전자부품관련 웹 정보 마이닝 에이전트 (WIMA-EE)

오석일[0] 변영태
홍익대학교 전자계산학과
{sioh, byun}@cs.hongik.ac.kr

A Web Information Mining Agent for Electrical Elements (WIMA-EE)

Seok-Il Oh[0] Young-Tae Byun
Dept. of Computer Science, Hong-Ik University

요 약

웹에서 공개하는 정보의 많은 부분이 문자에 의존해서 제공되고 있으며, 이렇게 단어의 여러 형태로 구성된 웹 문서에서 원하는 정보를 찾아 추출하기 위한 노력은 다양하게 시도되고 있다. 본 논문에서는 전자부품관련 정보 제공 사이트와 관련해서 텍스트 기반과 웹 문서가 갖는 특별한 형태의 태그를 포함하는 형태에서 테이블 형식의 정보 표현과 같이 반 구조적(semi-structured) 문서에서의 정보 추출 방법과 이를 적용한 시스템을 구성하여 정보 추출의 가능성을 제시하고자 한다.

1. 서론

인터넷은 이제 정보를 제공하고 얻는 방법으로 일반화되었으며, 이를 통한 정보의 제공은 모든 분야에서 다양한 형태의 구조와 함께, 텍스트, 이미지, 동영상, 사운드 등 복잡한 형태로 제공되고 있다. 이렇게 무수히 많이 제공되고 있는 정보들 중에 찾고자 하는 정보를 발견하기 위한 노력은 많은 인내를 필요로 하게 된다. 관련된 문서를 찾고자 하는 노력은 검색엔진 프로그램을 통해서 이루어지게 되지만 검색된 문서들을 통해서 원하는 정보만을 추출해서 정리하는 데는 더 많은 작업을 필요로 하게 된다. 여기에서 웹을 통한 정보 마이닝 에이전트 기술이 필요하게 되며, 이와 유사 관련된 연구들이 다각적으로 활발하게 이루어지고 있다.
웹을 통한 다양한 정보의 종류들 중에 텍스트가 대부분을 이루고 있으며 이러한 문자정보 중에 원하는 정보를 추출하는 시도로 본 논문에서는 확장된 전기실험관련 가상인공과학실험실(Artificial Science Laboratory for Electrical Experiments II)[1]과 관련내용으로 회로 구성에 사용되는 전기, 전자 부품과 관련된, 예를 들어, 부품 명, 제조회사, 사용용도, 가격정보 등의 내용을 웹을 통해서 추출하여 제공하는 것을 목적으로 전자부품관련 웹 정보추출 에이전트(A Web Information Mining Agent for Electrical Elements : 이하 WIMA-EE) 시스템의 설계 및 구축에 대한 내용으로 기술을 한다.

2장에서는 정보 마이닝에 대한 배경지식 및 WIMA-EE의 추출 규칙을 소개하며, 3장에서는 WIMA-EE의 전체 구조와 각각의 모듈에 대해서 설명하고, 4장에서의 실행형태로 복잡한 정보추출의 가능성을 살펴보며, 5장의 결론에서 천체 시스템에 대한 평가를 하는 것으로 끝을 맺는다.

2. 관련 연구

WIMA-EE를 구성하는데 있어서 관련된 몇 가지 배경지식을 소개한 후에 시스템의 개요 및 전체 구성을 설명하기로 한다.

본 연구는 뇌 과학 연구 사업의 지원으로 진행 되었음.

2.1 인터넷 정보 추출 에이전트

인터넷상에서 정보를 처리해주는 정보 에이전트는 정보검색 에이전트, 정보필터링 에이전트, 정보통합 에이전트, 정보추출 에이전트의 네 가지로 분류할 수 있다[2]. 정보추출 에이전트(information extraction agent)는 문서 내에서 원하는 정보가 내재된 특정 부분을 발취해 내는 작업을 수행하는 역할을 하며, Jango[3], Junglee[4]와 같이 상업용 인터넷 서비스 등과 같은 곳에 적용되고 있는 wrapper라고 하는 정보 추출 규칙에 의해서 처리되기도 한다.

2.2 Wrapper

웹을 구성하는 문서의 구조가 다르므로 해당 문서에 개별적으로 추출을 담당하는 추출규칙을 갖고 있는 wrapper[5][6]라는 것을 사용하게 된다. wrapper는 shopping mall과 같은 상품검색 결과 페이지와 같이 출력부분에서 같은 형태의 의미 없는 부분과 추출의 대상이 되는 변화되는 부분을 구분할 수 있는 개념적인 태그를 사용하는 단순한 추출 규칙에서부터 기계학습을 통한 추출규칙을 자동적으로 생성하는 복잡한 방법들도 사용된다. wrapper에서는 확장성이 중요한데 자동적/반자동적 추출규칙의 생성에 있어서 특정 대상이 되는 모든 문서에 대해서 적용이 가능해야 하지만 기계학습의 정도에 따라서 그 결과는 많이 달라 진다.

2.3 Web Information Mining Algorithm

웹에서 유용한 정보를 분석하거나 발견하는 것을 웹 마이닝[7]이라 하며, 그 중에서 정보를 추출하는데 중요 키워드를 사용하는 Keyword-based Mining, 문장 구성 단어의 어순 패턴을 사용하는 Pattern-based Mining, 그리고, 사용자가 정의한 예제를 이용한 Sample-based Mining의 KPS[8]라는 정보 마이닝 알고리즘이 있다.

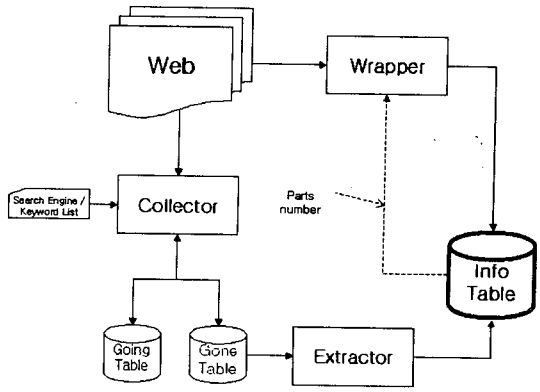
2.4 WIMA-EE 정보추출

전자부품과 관련된 웹 문서들은 일반적으로 테이블 형태의 구조로 정보를 제공한다. 웹에서 테이블 형태의 정보는 관련 태그를 이용해서 정보를 표현하는데 <table>과 </table>태그 사이에서 정보를 표현한다. 또한 하나의 행을 표현하기 위해서 <tr>과 </tr>을 사용하고 각각의 개별 필드는 <td>와 </td>나 <th>와 </th>의 쌍으로 사용해서 내용이 채워지게 된다. 그러나 자료의 표현 뿐만 아니라 웹 페이지 구성에 필요한 레이아웃을 잡기 위한 형태로 다양한 정보들이 채워지는 경우도 있으므로 실제로 많은 내용들이 필터링 되야 한다.

본 논문에서 사용한 추출 규칙에서 주요 관심대상은 자료의 행을 구성하는 <tr>와 </tr> 사이의 내용이며, 실제 정보 추출 대상은 <td>와 </td>사이 <th>와 </th>사이의 정보들이다. 이것을 추출하게 되면 데이터베이스와 같은 필드명과 필드데이터 테이블 형태의 내용의 정보가 추출되게 된다.

3. 시스템 구성

WIMA-EE시스템의 전체 구성은 [그림1]과 같으며 웹 문서들 중에 전자부품과 관련된 문서들만 추려서 조사된 정보만을 추출해 최종적으로 Info Table이라는 데이터베이스에 저장하도록 구성된다. 여기에서 Extractor와 Wrapper 모듈은 테이블 형태의 문서를 평가해서 정보를 가공할 수 있도록 필터링을 한 후에 정보를 추출하도록 한다.



[그림 1] WIMA-EE의 기본 구조

3.1 Collector Module

미리 정의된 전자부품에 관련된 키워드 리스트를 질의어로 검색엔진을 통해서 얻어지는(키워드 당 500개) URL들을 얻어서 웹 로봇과 같이 문서의 링크를 따라가면서 추가적으로 얻어지는 문서들에서 전자부품번호에 대한 정보가 있는지를 결정하여 정보추출 가능성 있는 문서로 판단을 해서 추출의 대상이 되도록 구분하여 기록하도록 하는 작업을 한다.

여기에서 부품번호에서 발견되는 유사패턴 규칙을 발견해 문서 내에서 같은 패턴이 있는가를 조사하여 정보가 있는지 판단하도록 하는 역할을 한다.

3.2 Extractor Module

Collector Module에서 수집된 추출정보대상 URL에서 테이블 형태의 문서 추출 규칙에 따라 테이블의 텍스트 내용을 데이터베이스와 같은 테이블과 같은 형태의 자료 구조에 저장되며 행과 열의 인덱스로 참조가 될 수 있는 동적인 문자배열에 저장한다.

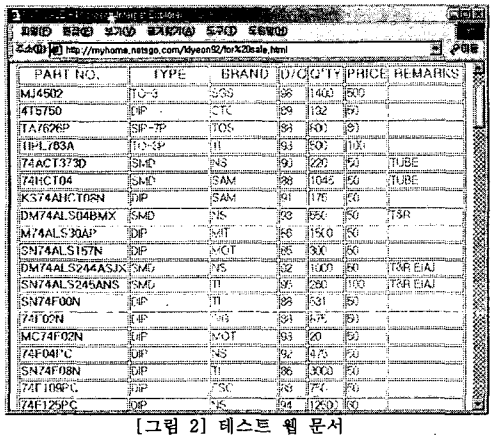
이렇게 필터링 된 문자 배열에서 관련된 정보를 추출하여 Info Table이라는 데이터베이스에 저장하게 된다. 이때 저장되는 내용 중에 부품번호는 wrapper에서 사용을 한다.

3.3 Wrapper Module

전자 부품에 대한 정보를 제공해주는 웹 사이트의 많은 부분이 쇼핑 몰(shopping mall) 형태로 사용자의 질의를 통해서 내부적으로 검색된 부품정보 결과를 제공하는 것으로 부품번호를 욕구하도록 되어 있다. Extractor Module를 통해서 추출된 부품번호를 사용해 결과 문서를 얻는다. 여기서 테이블형태의 추출 규칙을 적용해서와 중간단계의 자료구조로 변형한 후 원하는 정보를 추출, Info Table 데이터베이스에 저장하는 동작을 한다.

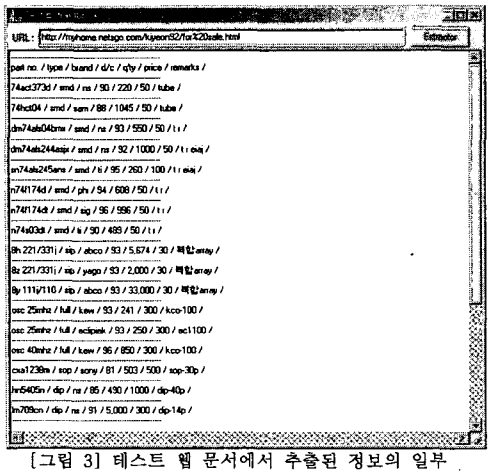
4. 실행 형태

다음은 WIMA-EE 실행도중 테이블 형태의 문서에서 추출한 중간 형태의 정보를 예시한 것으로 [그림 2]는 대상 웹 문서이며, [그림 3]은 데이터베이스와 같이 추출된 결과를 보여준 것이다.



PART NO.	TYPE	BRAND	D/C	Q'TY	PRICE	REMARKS
MJ4502	TO-3	SGS	86	1400	500	
4T5750	DIP	CTC	89	132	60	
TA7626P	SIP-7P	TOS	84	400	80	
TIP-L783A	TO-3P	TI	93	502	100	
74ACT373D	SMD	NS	90	220	50	TUBE
74HCT04	SMD	SAM	88	1045	50	TUBE
KS74AHCT08N	DIP	SAM	94	176	50	
DM74ALS04BMX	SMD	NS	93	550	50	T&R
M74ALS30AP	DIP	MIT	86	1500	50	
SN74ALS157N	DIP	MOT	85	300	50	
DM74ALS244ASJX	SMD	NS	92	1000	50	T&R EIAJ
SN74ALS245ANS	SMD	TI	95	260	100	T&R EIAJ
SN74F00N	DIP	TI	88	531	50	
74F02N	DIP	NS	91	475	50	
MC74F02N	DIP	MOT	93	20	50	
74F04PC	DIP	NS	92	475	50	
SN74F08N	DIP	TI	86	3000	50	
74F109PC	DIP	FSC	93	774	50	
74F125PC	DIP	NS	94	1200	50	

[그림 2] 테스트 웹 문서



part no. / type / brand / d/c / q'ty / price / remarks /
74act373d / smd / ns / 90 / 220 / 50 / tube /
74hct04 / smd / sam / 88 / 1045 / 50 / tube /
dm74als04bmx / smd / ns / 93 / 550 / 50 / t r /
dm74als244asjx / smd / ns / 92 / 1000 / 50 / t r eiaj /
sn74als245ans / smd / ti / 95 / 260 / 100 / t r eiaj /
74f174d / smd / ph / 94 / 608 / 50 / t r /
74f174dt / smd / sg / 96 / 996 / 50 / t r /
74s03dt / smd / ti / 90 / 489 / 50 / t r /
8h 221/331j / sip / abco / 93 / 5,674 / 30 / 특합array /
8z 221/331j / sip / yago / 93 / 2,000 / 30 / 특합array /
8y 111y/110 / sip / abco / 93 / 33,000 / 30 / 특합array /
osc 25mhz / full / kew / 93 / 241 / 300 / kco-100 /
osc 25mhz / full / sclipek / 93 / 250 / 300 / sc1100 /
osc 40mhz / full / kew / 96 / 850 / 300 / kco-100 /
cxa1238m / sop / sony / 81 / 503 / 500 / sop-30p /
nr6405m / dip / ns / 85 / 490 / 1000 / dip-40p /
lm709cn / dip / ns / 91 / 5,000 / 300 / dip-14p /

[그림 3] 테스트 웹 문서에서 추출된 정보의 일부

예로 든 문서의 URL은 http://myhome.netsgo.com/kiyeon92/for%20sale.html이며 추출된 형태의 자료에서 ' / ' 는 필드간의 구분을 나타내기 위해서 표현을 했다.
출력된 형태를 살펴보면 데이터베이스의 테이블 형태의 자료구조와 같이 표현되어 있음을 발견할 수 있고 이런 방법으로 테이블 형태의 웹 문서에서의 정보 추출 가능성을 보인다.

5. 결론 및 향후 과제

인터넷을 통한 정보 추출의 필요성은 정보의 양이 많아지면 많아 질수록 높아질 것이며 필요한 요소 기술들에 대한 연구도 활발해 질 것이다. 본 논문에서는 테이블 형태의 텍스트 기반 웹 문서에서 정보를 추출해 내는 에이전트의 구성에 있어서 테이블 태그를 분석한 정보 추출의 가능성을 제시 하고 있다. 정보를 추출함에 있어서 다양한 문서의 형태에 적응력이 있는 확장가능 한 wrapper의 작성에 있어서 많이 미치지는 못하고 있으나 전자부품과 같은 특정 정보를 제공하는 문서의 분석에 있어서 하나의 가능성을 보인다. 동일한 의미의 필드 명을 다르게 표현하는 경우의 처리, 누락 필드의 처리 등 추가 적인 연구가 계속되어야 하며 여러 형태의 테이블에 대해서 핵심화 된 wrapper의 자동 구성이 향후 과제로 남아 있다.

참고 문헌

[1] 오석일,변영태, "인공과학실험실의 확장 : 전기실험 + 화학실험", 한국정보과학회 봄 학술발표논문집, Vol. 26, No. 1, 1999
[2] 최중민, 인터넷 정보추출 에이전트, 정보과학회지 18권 5호, pp 48-53, 2000
[3] Jango, http://jango.excite.com
[4] Junglee, http://www.junglee.com
[5] N. Ashish, C. Knoblock. "Wrapper generation for semi-structured internet sources." In Proc. Workshop on Management of Semistructured Data, Tucson, 1997.
[6] Nickolas Kushmerick, Daniel S. Weld, Robert B. Doorenbos. "Wrapper induction for information extraction." In Intl. Joint Conference on Artificial Intelligencd(IJCAI), page 729-737. 1997.
[7] R. Cooley, J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997.
[8] Guan, T., Wong, K.F., "KPS - a Web Information Mining Algorithm", in Proc. of International World Wide Web Conference, May 11-14, 1999.