사용자 인터페이스 기반 범용 웹 정보 추출[1)]

허정수° 하상호
순천향 대학교 컴퓨터 학부
{majya°, hsh}@sch.ac.kr

GUI Based Generalized Fine-Grain Web Information Extractor

Jeongsoo Hur°, Sangho Ha
Computer Science and Engineering, Soonchunghyang University

요 약

인터넷이 보급되면서 사용자는 자신이 원하는 정보를 인터넷으로 접근하였으며, 정보에 대한 수요는 늘어나 검색이나 포털을 통한 정보의 접근이 이루어지고 있다. 사용자들이 원하는 정보를 통합하더라도 사용자들은 자신이 원하는 정보를 얻기 위해 불필요한 정보에 노출된다. 최근에 사용자가 필요한 웹 정보를 추출하는 연구가 진행되고 있으나, 이러한 연구는 추출 단위가 HTML 문서 수준이거나 일반적이지 못하다. 본 논문에서는 사용자가 원하는 임의의 웹 정보를 항목 단위의 수준에서 추출할 수 있는 사용자 인터페이스 기반 범용 웹 정보 추출기를 UML에 기반하여 설계하고 구현한다. 또한, 구현된 시스템에 대한 실행 예를 보인다.

1. 서 론

인터넷의 보급이 시작되면서 자신이 필요한 정보를 인터넷으로부터 구하려는 사용자들이 늘어나고 있다. 이런 사용자들을 위하여 Yahoo를 시작으로 많은 웹 검색 사이트가 생겨났으며, 여러 정보를 취합하여 사용자에게 전달하는 포털 사이트가 생겼다. 그래서 사용자들은 정보의 바다인 인터넷에서 원하는 정보를 검색을 통해 접근할 수 있는 지금에 이르렀다. 그러나 포털 사이트 등에 의해 정보가 취합되었다고는 하나 사용자는 자신이 원하는 정보에 접근하기 위하여 수많은 웹페이지를 지나치면서 자신이 원하지 않는 정보를 접하게 된다. 또한, Netscape나 Internet Explorer의 북마크 기능의 이용은 위의 문제를 해결하였으나 여전히 사용자는 불필요한 정보에 노출된다.

개인에 관심있는 웹 정보를 추출하는 연구가 최근에 이루어져 왔다[1,2,3,4]. [1]에서는 사용자가 관심 있는 웹 사이트를 방문하여 관심 사항을 반영하여 그 웹 사이트에 대한 scoping과 filtering을 통해서 개인화된 웹 뷰(Web View)를 생성한다. 그러나 웹 페이지로부터 정보 추출 단위는 HTML 문서이며 특정 데이터 항목이 아니다. 여기서 선택된 문서에 대한 하이퍼링크는 HTML 태그들로 표현된다. [2]에서는 사용자가 관심 있는 웹 페이지 접근 과정을 Smart Bookmark로 기록하고, 해당 웹 페이지에서 추출하고자 하는 항목들을 XPath[5]로 기술함으로써 개인화된 웹 뷰를 생성한다. 그러나 일반 사용자가 XPath를 사용하여 추출하고자 하는 항목들을 서술해야 하는 문제점이 있다. [3]에서는 각 웹 페이지에 대해서 사전에 정의된 도메인 지식에 기반하여 그 페이지로부터 항목을 추출하는데 사용되는 wrapper를 자동으로 생성하고, 사용자가 요청한 검색 결과의 웹 페이지에 해당 wrapper를 적용하여 필요한 항목을 추출한다. 그러나 이 방법은 레이블과 값의 쌍으로 표현된 데이터 항목 추출에만 적용 가능하며, 테이블과 같이 레이블을 갖지 않는 항목에 대해서는 적용되지 않는다. 또한, 웹 페이지가 다른 서술 패턴을 갖는 두 가지 이상의 유형 정보를 포함할 경우에, 그 웹 페이지에 대해서 모두 적용 가능한 패턴을 찾을 수 없다는 점에서 일반적이지 못하다. [4]에서는 데이터 항목 단위의 웹 정보를 추출하나 그 적용이 M-커머스[6]에 제한된다는 점에서 일반적이지 못하다.

본 논문에서는 인터넷상의 임의의 웹 문서로부터 데이터 항목 단위의 정보를 사용자 인터페이스에 기반하여 추출할 수 있는 시스템을 설계하고 구현한다. 사용자는 GUI에 기반하여 지속적으로 관찰하고자 하는 웹 정보를 추출할 수 있는 래퍼를 생성할 수 있고, 다음에 래퍼를 이용하여 원하는 시점에 해당 웹 정보를 웹 브라우저를 통하지 않고 GUI에 기반하여 간단 용이하게 추출할 수 있다. 래퍼는 웹 데이터 항목을 추출하는데 필요한 제반 정보들로 구성된다. 2장에서는 시스템에 대한 설계를 기술하고, 3장과 4장에서는 시스템의 구현 사항과 그 적용 예를 기술한다. 마지막으로 5장에서는 결론을 언급한다.

2. 시스템

본 논문의 시스템은 크게 경로 정보 생성과 웹문서로부터 정보추출의 두 부분으로 나뉘고, 각 기능은 일반화되어 컴포넌트로써 설계되었다. 컴포넌트로 구성된 정보 생성과 정보 추출은 간단한 사용 예를 보이기 위한 그림 1의 시스템으로 설계되었다. 사용자는 GUI기반의 래퍼 생성기를 이용하여 래퍼를 생성하며, 생성된 래퍼를 웹정보 추출기를 이용하여 원하는 정보를 얻을 수 있다.

그림 1 시스템 구조

2.1 래퍼

사용자가 원하는 정보의 경로를 담고 있는 문서로써 그림 2와 같다. 최상위에 Wrapper가 있으며 Name을 속성으로 갖는다. Wrapper의 Name은 사용자가 정한 래퍼의 이름을 나타낸다. Wrapper는 사용자가 원하는 정보로 접근하기 위한 경로 정보인 EPath의 집합으로 하나 이상의 EPath를 갖는다. EPath

1) 이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음. (R05-2004-000-12565-0)

는 하나의 정보를 표현하기 위한 단위로써 EPath의 이름인 Name과 정보를 담고 있는 웹문서의 주소를 나타내는 TargetAddress를 속성으로 갖고 있으며, 경로 정보의 한 단위인 Node를 하나 이상 갖는다. Node요소는 경로의 한 단계를 나타내는데, BODY 태그의 자식 태그부터 시작한다. TagName은 HTML 태그의 이름을 나타내고, TagIndex는 자식노드들 사이에서 순위를 나타낸다.

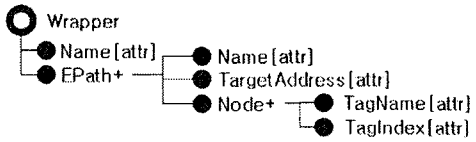

그림 2 래퍼의 구조

2.2 래퍼 생성기

래퍼 생성기는 사용자를 위한 GUI형태로 제공되며 사용자가 웹문서에서 원하는 부분의 정보로 접근하기 위한 경로 정보를 만들고 경로 정보를 담고 있는 래퍼를 생성한다. 그림 3은 "래퍼 생성기"의 구조를 보이는 UML 클래스 다이어그램이다. "래퍼 생성기"는 사용자에게 GUI를 제공하는 *UIManager*를 이용하여 사용자가 요구하는 일을 받아들인다. UIManager는 WrapperManager를 이용하여 래퍼를 생성한다.

WrapperManager는 UIManager로부터 사용자가 원하는 웹문서에 대한 정보를 받아들이고 사용자가 선택한 부분에 대하여 정보 추출을 위한 경로 정보를 생성하며 시스템의 알맞은 형태로 XML문서인 래퍼를 생성 및 사용자가 원하는 위치에 저장한다. WrapperCreator는 WrapperManager가 제공하는 웹문서를 이용하여 경로 정보를 만든다. 경로 정보는 EPath 객체로 구성되며 경로의 요소는 Node 객체에 저장한다. WrapperCreator는 하나의 독립된 컴포넌트로 제공되어 WrapperManager의 변경으로 다른 시스템에 적용이 용이하다.

2.3 웹 정보 추출기

2.2절의 "래퍼 생성기"에 의해 래퍼가 생성되면 사용자는 "웹 정보 추출기"를 사용하여 원하는 정보를 웹문서에서 추출할 수 있다. 그림 4는 "웹 정보 추출기"의 UML 클래스 다이어그램을 보인다. "웹 정보 추출기"는 사용자가 실행하는 부분인 Extractor 클래스와 래퍼를 받아들여 정보를 추출하는 처리에 대한 관리를 하는 ExtractManager가 있다.

Extractor는 사용자에게 제공하기 위한 어떠한 형태로도 UI를 구성할 수 있으며, ExtractManager는 UI에 대한 정보 추출을 위한 기능을 지원한다. ExtractManager는 사용자가 생성한 래퍼를 이용하면서 사용자에게 단순 제공을 위한 정보와 정보 추출을 위한 경로 정보를 구한다. 이때 경로 정보는 EPathForExtractor 객체를 이용하여 웹 문서에서 사용자가 원하는 정보를 추출하게 된다. EPathForExtractor는 래퍼 상의 경로 정보를 객체로 표현하며 경로에 따른 정보를 추출하기 위한 기능을 제공한다. 이 객체는 그림 3의 WrapperCreator와 같이 컴포넌트의 형태로 제공된다. EPathForExtractor는 Node를 이용하여 경로 정보의 요소를 표현하며 추출된 정보는 ExtractedElement에 저장된다. 또한 Publisher는 추출된 정보를 클라이언트가 이해할 수 있는 형태로 표현하는 인터페이스이다.

3. 구현

본 논문의 시스템은 .NET 환경에서 구현되었다. 국내 대부분의 웹페이지는 Microsoft Internet Explorer에서 디스플레이될 수 있도록 제작되었다. 따라서 웹 정보 추출 시 필요한 웹 문서의 파싱을 위해서 MS에서 제공하는 Internet Development SDK[7]가 사용되었다. 시스템의 구현 플랫폼은 Microsoft사의 Visual Studio 2005이며, Internet Development SDK의 WebControl과 MSHTML 컴포넌트가 "래퍼 생성기" 구현에 사용되었고, "웹 정보 추출기" 구현에는 MSHTML이 사용되었다.

4. 적용

여기서는 구현된 시스템을 적용한다. 사용자가 관심있는 회사의 주식 시세를 자주 조사한다고 가정한다. 사용자가 주식 시세를 조사하기 위해서 매번 웹 서치를 하는 것은 시간이 걸리고 귀찮은 일이다. 또한, 해당 웹 페이지는 다른 많은 정보도 포함하여 사용자가 그 페이지에서 필요한 항목을 찾아보아야 할 것이다. 여기서는 논문에서 개발한 시스템을 사용하여 사용자가 원하는 주식 시세를 추출하여 언제든지 간단 용이하게 그 정보를 조사할 수 있다는 것을 보여준다.

그림 5는 래퍼 생성기를 이용하여 관심있는 회사의 주식 시세를 보여주는 웹페이지를 방문하여 필요한 항목을 추출하는 과정을 보여준다. 먼저 래퍼생성기가 제공하는 GUI의 "추출" 메뉴에서 "항목 추가..."를 선택하고 사용자가 원하는 메뉴의 이름을 입력한다. 여기서는 삼성전자와 LG전자의 주가를 위한 이름을 입력하였다. 추출 항목에 대한 이름을 입력하면 추출 메뉴에 해당 항목이 생성된다. 이제 사용자는 자신만의 주가 정보를 위한 개인화된 래퍼를 생성할 수 있다. 사용자는 검색한 웹 페이지에서 주가를 포인터로 드레그해서 선택한다. 여기서 84,000이 주가로 선택되었으며, 래퍼생성기는 사용자가 추가한 메뉴를 이용하여 해당 항목에 대한 경로를 추출한다. 사용자는 필요에 따라서 추출하고자 하는 항목을 메뉴에 추가할

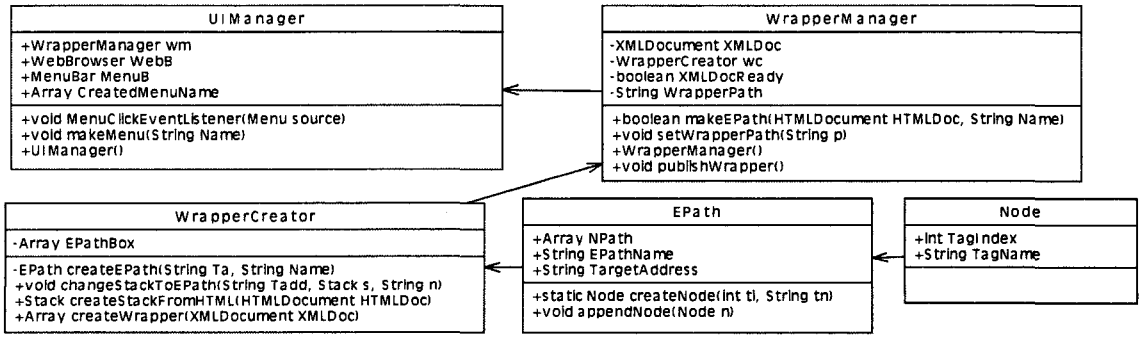

그림 3 "래퍼 생성기"의 UML 클래스 다이어그램

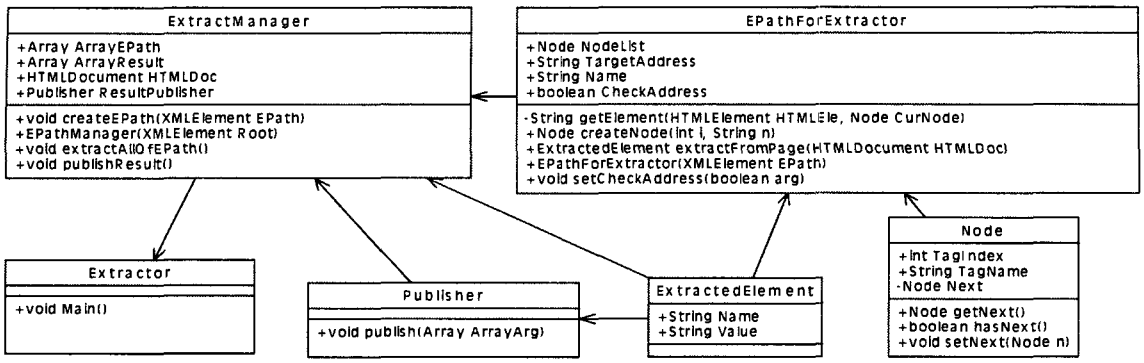

그림 4 "웹 정보 추출기"의 UML 클래스 다이어그램

수 있으며, 이러한 항목들은 여러 웹 페이지에 걸쳐서 추출될 수 있다. 경로 정보의 추출이 끝나면 "파일"메뉴에서 "래퍼 저장 위치..."를 선택하여 래퍼의 이름(여기서는 StockInfo), 래퍼가 저장되는 위치, 래퍼의 파일명을 입력하고, "래퍼 발행"메뉴를 선택하면, 사용자가 추출한 모든 항목들에 대한 경로들을 포함한 래퍼를 생성하고 저장한다.

그림 6은 사용자가 웹 정보 추출기를 사용하여 간단하게 추출된 웹 정보를 확인하는 과정을 보여준다. 사용자는 웹 정보 추출기 GUI에서 myFavoriateInfo를 클릭하면, 사용자가 추출한 웹정보 클래스 리스트를 보여준다. 여기서는 그림 4에서 추출한 StockInfo를 클릭하면, 사용자가 관심 있는 회사와 그 회사의 주식시세가 화면에 디스플레이된다. 웹 정보 추출기는 사용자가 선택한 래퍼 StockInfo를 가져와서, 래퍼에 포함된 각 항목에 대해서 웹 문서 주소를 이용하여 해당 웹 문서를 가져오고, 그 문서상의 항목의 경로를 이용하여 해당 항목을 웹 문서로부터 추출한다. 이러한 과정은 내부적으로 이루어지므로 사용자는 웹 브라우저를 실행시킬 필요도 없다. 그림 6에서 디스플레이되는 회사의 주식시세는 해당 웹 페이지에 가장 최근에 기록된 데이터임에 유의해라.

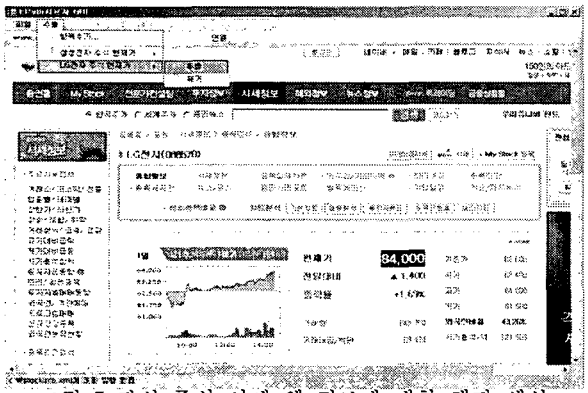

그림 5 관심 주식 시세 웹 정보에 대한 래퍼 생성

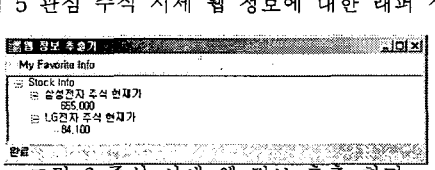

그림 6 주식 시세 웹 정보 추출 화면

5. 결론

논문에서는 사용자가 계속 살펴보고자 하는 관심 있는 웹 정보를 추출하여 사용자가 원하면 언제든지 간단 용이하게 해당 웹 정보를 확인해볼 수 있는 시스템을 컴포넌트에 기반하여 설계하고 구현하였다. 웹 정보 추출단위는 웹 페이지가 아닌 데이터 항목 수준으로 사용자에게 꼭 필요한 정보만을 추출하여 제공할 수 있다는 점이 시스템의 장점이다. 또한, 시스템은 컴포넌트에 기반하여 구현되었으며, 이러한 컴포넌트는 다양한 웹 응용 시스템 구성에 활용될 수 있다.

현재의 시스템은 데이터의 웹 정보 추출에만 제한된다. 그러나 앞으로 이미지의 웹 정보 추출도 고려하고 있다. 또한 추출된 웹 정보가 무선 환경을 포함하여 인터넷 상의 임의의 사용자에게 제공하여 효과적으로 사용자 단말기에 디스플레이될 수 있는 방법도 고려하고 있다.

6. 참고 문헌

[1] Zehua Liu, Wee-Keong Ng, Ee-Peng Lim, "Personalized Web Views for Multilingual Web Sources", IEEE INTERNET COMPUTING, p16~p22 ,July. August 2004

[2] Juliana Ferie, Bharat Kumer, DanielLieuwen, "WebViews: Accessing Personalized Web Content and Services", ACM 1-58113-348-0, p.576~p.586, May 1-5, 2001

[3] Jaeyoung Yang, Joongmin Choi, "Knowledge-Based Wrapper Induction for Intelligent Web Information Extraction", Springer, Web Intelligence, No.8, p.153~p.172

[4] 허정수, 하상호 , "모바일 단말기를 위한 인터넷 상품 정보 제공 시스템", 한국 정보과학회, 2005년 추계 학술 발표 논문집(2), p552-564

[5] XPath, http://www.w3.org/TR/xpath

[6] Norman Sadeh, "M-commerce:Technologies, Services, and Business Models, Reading, Wiley, 2002

[7] Internet Development SDK, http://msdn.microsoft.com/library/default.asp?url=/workshop/browser/prog_browser_node_entry.asp