# 비교 쇼핑 에이전트를 위한 Wrapper의 자동생성 설계 및 구현

강준규, 김병만, 김주연, 임은기, 오득환
금오공과대학교 컴퓨터공학부
{jgkang,bmkim,jykim,eklim,dhoh}@cespc1.kumoh.ac.kr

# Design and Implementation of Automatic Wrapper Generation for a Comparison-Shopping Agent

Joon Gyu Kang, Byeong Man Kim, Ju Youn Kim, En Ki Lim, Dukhwan Oh
Dept. of Computer & Software Engineering, Kumoh National University of Technology

## 요 약

본 논문에서는 비교 쇼핑 에이전트에 필수 요소인 wrapper를 자동으로 생성하는 방법에 대하여 제안한다. 상품 정보들을 추출하기 위하여 사용자로부터 URL을 입력받고, cgi URL, 질의 형식(Get 또는 Post), 입력 매개변수들, 검색된 HTML 페이지에서 출력 위치들(상품명, 모델명, 판매가...)을 추출한다. 또한, 상품명을 이용하여 검색을 실시하고, 검색 결과 문서들을 분석하여 상품가격을 추출하는 규칙을 생성하며, 생성된 규칙을 WIDL로 기술하여 데이터베이스에 저장한다.

## 1. 서론

인터넷 기술의 발전과 더불어 사이버 공간의 쇼핑몰을 이용하는 전자상거래가 급속도로 확산되고 있으며, 규모가 점차 거대해지고 쇼핑몰의 상품 정보가 다양해지고 있다[1]. 따라서 쇼핑몰을 이용하는 사용자의 요구 조건에 가장 적합한 상품에 대한 정보를 신속하고 정확하게 제공해주는 방법이 필요하게 되었으며, 이러한 시대적 요구에 따라 여러 온라인 쇼핑몰에서 제공하는 상품 정보를 소비자의 요구 조건을 고려하여 비교 검색할 수 있는 비교 쇼핑 에이전트가 필요하게 되었다[2].

비교 쇼핑 에이전트는 여러 온라인 쇼핑몰의 정보 제공 포맷을 분석하고 필요한 정보를 통합하여 사용자에게 제공함으로서 사용자가 상품에 관한 벤더를 비교 할 수 있도록 도와주는 에이전트이다. 그러므로, 이러한 쇼핑 에이전트에서는 특정한 형식으로 쓰여진 웹 페이지에서 필요한 정보를 추출할 수 있는 전용 프로그램인 wrapper가 필수 요소이며[3], 이러한 wrapper를 수동으로 작성할 경우 많은 시간적 낭비와 오류가 발생할 수 있으므로 자동으로 wrapper를 생성할 필요가 있다.

본 논문에서는 상품 정보를 추출하고자 하는 쇼핑몰의 URL을 이용하여 자동으로 *wrapper*를 생성하는 효과적인 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 wrapper 생성시 고려 사항에 대하여 기술하고, 제3 장에서는 자동 wrapper 생성에 관하여 설명하며, 제4 장에서는 결론 및 향후 연구 과제를 제시한다.

## 2. Wrapper 생성의 관련 연구 및 고려 사항

상품 정보를 추출하고자 하는 쇼핑몰의 URL을 이용하여 지속적으로 변경되는 상품 정보를 가장 효과적으로 탐색할 수 있는 방법은 쇼핑몰에서 제공되는 검색 서비스를 이용하는 것이며, 이러한 검색 서비스를 이용하여 상품 정보를 자동으로 추출하기 위해서는 아래와 같은 사항들을 고려하여야 한다.

1) 검색 정보 추출 : 검색 서비스를 이용하기 위해서는 시스템의 URL, Method, Parameter등을 자동으로 추출하는 방법

2) 질의 구성 : 각 쇼핑몰에서 제공되는 검색 시스템들은 질의를 구성하는 방법이 다양하므로, 시스템에 적합한 형태의 질의를 자동으로 구성하는 방법.

3) 검색 결과 분석 : 결과 문서들로부터 필요한 상품 정보들을 효과적으로 추출할 수 있는 방법.

Wrapper *생성에 관한 기존 연구들은 위에서 지적한 문제점 중에* 주로 세 번째 문제에 초점을 두어 연구되고 있으며, induction 기법을 바탕을 둔 학습 방법[4]을 통하여 추출 패턴을 생성하고 있다. 이러한 연구들의 대표적인 예로는 WIEN[5], STALKER[6]등이 있으며, 이러한 연구에서는 정형적인 방법을 사용함으로써 나름대로의 장점은 있지만 실제 학습 데이터를 구성하는데 많은 노력과 비용이 소요된다.

Induction 기법을 이용한 방법과는 달리 Wrapper 생성시에 도움을 줄 수 있는 도구[7]가 개발되고 있으며, Webmethod사의 Web Automation Toolkit이 그러한 예이다. 이 도구에서는 GUI를 통하여 추출할 정보를 입력 받고, 정보를 WIDL[8]라는 XML 기반 언어로 변환한 후 이를 바탕으로 Wrapper를 자동 생성하고 있다.

위에서 언급했듯이 기존 방법들이 상품 정보를 포함하는 페이지에서 필요한 정보를 추출하는 방향으로만 주로 연구하고 있으며, 그 단계, 예를 들어 상품정보 검색 입력 폼을 자동으로 파악하고 입력 값에 맞는 질의를 생성하는 등의 연구는 전무한 상태이다.

본 연구에서는 이러한 전 단계의 작업을 고려하여 상품 검색 입력 폼에 대한 정보와 검색 페이지에서 상품 정보를 추출하는데 필요한 정보를 자동으로 추출하는 방법을 제시하였다. 이러한 정보는 WIDL 형태로 기술되어 Webmethod사의 Toolkit을 통하여 자동으로 Wrapper가 생성되게 된다.

## 3. Wrapper의 자동생성 설계 및 구현

### 3.1 전체 구성

쇼핑몰의 상품 정보들은 지속적으로 변경되고, 이러한 변경된 정보들을 효과적으로 탐색하기 위해서는 해당 쇼핑몰에서 제공하는 검색 서비스를 이용하는 것이다. 따라서 본 논문에서는 해당 쇼핑몰에서 검색 서비스를 제공한다고 가정하였으며, 제안하는 시스템의 전체 구성은 그림 1과 같다.

상품 정보를 추출하고자 하는 쇼핑몰의 URL이 입력되면 질의 생성기에서는 URL의 모든 링크 주소를 저장하고, 이들 링크 주소의 일부 문서들로부터 해당 쇼핑몰에서 제공하는 상품명을 추출한다. 입력 분석기에서는 저장된 링크 주소를 이용하여 검색 서비스를 제공하는 문서를 탐색하고, 검색 서비스 이용에 필요한 정보들을 추출한다. 또한, 질의 생성기에서 생성된 상품명을 이용하여 해당 쇼핑몰에서 제공하는 검색 시스템에 질의를 수행하여 적합한 형태의 검색 결과인 경우, 결과 문서와 검색 정보를 규칙 생성부로 전달한다. 규칙 생성기에서는 입력 분석기를 통해 전달된 결과 문서들에서 상품 가격을 추출하는 규칙을 WIDL로 생성하여 DB에 저장하며, 통신처리기는 각 처리기에서 요구한 문서들을 탐색한다.

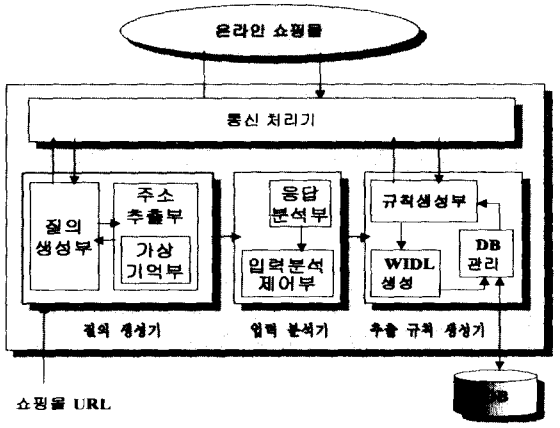


그림 1 시스템 전체 구조

### 3.2 질의 생성기

검색 서비스 이용에 필요한 정보 탐색 및 질의를 구성하기 위한 질의 생성기는 그림 1에서와 같이 질의 생성부, 주소 추출부, 가상 기억부로 구성된다.

주소 추출부는 상품 정보를 추출하고자 하는 URL에 연결된 주소들에서 광고주소와 실제 쇼핑몰내 링크 주소를 구분하기 위하여 입력된 쇼핑몰 주소와 동일한 도메인에 포함되는 링크 주소만을 추출하여 중복을 제거한 후, 질의 생성부와 입력 분석기의 분석 제어부에서 사용할 수 있도록 가상기억부에 순서적으로 저장한다.

질의 생성부에서는 링크된 일부 문서를 조사하여 가격 정보가 나타나는 테이블에서 상품명을 추출한 후 입력 분석기로 전달한다. 이때, 가격 정보를 판단하는 기준은 분석된 패턴 안에 텍스트가 숫자로만 구성되어있고 ','으로 구분되는 경우와 '원'이라는 문자로 끝나는 숫자, 또는 두 가지 경우 모두 해당하는 경우로 한다.

### 3.3 입력 분석기

쇼핑몰에서 제공하는 검색 시스템들은 지원되는 형태가 다양하므로 단일한 검색 정보만을 이용하여 질의를 수행할 수 없다. 따라서 그림 1에서와 같이 입력 분석기는 검색 정보를 추출하고 상품명을 질의로 사용하여 검색을 수행하는 입력 분석 제어부와 검색 결과가 질의된 상품에 관한 정보인지를 판별하는 응답 분석부로 구성하였다.

입력 분석 제어부에서는 각 시스템의 고유한 검색 정보들을 추출하기 위하여 상품에 관한 검색을 지원하는 문서를 탐색하고, Action, Method, Parameter등과 같이 시스템에 질의를 수행하는데 필요한 검색 정보들을 추출한다. 또한, 추출된 검색 정보들이 상품 검색에 적용될 수 있는지를 판단하기 위하여 추출된 검색 정보와 질의 생성기에 의해 생성된 상품명을 질의로 사용하여 검색을 수행한다. 이때, 검색 정보를 추출하기 위하여 본 논문에서는 아래와 같은 알고리즘과 제약 조건을 사용하였다.

**<수행 알고리즘>**

① 문서에서 Form Tag를 탐색
② Form에서 제약 조건을 비교
③ 만약, 제약 조건에 해당할 경우 다음 주소의 문서를 요구하고 ①을 반복.
④ 만약, 제약 조건에 해당되지 않을 경우 Action의 주소, 매개형식, Method 정보들을 추출
⑤ 검색 정보와 상품명을 질의로 사용하여 검색을 수행

**<제약 조건>**

① Input Type이 TEXT인 수가 두 개 이상인 경우
② TEXT TYPE이 PASSWORD인 경우
③ ACTION이 기존 검색 엔진과 동일한 경우
  http://www.naver.com
  http://www.simmani.com
  http://www.yahoo.co.kr
  http://www.lycos.co.kr
  http://search.netpia.com
④ ACTION이 다음과 같은 상품 검색 엔진인 경우
  http://www.shopbinder.com
  http://www.yavis.com
  http://www.martguide.co.kr
⑤ ACTION이 'mailto'로 시작하는 메일 관련인 경우
⑥ 게시판 검색인 경우

응답 분석부에서는 분석 제어부에서 수행한 검색 결과 문서를 분석하여, 상품에 관한 결과로 판단될 경우 검색 정보와 결과 문서를 규칙 생성기로 전달한다. 이때, 상품에 관한 정보는 테이블 형태로 기술된다고 가정하고, 상품에 관한 문서로 판단하는 규칙은 아래와 같다.

**<상품 문서 판단 규칙>**

① 가격 항목이 한 개 이상 반드시 존재하고, 상품명 항목이나 모델명 항목 중의 한 항목이 존재
② 가격으로 판단되는 항목이 규칙적으로 존재

**<각 항목 판단 규칙>**

① 가격 : 문자가 '가,격,액'으로 끝나는 경우(단, '규격'은 제외)
② 상품명 : 문자가 '명,목,품'으로 끝나는 경우(단, '설명', '사명', '업명'으로 끝나는 단어는 제외)
③ 모델명 : 문자가 '모델'로 시작하는 경우

### 3.4 규칙 생성기

규칙 생성기는 그림 1에서와 같이 규칙 생성부와 WIDL 생성부, DB 관리부로 구성된다.

규칙 생성부는 결과 문서를 분석하여 상품에 관한 정보가 기술된 규칙성을 파악하고, 추출 규칙을 생성하여 WIDL 생성부로 전달하는 기능을 수행한다. 이때, 추출 규칙은 상품 정보가 기술된 테이블에 헤드가 존재할 경우와 존재하지 않은 경우를 분리하여 생성하며, 상품명, 모델명, 판매가, 소비자가등의 정보가 포함된다.

**1) 헤드가 존재할 경우의 추출 규칙 생성 방법**

① 상품명, 모델명, 가격 정보가 기술된 위치 파악
② 만약, 가격 정보가 1개 존재할 경우 판매가로 설정
③ 만약, 가격 정보가 2개 존재할 경우 소비자가, 판매가로 설정
④ 만약, 가격 정보가 2개 존재하고 tag가 <STRIKE>로 표시된 가격은 소비자가로 설정

**2) 헤드가 존재하지 않을 경우의 추출규칙 생성 방법**

① 테이블의 행 정보를 나타내는 패턴 코드를 생성
- 공백일 경우 '0'으로 설정
- 문자열일 경우 '1'을 설정
- 가격 정보일 경우 '2'로 설정
② 만약, 패턴 코드에서 '1'이 1개만 존재할 경우 그 위치를 상품명으로 설정
③ 만약, 패턴 코드에서 '1'이 1개 이상 존재할 경우 ④-⑧을 수행
④ 임의의 필드에 존재하는 문자열을 추출
⑤ DB에 저장된 기존의 쇼핑몰 검색 시스템에 검색
⑥ 결과 문서에서 동일한 문자열을 탐색
⑦ 만약, 동일한 문자열이 탐색될 경우 추출 규칙 생성
⑧ 만약, 탐색되지 않을 경우 ④를 반복

WIDL 생성부는 추출 규칙들을 WIDL로 기술하고, 기술된 내용을 DB관리부로 전달하는 기능을 수행한다. 또한, DB관리부에서는 정보를 저장하거나, 필요한 정보를 탐색하는 기능을 수행한다.

### 4. 실험 및 결과 분석

본 논문에서는 성능을 평가하기 위하여 표 1에서와 같이 상품 정보를 담일 테이블 형태로 제공하는 쇼핑몰들 중에서 50개를 대상으로 실험하였다.

실험 결과 표 1에서와 같이 본 논문에서 제안하는 방법을 이용할 경우, 84%에 해당하는 42개의 사이트에서 wrapper를 자동으로 생성할 수 있었다. 또한, 실패한 경우를 분석해 보면 입력 분석기에서 JavaScript와 같은 스크립트를 분석하지 못하는 경우와 검색 시스템이 외부 쇼핑몰에 존재하여 검색 정보를 추출하지 못한 경우가 대부분을 차지하였다.

### 5. 결론 및 향후 연구 과제

본 논문에서는 상품정보를 추출 이전 단계에 해당되는 상품정보를 검색하기 위한 입력 폼을 자동으로 파악하고 입력 폼에 맞는 질의를 생성하는 것과 같은 작업까지 자동으로 수행한 후 필요한 정보를 추출하는 방법을 제시하고자 하였다. 비록, 방법이 직관에 의존하여 비정형적인 면은 있으나 다른 연구의 기초가 되는데 의미가 있다 하겠다.

향후의 연구과제로는, 일정한 형식이 아닌 일반 텍스트로 구성된 웹 문서에서 wrapper를 자동으로 생성할 수 있는 방법이 연구되어야 하며, 쇼핑몰에서 제공하고 있는 광고 주소와 실제 쇼핑몰내 링크 주소를 구분할 수 있는 방법이 연구 되어져야 한다. 또한, 상품과 관련된 다양한 정보들을 추출할 수 있는 방법이 연구되어야 한다.

### 참고 문헌


[1] Jae Kyu Lee, Young Uk Song, Jae Won Lee, "A Comparison Shopping Architecture over Multiple Malls : the Meta-Malls Architecture", ICEC '98, pp149-154, 1998.

[2] 비교쇼핑에이전트, http://cse.hanyang.ac.kr/~jmchoi/agents/comp-shopping.html

[3] William W. Cohen, "Recognizing Structure in Web Pages using Similarity Queries", American Association for Artificia Intelligence, 1999.

[4] William W. Cohen, "Recognizing Structure in Web Pages using Similarity Queries", American Association for Artificial Intelligence, 1999.

[5] Nicholas Kushmerick, Daniel S.Weld, Robert Doorenbos,"Wrapper Induction for Information Extraction", IJCAI-97(Nagoya), 1997.

[6] William W. Cohen, "Recognizing Structure in Web Pages using Similarity Queries", American Association for Artificial Intelligence, 1999.

[7] webMethods, "Web Automation Toolkit 2.1 User's Guide", 1977.

[8] Rohit Khare, Adam Rifkin, "XML: A Door to Automated Web Applications", IEEE INTERNET COMPUTING, pp78-87, July, August ,1997.


표 1 실험 결과

| 주소 | 결과 | 실패 원인 |
|---|---|---|
| http://www.altoran.com/ | | |
| http://www.youngflower.com/ | 실패 | 입력 분석기 오류 |
| http://www.autoworld.co.kr/cgi-bin/cybershop | | |
| http://www.webtown.co.kr/ | | |
| http://www.kookmincard.co.kr/shopping/ | | |
| http://www.easyclub.co.kr/ | | |
| http://www.joylet.com/ | | |
| http://www.kmall.com/ | | |
| http://www.hotmart.co.kr/ | | |
| http://www.himart.co.kr/ | | |
| http://www.koreacenter.co.kr/ | | |
| http://www.babymall.co.kr/babymall/ | | |
| http://www.dm-zone.com/ | 실패 | 입력 분석기 오류 |
| http://www.epost.go.kr/ | 실패 | 입력 분석기 오류 |
| http://www.booknet.co.kr/ | | |
| http://www.cardeco.com/ | | |
| http://www.ectrack.com/ | | |
| http://www.epostop.com/ | | |
| http://www.unibook.co.kr/ | | |
| http://www.i778.co.kr/mf.htm | 실패 | 입력 분석기 오류 |
| http://www.igift.co.kr/htm/igift.htm | | |
| http://www.yongho.co.kr/ | 실패 | 입력 분석기 오류 |
| http://www.lghs.co.kr/ | | |
| http://www.pisa.co.kr/shop/shop.dll | | |
| http://www.toy.co.kr/html/thome.html | | |
| http://www.i39.co.kr/ | | |
| http://www.withcomputer.com/main.asp | | |
| http://www.auction.co.kr/auction.htm | | |
| http://shop.metaland.com/ | | |
| http://www.refills.co.kr/ | | |
| http://www.bookpark.com/bookindex.html | | |
| http://www.et-land.co.kr/ | 실패 | 입력 분석기 오류 |
| http://www.khimall.com/ | | |
| http://www.gamemart.co.kr/ | | |
| http://www2.provin.kangwon.kr/frame_2.shtml | | |
| http://www.sports999.co.kr/index.html | 실패 | 입력 분석기 오류 |
| http://www.golagola.co.kr/ | | |
| http://www.cyberskg.co.kr/ | 실패 | 검색 결과 가격 오류 |
| http://www.newbook.co.kr/ | | |
| http://www.mukmul.com/ | | |
| http://www.mediaclub21.co.kr/shop/shop.dll | | |
| http://www.musicplaza.co.kr/ | | |
| http://www.bigsale.com/default.asp | | |
| http://mall.buddhapia.co.kr/ | | |
| http://sandulip.co.kr/index.cgi | | |
| http://www.cworld.co.kr/ | | |
| http://www.sejin.co.kr | | |
| http://www.supply.co.kr/ | | |
| http://www.shinbimall.com/ | | |
| http://www.dialog.co.kr/ | | |