Smart Monitoring System

제출일	2017.4.13	**소속학부**	컴퓨터학부
과목명	전공종합설계1	**조명**	Mr.Robot
담당교수	이상준	**조원**	20112315 강준호 20112339 깅성진 20112423 유상현

1. 프로젝트 소개

우리 프로젝트는 웹 환경에서 RSS피드를 제공하지 않는 웹 페이지로부터 rss피드를 제작하여, 일반적인 사용자에게 RSS서비스를 제공하는 것에서 시작한다. 기존의 RSS서비스와 달리 사용자가 원하는 페이지의 원하는 섹션을 지정하면, 그 부분의 RSS피드를 우리가 제작하여 보다 사용자 욕구에 가까운 RSS서비스를 누릴 수 있게 하는 것이다. 그리고 궁극적으로는 우리 프로젝트가 만들어낸 RSS피드를 바탕으로 크롤러 엔진을 제작하여, 크롤러 제작을 용이하게 하는 것을 목표로 한다.

우리가 살고 있는 시대는 정보의 바다를 넘어 정보의 우주라고 할 정도로 데이터가 과다한 시대에 살고 있다. 일반적인 사용자는 사용해본 적도 없는 제타바이트(ZB) 단위의 데이터가 웹 상에서 돌아다니고 있다. 몇 년째, 언론에서는 빅데이터를 외치며 데이터의 중요성을 강조하고 있다. 뿐만 아니라 지난 해 알파고가 초래한 인공지능이라는 화두는 더욱 더 데이터의 중요성을 우리에게 상기시킨다. 이러한 현실 속에서 일반 사용자들에게는 RSS서비스를 응용하여 사용자 맞춤형 데이터 구독 서비스를 제공하고, 동시에 데이터가 필요한 개발자들과 관계자들에게 RSS기반 크롤러 엔진을 제공하려 한다.

2. 기존 시스템 분석

현재 존재하는 RSS서비스들은 구독하고자 하는 웹 페이지의 관리자가 만들어준 피드를 사용하는 것이 대부분이다. 또한 RSS피드를 제공하지 않는 웹 페이지가 대다수이다. 사용자 입장에서는 RSS서비스를 누리고 싶은데 누리지 못하는 경우가 태반이며, 본인이 지정하지 않은 데이터도 불필요하게 받아야 하는 불합리함이 있다. 따라서 RSS피드를 제공하지 않는 웹 페이지의 피드를 제공하고, 사용자가 원하는 섹션의 정보만을 지정하여 받아볼 수 있게 하는 시스템이 필요하다.

그리고 크롤러는 일반적인 사용자나 초급 개발자에게는 진입장벽이 있는 편이다. 또한 한번 잘 제작해놓은 크롤러라 하더라도 사이트 구조가 바뀌는 등의 이슈가 있다면 제대로 동작하지 않을 가능성이 매우 농후하다. 웹 페이지와 크롤러 사이에 중간 단계로 RSS 피드 제작기를 두어, 대상 사이트 구조에 대한 의존성을 줄이고 우리가 제작하는 RSS피드를 바탕으로 크롤링할 수 있게 한다.

3. 특징 및 구상도

-특징

일반적으로 사용되는 RSS 서비스 프로세스는 데이터를 직접 소유하고 있는 회사나 단체에서 통일된 형식을 갖춘 XML형식의 FEED를 만들어 제공하고 사용자는 제공받은 FEED를 RSS 리더기를 통해 확인할 수 있다. 마찬가지 방식으로 웹 크롤러

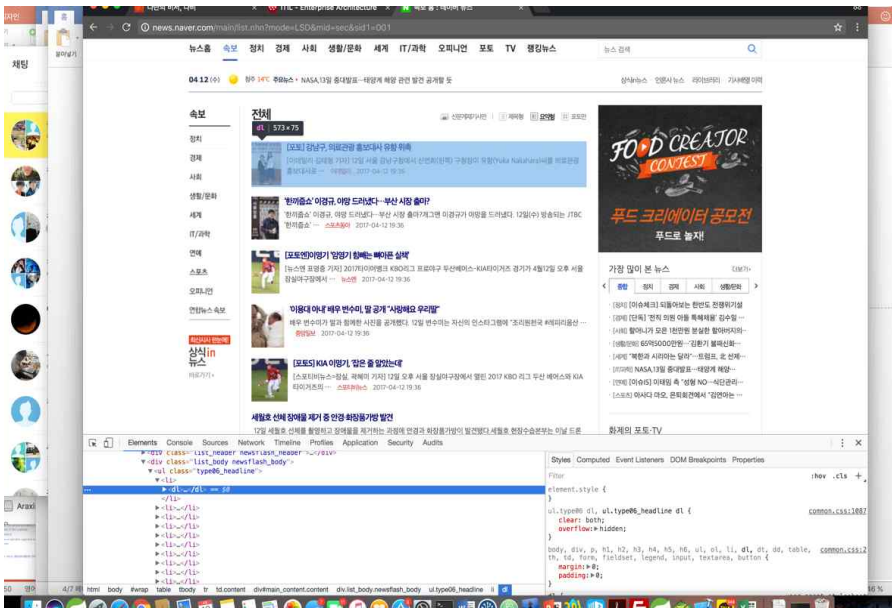
(웹 수집기)의 경우도 제공자로부터 CSV,JSON,XML 등 가공된 형태의 데이터를 받아 수집 하거나 경우에 따라 수집하고자 하는 특정 웹페이지 소스를 개발자가 직접 파싱하여 수집하는 경우도 있다.

앞서 언급한 일반적인 RSS 서비스와 웹크롤러 서비스가 가공된 형태의 데이터를 제공하는 것과 달리 눈에 보이는 페이지로부터 간단한 조작을 가해 수집하고자 하는 영역이나 요소를 등록하면 서비스가 자동으로 반복되는 요소를 웹 소스 내에서 검색하여 JSON,XML 등의 형태로 가공해 제공한다. 이러한 방법으로 가공된 데이터는 RSS FEED로 제공하거나 수집할 수 있도록 제공한다. 이 서비스를 통해 사용자는 RSS 서비스를 원하거나 수집하고자 하는 특정 웹사이트의 데이터를 쉽게 제공받아 재가용하여 새로운 서비스로의 확장이 용이하다.

- 기능 및 구상도

1. 웹상에서 웹요소 클릭을 통한 선택

크롬의 개발자 도구 기능 중 엘리먼트 선택기능과 유사하게 사용자가 웹요소를 보고 클릭하면 서비스 내부에서 반복되는 요소들을 판단한다.



2. JSON,XML 등 다양한 형식으로 데이터 제공

기존 RSS 리더기와 호환이 가능한 XML 및 개발에 용이한 JSON 형식 등 다양한 형태로의 데이터를 제공한다.

4. 간략수행계획서

순차적으로 프로젝트 중 다음의 핵심 기능들을 구현하도록 한다.

① Element Selector를 이용한 반복되는 데이터 선택
② 선택된 태그들을 정규식으로 변환 하여 서버에 저장
③ RSS 피드 형식으로 사용자들에게 제공
④ 수집한데이터를 서버에 저장

5. 관련기술

- Element Selector

기존에 사이트에서 제공하는 Format 외에는 원하는 데이터가 있어도 RSS를 통해서 정보를 받아볼 수 가 없다. SMS는 사용자로부터 원하는 부분을 HTML 태그 단위로 쪼개어 선택할 수 있는 기능을 제공해야 한다.

Chrome 요소검사의 Element Selector

Chrome 브라우저에서 제공하는 요소검사 기능 중 HTML태그를 선택하는 기능에 착안하여 Element seletor를 구현하고자 한다. 수시로 업데이트 되는 부분을 선택해야 하기 때문에 그림에서 보이는 것처럼 마우스로 클릭한 부분만을 선택하는 것이 아닌 반복되는 Tag를 통해 출력 되고 있는 모든 요소들을 한꺼번에 선택 할 수 있도록 구현할 예정이다.

태그 선택의 예시화면

예시화면에서 보이는 것처럼 "한국 경제"부분을 클릭하면 같은 구조로 이루어져 있

는 모든 태그를 선택할 수 있도록 하는 기능을 구현할 예정이다.

-RealTime Communication System(RTCS)

사용자에 의해서 선택된 부분에 대해서 데이터를 지속적으로 갱신해주는 것이 무엇보다도 우선시 되어야 할 기술이다. 기존에 사이트에서 RSS를 제공해주었다면 사이트에서 RSS 피드를 갱신을 해주어야만 업데이트된 정보를 받아볼 수가 있다. 하지만 SMS은 RSS 피드를 독점적으로 생산, 관리를 하기 때문에 실제 사이트에서 갱신된 정보를 받아오는 과정과 피드를 갱신시켜주는 과정을 구현해야 한다.

웹은 일반적으로 비동기방식 통신을 기반으로 하기 때문에 정보를 RealTime으로 보이기 위해서는 동기방식 기술구현이 필요하다. 현재 웹에서 사용되는 실시간 통신 방식은 몇가지가 있다.

가장기본적인 통신 방식은 Polling 방식인데 서버에 주기적으로 요청하는 방식이기 때문에 서버에 오버헤드가 발생할 수 있어 SMS에서는 WebSocket을 기반으로하는 클라이언트의 한번의 요청만으로 지속적으로 통신이 가능한 방법으로 구현하고자 한다.

6. 기대효과

- RSS 피드를 제공하지 않는 사이트에서도 RSS리더기를 통해서 마치 사이트에서 RSS피드를 제공하는 것처럼 사용할 수 있다.
- RSS 피드를 자체적으로 소유하고 관리하기 때문에 데이터수집엔진의 바탕을 만들 수 있다.