웹 크롤링 기반 SNS웹사이트 설계 및 구현

윤경섭*, 김연홍O
O*인하공업전문대학 컴퓨터정보과
e-mail: ksyoon@inhatc.ac.kr*, yeounhong@nate.comO

Designing and implementing web crawling-based SNS web site

Kyung Seob Yoon*, Yeon Hong KimO
O*Dept. of Computer Science, Inha Technical College

● 요 약 ●

기존 Facebook 페이지의 경우에는 수많은 제보 글이 올라와 사용자가 원하는 글을 찾기 어렵다는 문제점이 발생하고 있다. 본 논문에서는 이를 위해 다양한 Facebook 페이지 내용을 크롤링하여 사용자가 원하는 Facebook 페이지 내용을 검색하여 사용자에게 제공할 수 있도록 데이터베이스 서버에 저장 한 후 크롤링 된 Facebook 페이지 내용을 제공할 수 있는 웹사이트를 설계하고 구현한다.

키워드: 웹크롤링기술(Web crawling technic) , SNS(Social Network Service), Facebook, 웹사이트(Web Site)

I. 서론

최근에는 수많은 사람들이 SNS(Social Network Service)를 통해서 안부를 묻고 일상 생활을 공유하고 있다. SNS 란, 특정한 관심이나 활동을 공유하는 사람들 사이의 관계망을 형성하고 구축해주는 온라인 서비스를 일컫는다.

SNS는 우리의 일상생활에 수많은 파급력을 가져오면서 관심대상이 되고 있다. 파급력을 가져 오는 이유 중 하나는 조그만 한 내용일지라도 SNS를 통해 제보를 받고 서로 소통을 하기 때문이다. SNS의 대표적인 예로는 Facebook, Instagramm등이 존재한다.

하지만 Facebook등과 같은 SNS는 소통이 활발한 만큼 수많은 제보 글이 올라오기 때문에 많은 양의 데이터가 존재한다. 그러므로 사용자가 원하는 데이터를 검색하기 힘들다는 단점이 발생 한다.

본 논문에서는 웹 크롤링 기술을 이용하여 Facebook 페이지의 다양하고 방대한 글 내용을 효율적으로 제공 할 수 있도록 제공된 검색 엔진을 통해 보다 쉽고 빠르게 볼 수 있도록 하는 웹사이트를 설계하고 구현한다.

II. 관련 연구

그동안 웹 크롤링 기술 방법으로 연구되었던 것 중 하나는 웹사이트의 웹페이지를 호출하여 웹페이지의 내용을 HTML 스트림으로 데이터를 받아오는 형식이다. 하지만 HTML스트림으로만 바로 받아올 경우, 데이터가 무분별하게 섞여 있기 때문에 원하는 데이터를 바로 추출해 내기 힘들다는 단점이 존재한다. 그렇기 때문에 추후에 XML형태나 JSON 형태로 원하는 데이터를 가공하여야 한다. 그러나 무작위로 존재하는 HTML스트림 데이터를 XML형태나 JSON 형태로 바꾸기 위해서는 주제별, 내용별, 상황별로 각각의 수많은 분류 알고리즘이 필요하기 때문에 알고리즘 구현에 많은 시간을 소비해야 되므로 효율성이 떨어진다. 이는 보통 웹 크롤링을 할 때 주로 쓰이는 방식으로 프로그래밍 언어와 상관없이 공통으로 적용되는 방식이다[1].

본 논문에서는 이를 보완할 수 있도록, 통계 및 데이터 분석을 위한 오픈 소스 프로그래밍 언어인 R프로그래밍 언어를 이용한다[2]. R에서 제공하는 library인 Rfacebook library와 Facebook API를 활용하여 XML형태나 JSON형태로의 변환 없이 쉽게 크롤링하고 가공, 추출할 수 있도록 하는 검색엔진을 설계한다. 이를 바탕으로 크롤링 한 데이터를 제공하고 사용자가 이를 활용할 수 있도록 한 SNS 검색 웹 사이트를 설계하고 구현 한다.

III. 설계

본 논문에서 개발하고자 하는 웹 크롤링 기반 웹사이트는 크롤링을 이용해 가공된 데이터를 통해 쉽게 검색할 수 있도록 웹사이트 구현하는 것을 목적으로 한다.

R프로그래밍언의 liberay인 Rfacebook은 R 사용자가 Facebook의 API에 접근하여 공개 페이지, 그룹 및 게시물에 대한 정보뿐

아니라, 인증된 사용자의 데이터에 대한 정보를 얻을 수 있는 기능을 제공한다.

우선 Facebook의 API에서 데이터를 받아오기 위해서는 액세스 토큰 즉, 개발자 토큰이 필요하다. 이러한 토큰은 Facebook 계정 로그인을 통해 facebook for developers 사이트에서 그래프 API를 이용하여 부여받는다.

그래프 API란, Facebook의 소셜 그래프에서 데이터를 가져오고 내보내는 기본방법으로, 데이터를 검색하고, 새 소식을 게시하는 등 다양한 작업을 실행하기 위해 사용할 수 있는 낮은 수준의 HTTP 기반 API이다[3]. 토큰 생성 후 R언어를 이용해 Rfacebook package를 다운받아서 library인 Rfacebook을 불러온다. Rfacebook에서 제공해주는 getPage 함수는 해당하는 페이지의 글 내용을 크롤링을 통해 데이터로 저장 할 수 있도록 제공한다. getPage 함수를 실행시키면 크롤링한 데이터를 반환하고 반환한 데이터는 .data형식으로 저장한 후 DB에 연동한다.

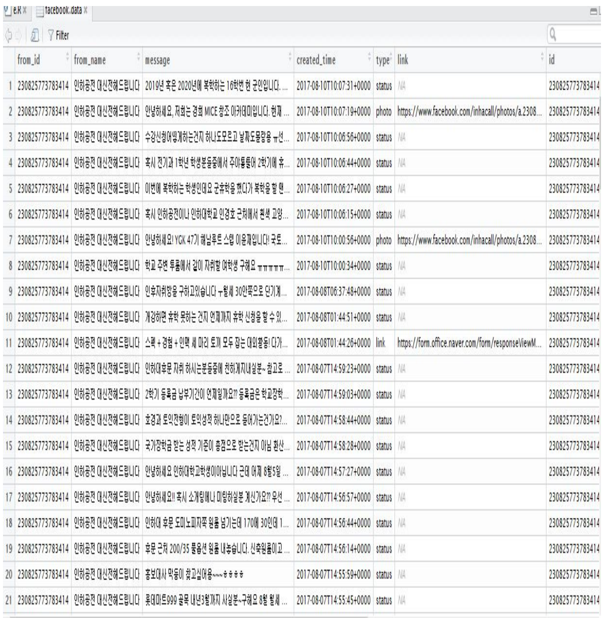

Fig. 1. Web crawling data based R

크롤링을 통해 DB 연동 후 저장한 데이터를 이용해 웹사이트 검색엔진을 구축한다.

검색엔진은 기본 검색어와 상세 검색어가 존재한다. 상세 검색어는 검색 옵션을 여러 개 설정이 가능하고, 설정한 옵션들을 연결해서 검색 할 수 있다. 그림2는 검색엔진 알고리즘을 나타낸다.

검색 알고리즘을 통해 검색 옵션 상태에 따라 서로 다른 Query문을 이용한다. 해당 데이터 Facebook 페이지, 게시물 주소를 변수로 지정해 <iframe>태그를 이용해서 해당 게시물 글을 보여준다. <iframe>태그란, Facebook에서 웹사이트에서도 적용이 가능하도록 제공하는 게시물 퍼가기 형태로 해당 게시 글을 보여줄 수 있도록 프레임을 제공해준다.

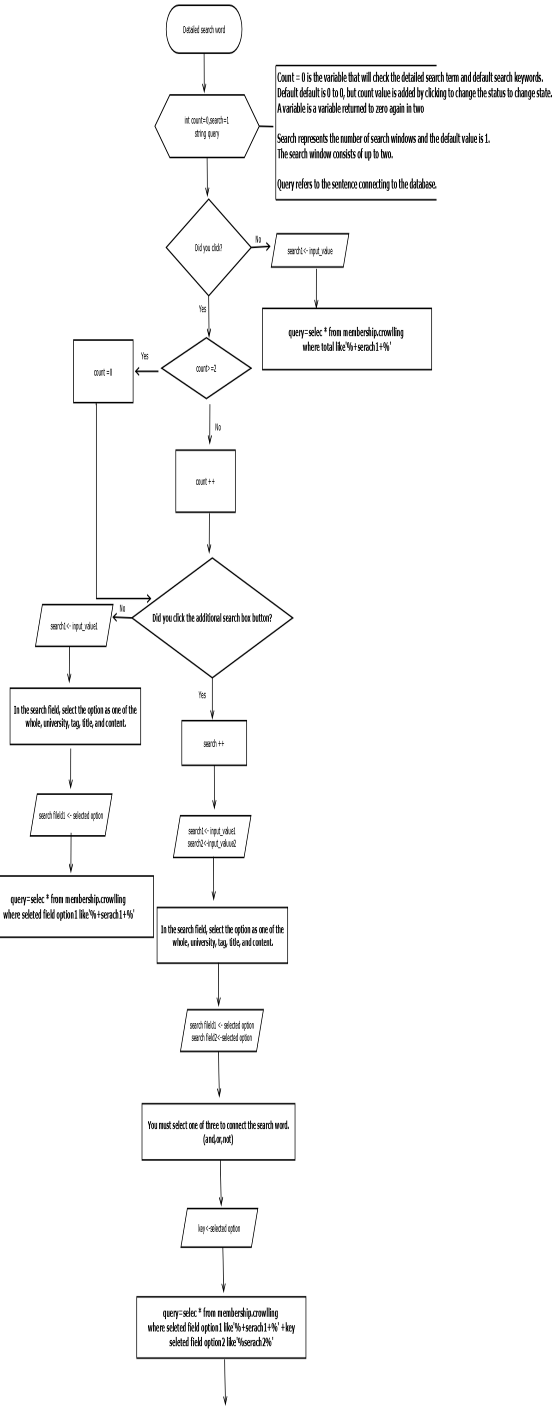

Fig. 2. Search algorithm

VI. 구현

본 웹사이트를 구현하기 위해서 JSP 프로그래밍 언어를 이용해 4개 대학교 페이스북 페이지에 개발한 검색 알고리즘을 적용한 웹사이트를 구현하였다. 그림3와 그림4은 해당 웹사이트에 들어가서 검색엔진을 통해 검색을 구현한 화면이다.

Fig. 3. Implementing the website default search page

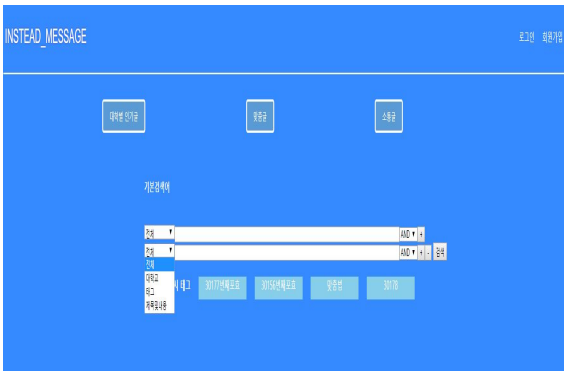

Fig. 4. Implementing the default search page

그림4 상세 검색어 같은 경우에는, 검색을 상세하게하기 위해 전체 , 대학교, 태그 , 제목 및 내용 중 하나 옵션을 선택해서 검색할 수 있다. 해당하는 옵션 설정을 추가하고 싶을 때는 +, - 버튼으로 추가 및 삭제가 가능해 여러 가지의 옵션들로 검색이 가능하다. 또한, 여러 개로 설정한 옵션들은 AND , OR , NOT 으로 연결이 가능하다. A AND B일 경우에는, A, B의 옵션 설정 내용이 둘 다 있는 내용만 검색한다. A OR B일 경우에는, A, B의 내용 중 하나만 존재해도 검색한다. A NOT B 는 A의 내용만 검색하되 B의 내용이 포함하지 않는 내용에 한해서 검색하도록 한다.

그림 5는 키워드 '안녕'을 검색했을 경우, 검색엔진을 통해서 나온 결과 화면을 구현한 예이다.

Fig. 5. Implementation Example 1

Fig. 6. Implementation Example 2

그림 6은 해당하는 카드UI의 글을 클릭했을 경우 존재 글의 Facebook 페이지의 글로 이동하게 된다.

VII. 결론

본 논문은 웹크롤링 기반으로 Facebook 페이지글을 크롤링하여 크롤링한 데이터를 데이터베이스와 연동하고 해당 데이터를 이용한다. 그 후 기본검색어와 상세 검색어로 분류해 사용자가 조금 더 쉽고 편리하게 검색할 수 있도록 검색 알고리즘을 설계하였다.

설계된 웹사이트는 쉽게 Facebook 페이지 글의 내용을 검색해서 보여줄 뿐만 아니라, 해당 글 주소의 페이지로까지의 연동을 가능하게 만들어 직접적인 접근이 가능해 사용자에게 편리성을 증진시킬 것이라 기대한다.

향후 연구는 실시간으로 데이터를 크롤링하여 웹사이트에 적용할 수 있도록 하는 방안을 모색할 것이다.

References

[1] http://oasis.dcollection.net/common/orgView/ 000001094618

[2] http://www.riss.kr/link?id=T13270895

[3] http://www.riss.kr/link?id=T12892863