전자상거래에서 정보추출 규칙과 Ontology생성을 위한 인터페이스 에이전트

서희경*, 양재영, 구남숙, 최중민
한양대학교 전자계산학과

An Interface Agent for Creating Information Extraction Rules and Ontology in Electronic Commerce

Heekyoung Seo, Jaeyoung Yang, Namsuk Koo, Joongmin Choi
Dept. of Computer Science and Engineering, Hanyang University

요 약

인터넷의 증가로 온라인 상점들의 수는 매우 빠르게 증가하고 있다. 상점의 수가 늘어날 수록 사용자가 이러한 상점들에서 원하는 정보를 찾는 일은 쉽지 않다. 사용자의 어려움을 줄이고자 여러 쇼핑몰의 정보들을 통합해서 보여주는 전자상거래 통합 시스템들이 생겨나고 있지만, 새로운 쇼핑몰이 추가될 때마다 관리자가 추가되는 쇼핑몰의 정보를 추출하기 위한 규칙이나, Ontology등을 수동으로 만들거나 확장해야 하기 때문에 사람이 소비해야 하는 시간과 노력이 많고, 시스템을 관리하는 사람에 따라 정보추출의 정확도도 다르다. 따라서 사람이 소비하는 시간을 줄이고, 좀더 정확한 정보추출을 위해 쇼핑몰마다 만들어야 하는 규칙과 그러한 규칙 생성에 필요한 Ontology를 자동으로 생성하는 방법과 이 방법에서 요구되는 사용자의 입력을 최소한 줄인 인터페이스 에이전트를 제안한다.

1. 서 론

최근 통신의 발달로 인터넷 사용이 증가하고, 인터넷상의 정보의 양도 매우 빠르게 증가하고 있다. 인터넷 사용자가 증가함에 따라 전자상거래의 수요가 많아지고 중요성이 부각되면서, 많은 수의 전자상거래 상점들이 생성되고 있다. 결과적으로 사용자가 수백 개의 전자상거래 상점들에서 취향에 맞는 상품을 찾는 일은 인터넷상의 많은 사이트들에서 원하는 정보를 찾는 일만큼 시간이 소비된다. 이러한 이유로, 원하는 상품의 정보를 찾기 위해 소비하는 사용자의 시간을 줄이기 위해 여러 전자상거래 상점들을 통합해서 그 결과를 사용자에게 보여주는 전자상거래 통합 시스템들이 생겨나고 있다. 이러한 통합 시스템들은 여러 상점들 각각에 대해서 정보추출 규칙을 생성해야 한다.

대부분의 통합 시스템들은 정보추출 규칙을 수동으로 생성하거나, 도메인 지식(Domain knowledge)을 사용해서 자동 생성[3][4]하기도하고, 다른 방법을 이용해서 자동 생성하기도 한다. 하지만, 도메인 지식을 이용해서 자동으로 정보추출 규칙을 생성하는 경우에 도메인 지식은 수동으로 생성하는 경우가 대부분이다. 다른 방법을 이용하는 경우에도 규칙이 제대로 생성이 되지 않는 경우 마지막 방법은 결국 수동으로 생성한 도메인 지식을 이용해야 한다. 따라서, 매번 상점이 추가될 때마다 시스템 관리자는 정보추출 규칙을 생성하거나, 도메인 지식을 추가해 주어야만 한다. 현재 수백 개의 전자상거래 상점들이 존재하고, 앞으로도 많은 수의 상점들이 만들어 질 것인데, 시스템 관리자가 이러한 일들을 수동으로 한다면 많은 시간과 노력이 요구될 것이다.

본 논문에서는 일정한 형식으로 구성된 전자상거래 상점들의 통합 시스템을 생성하거나, 관리하는데 있어서 드는 노력과 오류 발생을 줄이기 위해 Ontology를 이용하여 상품정보의 위치를 알아낸 후 정보추출 규칙을 생성하는 시스템을 제안한다. 본 논문에서 도메인 지식은 Ontology로 구성된다. 이러한 Ontology의 확장 및 관리는 지금까지 시스템 관리자들에 의해 수동으로 구축되었다. 이러한 문제의 극복을 위해 사용자로부터 최소한의 입력을 통한 Ontology의 자동생성 방법을 적용한 인터페이스 에이전트를 구현한다.

2. 관련연구

문법이 정확하고 간략한 텍스트로 구성된 문서에서 정보를 추출하는데 필요한 패턴을 학습해서 생성하는 시스템들에는 AutoSlog, LIEP, PALKA, HASTEN이 있지만, 자연어를 연구한 이러한 시스템들은 온라인상의 문서에 적용하기에 적합하지

않다. 따라서 온라인상의 문서를 추출하고 통합하기 위해 나타난 연구분야가 wrapper 생성분야이다. wrapper 생성에서 연구되는 시스템들은 HTML의 특성상 구분기호(Delimiter) 기반의 추출 패턴을 생성한다[2][5]. 이런 시스템으로 처음 개발 된 것이 WIEN[1]이다. WIEN은 많은 양의 가공되지 않은 인터넷 정보 소스 중 필요 없는 부분은 삭제하거나 필요한 부분의 정보를 추출하기 위한 wrapper를 생성하는 시스템이다. 이 시스템은 사용자가 HTML 문서의 텍스트를 선택하고 표시하기 위한 GUI도 제공한다. 이외에도 WIEN을 확장한 SoftMealy 시스템이 있다. 이런 시스템은 Training Set이 사용자에 의해 입력되어야만 올바른 추출 패턴을 생성할 수 있다. 반면, 본 논문에서는 사용자가 인터페이스에서 정보의 위치와 정보의 의미만을 시스템에게 알려주면 자동으로 규칙을 생성할 뿐 아니라, Ontology도 생성해주는 시스템을 제안한다.

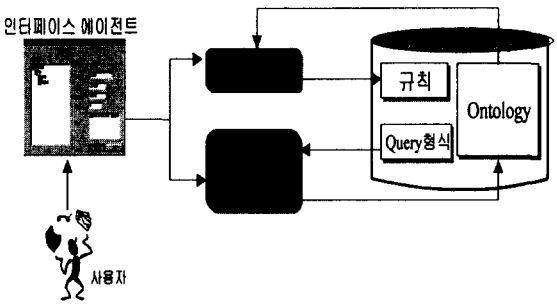

그림 1 시스템 구조도

3. 시스템 구조

- 사용자 인터페이스 : 사용자 인터페이스는 사용자로부터 직접 입력을 받아들이는 부분이다. 본 논문에서 제안하는 시스템에서 사용자 인터페이스의 기능은 사용자로부터 상품 정보를 입력받음과 동시에 상품위치 정보를 규칙 생성기에 기억하도록 한다. 따라서, 사용자로부터 최소한의 입력을 받으면서 위의 작업을 수행하기 위해서 사용자가 상품 정보를 마우스로 드래그한 후 오른 쪽 버튼을 누르게 되면, 그림 4의 인터페이스를 통해 정보를 입력받고 동시에, 드래그를 하는 사용자의 행동에서 상품정보 위치를 알아낸다.
- 구조분석기 : 인터페이스 에이전트에 의해서 규칙 생성기가 실행이 되면, 구조 분석기는 지정한 상점의 키워드 검색한 결과 페이지에서 상품위치 정보를 알아낸 후 정보추출 규칙을 생성한다. 구조분석기에서 생성한 결과는 도메인 지식의 규칙 부분에 추가된다. 규칙생성부분이 완료되면, Ontology생성기가 규칙생성기에서 도메인 지식에 추가한 규칙을 이용한다.
- Ontology 생성기 : Ontology 생성기는 Ontology의 기본구조를 사용자로부터 입력을 받아서 구성하고, 도메인 지식에 포함되어 있는 규칙과 Query형식을 이용해서 Ontology를 확장한다. 이렇게 구성 된 Ontology는 사용자가 일정 형식을 가진 전자상거래 상점을 추가하는 경우 상품위치 정보를 알아내 자동 규칙 생성에 이용된다.

4. 규칙 및 Ontology 생성

4.1 규칙생성

인터페이스 에이전트는 사용자로부터 상품정보를 입력받는 것과 동시에 규칙생성기를 실행해서 규칙생성기가 상품 위치 정보를 기억하게 한다. 규칙생성기는 기억된 상품 위치 정보를 바탕으로 그 상점에 대한 정보추출 규칙을 생성한다.

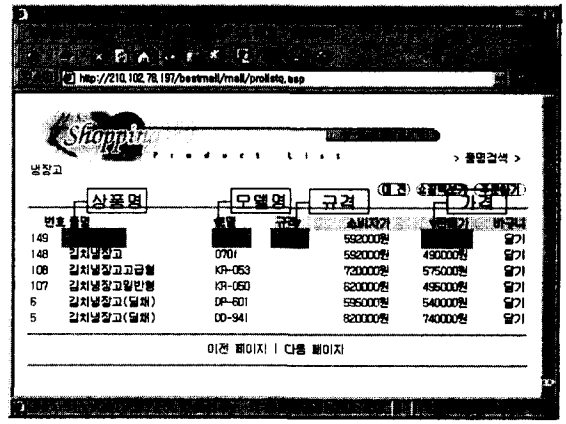

그림 2 베스트몰

위의 그림 2는 국내 전자상거래 상점 베스트몰에서 냉장고라는 상품명으로 검색 한 결과이다. 화살표로 표시된 부분은 사용자 입력으로 규칙생성기가 기억하고 있는 상품 위치 정보이다. 이러한 상품 위치정보를 토대로 베스트몰 정보추출 규칙을 생성하게 되면 다음과 같이 된다.

start <table[3] : 결과로 가져온 페이지 중에서 처음 추출해야할 정보가 나타나는 시작 부분을 표현한 규칙이다. <table[3]이라고 나타낸 것은 <table tag가 여러 번 존재하기 때문에 세 번째 <table부터 시작부분이라는 것을 나타낸다.

search2<table#/table>& : 추출 정보의 시작부분부터 끝나는 부분까지를 추출하라는 규칙이다. 즉 <table부터 /table> 사이의 정보를 추출한다.

search1<tr>#& : 각각의 상품 정보를 구분하는 규칙이다. 여기서는 <tr>로 시작해서 다음 <tr>이 나올 때까지가 한 상품에 대한 정보이다.

item_search[7] <td>#<td>#<td>#<td>#<td>#<td>#<td>#<td>& : 상품에 대한 각각의 정보별로 구분해서 추출한다. item_search[다음에 오는 숫자가 그 쇼핑몰에 존재하는 상품에 대한 정보의 수이고, #로 표시된 부분이 추출돼야 하는 부분이다.

item_number[2:상품명;3:모델;4:규격;6:가격]&
item_search[7]에서 추출된 결과 중에서 2, 3, 4, 6 번째 정보가 사용자가 입력한 상품위치 정보라는 규칙이다.

이 규칙은 table 형태일 때를 나타내는 규칙이고, table이 아니지만 일정한 규칙을 가지는 쇼핑몰에 대해서는 위의 것과는 다른 규칙이 생성된다.

Ontology가 어느 정도 확장이 된 상태일 때, 사용자가 상점을 지정해주고 이 상점이 일정한 형식으로 구성되어 있다면, 규칙 생성기는 Ontology만으로 상품위치 정보를 파악해서 해당 홈페이지에 알맞은 정보추출 규칙을 생성한다.

4.2 Ontology 생성

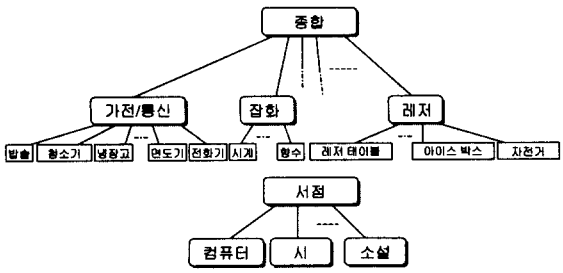

그림 3 Ontology 구조

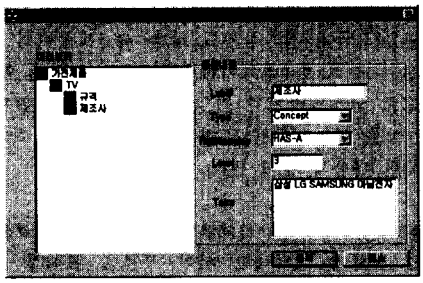

그림 4 Ontology 등록

• 초기 Ontology 생성 및 확장: Ontology를 자동으로 생성하기 전에 인터페이스 에이전트를 통해 사용자로부터 Ontology에 기본적인 구조를 입력받는다. 사용자는 특정 상점에서 제공하는 하나의 상품에 대해서만 이러한 정보를 입력한다. 이 정보를 바탕으로 구조분석기에서 특정 상점에서 Ontology 자동 생성을 위한 구조를 분석하여 규칙을 생성한다. 구조분석기에 의해 Ontology의 자동생성이 실행되면 사용자가 정보를 입력한 특정 상점에 다양한 예제로 질의를 한다. 상점에서 되돌려주는 검색결과를 구조분석기가 제공하는 규칙을 통해서 Concept이나 Term들을 획득하고 확장한다.

• Ontology 구성: 본 논문에서 생성한 Ontology는 Concept과 Concept간의 관계를 통해 Ontology를 확장한다. Concept간의 관계는 IS-A와 HAS-A관계로 나눈다. IS-A관계는 현재의 Concept과 동의어 정도로 볼 수 있고 HAS-A관계는 현재 Concept에 종속되는 용어들로 볼 수 있다. 예를 들면, 그림 4에서 제조사 Concept의 HAS-A 관계는 "삼성, LG, SAMSUNG, 아남전자,…"이고 IS-A 관계는 "메이커, Maker, Manufactured, …"등이다. 이러한 Ontology 구성은 간단한 관계만으로도 효율적으로 전자상거래를 위한 Ontology를 구축할 수 있다.

5. 결론 및 향후 연구 계획

본 논문에서는 전자상거래 통합 시스템에서 새로운 상점이 생성될 때 이 도메인 지식도 확장되어야 하는데, 이러한 도메인 지식의 확장을 Ontology와 규칙 생성기를 통해 자동으로 수행한다. 이렇게 확장된 도메인 지식은 상점 통합 시스템에서 사용자가 원하는 정보의 위치를 찾아내기 위해 사용된다.

향후 연구계획은 상점 통합 시스템에 자동으로 확장된 도메인 지식을 이용하여 사용자의 취향을 학습하여 상품 추천이 가능한 에이전트에 대해 연구하고자 한다.

[참고문헌]

[1] Kushmerick N., Weld D., Doorenbos B., "Wrapper Induction for Information Extraction", IJCAI-97 (Nagoya), 1997

[2] Ion Muslea, "Extraction Patterns for Information Extraction Tasks: A Survey", The AAAI-99 Workshop on Machine Learning for Information Extraction, 1999

[3] Robert B. Doorenbos, Oren Etzioni, Daniel S. Weld, "A Scalable Comparison-Shopping Agent for the World-Wide Web", ACM Autonomous Agents97, 1997

[4] Steve Lawrence, C.Lee Giles, Kurt Bollacker, "Digital Libraries and Autonomous Citation Indexing", IEEE Computer, Volume 32, Number 6, pp. 67-71, 1999

[5] Stephen Soderland, "Learning Information Extraction Rules for Semi-structured and Free Text", Machine Learning, Kluwer Academic Publishers, pp. 1-44, 1999