래퍼 기반 경제 데이터 수집 시스템 설계 및 구현

박철호 · 구영현 · 유성준*

세종대학교

Wrapper-based Economy Data Collection System Design And Implementation

Zhegao Piao · Yeong Hyeon Gu · Seong Joon Yoo*

Sejong University

E-mail : piaozhegao5@gmail.gom, yhgu@sju.ac.kr, sjyoo@sejong.ac.kr

요 약

경제의 흐름, 주가 등을 분석, 예측을 위해 경제 뉴스, 주가 등 데이터 수집이 필요하다. 일반적인 웹 크롤러는 자동적으로 웹서버를 방문하면서 웹페이지 내용을 분석하고 URL들을 추출하면서 웹 문서를 수집한다. 반면 특정한 주제의 문서만을 수집할 수 있는 크롤러 형태도 있다. 특정 사이트에서 경제 뉴스 정보만 수집하기 위하여 사이트의 구조를 분석하고 직접적으로 데이터를 수집해올 수 있는 래퍼 기반 웹 크롤러 설계가 필요하다. 본 논문에서는 빅데이터를 기반으로, 경제뉴스 분석 시스템을 위한 크롤러 래퍼를 설계, 구현하여 경제 전문 분야의 뉴스 데이터를 수집하였다. 2000년부터 현재까지 미국 자동차 시장의 주식 데이터를 래퍼 기반으로 가져오고, 사이트 상에서의 데이터가 업데이트되는 주기를 판단하여 주기적으로 업데이트 함으로써 중복되지 않게 하였다. 그리고 미국, 한국의 경제 기사를 래퍼 기반의 웹 크롤러를 사용하여 수집하고, 향후 분석이 쉽게 데이터를 정형화 시켜 저장한다.

ABSTRACT

For analyzing and prediction of economic trends, it is necessary to collect particular economic news and stock data. Typical Web crawler to analyze the page content, collects document and extracts URL automatically. On the other hand there are forms of crawler that can collect only document of a particular topic. In order to collect economic news on a particular Web site, we need to design a crawler which could directly analyze its structure and gather data from it. The wrapper-based web crawler design is required. In this paper, we design a crawler wrapper for Economic news analysis system based on big data and implemented to collect data. we collect the data which stock data, sales data from USA auto market since 2000 with wrapper-based crawler. USA and South Korea's economic news data are also collected by wrapper-based crawler. To determining the data update frequency on the site. And periodically updated. We remove duplicate data and build a structured data set for next analysis. Primary to remove the noise data, such as advertising and public relations, etc.

키워드

정제 데이터 수집, 래퍼 크롤러, 경제 데이터, 주가 데이터

Ⅰ. 서 론

본 논문**에서는 경제의 흐름, 주가 예측 등에 필요한 데이터를 수집하는 시스템 설계 및 구현에 관해 기술하고자 한다. 현재 많은 사이트에서 통계 데이터를 보여 주지만 데이터의 수량이 많고 종류가 다양한 동시에 주기적으로 업데이트됨으로 자동 데이터 수집 시스템이 필요하다.

본 논문에서 제시하는 경제 데이터 수집 시스템은 지정된 사이트에서 태그를 분석하고 필요한 데이터만을 가져오는 Wrapper 기반으로 설계하였

* 교신저자

** 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 해외 전문인력활용촉진사업의 연구 결과로 수행되었음(IITP-2014-H0905-14-1005)

다. 시스템은 미국 도시의 날씨 데이터, 자동차 회사의 미국 시장 주가, 미국 Dollar의 환율, 미국의 GDP, PPI 등 경제수치, 금 가격, 유가 등 데이터를 수집하고 업데이트 주기를 체크하여 데이터 자동 업데이트하는 목적으로 설계된다.

II. 관련연구

일반적인 웹 크롤러는 자동적으로 웹 서버를 방문하면서 웹 페이지 내용을 분석하고 URL들을 추출하며 웹 문수를 수집한다[1][2]. 이와 같은 크롤러는 데이터 수집 능력이 뛰어나 많은 분야에서 사용되고 있지만 특정 데이터에 대한 정확도가 떨어진다는 특징을 가지고 있다.

반면, 특정된 데이터 수집에 사용되는 크롤러도 많다. 그 중 하나인, Crawler4j는 메모리를 많이 사용하고 오류가 많다[3][4]. Topical Crawler는 작은 범위에서 주제 검색으로 범위가 제한되어 속도가 빠르게 특정한 데이터를 수집 할 수 있다[5]. 하지만 데이터의 완전성을 보장하지 못하고 우리가 원하는 경제 데이터 수집에 적용하기가 어렵다[6][7]. 래퍼 기반으로 만든 크롤러들은 각자에 지정된 데이터는 정확히 수집 할 수 있지만 HTML 구조상 형식이 같지 않고 수집 내용이 다르기 때문에 경제 데이터 수집에는 적용하지 못한다[8-10]. 그리하여 경제 데이터 수집에 사용되는 래퍼 기반 크롤러를 설계할 필요가 있다.

III. 요구사항 분석

사용자들의 요구 사항을 분석한 결과는 다음과 같다. 첫 번째는 효율적이고 신뢰성 있는 데이터 분석을 위해 우리는 사용자에게 정확하고 정결한 데이터를 제공해주어야 한다. 두 번째는 데이터 분석 측면에서, 현재의 데이터로 과거의 예측을 검증하고 미래에 대한 재예측이 제대로 이루어질 수 있도록 데이터의 주기적인 업데이트를 보증해 주어야 한다. 세 번째 사용자가 원하는 데이터들은 크게 나누면 미국의 날씨데이터, 미국 자동차 회사의 미국 주가 데이터, 금의 국제 가격, 석유의 국제 가격, 미국 Dollar과 타국 화폐 사이의 환율, 미국의 각종 경제 지표 수치 등을 필요해 한다. 네 번째로는 이러한 데이터들을 사용자가 쉽게 사용 할 수 있게 구조를 잡아 주는 것이다.

IV. 시스템 설계 및 구현

4.1 시스템 설계

그림 1과 같은 시스템은 URL 목록을 제공해주면 그 페이지를 크롤링하여 HTML Source를 가져

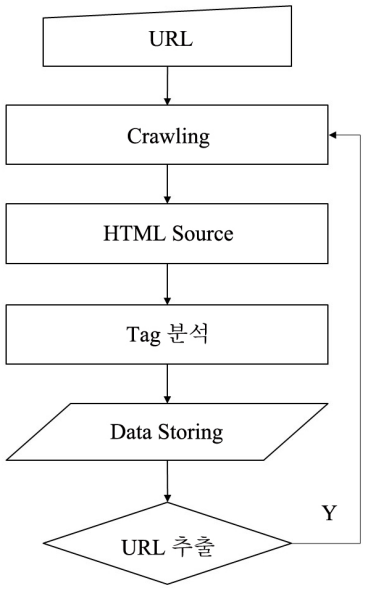

그림 1. 시스템 흐름도

와 안에 Tag를 분석한다. 그리고 Tag 분석을 통하여 얻으려는 특정 데이터만 추출하여 Data Storing를 형성하고 다음 페이지로 이동하는 링크를 자동으로 추출하여 낸다. 이 시스템 과정에서 수집한 데이터를 저장하기 위해서는 데이터베이스를 설계하여야 한다. 이는 아래와 같은 구조를 갖는다.

$$Weather = \{a_1, a_2, a_3, a_4, \ldots, a_{24}\}$$

Weather 날씨에는 순서대로 "최고 온도", "평균 온도", "최저 온도", "최고 노점", "평균 노점", "평균 노점", "최저 노점", "최고 습도", "평균 습도", "최저 습도", "최고 해면 기압", "평균 해면 기압", "최소 해면 기압", "최고 가시성", "평균 가시성", "최저 가시성", "최고 바람의 세기", "평균 바람의 세기", "최저 바람의 세기", "강수량", "구름", "사건", "바람 디렉터리", "도시 이름", "날짜", "입력된 시간"이 있다.

$$Stock = \{a_1, a_2, a_3, a_4, \ldots, a_{11}\}$$

Stock 주식에는 "오펀 가격", "최고 가격", "최저 가격", "마감 가격", "거래 총량", "adj_close", "날짜", "주식 이름", "피제수(Dividend)", "입력된 시간"이 있다.

4.2 시스템 구현

데이터를 수집할 7개 분야의 사이트를 분석해 보면 사이트 소스코드 형태 데이터를 수집하는 방법을 통해 래퍼로 크롤링하는 방법, 파일을 읽어 오는 방법 두 가지 방법으로 분류할 수 있다.

수집할 내용은 미국 자동차 회사 주식 데이터, 미국 날씨 데이터, 금 가격 데이터, 유가 데이터, 환율 데이터, 미국의 각종 경제수치 데이터 등 여러 가지가 있는데 여기에서는 Stock(미국 자동차 회사 주식) 데이터 만 예를 들어 설명한다.

4.2.1 Type 1

Stock 데이터는 주식 데이터 사이트 URL에 주식 이름 해당 코드를 준다. 문자열 u는 수집할 주식 사이트 홈페이지 URL 주소이다.

$$u = \text{“http://finance.yahoo.com/q/hp?s=”}$$

문자열 t_1는 단계별로 추출하려는 데이터의 시작 위치이다.

$$t_1 = \text{“<table class=\"yfnc_datamodoutline1\"”}$$

문자열 t_2는 단계별로 추출하려는 데이터의 끝 위치이다.

$$t_2 = \text{“</td></tr></table>”}$$

문자열 a_1은 t의 하위 단계 데이터 종류의 시작 위치이다.

$$a_1 = \text{“<tr>”}$$

문자열 a_2는 t의 하위 단계 데이터 종류의 끝 위치이다.

$$a_2 = \text{“</tr>”}$$

문자열 b_1은 실제 데이터의 시작 위치이다.

$$b_1 = \text{“<tr>”}$$

문자열 b_2은 실제 데이터의 끝 위치이다.

$$b_2 = \text{“</tr>”}$$

문자열 n_1은 다음 페이지가 URL을 찾는 시작 위치이다.

$$n_1 = \text{“<a rel=\"next\" href=\"”}$$

문자열 n_2는 다음 페이지가 URL을 찾기 위한 끝 위치이다.

$$n_2 = \text{“\"”}$$

Weather 데이터는 URL에 GET 방법으로 city, year, month, day를 파라미터로 주어 지정된 미국 도시의 지정 시간의 날씨 데이터를 보여주는 페이지 HTML Source에서 위 방법을 사용하여 데이터를 수집한다.

4.2.2 Type 2

```
<html>
<style><style>
<body>
...
<table class="yfnc_datamodoutline1" ...>
<tr>
<td class="yfnc_tabledata1">stock high price</td>
...
<td class="yfnc_tabledata1">stock low price</td>
</tr>
...
<tr>
<td class="yfnc_tabledata1"> stock high price </td>
...
<td class="yfnc_tabledata1"> stock low price </td>
</tr>
</table>
</body>
</html>
```

그림 2. 주식가격 페이지의 예(HTML)

유가 데이터 수집은 사이트에서 제공하는 URL(http://www.eia.gov/dnav/pet/hist_xls/RWTCd.xls)에서 제공되는 파일을 받아 parsing 한다. 미국의 경제수치 데이터는 URL에 사전에 준비해 놓은 경제 수치 이름에 해당하는 코드를 파라미터 주어 웹에서 제공되는 txt 파일을 parsing 한다. 그리고 금 가격 데이터는 사이트에서 제공되는 CSV 파일을 찾아 다운로드한다.

V. 결 론

위에서 제시한 방법으로 표 2에서와 같이 2010년부터의 15개 자동차 회사 주식 데이터를 일 단위로 45,963건, 2000년부터의 뉴욕과 로스앤젤레스의 날씨 데이터를 일 단위로 10,880건, 1995년부터의 금가를 월 단위로 239건, 1988년부터의 유가를 일(매주 5일) 7,381건, 2010년부터의 12개 나라와의 환율을 일 단위로 53,706건, 1960년부터 지금까지의 14개 경제수치 데이터를 상은 한 기간을 단위로 8,538건 수집하였다. 향후 이 데이터는 주기적으로 업데이트 할 것이다.

표 1. 데이터 수집 결과

구분	데이터 수량	데이터 주기
주식	45,963	일
날씨	10,880	일
금 가격	239	월
유가	7,381	일
환율	53,706	일
경제지표	8,538	월

참고문헌

[1] D. M. Seo and H. M. Jung, “Intelligent Web Crawler for Supporting Big Data Analysis Services”, The Journal of the Korea Contents Association, vol. 13, no. 12, (2013), pp. 575-584.

[2] Soumen Chakrabarti, Martin van den Bergb, Byron Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery", Computer Networks, Vol.31, No. 11-16, pp.1623-1640,1999

[3] Han-Gil Kim, Jong-Won Lee, Tae-Hak Ban and Hoe-Kyung Jung*, "A Study on Distributed Crawling-Based Overhead Optimization" ,International Journal of Software Engineering and Its Applications, Vol. 9, No. 3 (2015), pp. 175-182

[4] P. Srinivasan, F. Menczer and G. Pant, “A general evaluation framework for topical crawlers”, Information Retrieval, vol. 8, no. 3, (2005), pp. 417-447.

[5] Ziyu Guan, Can Wang, Chun Chen, Jiajun Bu, Junfeng Wang, "Guide Focused Crawler Efficiently and Effectively Using On-line Topical Importance Estimation", ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 757-758, 2008

[6] J. Cho, H. Garcia-Molina, and L. Page, “Efficient Crawling through URL Ordering”, Computer Networks 30(1-7), pp. 161-172, 1998

[7] H. T. Yani Achsan and W. Catur Wibowo, “A Fast Distributed Focused-Web Crawling”, Procedia Engineering, vol. 69, (2014), pp. 492-499

[8] Hanhoon Kang, Seong Joon Yoo and Dongil Han, "Design and Implementation of Web Crawler Wrappers to Collect User Reviews on Shopping Mall with Various Hierarchical Tree", Journal of Korean Institute of Intelligent Systems 06/2010; 20(3). DOI: 10.5391/JKIIS.2010.20.3.318

[9] Jaeyoung Yang, Tae-Hyung Kim, Joongmin Choi, "An Interface Agent for Wrapper-Based Information Extraction", the International Confenrence on Principles of Practice in Multi-Agent Systems, pp.291-302, 2004

[10] Claudio Bertoli, Valter Crescenzi, Paolo Merialdo, "Crawling programs for wrapper-based applications" IEEE International Conference on Information Reuse and Integration, pp.160-165, 2008