비교 쇼핑 정보 수집을 위한 멀티 에이전트 시스템

신주리[0] 전중남 이건명
충북대학교 컴퓨터과학과, 첨단정보기술 연구센터
edutopia@aicore.chungbuk.ac.kr

A Multi-Agent System for Collecting Comparative Shopping System

Ju-Ri Shin[0] Joong-Nam Jeon Keon-Myung Lee
Dept. of Computer Science, Chungbuk national University and AITrc

요 약

인터넷 상의 많은 전자 상거래 쇼핑몰에 있는 상품 정보에 대한 비교 서비스를 제공하는 시스템들이 개발되고 있다. 이러한 서비스를 위해서는 분산된 전자 상거래 쇼핑몰들의 정보를 수집하여 통합하는 노력이 필요하다. 이 논문에서는 멀티 에이전트 구조로 설계한 인터넷 상의 쇼핑몰들로부터 상품 정보를 수집하여 서비스하는 시스템에 대해서 소개한다. 이 시스템에서는 랩퍼 생성 서브시스템, 정보 수집 서브시스템, 카테고리 분석 서브시스템, 데이터 정제 서브시스템 등의 구성 요소들이 유기적으로 결합되어 동작한다. 이 논문에서는 전체적인 시스템의 구성에 대해서 살펴보고, 각 서브시스템의 기능 및 구조에 대해서 기술한다. 또한 쇼핑몰로부터 정보를 추출하기 위한 랩퍼 생성 기법과 상품 정보의 카테고리를 결정하는 방법에 대해서 소개한다

1. 서론

인터넷 사용의 보편화와 더불어 인터넷 기반의 전자 상거래를 통한 경제 활동이 급격히 확산됨에 따라 전자 상거래를 위한 정보 제공자들의 노력 또한 활발해지고 있다. 국내의 인터넷 쇼핑몰(shopping mall)을 통한 정보 제공자들도 늘고 있는 추세이다. 이로 인해 인터넷에서 상품을 구매하려는 구매자들은 상품을 구매하기 위해서는 각 사이트에서 상품의 검색과 나름대로의 검색한 결과에 대해 비교 종합하여 상품을 구입한다. 소비자가 여러 쇼핑몰을 검색하여 상품을 비교하여 구매하는 불편을 덜 수 있도록 하기 위해서 상품 비교를 자동으로 해줄 수 있는 여러 비교 쇼핑 시스템들이 개발되어 서비스 중에 있다.

비교 쇼핑 시스템들은 쇼핑몰들로부터 상품 정보를 수집하여 자체 데이터베이스를 구축하고, 사용자의 질의에 대해서 자체 데이터베이스로부터 정보를 검색하여 제공한다. 이러한 비교 쇼핑 시스템의 구축에 가장 큰 문제가 되는 것은 쇼핑몰들로부터 효과적으로 상품 정보를 수집하는 것이다. 대부분의 상용 서비스 중인 비교 쇼핑 시스템들은 각 쇼핑몰 별로 상품 정보를 추출하는 프로그램을 개발하거나, 쇼핑몰과 계약하여 데이터베이스를 직접 입수하는 방법을 사용하고 있다. 쇼핑몰과 계약을 통하는 경우에는 쇼핑몰에서 해당 비교 쇼핑 시스템만 접속할 수 있는 포트나 URL을 제공하는 방식으로 정보를 수집한다.
비교 쇼핑 시스템이 정보 수집 대상으로 하는 쇼핑몰들은 고객 유치를 위해 빈번하게 홈페이지 디자인 및 구조를 변경하고 있다. 또한 비교 쇼핑 자체가 쇼핑몰의 경쟁력을 악화시킬 수 있다는 우려 때문에 많은 대형 쇼핑몰들은 비교 쇼핑 시스템에 상품 데이터베이스를 제공하지 않으려는 추세로 가고 있다. 따라서 비교 쇼핑 서비스를 제공하기 위해서는 대부분의 경우 각 쇼핑몰이 구조 또는 디자인이 변경될 때마다 해당 쇼핑몰에 대한 정보 수집을 위한 프로그램을 개발하는 부담을 갖게 된다. 이러한 문제에 효과적으로 대처하기 위한 방안으로 랩퍼(wrapper)를 자동으로 생성하여 변하는 쇼핑몰의 정보를 효과적으로 수집하려는 연구들이 활발히 수행되고 있다.

이 논문은 첨단 정보기술 연구센터(AITrc)를 통해서 과학재단의 지원을 받은 것임.

이 논문에서는 비교 쇼핑 서비스를 위해 제안한 멀티 에이전트 기반의 상품 정보 수집 시스템에 대해서 기술한다. 제안한 시스템은 랩퍼 생성 서브 시스템, 정보 수집 서브시스템, 카테고리 분석 서브시스템, 데이터 정제 서브시스템 등 네 개의 서브 시스템으로 구성되고 있고, 이들 각각 서브 시스템은 에이전트 구조로 구성되어 있다. 이 논문에서는 제안된 시스템의 구조와 각 서브 시스템의 내부 구조에 대해 설명하고, 제안한 시스템에서 채택한 랩퍼 학습 방법과 카테고리 분류 방법에 대해 설명한다.

2.시스템 구조와 특성

2.1 랩퍼 생성 서브시스템

랩퍼 생성 서브 시스템은 주어진 페이지에 대한 랩퍼를 생성하는 시스템이다. 여기에서 사용된 랩퍼 생성 방법은 추출 정보에 레이블이 붙은 페이지로부터 정보 추출을 위한 패턴을 추출하는 학습 방법을 이용한다 패턴의 형태는 HTML TAG와 휴리스틱적인 요소를 포함시키기 위한 프로시져로 구성된다. 한편, 추출 정보에 대한 레이블 정보를 부여하기 위해 단순한 GUI 환경을 제공한다.

2.2 정보 수집 서브 시스템

정보 수집 서브시스템은 랩퍼 생성 시스템에서 학습한 랩퍼를 사용하여 쇼핑몰로부터 상품 정보를 추출하는 역할을 한다. 정보 수집 서브시스템은 여러 개의 에이전트로 구현되는데, 각 쇼핑몰 별로 정보를 크롤링하여 수집하는 랩퍼 에이전트가 있고, 이러한 랩퍼 에이전트를 스케쥴에 따라 활성화시키고 종료시키는 랩퍼 관리 에이전트가 있고, 랩퍼 에이전트들로부터 수집된 정보를 통합하여 데이터베이스에 저장하는 정보 통합 에이전트가 있다.

2.3 카테고리 분석 서브시스템

비교 쇼핑 정보 서비스를 위해서는 상품 정보에 대한 수집뿐만 아니라 상품 정보가 어떤 카테고리에 속하는지 판별할 수 있어야 한다.

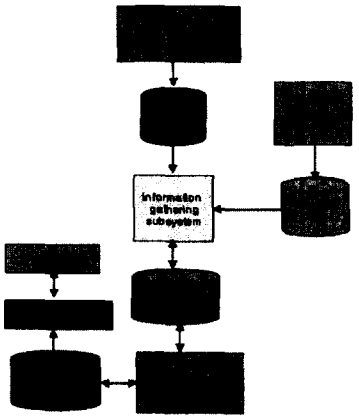

그림 1. 비교 쇼핑 정보 수집 시스템 구조

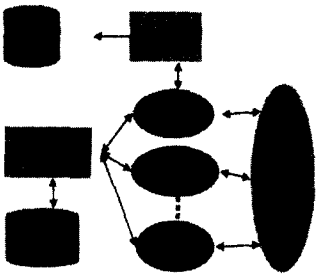

그림 2. 정보 수집 서브 시스템 내부 구조

또한 쇼핑몰 사이트를 크롤링하면서 어떤 페이지에 상품 정보가 있는지 판정하는 것이 필요하다.카테고리 분석 서브시스템에서는 주어진 쇼핑몰을 크롤링하면서 정보가 있는 페이지를 판별하고, 해당 페이지에 있는 상품 정보가 어떤 카 테고리에 속하는지 판정하는 역할을 한다. 카테골리 분류 서브 시스템은 [그림 3]과 같이 쇼핑몰 사이트를 크롤링하여 상품정보가 있는 페이지를 찾는 역할을 하는 크롤링 에이전트와 주어진 페이지에 있는 상품 정보로부터 해당 상품들의 카테고리를 학습하는 카테고리 학습 에이전트가 있다.

2.4 데이터 정제 서브시스템

상품에 대한 비교 쇼핑 정보를 제공하기 위해서는 동일 또는 유사 상품을 분류하여 정보를 제공하는 것이 바람직하다. 동일한 상품이더라도 대부분의 쇼핑몰들에서 추출되는 정보는 일정하지 못하다. 상품명의 표현방법, 제조회사의 표현방법 등이 사이트별로 차이가 나기 때문에 비교 쇼핑 정보를 제공하기 위해서는 상품명, 모델, 제조회사명 등을 통일하는 작업이 필요하다. 이러한 역할을 하는 부분이 데이터 정제 서브시스템으로 랩퍼 에이전트들이 추출해놓은 상품 정보를 가공하는 역할을 한다. 여기에서는 표준명칭을 생성관리하고 이를 데이터베이스화하여 다음 상품 정보 갱신 단계에서 사용할 수 있도록 하는 일종의 온톨로지 관리 역할을 한다. 또한 상품 데이터로서 잘못 추출된 것을 제거하고, 이를 로그 파일로 관리하여 랩퍼의 유효성을 검증하는데 사용하도록 한다. 한편, 카테고리 불량 데이터에 대해서는 오류 카테고리 리스트를 작성하여 이 정보를 카테고리 분석 서브시스템에서 사용할 수 있도록 한다. 데이터 정제 서브시스템에서는 휴리스틱을 사용하여 오류 데이터를 여과하고, 표준 명칭 데이터베이스를 활용하여 명칭를 통일하는 작업을 수행하지만, 많은 부분에서 실제 관리자가 판여를 해야 하는 부분이다. 이를 위해서 그래픽 기반의 인테페이스를 제공한다

2.5 질의처리기와 사용자 인터페이스

사용자 인터페이스는 웹을 통해서 비교쇼핑 정보 검색을 서비스하는 부분이고, 질의처리기는 랩퍼 등을 통해서 수집되고 가공된 데이터를 가지고 있는 상품 정보 데이터베이스로부터 사용자가 질의한 단순 상품 검색과 비교 상품 검색을 지원하는 부분이다

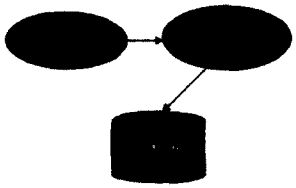

그림 3. 카테고리 분석 서브시스템 구조

2.6 랩퍼 학습 방법

랩퍼는 특정 사이트에서 관심있는 정보를 추출하는 프로시쥬로서, 패턴의 형태로 표현되거나 실제 기본적인 프로시쥬들로 구성된 프로시쥬 형태를 갖는다. 여기에서 제안한 방법은 전처리 부분에 휴리스틱에 기반한 프로시쥬를 사용하고, 실제 학습된 랩퍼는 패턴으로 표현되는 것이다. 대부분의 주요 인터넷 쇼핑몰의 페이지 형태를 보면 기본적으로 테이블을 이용하여 구성된다. 이는 실제 화면 레이아웃을 설계할 때 테이블을 사용하는 것이 편리하기 때문이다. 따라서 자유 문장(free text)로 되어 있는 페이지로부터 정보를 추출하는 것보다는 비교적 쉽게 랩퍼를 생성할 수 있다. 쇼핑몰과 같이 반구조화된(semi-structured) 페이지에 대한 랩퍼 방법들이 최근 활발히 연구되고 있다.[1,2,3,5,6]

테이블 형태를 갖는 쇼핑몰의 페이지에서 상품정보를 가지고 있는 단위의 형태를 보면 [그림 5]와 같다. [그림 5]에서 빗금이 있는 부분이 하나의 상품정보를 기술하는 부분에 해당한다. [그림 5]의 (a)는 하나의 행이 하나의 상품 정보를 표현하는 전형적인 형태이고, (b)는 여러 개의 행에 걸쳐 하나의 상품 정보가 표현된 형태이다. (c)는 하나의 셀(cell)안에 하나의 상품 정보를 표현하고 있는 것으로, 상품 정보의 추출을 위해서는 테이블 정보 뿐만 아니라
 <P> 태그등과 같은 포맷 태그와 구두점에 대응하는 정보가 함께

그림 5. 상품 정보 단위의 형태

필요한 형태이다. 제안된 방법에서는 GUI를 통해서 관리자가 상품 정보에 관련된 속성 정보 위치를 마크해 주면, *HTML* 파서(*parser*)를 통해 문장 구조를 분석하고, 태그와 텍스트를 일반화된 태그와 마크로 변경한 다음, 반복되는 패턴들을 일반화시켜 랩퍼 패턴을 생성한다

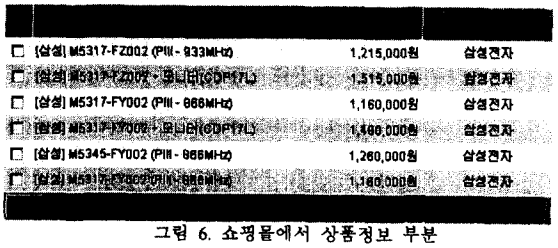

그림 6. 쇼핑몰에서 상품정보 부분

[그림 6]은 쇼핑몰에서 상품정보를 포함한 부분을 예를 보인 것으로, 이와 같은 쇼핑몰에 대해서 GUI를 통해서 상품명, 판매가, 제조업체 등의 위치정보를 저장하게 된다.

[그림 7]은 [그림 6]에 대한 실제 HTML 코드를 보인 것이다. 이러한 코드에 대해서 태그와 텍스트를 일반화시켜 반복되는 패턴 패턴들의 일반화된 형태를 추출하면 [그림 8]과 같은 랩퍼 패턴이 구해진다.

```
<a href="/jsp/mall/ViewPrdItm.jsp?ecpid=113196&ecsid=9948&
  ecstid=PersCom&hana=9948&image=sub_com.jpg&slide=SlideCom.
  jsp">[삼성] M5317-FZ002 (PIII - 933MHz)</a></td>
  <td align=right><font class="main">1,215,000원</font></td>
  <td align=center><font class="main">삼성전자</font></td> </tr> <tr
bgcolor="#edf3e5"> <td><input type="checkbox" name="ecpid"
value="113010"></td>
```

그림 7. HTML 소스 코드의 일부

```
<TD><A><URL><PRODUCT></A></TD><TD><PRICE></TD><T
D><FONT><COMPANY></FONT></TD></TR>
```

그림 8. 학습된 랩퍼 패턴의 예

2.7 카테고리 분류 방법

카테고리 분석은 상품 정보가 있는 페이지가 상품 카테고리 중에서 어디에 속하는지 결정하는 것을 말한다. 제안한 시스템에서는 카테고리를 대분류, 중분류, 소분류 카테고리로 나누어서 표현하고 있다. [그림9]는 이러한 카테고리 분류 예를 보인 것이다. 특정 페이지의 상품들에 대한 페이지를 부여하기 위해, 제안한 방법에서는 이미 구축되어 있는 상품 정보 데이터베이스를 사용하는 일종의 사례기반(instance-based) 학습방법을 이용한다. 상품 정보 데이터베이스에는 이미 카테고리가 부여되어 있는 많은 상품정보가 있다는 것을 전제한다. 카테고리를 결정하기 위해, 우선 해당 페이지의 상품정보를 랩퍼를 이용하여

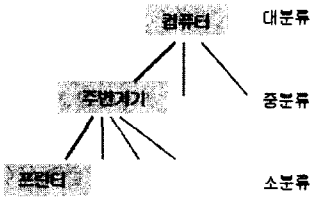

그림 9. 상품 카테고리 분류

추출한다. 추출된 상품이 상품 정보 데이터베이스에 존재하는지 검색하여 해당 상품에 대한 카테고리 정보를 추출한다. 추출된 카테고리 정보들에 대해서 대분류, 중분류, 소분류 카테고리 각각에 대해서 가장 많은 상품이 포함된 카테고리를 찾는다. 카테고리 정보가 없는 상품의 개수, 최다 상품을 갖는 카테고리에 포함되는 상품의 개수, 최다 상품의 카테고리와 다른 카테고리를 갖는 상품의 개수 등을 고려하여 페이지의 특정 카테고리에 대한 소속정도를 계산하고, 그 값이 임계값 이상의 카테고리가 존재하면 이를 해당 카테고리로 분류한다. 지정한 임계값에 미달하는 경우에는 대분류 또는 중분류까지 만 카테고리를 부여하고, 나중에 관리자가 수작업을 통해 카테고리를 부여할 수 있도록 한다. 이러한 카테고리 분류 방법은 초기 수작업이 다소 요구되지만 어느 정도 상품 데이터베이스가 축적되고 나면, 효과적으로 카테고리를 분류할 수 있도록 해준다.

3. 시스템 구현

제안한 시스템의 Windows 2000에서 MFC 라이브러리, ActiveX 컨포넌트 등을 사용하여 구현하고 있다. [그림 10]은 현재 프로토타입으로 구현된 랩퍼 생성 서브시스템을 보인 것이다. 랩퍼 생성 서브시스템은 [그림 10]의 오른쪽 상단 화면에서 보는 바와 같이 웹브라우저 컨포넌트를 이용하여 응용 프로그램안에서 쇼핑몰을 브라우징하면서, GUI를 통해 랩퍼 학습에 필요한 정보를 입력하고, 이 정보를 이용하여 랩퍼를 생성한다. 생성된 랩퍼를 이용하여 그림의 오른쪽 하단과 같이 실제 상품 정보를 추출하여 검증할 수 있는 기능을 가지고 있다. [그림 11]은 정보 수집 서브시스템의 인터페이스를 보인 것이다. 여기에서는 랩퍼 생성 서브시스템에서 생성한 랩퍼를 이용하여 쇼핑몰로부터 상품 정보를 수집하는 역할을 한다. 현재는 네트웍의 속도 때문에 한번에 하나의 쇼핑몰에서 정보를 수집하지만, 구조적으로는 여러 개의 사이트에서 에이전트들이 동시에 정보를 수집할 수 있도록 구현되어 있다.

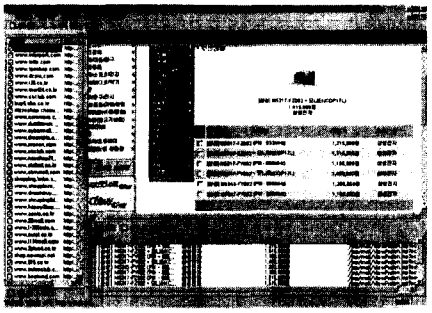

그림 10. 랩퍼 생성 서브시스템

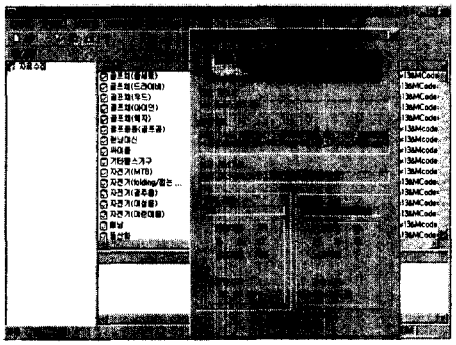

그림 11. 정보 수집 서브시스템의 인터페이스

4. 결론

인터넷 상의 정보를 통합하여 유용한 서비스를 하기 위한 여러가지 시도가 있어왔다. 이 논문에서는 이러한 정보 통합의 한가지 응용 분야인 비교 쇼핑 정보 서비스를 위한 시스템을 설계하고 구현한 내용을 소개하였다. 시스템은 기능적 모듈간의 독립성과 자율성을 향상시키기 위해 에이전트들로 설계하였다.

한편, 랩퍼 학습을 위해 반구조화된 테이블 형태의 웹 페이지 정보 추출에 유용한 태그 기반의 랩퍼 패턴 학습 방법을 제안하였다. 또한 특정 웹 페이지의 상품 정보에 대한 카테고리를 부여하는 방법으로 기 구축된 상품 데이터베이스를 이용하는 방법을 제안하였다. 향후 현재 구현된 프로토타입 시스템에서 효과적으로 대처하지 못하고 있는 자바 스크립트 등과 같은 스크립트를 포함한 사이트에 대한 효과적인 크롤링 방법과 비구조화된 정보도 추출할 수 있는 랩퍼에 대한 연구가 필요하다.

5. 참고 문헌

[1] I. Muslea, S. Minton, C. Knoblock, A hierarchical approach to wrapper induction, *Proc. of the 3rd Annual Conf. on Autonomous Agents*, pp.190-197, Seattle, 1999.

[2] N. Kushmerick, D. S. Weld, and R. Doorenbos, Wrapper Induction for Information Extraction, *Proc. of International Joint Conference on Artificial Intelligent* (IJCAI), Nagoya, Japan, 1997].

[3] L. Xiaoying, AutoWrapper: Automatic wrapper generation for multiple online, *Proc. of Asia Pacific Web Conference*, Hong Kong, 1999.

[4] T. Mitchell, *Machine Learning*, (ch.2), McGraw -Hill Co., 1997.

[5] M. E. Cliff, R. Mooney, Relational learning of pattern-match rules for information extraction, *Proc. of the 6th National Conf. on Artificial Intelligence* (AAAI-99), pp.328-334, Orlando, 1999.