한국지능시스템학회 논문지 2010, Vol. 20, No. 3, pp. 318-325

다양한 계층 트리 구조를 갖는 쇼핑몰 상에서의 상품평 수집을 위한 웹 크롤러 래퍼의 설계 및 구현+

Design and Implementation of Web Crawler Wrappers to Collect User Reviews on Shopping Mall with Various Hierarchical Tree Structure

강한훈 · 유성준++ · 한동일

Hanhoon Kang, Seong Joon Yoo and Dongil Han

세종대학교 컴퓨터공학과

요 약

본 논문에서는 다 계층 구조와 다양한 웹 언어로 구성된 한국내 쇼핑몰로부터 상품평 수집을 위한 래퍼 데이터베이스 기술 언어 및 모델을 제안한다. 기존에 제안된 래퍼 기반 웹 크롤러는 HTML 문서를 수집할 수 있고, 수집 대상으로 하는 문서의 계층 구조는 2~3계층이다. 그러나 한국형 쇼핑몰 사이트는 HTML 문서뿐만 아니라 다양한 웹 언어(JavaScript, Flash, AJAX)로 구성되어 있고, 그 계층 또한 5계층으로 이루어졌다. 웹크롤러가 이 5 계층 사이트에 있는 상품평만을 수집하려고 하면 상품평이 있는 위치를 정확히 알고 있으면 된다. 우리가 제안하는 래퍼에는 이러한 정보를 포함하고 있도록 하였고, 이러한 정보를 기술하기 위해 필요한 래퍼 데이터 기술 문법도 제안한다.

키워드 : 래퍼, 쇼핑몰, 상품평, 오피니언 마이닝

Abstract

In this study, the wrapper database description language and model is suggested to collect product reviews from Korean shopping malls with multi-layer structures and are built in a variety of web languages. Above all, the wrapper based web crawlers have the website structure information to bring the exact desired data. The previously suggested wrapper based web crawler can collect HTML documents and the hierarchical structure of the target documents were only 2-3 layers. However, the Korean shopping malls in the study consist of not only HTML documents but also of various web language (JavaScript, Flash, and AJAX), and have a 5-layer hierarchical structure. A web crawler should have information about the review pages in order to visit the pages without visiting any non-review pages. The proposed wrapper contains the location information of review pages. We also propose a language grammar used in describing the location information.

Key Words : wrapper, shopping mall, review, opinion mining

1. 서 론

본 논문은 다 계층 구조와 다양한 웹 언어로 구성된 한국내 쇼핑몰로부터 상품평을 수집하기 위한 래퍼 데이터베이스 기술 언어 및 모델을 제안한다. 최근 들어 오피니언 마이닝 연구[1~5]가 활발하게 진행되고 있다. 이 연구의 궁극적인 목표는 쇼핑몰에 게재된 상품평을 자동으로 분석하고 그 결과를 사용자에게 제공함으로써 사용자가 제품의 구매 결정을 쉽게 할 수 있도록 하는 것이다. 각각의 연구가 그 목표를 이루기 위해 접근하는 방법은 다양하지만, 대부분의 연구에서 공통적으로 필요한 부분은 실제 환경에서 사용되는 상품평 문서를 정확하고 빠르게 수집하는 방법이다.

프로그램 에이전트에 의해 자동으로 빠르게 상품평을 수집하기 위해서는 웹 크롤러(Web Crawler)를 이용한다. 그러나 웹크롤러의 종류에 따라 수집할 수 있는 데이터의 범위가 다양하기 때문에 특정 사용자에게 적합한 데이터를 쉽게 수집하기 위해서는 웹 크롤러의 기능을 고도화할 필요가 있다.

기존에 제안된 웹크롤러는 일반적인 크롤러[5], 포커스드 크롤러(focused crawler)[6], 토피컬 크롤러(topical crawler)[7][8], 래퍼 기반 웹 크롤러[10~13]가 있다. 이 중 상품평과 같이 특정 주제의 웹 문서를 정확하게 수집하기 위해서는 래퍼 기반 웹 크롤러를 사용한다. 이 래퍼 기반 웹 크롤러는 해당 사이트의 구조를 미리 분석한 정보를 가지고 있어, 대상 데이터가 존재하는 부분만 접근할 수 있다. 그러나 기존의 래퍼 기반 웹 크롤러[10~13]는 한국내 쇼핑몰의

접수일자 : 2009년 12월 2일
완료일자 : 2010년 5월 10일
+ 이 논문은 2008년도 세종대학교 교내 연구비 지원에 의한 논문임
++ 교신저자

상품평을 수집하는 데 있어 직접적으로 적용할 수 없다. 수집 대상으로 하는 웹사이트의 특징이 다르기 때문이다. [10~13]에서 대상으로 한 사이트는 URL 접근이 용이한 HTML 문서만 수집 대상으로 하였고, 목표 데이터를 수집하기 위한 계층 구조가 비교적 단순한 2~3단계로 이루어진 사이트만 대상으로 하였다. 그러나 한국내 쇼핑몰은 HTML 문서 뿐만 아니라 JavaScript, Flash, AJAX 등의 다양한 웹 언어로 표현되어 있어 기존에 제안된 래퍼 기반 웹 크롤러로는 페이지 접근이 불가능하다. 또한 그 계층구조도 5단계 정도로 이루어져 있어, 각 단계별로 필요한 URL 정보를 분석하여 래퍼를 구성해야 한다. 이에 따라 본 논문에서는 한국내 쇼핑몰에 적용할 수 있는 래퍼 기반 크롤링 모델을 제안하고 구현을 통해 성능 평가를 실행한다.

본 논문의 2장에서는 기존의 웹 크롤러 관련 연구를 분석한다. 3장에서는 한국형 쇼핑몰 사이트의 구조를 분석한다. 4장에서는 한국형 쇼핑몰에 적용 가능한 래퍼 를 설계한다. 5장에서는 래퍼를 기반으로 크롤러를 구현하고 성능 평가를 수행한다. 마지막으로 6장에서 결론을 맺는다.

2. 관련 연구

일반적인 크롤러는 모든 웹 문서를 수집 대상으로 하며, 포커스드 크롤러(focused crawler)와 토피컬 크롤러(topical crawler)는 특정한 주제의 문서만을 수집하기 위해 일반적인 크롤러에 문서 분류기(document classifier)나 규칙(rules)을 포함하고 있다. 이 때, 포커스드 크롤러나 토피컬 크롤러에서 분류기의 성능이 좋지 못하고 규칙의 양이 불충분할 경우 사용자가 원하는 데이터를 쉽게 수집할 수 없다. 이러한 이유로 수집 대상 사이트의 구조를 분석하여 직접적으로 데이터를 수집해올 수 있는 래퍼 관련 연구가 진행되어 왔다.

[10]은 래퍼를 자동으로 구성하는 방법과 수동으로 구성하는 방법에 대해 설명하고 있다. 수동으로 래퍼를 구성하면 주기적으로 변화하는 특정 사이트에서는 매번 래퍼를 다시 구성해야 하는 단점을 설명하고 있다. 이 때문에 자동으로 래퍼를 구성하는 방법도 설명하고 있다. 이와 더불어 인터넷 상의 수많은 웹 데이터를 추출하는 방법을 비교 분석하였다.

[11]은 XML을 이용하여 규칙을 만들고, 규칙을 기반으로 하여 래퍼를 구성하는 방법을 제안하였다. 규칙을 생성하는 데 있어 수동적인 방법과 자동적인 방법을 둘 다 고려하고 있다. [11]에서는 쇼핑몰과 같이 구조가 복잡한 사이트를 대상으로 하지 않았고, 그 계층 구조가 비교적 간단하고 준 구조화(semi-structured) 되어 있는 사이트를 대상으로 래퍼를 구성하였다.

[12]는 수동으로 래퍼를 구성했을 때 단점을 지적하고 있다. 이러한 단점을 보완하기 위해 특정 사이트?의 구조를 분석하여 데이터 모델을 구축하였고, 이를 이용하여 자동 래퍼를 구성하는 방법을 소개하고 있다. 수집 대상으로 하는 사이트는 한 곳이며, 원하는 정보를 찾기 위한 계층 구조는 2-3계층으로 이루어져있다.

[13]은 래퍼 기반 정보 추출 방법을 설명하고 있다. 이 기법은 구조화되어 있는 웹문서부터 구조화 되지 않은 자연어 문장의 범위까지 원하는 정보를 추출하기 위해 규칙(rule)을 구축하는 방법을 소개하고 있다.

3. 한국내 쇼핑몰 사이트의 구조 분석

본 절에서는 한국내 쇼핑몰의 구조를 분석한다. 그 결과 한국내 쇼핑몰에서 상품평을 발견하기 위해 5단계의 이동 경로를 거쳐야 한다는 것을 발견하였다.

(1)카테고리 페이지->(2)리스트 페이지->(3) 상세 페이지->(4)리뷰 리스트 페이지->(5) 상세 리뷰 페이지(상품평 페이지)

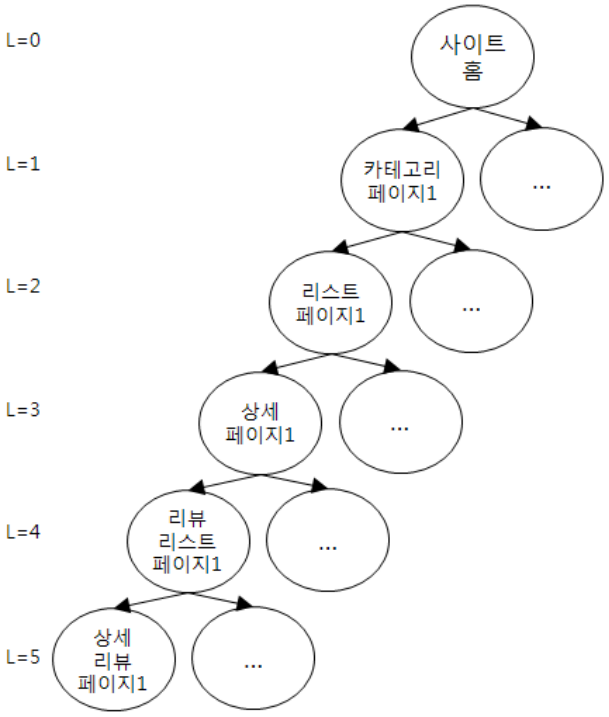

그림 1. 쇼핑몰 사이트의 트리 구조
Fig. 1. Tree structure of shopping mall site

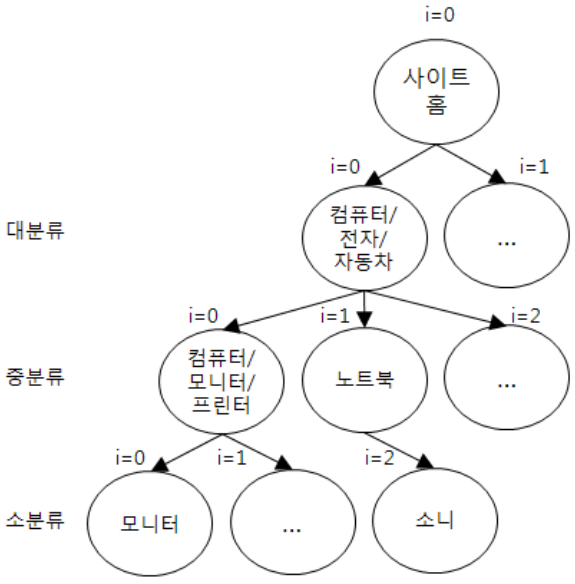

그림 2. 쇼핑몰 사이트의 카테고리 트리
Fig. 2. Category tree of shopping mall site

그림 1의 카테고리 페이지(L=1)는 그림 2처럼 3개 정도의 카테고리(대,중,소분류) 레벨로 구성 되지만 이 3개를 묶어 하나의 레벨로 취급하도록 한다.

카테고리로부터 상품평까지 접근을 시도할 때 내부적으로는 하이퍼링크에 의해 이동한다. 따라서 링크를 따라 다니면서 웹 문서를 수집하는 웹 크롤러를 이용하면 자동으로 상품평을 수집할 수 있을 것이라 예상할 수 있다. 그러나 한국내 쇼핑몰 페이지는 HTML 문서뿐만 아니라 다양한 웹 언어(AJAX, JavaScript, Flash)로 구성되어 있어 링크를 추출하기가 쉽지 않다. 또한 그림 1에서 표현한 단계의 링크 외에 광고 링크, 광고와 연결된 회사 사이트의 링크 등이 혼재해 있다?. 따라서 상품평을 수집하기 위해서는 카테고리 정보로부터 상품평이 존재하는 링크까지 정확하게 따라갈 수 있도록 사이트의 구조 정보와 URL 링크 정보를 분석하여 래퍼를 설계해야 한다. 래퍼에 명시해야할 데이터는 그림 1의 구조로부터 각 단계별로 추출해야 할 정보의 위치와 다음 단계로 이동하기 위한 링크의 추출 위치이다.

그림 3은 그림 1의 최상위에 있는 '사이트 홈' 페이지의 구조를 표현한다. 여기에서 '상위 정보'는 메타 태그(meta tag), 스타일 태그(style tag), 자바스크립트(javascript), AJAX 등으로 구성되어 있다. '본문 정보'는 상품의 대분류 카테고리로 접근하기 위한 카테고리명과 링크 등을 포함한다. 이외에 인기 상품 리스트, 신 상품 리스트, 광고 페이지 등을 포함할 수 있다. '하위 정보'는 저작권 관련 정보와 회사 소개 등을 포함할 수 있다.

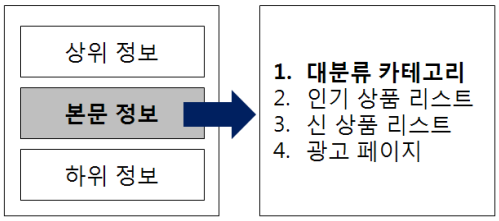

그림 3. 쇼핑몰 페이지의 구조
Fig. 3. The Structure of a shopping mall page

그림 3의 '대분류 카테고리' 부분은 상품평을 찾기 위한 5단계 중 첫 단계에 접근하기 위한 데이터를 포함한다. 래퍼의 궁극적인 목표는 각 단계로 접근하면서 5번째 단계에서 표현하는 상품평을 추출하는 것이다. 이를 위해 첫 번째 단계부터 마지막 단계까지 각 단계별로 필요한 데이터의 시작 위치와 끝 위치, 다음 단계로 이동하기 위한 링크의 시작 위치와 끝 위치를 래퍼의 규칙으로 정의해야 한다.

다음 절에서는 본 절에서 분석한 사이트 구조를 기준으로 한국형 쇼핑몰 내의 상품평 페이지만을 방문할 수 있도록 해당 위치와 관련 태그 정보를 포함할 래퍼(이하 K-Wrapper)에 대해 기술한다.

4. K-Wrapper

본 절에서는 3절에서 분석한 한국내 쇼핑몰 사이트의 구조로부터 상품평을 추출하기 위한 래퍼를 제안한다. 래퍼는 에이전트 또는 웹 로봇을 이용하여 정보를 추출하기 위해 각 문서에 대해서 추출하고자 하는 정보의 위치와 구조, 포맷 등에 관한 데이터를 포함하고 있어야 해당 페이지만 정확히 방문할 때 사용할 수 있다. 래퍼 데이터는 Context-free 문법을 기반으로 한 언어로 표현한다.

4.1 K-Wrapper의 데이터베이스 기술 문법

한국 쇼핑몰 내에서 상품평 데이터를 추출하기 위해서는 다양한 웹 언어(AJAX, JavaScript, Flash)로 표현된 각 단계를 거쳐야 하고, HTML 문서로 표현된 5번째 단계에서 상품평 데이터를 찾아야 한다. 일반적인 웹크롤러는 링크 주소가 포함된 HTML 문서에는 접근할 수 있지만, 링크 주소가 감춰 있는 다양한 웹 언어로 표현된 문서에는 접근이 불가능하다. 이러한 문서에 접근하기 위해서는 쇼핑몰 페이지의 코드를 분석하고 링크를 추출하여야 한다. 여기에서 추출한 링크 정보가 래퍼 데이터베이스로 저장된다.

5번째 단계에 존재하는 상품평 데이터는 HTML 문서로 표현되기 때문에 데이터의 위치는 주로 태그와 태그 내 데이터 값의 연속으로 구성된다. 이 데이터 명세는 Context-free 문법 G를 기반으로 한 언어로 표현한다.

정의 1 : 알파벳 T

심벌들의 유한 집합으로 아래와 같은 T_1, T_2, T_3가 알파벳으로 사용될 수 있다. T_1은 한글 자음과 모음의 집합이다. T_2는 영어 알파벳 대소문자와 숫자로 이루어진 집합이다. T_3은 HTML에서 사용되는 특수문자의 집합이다.

$$T_1 = \{ㄱ, ㄴ, ㄷ, ..., ㅎ, ㅏ, ㅑ, ..., ㅡ, ㅣ\}$$

$$T_2 = \{A, B, C, ..., Z, a, b, c, ..., z, 0, 1, 2, ..., 9\}$$

$$T_3 = \{!, @, \$, \%, :, /, ..., <, >, ?\}$$

정의 2 : 스트링 W

알파벳 T_1, T_2, T_3에 속하는 하나 이상의 심벌들을 나열한 것이다. 정의 1의 T_2로부터 만들 수 있는 스트링의 예는 다음과 같다.

$$W = a, b, c, ..., aa, ab, ba, ...$$

정의 3 : 문법 G

Context-free 문법 G는 V_N, V_T, P, S로 구성된다.

$$G = \{V_N, V_T, P, S\}$$

V_N은 nonterminal 심벌의 유한 집합으로 문법에서 스트링을 생성하는데 사용되는 중간 과정의 심벌이다. V_T는 terminal 심벌의 유한 집합이며 알파벳으로 구성된 스트링 심벌이다. P는 언어의 생성 규칙이고 S는 생성 규칙의 시작 심벌이다.

정의 1, 2, 3을 바탕으로 K-Wrapper 데이터베이스 기술 언어 문법 G_K를 정의한다.

$$G_K = (\{S, A, B, C\}, \{u, c_1, c_2, n_1, n_2, t_1, t_2, t_3, t_4, t_5\}, P, S)$$

문법 G_K에서 V_N은 $\{S, A, B, C\}$으로 구성되고, V_T는 $\{u, c_1, c_2, n_1, n_2, t_1, t_2, t_3, t_4, t_5\}$으로 구성된다.

문법 G_K의 생성 규칙 P는 다음과 같이 정의한다.

P: $S \rightarrow A$
$A \rightarrow t_1 u B A \mid \varepsilon$
$B \rightarrow t_2 c_1 C t_3 c_2 B \mid \varepsilon$
$C \rightarrow t_4 n_1 t_5 n_2 C \mid \varepsilon$

S는 문법 G의 정의에 따라 시작 심벌을 의미한다.

V_T집합에 포함된 각 원소는 특정 데이터를 추출하기 위한 시작 위치와 끝 위치를 정의하는 스트링으로 래퍼의 핵심이다. 이에 대한 자세한 내용은 다음에서 기술한다.

스트링 u : 수집할 상품평이 존재하는 사이트의 홈페이지 URL 주소이다. URL 주소는 주로 영문과 숫자로 이루어지기 때문에 알파벳 T_2의 심벌로 구성된다.

$$u = \text{http://www.shop.co.kr}$$

스트링 c_1 : 단계별로 추출하려는 데이터의 시작 위치이다. 웹 페이지는 HTML 문서로 이루어지고 데이터의 시작 위치는 주로 태그 값으로 이루어진다. 따라서 스트링 c_1은 알파벳 T_1, T_2, T_3의 심벌로 구성된다.

$$c_1 = \text{<h1>}$$

스트링 c_2 : 단계별로 추출하려는 데이터의 끝 위치이다. 구성되는 스트링 값의 형태는 c_1과 동일하다.

$$c_2 = \text{</h1>}$$

스트링 n_1 : 하위 단계로 이동하기 위한 링크의 시작 위치이다. HTML 문서 상에서 링크는 '<A HREF=' 로 시작하여 '>'로 끝나거나 공백, 기호 등으로 끝나기 때문에 알파벳 T_2, T_3의 심벌로 구성된다.

$$n_1 = \text{<a href=}$$

스트링 n_2 : 하위 단계로 이동하기 위한 링크의 끝 위치이다. 구성되는 스트링 값의 형태는 n_1과 동일하다.

$$n_2 = \text{>}$$

스트링 $t_1 \sim t_5$: 각 스트링 값의 구분 스트링이다. 본 논문에서 정의한 u, c_1, c_2, n_1, n_2는 연속된 스트링으로 표현된다. 또한 c_1과 c_2, n_1과 n_2 는 각각 쌍으로 하여 반복적으로 나타난다. 따라서 각 스트링 값을 구분할 수 있는 스트링이 필요하다. $t_1 \sim t_5$는 알파벳 T_3의 심벌로 구성되며, 각 스트링 값을 구분할 수 있도록 각 스트링 값의 앞에서 시작 위치를 표시한다. $t_1 \sim t_5$에 정의된 스트링 값은 다음과 같다.

$$t_1 = \$,\ t_2 = \#,\ t_3 = @,\ t_4 = @@,\ t_5 = \#\#$$

t_1의 스트링 값인 $는 u의 시작 위치를 표시한다. t_2의 스트링 값인 #은 c_1의 시작 위치를 표시한다. t_3의 스트링 값인 @는 n_1의 시작 위치를 표시한다. t_4의 스트링 값인 @@는 n_2의 시작 위치를 표시한다. 마지막으로 t_5의 스트링 값인 ##은 c_2의 시작 위치를 표시한다.

그림 4는 그림 1의 단계 1(L=1)에 해당하는 페이지를 HTML 코드로 표현한 것이다. 이를 분석하여 구축한 래퍼 데이터베이스는 그림 5에서 보여준다. 이는 문법 G_K를 이용하여 표현한 것이다. 본 논문에서는 지면 관계상 단계 1부터 단계 5에 해당하는 문서로부터 래퍼 데이터베이스 베이스를 구축하는 과정을 설명하지 않고, 단계 1에 해당하는 문서로부터 래퍼 데이터베이스를 구축하는 과정만 설명한다.

그림 5와 같은 래퍼 데이터베이스에는 쇼핑몰 페이지의 URL 주소와 각 단계별로 하위 단계에 접근하기 위해 필요한 정보의 위치를 가지고 있어야 한다. 그림 4로부터 그림 5와 같은 래퍼 데이터베이스의 구축 과정에서 쇼핑몰 페이지의 URL 주소(예 http://www.shop.co.kr)는 직접 입력하면 되지만, 하위 단계로 접근하기 위한 위치 정보는 그림 4의 코드를 분석함으로써 래퍼 데이터베이스의 구축이 가능하다. 그림 4는 단계 1에 해당하는 페이지이므로 여기서 추출해야 할 정보는 단계 2의 문서로 접근하기 위한 링크 정보이다. 분석 과정을 통해 그림 4에서는 '<!--카테고리 시작-->'의 주석 태그 밑에 존재하는 링크 주소가 단계 2의 문서로 접근하기 위한 것임을 알 수 있다. 따라서 이 스트링 값을 그림 5의 래퍼 데이터베이스에 추가한다. 이는 문법 G_K의 c_1에 해당한다. 그 다음으로 단계 2의 문서에 접근하기 위한 링크 주소의 시작 위치인 '<A HREF='와 링크 주소의 끝 위치인 '>'를 래퍼 데이터베이스에 추가한다. 이는 각각 G_K의 n_1과 n_2에 해당한다. 마지막으로 단계 2의 문서로 접근하기 위한 링크만 추출할 수 있도록 그 범위를 한정시킬 수 있는 끝 위치인 '<!--카테고리 끝-->'을 래퍼 데이터베이스에 추가한다. 이는 문법 G_K의 c_2에 해당한다.

문법 G_K에서 정의하고 있는 c_1과 c_2는 특정 범위의 링크만 추출할 수 있도록 한정시키는 역할을 한다. 만약, 이 범위가 존재하지 않는다면 그림 4의 '<!--인기 제품 리스트 시작-->' 밑에 존재하는 링크까지 접근하여 원하지 않는 정보까지 수집할 수 있다.

```
<HTML>
<STYLE></STYLE>
<SCRIPT></SCRIPT>
<BODY>
...
<!--카테고리 시작-->
<A HREF=category.php?level=1&code=1>컴퓨터</A>
<A HREF=category.php?level=1&code=2>가전</A>
<!--카테고리 끝-->

<!--인기 제품 리스트 시작-->
...
<A HREF=list.php?code=112233>의류</A>
<A HREF=list.php?code=112235>자동차</A>
<!--인기 제품 리스트 끝-->
Copyright by ABC Company
<BODY>
</HTML>
```

그림 4. 쇼핑몰 페이지의 예(HTML)
Fig. 4. An example of a shopping mall page(HTML)

```
$http://www.shop.co.kr#<!--카테고리
시작-->@<A HREF=@@>##<!--카테고리 끝-->
```

그림 5. G_K 문법으로 기술한 래퍼의 예
Fig. 5. An example of wrapper represented by G_K

4.2 래퍼를 적용한 상품평 수집 프로세스

본 절에서는 K-Wrapper를 이용해 상품평을 추출하는 크롤링 과정을 설명한다.

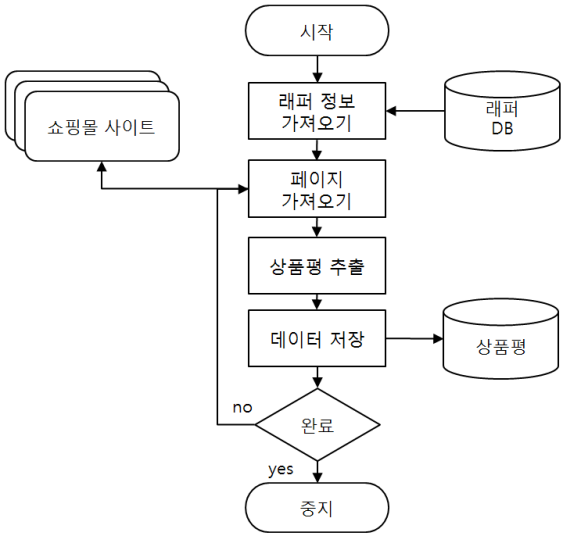

그림 6. 래퍼 기반 크롤러의 프로세스
Fig. 6. Process of wrapper based crawling

최초에 크롤러는 래퍼 DB로부터 그림 5에서 정의한 래퍼 정보를 로드한다. 그런 후에 래퍼에 정의 되어있는 특정 사이트를 방문해 '페이지 가져오기'를 수행한다. '상품 평 추출' 단계는 가져온 페이지로부터 래퍼의 규칙을 기반으로 상품평을 추출한다. 이때, 함께 추출되는 다음 계층의 이동 링크는 별도로 데이터베이스에 저장 되지 않는다. 궁극적으로 크롤러는 상품평을 수집하기 위한 것이고,중간 중간에 추출되는 다음 링크 정보는 상품 평을 추출하기 위한 중간 단계이기 때문이다. 또한 링크가 바뀔 수도 있기 때문에 그 상황에 맞게 링크를 추출하도록 한다. 추출된 상품평은 '데이터 저장' 단계에서 상품평 DB에 저장된다. '완료' 단계에서는 래퍼에 정의된 사이트를 모두 방문하여 수집하면 종료한다.

5. 구현 및 성능 평가

래퍼 기반 웹 크롤러는 자바 1.6, Eclipse SDK 3.4 환경에서 개발되었다. 크롤러를 통해 수집된 상품평은 사이트 종류, 카테고리 종류를 구분지어 MySQL 5.0 Database에 저장하였다.

그림 7은 본 논문에서 제안하는 래퍼 기반 크롤러 프로그램의 데이터 구조체를 나타낸다.

그림 8은 그림 7의 데이터 구조체를 바탕으로 특정 사이트에 대한 1 단계에서 추출해야 하는 데이터의 위치와 다음 단계로 이동하여야 할 링크의 위치 값을 할당하는 코드를 보여준다. 이는 시스템 메모리에 저장하는 예이다. 만약, 래퍼로 구축하여야 할 데이터의 양이 증가하면 그림 5의 형태로 래퍼 DB(그림 6)에 저장하여 활용해야 한다.

LoadModel() 에서는 그림 7에서 정의한 클래스를 메모리에 할당하고, V_T집합의 원소로 이루어진 c_1, c_2, n_1, n_2에 위치 정보를 할당한다.

```
class Data {
        String  c1;
        Vector n1;
        Vector n2;
        String  c2;
}

class Model {
        Site Site_A;
        Site Site_B;
        Site Site_C;
        Site Site_D;
}

class Site {
        String u;
        Data Level_1;
        Data Level_2;
        Data Level_3;
        Data Level_4;
        Data Level_5;
}
```

그림 7. 래퍼의 데이터 구조체
Fig. 7. The data structure of a wrapper

```
public Model LoadModel() {

    Model wrapperModel = new Model();

    wrapperModel.Site_A = new Site();
    wrapperModel.Site_A.Level_1 = new Data();
    wrapperModel.Site_A.Level_1.n1 = new Vector();
    wrapperModel.Site_A.Level_1.n2 = new Vector();
    wrapperModel.Site_A.Level_1.c1 = new String();
    wrapperModel.Site_A.Level_1.c2 = new String();

    wrapperModel.Site_A.Level_1.c1
       = "<!--카테고리 시작-->";
    wrapperModel.Site_A.Level_1.c2
       = "<!--카테고리 끝-->";
    wrapperModel.Site_A.Level_1.n1.add("<A HREF=");
    wrapperModel.Site_A.Level_1.n2.add(">");

    /* 이하 생략 */
    return wrapperModel;
}
```

그림 8. 래퍼 코드의 예
Fig. 8. An example of wrapper code

표 1. 수집된 상품 리뷰
Table 1. Product reviews collected

		모니터	노트북	디지털 카메라	MP3 플레이어
A	제품 수	245	165	148	113
	상품평 수	8,724	3,010	4,365	19,732
	소요 시간	36	17	19	64
B	제품 수	176	135	31	114
	상품평 수	3,919	2,413	1,954	11,329
	소요 시간	18	11	6	30
C	제품 수	104	264	40	504
	상품평 수	250	8,681	1,531	23,702
	소요 시간	14	114	70	198
D	제품 수	109	59	66	60
	상품평 수	12,535	6,785	7,590	6,900
	소요 시간	28	15	18	17

표 2. 기존 연구와의 특징 비교
Table 2. Comparing features of K-Wrapper

	포커스드 크롤러[6]	포커스드 크롤러[7]	래퍼기반 크롤러[11]	래퍼기반 크롤러[12]	K-Wrapper
수집 대상 문서의 형식	HTML	HTML	HTML	HTML	HTML, Flash, AJAX, JavaScript 적용 문서
수집 대상 문서의 계층*	1계층	1계층	1계층	2~3계층	5계층
실험에 사용한 문서	yahoo.com에서 제공하는 20개의 주제와 동일한 문서	ODP**제공하는 30개의 주제와 동일한 문서	1) 학회 홈페이지의 정보 2) 구인 광고 웹문서 3) 온라인 서점에서 책의 상세 정보	yahoo.com에서 제공하는 '자동차' 카테고리에 있는 신상품 정보	국내 대형 쇼핑몰 4곳의 상품 평 문서
수집 시간	초당 1.6개의 URL 수집	N/A	N/A	N/A	초당 2.98건의 상품평 수집
재현율	실험 문서에 대한 재현율 평균30~70%	실험 문서에 대한 재현율 평균 35%	실험 문서에 대한 재현율 100%	실험 문서에 대한 재현율 100%	실험 문서에 대한 재현율 100%

* 목표 데이터를 얻기 위해 계층적으로 접근하는 의존적 구조를 의미함. 예를 들어 쇼핑몰에서 특정 상품의 상세 정보를 알고 싶다면 '상품 카테고리->상품 리스트->상품 상세 정보' 순으로 접근해야 한다. 일반적인 웹 페이지에 존재하는 URL 링크도 계층적 구조를 갖는다고 할 수 있으나, 포커스드 크롤러는 계층에 관계 없이 문서 분류기를 이용하여 원하는 문서만 수집하기 때문에 이를 1계층이라 볼 수 있다.

** Open Directory Project(www.dmoz.org)

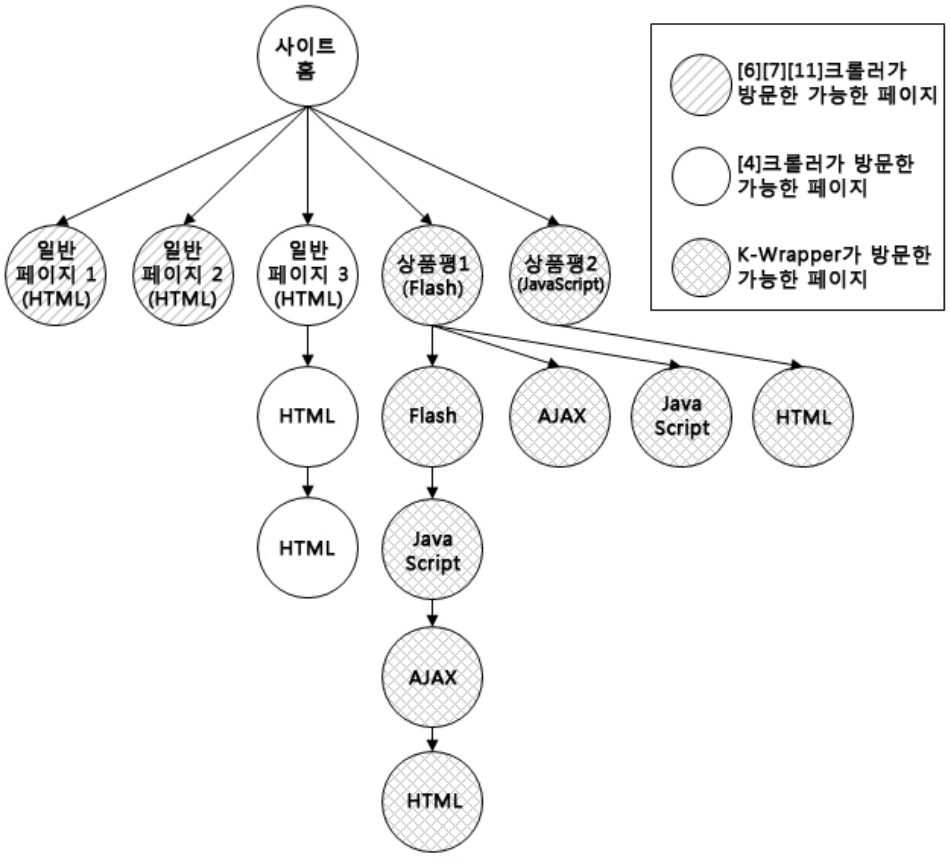

그림 9. 각 크롤러의 접근 가능한 구조
Fig. 9. The structure of accessible page for each crawler

표 1은 쇼핑몰 사이트(A, B, C, D)별 카테고리별로 수집한 상품평 현황을 보여준다. 표 1에서 제품 수는 수집한 제품의 개수이고, 상품평 수는 개별 제품에 대한 여러 개의 상품평 수를 나타낸다. 소요시간은 상품평을 수집하기 위해 소요된 시간으로 단위는 '분(min)'이다. 표 1에 나타난 수집 현황은 수집할 당시의 현황이기 때문에 새롭게 수집을 시도할 경우 표 1의 내용과 다를 수 있다.

제안하는 래퍼의 우수성을 입증하기 위해, 기존에 제안된 일반적인 웹크롤러와 포커스드/토피컬 크롤러를 이용하여 상품평 문서를 얼마나 정확하고 빠르게 수집할 수 있는지 정량적인 비교가 수행되어야 한다. 그러나 본 논문에서 관심 대상으로 하는 한국형 쇼핑몰(A-D)사이트는 다양한 웹 언어로 구성되어있기 때문에 일반적인 웹크롤러나 포커스드/토피컬 크롤러로는 상품평을 수집할 수 없다. 따라서 표 2와 같이 수집 대상 문서의 형식,수집 대상 문서의 계층, 실험에 사용한 문서, 수집시간, 재현율을 비교하였다. 그 결과, 본 논문에서 제안하는 래퍼는 기존에 제안된 크롤러로는 접근 불가능한 한국형 쇼핑몰 사이트에 접근하여 상품평을 수집할 수 있었다. 이는 한국형 쇼핑몰 사이트를 분석하고 일반화한 것으로 한국형 쇼핑몰에 적용할 수 있는 래퍼를 기반으로 한 것이다. 그림 9는 기존에 제안된 포커스드 크롤러, 래퍼 기반 크롤러, 그리고 제안하는 K-Wrapper가 방문 가능한 페이지를 보여준다. K-Wrapper는 다양한 웹 언어로 구성된 5계층의 웹 문서에 접근할 수 있음을 알 수 있다.

6. 결 론

본 논문에서는 다 계층 구조와 다양한 웹언어로 구성된 한국형 쇼핑몰로부터 상품평 수집을 위한 래퍼 기반 크롤링 모델을 제안하였다.

기존에 제안된 포커스드 크롤러는 원하는 문서를 찾기 위해 URL을 수집하고, 수집된 문서의 적합성을 판단하기 위해 분류기를 이용하였다. 이러한 이유로 문서 수집에 비교적 많은 시간을 필요로 하며, 문서 분류기의 정확도가 떨어지는 경우는 원하지 않는 문서도 수집할 수 있다. 이러한 문제를 해결하기 위해 제안된 것이 래퍼기반 크롤러인데, [6][7]에서는 주로 HTML로 작성된 문서만을 수집 대상으로 하였다. 또한 그 계층 구조는 2~3계층 정도에 그쳤다.

그러나 한국 내 쇼핑몰은 그 구조가 5 계층 정도로 구성되어 있고, 다음 계층으로 이동할 수 있는 링크 경로가 JavaScript, AJAX, Flash 등의 다양한 웹 언어를 사용하여 감춰져 있기 때문에 기존에 제안한 크롤러의 적용은 불가능하다. 이에 따라 한국형 쇼핑몰에서 상품평을 수집할 수 있는 일반화된 래퍼 모델(K-Wrapper)을 제안하게 되었다. 기존의 크롤러는 접근이 불가능한 한국형 쇼핑몰과 같은 구조를 접근할 수 있는 것을 확인하였다. 실제로 K-Wrapper를 이용하여 2,333개의 상품에 대한 123,420건의 상품평을 수집하는데는 약 11시간 30분 정도의 시간이 소요되었다. 실제로 여타 방식의 크롤러는 해당 사이트에 접근이 어려우므

로 시간적으로 비교는 할 수가 없었다.

참 고 문 헌

[1] P.Tuerny, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” *In Proc. of the Meeting of the Association for Computational Linguistics(ACL'02)*, pp.417-424, 2002

[2] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques,” *In Proc. of the Conference on Empirical Methods in Natural Language Processing(EMNLP'02)*, pp.79-86, 2002

[3] M. Hu and B. Liu, “Mining and Summarizing Customer Reviews,” In Proc. of ACM SIGKDD '04, pp.168-177,2004

[4] M. Hu and B. Liu, “Mining Opinion Features in Customer Reviews,” *In Proc. of the 19th National Conference on Artificial Intelligence(AAAI'04)*, pp. 755-760, 2004

[5] Bing Liu, *Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data*, Springer, pp. 273-289, 2007

[6] S. Chakrabarti,M. van den Berg, and B. Dom, “Focused Crawling : A New Approach to Topic-Specific Web Resource Discovery,” *Computer Networks*, Vol.31, No. 11-16, pp.1623-1640, 1999

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

[8] S. Chakrabarti, *Mining the Web. Discovering Knowledge from Hypertext Data*, Morgan Kaufmamm, pp. 257-287, 2003

[9] J. Cho, H. Garcia-Molina, and L. Page, “Efficient Crawling through URL Ordering,” *Computer Networks*, Vol.30, No.1-7, pp. 161-172, 1998

[10] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, and Khaled Shaalan, “A Survey of Web Information Extraction Systems,” *IEEE Transaction on Knowledge and Data Engineering*, Vol.18, No. 10, pp.1411-1428, 2006

[11] Jaeyoung Yang, Tae-Hyung Kim, and Joongmin Choi, “An Interface Agent for Wrapper-Based Information Extraction,” *In Proc. of the International Confenrence on Principles of Practice in Multi-Agent Systems(PRIMA'04)*, pp.291-302, 2004

[12] Claudio Bertoli, Valter Crescenzi, and Paolo Merialdo, “Crawling programs for wrapper-based applications,” *In Proc. of IEEE International Conference on Information Reuse and Integration(IRI'08)*, pp.160-165, 2008

[13] Stephen Soderland, Claire Cardie, and Raymond Mooney, "Learning information extraction rules for semi-structured and Free Text," *Machine Learning*, Vol. 34, No.1-3, pp.233-272, 1999

[14] Hanhoon Kang, Seong Joon Yoo, Dongil Han, “Modeling Web Crawler Wrappers to Collect User Reviews on Shopping Mall with Various Hierarchical Tree Struture,” *In Proc. of the International Conference on Web Information Systems and Mining(WISM '09)*, pp. 69-73, 2009

[15] http://autos.yahoo.com/new_cars.html

저 자 소 개

강한훈 (Hanhoon Kang)

2006년 : 세종대 컴퓨터소프트웨어학과(학사)
2008년 : 세종대 컴퓨터공학과(석사)
2008년～현재 : 세종대 컴퓨터공학과 박사 과정

관심분야 : 데이터 마이닝, 오피니언 마이닝, 기계학습, 웹 크롤러
E-mail : kangcom@paran.com

유성준 (Seong Joon Yoo)

1982년 : 고려대 전자공학과(학사)
1990년 : 고려대 전자공학과(석사)
1996년 : 시라큐스대 전산학과(박사)
2002년～현재 : 세종대학교 컴퓨터 공학과 교수

관심분야 : 러닝 시스템, 패턴 인식, 데이터마이닝, 이미지 프로세싱
Phone : 02-3408-3755
E-mail : sjyoo@sejong.ac.kr

한동일 (Dongil Han)

1988년 : 고려대 전자전산공학과(학사)
1990년 : 한국과학기술원 전기 및 전자공학과(석사)
1995년 : 한국과학기술원 전기 및 전자공학과(박사)
1995년 2월 ～ 2003년 2월 : LG전자 디지털TV연구소 책임연구원
2003년 ～ 현재 : 세종대학교 컴퓨터공학과 교수

관심분야 : 영상 처리, 신호 처리, 컴퓨터 비전, 데이터 마이닝
E-mail : dihan@sejong.ac.kr