum.net/653245473",
        "categoryCode": "5 68 18086 22480",
        "category": "가정,생활 > 문구,사무용품 > 디자인문구 > 카카오프렌즈",
        "newAddress": "서울 서초구 강남대로 429",
        "title": "카카오프렌즈 강남플래그십스토어",
        "id": "653245473",
        "phone": "02-6494-1100",
        "imageUrl": "http://t1.daumcdn.net/place/C79325E3CD2E4D5CA762E6907FC8207A",
        "address": "서울 서초구 서초동 1305-7",
```

그림 3. LocalAPI를 이용한 장소 정보.

수집된 장소정보를 Selenium PhantomJS[3]를 이용하여, 사람들이 주로 쓰는 포털사이트인 네이버와 다음에서 해당장소명에 대한 정보를 수집하였다. 그림 4는 크롤링의 대상이 되는 네이버 및 다음지도의 내용을 보여준다.

그림 4. 네이버 및 다음지도

장소의 사진 정보의 경우도 Selenium PhantomJS를 이용하여 구글에서 검색하여 수집하였다. 해당 장소가 저장될 테이블을 결정하기 위해서는 로컬API를 이용하여 수집된 정보에서 “Category”를 네이버 기계번역API[4]를 이용하여 Category를 영문으로 변경한 내용을 데이터베이스 테이블명으로 사용하였다. 포털사이트에서 추출된 정보 및 로컬API를 통해 수집된 정보를 통합하여, 형식에 맞게 데이터베이스에 저장하였다.

4. 시스템 구현

4-1 위치정보 표시

홈페이지가 처음 시작될 때, 그림 5와 같이 데이터베이스에 저장되어 있는 정보들이 가지고 있는 GPS 좌표정보를 이용하여 구글맵 AP에서 제공하는 Marker객체를 이용하여 전부 표시한다.

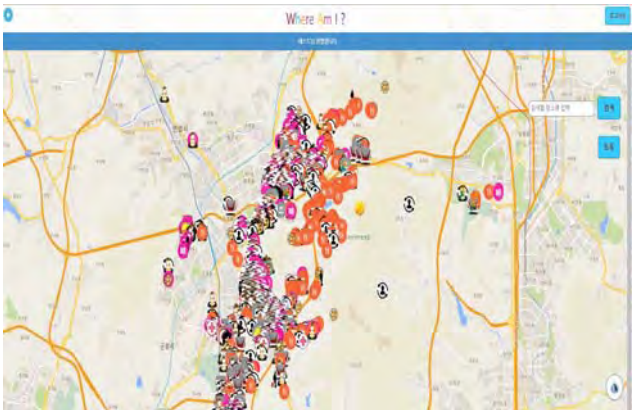

그림 5. 위치정보 표시

4-2 위치정보 분류

왼쪽 상단의 툴팁 버튼을 누르게 되면 분류에 따른 정보

가 지도상에 표시되도록 해주는 기능이다. 그림 6은 “가정”이라는 분류를 가진 정보들만 지도상에 표시해 준 예를 보여준다.

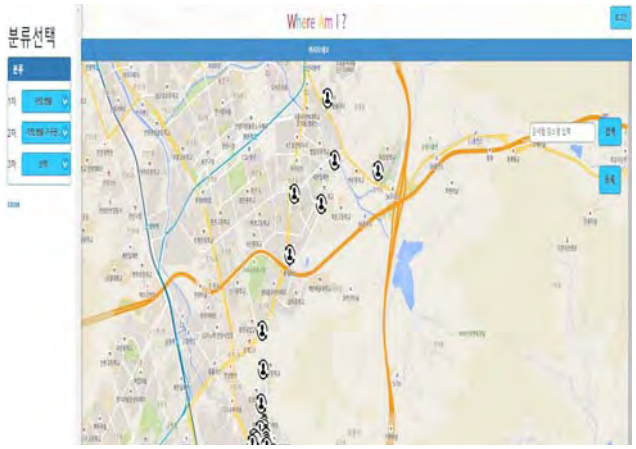

그림 6. 분류를 통한 표시

4-3 위치정보 등록기능

시스템에 등록된 사용자들이 지도상의 장소의 위치를 클릭 후, 등록 버튼을 누르게 되면, 해당 장소에 대한 정보를 등록하게 해주는 기능이다. 등록되는 정보로는, 장소의 분류, 장소명, 주소, 전화번호, 부가정보, 장소대표사진, 장소의 정보, 링크, 등록자의 한마디를 입력하게 한다. 그림 7은 등록화면을 보여준다.

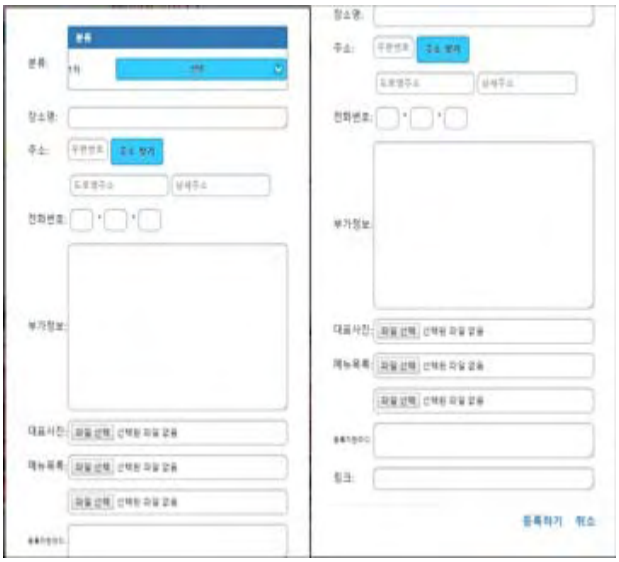

그림 7. 위치정보 등록

4-4 위치 정보에 대한 댓글

지도상에 표시된 마커들을 누르게 되면, 사용자들은 해당 장소에 대한 상세한 정보들을 볼 수 있게 된다. 시스템 사용자들은 해당 장소에 대한 후기 및 평가를 댓글로 남길 수 있게 했다. 댓글을 남기게 됨으로써 다른 사람들과 장소에 대한 정보를 공유할 수 있게 한다. 그림 8은 장소정보에 대한 댓글의 예시를 보여준다.

그림 8. 위치정보에 대한 세부정보 출력.

4-5. 위치정보 자동 추가 및 갱신

장소에 대한 정보가 하루가 다르게 변할 수 있기 때문에 일정시각에 기존 데이터베이스에 저장되어있는 전체정보가 최신정보를 유지할 수 있도록 한다. 그림 9는 기존 정보의 업데이트에 쓰이는 링크필드와 포털정보의 최근에 업데이트된 날짜를 보여준다.

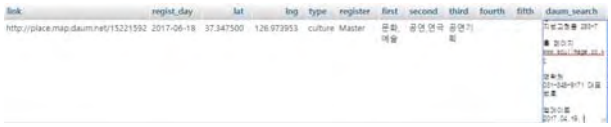

그림 9. 업데이트에 필요한 데이터

데이터베이스에 저장된 링크를 Selenium PhantomJS를 통해 데이터를 수집하게 된다. 수집된 데이터의 날짜가 기존 업데이트 날짜보다 늦는다면, 해당 정보를 업데이트 한다. 장소정보는 일(日)에 따라 변동됨으로, 도로명 주소 API를 이용하여 변동자료를 수집한다. 그림10은 수집된 일변동 자료를 보여준다.

그림 10. 일변동 자료

수집된 일변동 자료에 포함돼 있는 변동코드를 통해 해당 장소의 정보를 시스템이 자동적으로 데이터베이스에

추가, 삭제, 갱신할 수 있도록 한다. 데이터의 자동갱신은 해당 시스템의 운영체제에 포함돼 있는 Crontab[4]기능을 이용하여, 작업에 필요한 명령어들을 배치파일로 만들어 정해진 시간에 실행한다. 그림 12는 데이터베이스의 날짜가 갱신되어있는 것을 보여준다.

그림 11. 갱신날짜가 표시되어있는 세부정보창 윗부분

4-6. 위치정보 검색 기능

사용자들이 검색창에 검색어를 입력하게 되면, 해당 검색어가 장소명에 포함돼 있는 정보들을 그림 11과 같이 표시한다. 이를 이용하면 보다 사용자가 원하는 장소를 편하게 찾을 수 있게 한다.

그림 12. 위치정보 검색

4-7. 현재위치 표시 기능

사용자가 현재 자신의 위치를 파악하기 위해, HTML5의 Geolocation API를 이용하여 자신의 위치를 그림 12와 같이 나타내게 한다[6].

그림 13. 사용자의 현재위치 표시

5. 결론

제안된 시스템을 이용하면 사용자들은 보다 신뢰성 있는 위치 정보를 얻을 수 있다. 위치정보 자동 등록을 통해 사용자들이 기존 포털사이트에서 제공하는 정보에 비해 더 많은 정보를 얻을 수 있게 되었다. 그리고 정보표시 기능을 통해 해당 장소에 대한 사용자들의 평가, 데이터베이스에 등록되어 있는 세부적인 정보 등을 쉽게 파악할 수 있게 되었다. 그리고 댓글기능을 통해 사용자들이 정보공유를 할 수 있게 됨으로써 정보의 질적인 부분과 양적인 부분에서 향상되도록 기여하였다.

하지만, 본 시스템에서 제공되는 사진 정보에 있어서는 검색결과를 토대로 나온 사진의 해상도가 일정하지 않아 화질이 흐려지게 되었다. 향후 우리는 수집된 사진의 이용을 확대하기 위해서 보다 좋은 화질의 사진을 제공하여야 하며 이를 위해서는 영상처리 기법 등을 사용하여 화질의 개선 등이 필요하다. 그 외, 장소 검색결과를 통해 나온 상위 5개의 사진을 사용함에 있어서, 장소 정보와 불일치하는 사진을 수집하게 되는 것을 방지하기 위해, 해당 장소에 대한 정확한 사진을 수집하고 학습데이터로 구축한 후, 기계학습 기법을 통해 자동 수집 사진들 중 해당 장소와 일치하는 사진을 보다 정확하게 수집할 수 있게 할 것이다.

5. 참고문헌

[1] 양성철. (2013). 도로명주소기본도를 이용한 부동산개발안내도 구축 방안 연구. 한국지형공간정보학회지, 21(3), 47-54.

[2] “Daum Local API - 키워드로 장소검색” https://developers.daum.net/services/apis/local/v1/search/keyword.format 2017-04-01

[3] “Selenium” http://selenium-python.readthedocs.io/ 2017-04-10

[4] “기계번역API” https://developers.naver.com/docs/labs/translator/ 2017-04-21

[5] “Crontab” http://www.adminschoice.com/crontab-quick-reference 2017-05-30

[6] hollawer (2015). “지오로케이션 사용하기” https://developer.mozilla.org/ko/docs/WebAPI/Using_geolocation 2017-05-21