주제 중심의 크롤링 방법

이원섭*, 신재문**, 임지호***, 김단이****, 조경일*****,
*건국 글로컬대학교 컴퓨터공학과,
**인천대학교 컴퓨터공학과,
***한양대학교 식품경영학과,
****계원디자인예술대학교 디지털미디어학과,
*****수원대학교 행정학과,
e-mail:beginner74@naver.com

Subject oriented crawlling method

Won-Seob Lee*, Jea-Moon Shin**, Ji-Ho Lim***, Dan-I Kim****, Kyung-Ill
*Dept of Computer Engineering, Konkuk University
**Dept of Computer Engineering, Incheon University
***Dept of Food Management, Hanyang University
****Dept of Digital Mediat, Kaywondaehangno University
*****Dept of Public Administration, University of Suwon

요 약

크롤링을 통해서 주제와 관련된 데이터를 판단한다. 주제와 관련성을 위해서 가중치를 사용하고 정확도와 크롤링 속도를 위해 응집력과 중복성 검사 등을 사용한다.

1. 서론

최근 자료가 다양하고 복잡해지면서 사람들은 자신이 원하는 정보를 정확하게 찾는데 어려움을 겪고 있다. 또한, 많은 사람들은 데이터나 정보를 이용하여 다양한 사업을 하거나 연구를 하고 있지만, 대부분의 사람들은 직접 데이터를 수집하는 경우가 많아 인력과 시간이 낭비 되고 있다. 이러한 문제를 해결하고자 검색의 도움을 주는 많은 소프트웨어(크롤링)들이 개발되었다. 대표적인 예로, 네이버의 검색엔진이 있다. 네이버의 크롤링 방법의 경우 해당 키워드가 포함되어 있는 정보를 모두 수집해 추출한다. 따라서 연관성이 적은 데이터 까지 추출되어 보여 진다. 하지만, 이러한 크롤링 방식의 경우 정확도가 떨어지고 자료가 방대하다는 문제가 있다. 조사해본 결과, 크롤링과 관련된 모든 API는 네이버와 같은 방식을 사용하고 있음을 발견했다.

이러한 문제점을 해결하기 위해, 우리는 주제와 관련된 데이터를 수집하는 과정에서 주제와 관련된 정보 중 유의미한 데이터만을 판단하여 추출하는 크롤링 방식을 개발하였다.

2. 주제 관련 크롤링 방법

크롤링 방식은 여러 가지 방법은 있으나 기본적으로 탐색, 분류, 저장 순으로 나열할 수 있다. 지금 설명하고자 하는 방법도 탐색, 분류, 저장 순으로 만들었다.

2.1 자동 크롤링

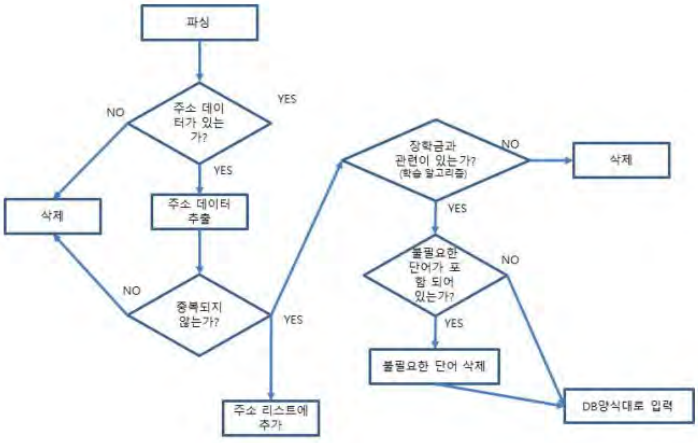
수집은 크롤링에서 가장 기본인 되는 기능이다. 자동으로 수집이 되지 않는다면 크롤링을 하는 의미가 없다.

크롤링의 방법에는 다양한 방법이 있지만 주제에 관련 크롤링을 하기 때문에 파싱으로 검색된 주소를 기반으로 한 자동 크롤링을 개발하였다.

첫 번째 크롤링 페이지의 경우 사용자가 원하는 페이지를 지정하면 해당 페이지를 출력해서 보여주고 만약, 지정을 하지 않을 경우 기본적으로 설정된 페이지를 크롤링하는 방식이다. 예를 들어 장학금이면 교육청과 같이 장학금과 관련된 페이지로 이동할 수 있는 링크가 많은 페이지를 검색리스트에 추가한다. 검색리스트에 추가된 주소를 파싱하여 파싱된 데이터에서 주소를 추출하여 다시 주소리스트에 추가한다. 이 과정을 반복문으로 묶어준다면 두 번째 파싱이 시작되기 전에 첫 번 째 파싱에서 찾은 주소가 추가된다. 이 반복은 파싱된 데이터에서 더 이상 주소 관련 데이터가 없을 때까지 무안히 반복된다. 즉, 대부분의 데이터를 검색하게 되는 것이다. 자동탐색에서 똑같은 주소가 계속 탐색되어 무한히 똑같은 페이지를 탐색하는 문제가 발생 할 가능성이 있다. 이러한 무한루프를 방지하기 위해서 중복성 검사를 추가한다. 중복성 검사를 통해서 탐색했던 페이지라면 탐색하지 않음으로서 무한루프를 방지한다. 또한 중복성 검사에서 중복된 페이지이외에 탐색하지 말아야할 페이지 또한 검사하여 파싱하는 시간을 줄여 프로그래밍 속도를 향상시킨다.

자동 크롤링에 학습기능이 들어가 있다. 자동크롤링 중

중복성 검사에 사용되는 데이터를 이전에 사용되었던 자료를 불러와서 사용하고 그 결과를 다시 데이터에 저장한다. 결과적으로 자동검사를 많이 할수록 탐색해야 할 페이지와 탐색하지 말아야 할 페이지를 분류하는 페이지 목록을 업데이트해 나아간다.



<사진 1> 자동탐색 알고리즘

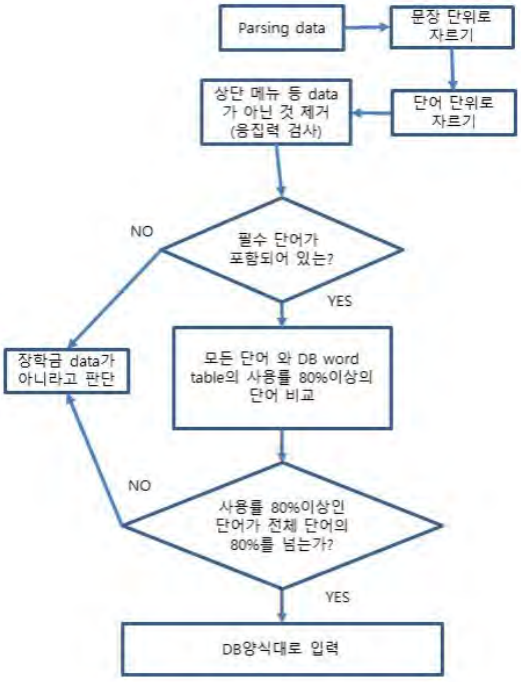
2.2 분류

분류는 주제에 대한 내용인지 판단하는 가장 중요한 단계이다. 분류에는 가중치와 응집도가 사용된다.

주제와의 관련성을 확인하기 위해서 파싱된 단어의 각각의 가중치가 82% 이상인 단어가 전체적으로 79.3%이상인 페이지를 주제와과 관련된 페이지라고 판단한다.

가중치의 수치는 학습기능과 초기입력 데이터에 의하여 결정된다. 초기입력데이터란 주제와 관련된 데이터를 미리 입력하여 단어의 가중치를 판단한다. 단어의 가중치는 $\frac{\text{단어가사용된페이지갯수}}{\text{전체파싱한횟수}}$ 이다.

분류에서 사용된 데이터는 모두 학습기능을 통해서 정확도를 높인다. 분류에서의 학습기능은 파싱할 때마다 전체 파싱한 횟수를 증가시키고 파싱했을 때 사용된 단어의 모든 횟수를 1회 증가 시킴으로서 구현하였다. 그리고 새로운 단어가 발견되었을 때는 사용자에게 메시지를 띄워서 알려준다.

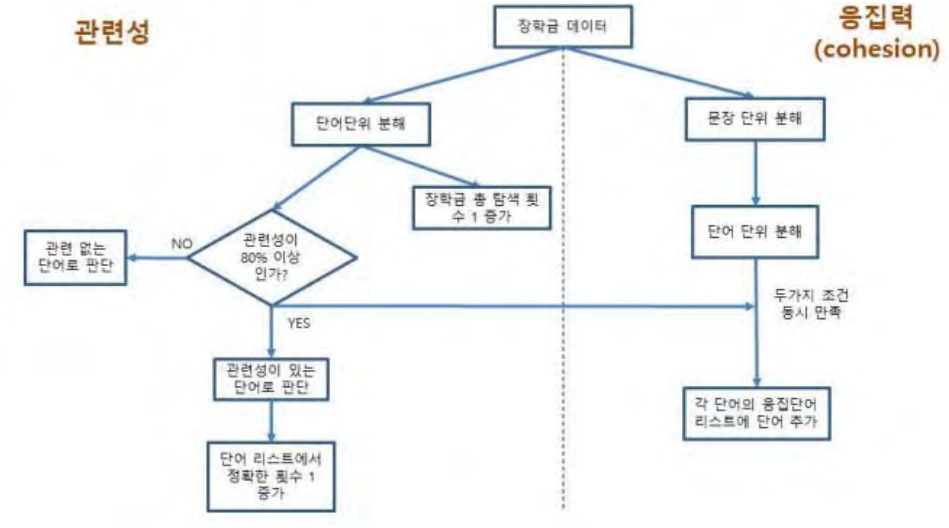


<사진 2> 분류 알고리즘

2.3 저장

저장기능으로는 분류된 내용을 사용자가 사용하기 좋도록 양식에 맞는 형식으로 저장한다. 저장에서 양식에 맞춰서 데이터를 저장하는데 응집력을 사용한다. 문장은 단어로 연결되어 있고 주제에 맞는 문장은 비슷한 단어들끼리 사용된다. 즉, 각 단어마다 단어끼리의 응집된다. 이러한 응집하는 성질을 수치화 하여 문장별로 양식 중 어느 데이터와 맞는 데이터인지 판단한다. 또한 응집력이 양식과 관련이 없다고 판단되면 불 필요한 문장이라고 판단한다.

저장에는 저장하는 기능도 있지만 학습기능을 구현하는 역할도 한다. 학습기능은 저장 과정에서 한 가지 과정을 더 하면서 구현할 수 있다. 구현방법은 크롤링을 통해서 파싱된 데이터 중 응집력이나 관련성 또는 정확성 검사에 사용하는 데이터가 있다면 데이터에 반영한다. 이 과정이 계속 반복된다면 분류에 사용되는 데이터는 단순히 하나의 데이터에 의한 결과물이 아니라 수천 수만 가지의 데이터에서 추출한 데이터를 토대로 한 작업을 하게 된다.



<사진 3> 분류 알고리즘

3. 결론

크롤링을 통해서 주제와 관련된 데이터를 판단한다. 자동탐색을 통해서 모든 페이지를 찾고 자동탐색 과정에 정확도 검사를 넣어서 무한루프와 같은 기타 오류 발생 요인을 제거한다.

주제와의 관련성을 판단하기 위해서 가중치를 사용 한다. 가중치가 일정 퍼센테이지(%)이상이라면 탐색된 데이터가 주제와 관련이 있다고 판단하여 저장한다.

저장에서는 응집력을 사용하여 각 문장과 양식이 맞는지 판단하고 불필요하다면 제거한 뒤 저장한다.

이러한 과정을 통해서 약 88%정확도를 자랑하는 크롤링을 할 수 있다.

참고문헌

[1] 한빛미디어 "파이썬으로 웹 크롤러 만들기" 라이언 미첼