2000 년 한국정보처리학회 추계 학술발표논문집 제 7 권 제 2 호

웹 정보 통합시스템상의 래퍼(Wrapper) 생성에 대한 연구

김 홍 석, 신 동 천
중앙대학교 정보시스템학과
e-mail : dmaj7@hananet.net
dcshin@cau.ac.kr

A Study on Wrapper Generation for Web Mediation System

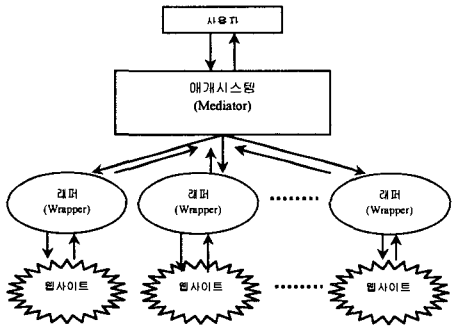
Hongsok Kim, Dongchun Shin
Dept. of Information System, ChoongAng University

요 약

웹을 하나의 정보 저장소로 바라보고 웹에 대해 데이터베이스의 방법론을 적용하려는 시도들이 많은 데이터베이스 분야에서 이루어지고 있으며 그 중 한 분야로써 웹을 정보 원천으로 보고 여러 웹 사이트에 산재해 있는 이질의 정보를 통합하는 정보 통합시스템에 대한 연구도 활발히 이루어지고 있다. 본 논문에서는 웹 정보 통합시스템의 필수 요소인 래퍼(**wrapper**)를 구현함에 있어 웹 사이트의 구조와 사이트 내의 웹 문서들의 구조를 파악하여 요약한 사이트 카탈로그를 이용함으로 래퍼 (**wrapper**)를 보다 효율적으로 생성 및 운영할 수 있는 새로운 방법과 모형을 제시하고자 한다.

1. 서론

월드 와이드 웹은 HTML 과 TCP/IP 기반의 HTTP 를 이용하여 플랫폼에 관계 없이 일관된 방식으로 정보를 표현하고 배포 및 공유할 수 있는 획기적인 방법을 우리에게 가져 다 주었다. 이를 바탕으로 웹 형식의 정보가 기하급수적으로 증가하였으며 웹으로 표현된 인터넷 상의 정보들을 효율적으로 이용하기 위해, 웹 자료의 비정형 구조를 표현할 수 있는 새로운 데이터모델 및 데이터베이스 시스템 개발, 웹의 자료를 일관된 방법으로 접근 및 생성, 삭제 가능하도록 하는 질의 언어 개발등과 같은 데이터베이스 분야의 방법론을 적용하고자 많은 부분에서 연구가 이루어지고 있다[1]. 이런 시도중의 하나로써 기존의 이질의 정보를 통합하는 자료 통합시스템(Mediation system)에 웹의 자료를 포함시키고자 하는 연구가 활발하게 진행되고 있다[1][2][10]. 그림(1)은 이러한 웹 자료 통합시스템의 구조를 나타낸 것이다. 이 시스템에서 사용자가 텍스트 기반의 질의 도구나 웹 브라우저 또는 특정 응용프로그램을 사용하여 질의를 매개시스템에 전달한다. 매개시스템(Mediator)은 가상의 뷰 역할을 하며 이를 위해 공통의 데이터 모델이 필요하다. 매개시스템은 사용자의 질의를 분석 및 분해하여 해당 데이



(그림 1) 웹 자료 통합 시스템의 구조

터가 존재하는 정보 원천을 담당하는 래퍼(Wrapper)로 전달하며 후에 래퍼(Wrapper)로부터 전달된 결과값을 정제 및 통합하여 사용자에게 최종 결과를 전송하여 준다. 래퍼(Wrapper)는 매개시스템으로부터 받은 질의를 분석 정보 원천이 처리할 수 있는 질의형태로 변환하고 정보원천으로부터 돌려 받은 결과값을 매개시스템으로 반환하여 준다. 본 논문에서는 웹 정보 통합 시스템을 이루는 한 요소인 래퍼(Wrapper)를 생성(Generation)시에 웹 사이트 카탈로그를 이용하는 새로운 방법을 제시하고자 한다. 본 논문에서 다룰 내용은 다음과 같다. 2 장에서는 기존의 이질의 자료통합 시스템에서 소개된 일반적인 래퍼의 구성요소와 그 기능 및 웹에 적용시의 문제점을 살펴보고 3 장에서, 기존의 래퍼를 웹에 적용시키고자 했던 기존의 연구들을 간략하게 고찰하며 4 장에서는, 새로운 래퍼의 모형을 제시하고 5 장에서는 제안 래퍼 모형의 가장 큰 특징인 사이트 카탈로그를 어떻게 얻을 것인지 살펴보며 마지막으로 6 장에서 향후 연구 방향과 제안 래퍼의 구현시 기대되는 이점을 간략하게 논하고자 한다.

2. 일반적인 래퍼(Wrapper)의 구성요소와 기능

본 논문에서 제시하는 모형을 살펴보기 전에 이질 자료 통합시스템에서 사용되는 래퍼(Wrapper)의 일반적인 구성요소와 그 기능을 살펴 보면 다음과 같다.

- 파서 : 매개시스템으로부터 전송된 질의를 해석하고 정보 원천에 적합한 질의 언어로 바꾼다.
- 정보 원천 : 파서의 정보를 바탕으로 재작성 된 질의 언어를 실행하여 질의 결과를 얻는다. 웹 사이트의 경우 질의어가 없으므로 추출기를 작성하여야 한다.
- 변환기 : 정보 원천으로부터 받은 결과값을 통합 시스템에서 사용하는 공통의 데이터 모델로 변환한다.
- 제어기 : 위의 기능들을 계획에 따라 관리하며 변환된 최종 결과를 매개시스템에 반환한다

이러한 기능을 제공하는 기존의 래퍼를 웹을 대상으로 적용하는 데에는 다음과 같은 어려움이 있다. 우선 웹 사이트는 기존의 데이터베이스 시스템에서 제공하는 자료사전(Data Dictionary)를 제공하지 않으며 다음으로 웹 사이트와 페이지의 구조는 반(半) 구조적(Semi structure)인 성격을 보이고 또한 웹사이트는 데이터베이스 시스템과 비교하여 볼 때 그 구조적인 특징이 자주 변하며. 마지막으로 웹 사이트에는 데이터베이스 시스템과 이를 이용하는 응용프로그램처럼 모델링 과정을 통해 현실 세계의 의미가 반영된 스키마가 존재하지 않거나 존재하더라도 그 의미가 태그 안에 숨어있다. 이러한 특징을 보이는 웹을 대상으로 래퍼를 생성하고자 하는 연구들은 3 장에서 살펴보도록 한다.

3. 관련 연구

관련 연구들에서 연구의 초점은 래퍼 생성시에 생성 과정을 얼마나 편리하게 그리고 자동화 할 수 있을 것인가 하는 것과 얼마나 정확하게 자료를 추출할 것인가에 맞추어져 있다. [3,4]는 새로운 정보원천 추가 시 템플렛을 이용, 래퍼 생성을 간소화 하고자 한다. [5]에서는 래퍼의 필요 요소들을 정의하고 있지만 실질적인 연구는 자료 추출 방법에 치우쳐져 있다. 여기서는 관심 있는 자료가 존재하는 페이지의 태그에 LEX/YACC 를 이용 토큰을 설정하고 태그가 포함하고 있는 폰트의 크기나 여백 등을 비교하여 텍스트 문서의 구조를 결정, 추출하는 방법을 제시한다. [6]에서는 휴리스틱 방법을 적용하여 래퍼 클래스들을 생성하고 이로부터 귀납적인 추론을 적용하여 래퍼를 자동으로 생성하는 방법을 제시한다. [7]에서는 특정 사이트를 대상으로 사용자로부터 키워드를 입력 받고 이 입력값을 URL 로 바꾸어 웹 페이지를 얻고 이에 적용되는 추출기를 호출하는 JDBC(Java DataBase Connectivity) 호환 래퍼 생성 방법을 제시하고 있으며 이를 시각적으로 구축할 수 있는 도구를 제시한다 마지막으로 [8]에서는 복잡한 구조의 HTML 문서로부터 자료를 추출할 수 있도록 정의된 언어와 추출된 자료를 DOM (Documents Object Model)을 이용 XML 및 Java Object 로 변환할 수 있는 매핑 방법과 전체적인 래퍼 생성 과정을 시각적으로 구축할 수 있는 도구를 제시하고 있다.

4. 제안하는 래퍼(Wrapper)의 모형

관련 연구들에서 제시하지 않은 한계점들을 요약하면 다음의 두 가지로 나누어 볼 수 있다. 먼저 위의 연구들에선 웹 사이트 내에 어떤 정보가 어떤 페이지에 내재되어 있는지 그리고 그 정보에 어떻게 접근할 수 있는지에 대한 방법이 소개되지 않았다. 기존 연구들이 제시한 방법들에선 예를 들어 "오늘의 정치면 기사들을 보여달라 !"와 같은 질의에 응답할 기능이 없거나 이에 응답하기 위해선 해당기사마다 하나의 래퍼를 생성하거나 신문사 웹 사이트의 페이지마다 자료를 추출하고 그에 해당하는 기사들의 내용을 비교하여야 만 한다. 이는 전체 통합시스템의 질의응답성능을 저하시킬 것이다. 두 번째로 추출기(Extractor)에 대한 것으로 HTML 의 구조가 복잡해 짐에 따라 이를 분석하고 태그내의 내용들을 추출할 수 있는 추출기의 개발에 중점을 두고있다. HTML 의 발전방향이 표현형식과 데이터의 속성을 나타내는 태그가 분리되는 방향으로 발전하고 있어서 구조적인 분석은 정확하게 할 수 있다. 하지만 대부분의 사이트에 똑 같은 추출기를 적용하는 것은 문제가 있다. 이는 모든 사이트에서, 같은 성격과 내용을 다루는 사이트라 할 지라도 형식과는 별개로 페이지 내용의 의미가 다를 수 있기 때문이다. 기존의 추출기를 통해 구문적인 분석은 정확하게 이루어질 수 있으나 어디서부터 어디까지의 내용이 어떤 애트리뷰트로 지정되어야 할지는 태그의 구조적인 분석만으로는 항상 동일하게 적용할 수 없기 때문이다. 이상과 같은 문제점들을 해결하고자 다음과 같은 기능들을 제시함으로써 소개한 한계점들을

해결해 보고자 한다.

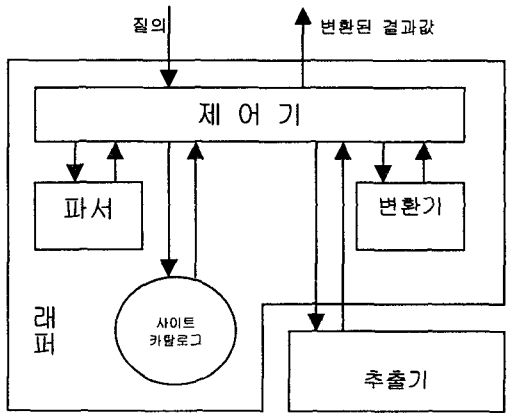
1) 사이트 카탈로그 : 사이트 카탈로그의 내용은 크게 두 가지로 나누어 볼 수 있다. 하나는 해당 웹 사이트의 구조이며 다른 하나는 개별 웹 페이지의 구조 및 내용에 대한 것이다. 이를 구분할 수 있는 이유는 웹 사이트는 하나의 그래프로 표현될 수 있으며 또한 이를 의미적인 트리로 볼 수 있다. 그리고 사용자가 원하는 자료는 대부분 이 의미적인 트리의 단말노드로 표현되는 페이지들에 존재하게 된다. 단말노드 이외의 노드들은 일련의 경유노드의 역할을 하며 단말노드를 일련의 의미로 묶어주는 범주의 역할을 하게 된다. 다음은 이러한 사이트 카탈로그의 내용을 신문사 사이트를 예로 들어 간략하게 표현한 것이다.

범주	하위범주	...	URL	주제어	페이지구조
sports	baseball	...	../aa.html	이승엽...	A001
sports	soccer	...	./bb.html	차범근...	AOO2
.....	.....	...	.....	.....	.....

(표 1) 사이트 카탈로그의 내용

사이트 카탈로그의 내용에는 각 범주및 하위범주로 그리고 각 하위 범주에 속하는 개별 페이지들의 URL 과 각 범주 및 개별 페이지를 특징 짖는 주제어들로 이루어져 있다. 여기서 각 URL 은 해당 페이지들을 가리키는 Hyperlink 태그를 분석하여 얻어낼 수 있으며 주제어는 각 페이지에 포함된 <title> 태그나 이 페이지를 연결하는 상위 페이지의 < a href=> 태그 사이의 문자열을 이용해 얻어낸다. 다음으로 페이지의 구조로 이는 경량의 추출기와 연관이 있다.

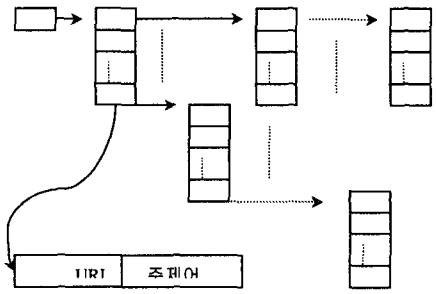
2) 경량 추출기(Light weight Extractor) : 모든 경우에 정확하게 의미까지 반영하여 웹 페이지로부터 자료를 추출하는 것에 대한 어려움은 앞서 살펴보았다. 이를 해결하게 위해 웹 사이트들을 종류별 혹은 의미별로 묶어 각 사이트에 나타나는 공통의 특징을 파악하고 이를 바탕으로 모든 태그를 분석할 수 있는 추출기를 작성하는 것이 아니라 각 공통의 특성으로 파악된 태그만을 파악하여 자료를 추출할 수 있는 추출기를 작성하는 것이 보다 경제적일 것이다. 예를 들면 수 많은 신문사 사이트들이 인터넷 상에 존재하지만 그들이 보유하고 있는 웹 페이지들은 헤드라인, 기사, 기자등과 같이 공통의 특징을 가지고있다. 따라서 이런 특징을 나타내는 태그만을 파악하여 이를 하나의 패턴으로 표현할 수 있다면 보다 간단하게 추출기를 작성할 수 있으며 같은 범주의 새로운 정보원천의 추가시 보다 손쉽게 이를 재적용할 수 있을 것이다. [2][3]에서 제시한 바와 같이 위의 구성요소 중 정보 원천과 관련된 부분을 제외한 요소는 래퍼(wrapper) 생성시에 한 번 작성해 두면 새로운 정보 원천을 추가 할 때에도 대부분의 코드를 재사용할 수 있다. 그림[2]는 제안하는 래퍼의 모형을 나타낸 것이며 이의 구동 메커니즘은 다음과 같다. 제안된 래퍼의 구동 메커니즘을 살펴보면 매개시스템으로부터 전달된 질의를 받으면 제어기가 파서를 호출하여 이를 분석하고 관심 자료가 무엇인지를 파악한다. 파서에서 생성된 정보를 바탕으로 제어기는 사이트 카탈로그를 호출 해당 정보들이 웹 사이트내 어느 페이지에 있는지 알아내고 해당 페이지의 URL 과 자료형식을 반환한다. 제어기는 이 정보를 이용해 추출기를 호출하고 호출된 추출기는 URL 정보를 이용해 해당 페이지(들)을 얻고 저장형식 정보를 이용해 적용 가능한 추출 메쏘드를 이용, 자료를 추출한다. 추출된 자료는 다시 제어기에 전달되고 제어기는 이를 변환기를 이용하여 웹 통합 시스템에서 사용되는 공통의 데이터모델로 변환하고 매개시스템에 결과값을 전송한다.



(그림 2) 제안 래퍼(Wrapper)의 구조

5. 웹 사이트 카탈로그(Web Site Catalog)

4 장에서 제안한 래퍼 모형의 핵심 요소인 사이트 카탈로그를 얻으려면 해당 자료를 담고있는 페이지에 접근할 수 있는 경로를 보다 명확하게 하기 위해서 먼저 그래프 구조의 웹 사이트를 다음과 같은 트리 구조로 변환하여야 한다.



(그림 3) 트리화 된 웹 사이트 구조

위의 트리에서 루트는 항상 URL 이 정해지며 단말 노드에 이르기까지의 경유 노드들은 각 단말노드를 특정 의미로 묶어주는 범주의 역할을 한다. 이를 얻기 위해서는 해결해야 할 문제들이 있는데 먼저 위의 그림에서 같은 배열의 구성원으로 표현되는 연결 페이지들 사이의 링크를 무시해야 하며 다음으로 한 배열의 상위 배열로 향하는 링크들을 무시해야 한다. 또한

이를 얻기 위해서 웹 사이트의 구조가 한 연결 페이지가 다른 연결페이지의 하위 페이지로는 링크를 제공하지 않을 것을 가정한다. 위와 같은 트리를 얻기 위한 알고리즘의 개요를 소개하면 다음과 같다.

1) 루트 노드를(처음 URL) 얻는다.
2) 얻어진 URL 을 Unreachable 리스트에 저장한다.
3) 얻은 노드(루트)에 나타난 모든 하이퍼링크의 URL 을조사하고 이들을 구조체 배열로 저장한다.
4) 카운트 리스트를 생성하고 구조체 배열에 저장된 URL 을 카운트 리스트에 복사 한다. 단 Unreachable 리스트에 저장된 URL 은 카운트리스트에서 제외시킨다.
5) 구조체 배열의 멤버로 저장된 URL 이 지칭하는 페이지들을 하나씩 탐색하여 그 속에 나타난 하이퍼링크가 가리키는 URL 중 카운트 리스트에 속한 URL 이 있으면 카운트 리스트에 카운트 횟수를 기입한다.
6) 카운트 리스트의 카운트한 수를 비교하여 카운트 수가 다른 멤버들과 다른 것은 구조체 배열에서 삭제 하고 카운트 리스트를 없앤다.
7) 구조체 배열에 저장된 URL 을 Unreachable 리스트에 저장한다.
8) Root 포인터를 생성된 구조체 배열의 첫번째 멤버로 이동시킨다.
9) 3 에서 8 까지의 과정을 더 이상 탐색할 URL 이 없을 때까지 반복한다.

위와 같은 트리를 얻은 후에 이를 바탕으로 원하는 자료가 존재하는 페이지를 빠르게 검색하는 기능을 마련해 주어야 한다. 물론 주어진 트리를 깊이우선탐색이나 너비우선탐색을 이용해 주어진 키워드와 매칭되는 주제어를 비교하고 원하는 페이지에 도달할 수 있지만 규모가 큰 사이트의 경우 이에 많은 시간이 소모될 것이므로 이를 주어진 트리의 내용을 이용하여 빠르게 해당 페이지에 도달할 수 있는 방법이 필요하다. 이를 위해선 트리의 관계를 적절히 이용하고자 한다. 위와 같은 트리가 얻어진 후에는 각 노드 사이에 부모와 자식 그리고 형제노드와 같은 관계를 쉽게 유추해 낼 수 있으며 이를 이용해 질의 파악 후 사용자의 관심 데이터가 있는 페이지에 얻어진 주제어를 이용해 바로 접근할 수 있는 기능을 마련할 수 있다. 검색 알고리즘 작성시 참고해야 할 이런 기능이 가져야 할 특징으로는 특정 노드의 하위 노드들에 한꺼번에 접근할 수 있는 집합적인 접근 방법과 개별적인 페이지들에 직접 접근할 수 있는 방법이 지정되어야 한다.

6. 결론 및 향후 연구 방향

본 논문에서는 웹 통합시스템 구현 시에 필요한 새로운 래퍼의 모형과 생성 방법을 제시하였다. 제안한 래퍼가 구현될 경우 기대되는 이점으로 먼저 관심 웹 사이트의 자료가 어느 페이지에 존재하는지 접근 경로를 제공함으로 웹 통합 시스템의 전체적인 질의 성능 향상이 기대되며 다음으로 사이트가 가지고 있는 자료와 그에 대한 형식을 제공함으로 웹 통합 시스템의 매개시스템을 보다 정확하게 할 수 있을 것으로 기대된다. 위와 같은 래퍼를 구현하기 위해서 향후 연구할 사항으로는 먼저 본 논문에서 제시한 경량의 추출기를 작성하기 위해서 사이트를 종류별로 범주화하고 그에 따른 특성을 발견해야 하며 다음으로 전체 웹 통합 시스템에서 사용할 데이터모델을 결정하여야 한다 데이터모델이 래퍼를 구성하는 요소들에 중요한 영향을 미치기 때문이다. 따라서 사이트 카탈로그를 얻기 위해 제시한 알고리즘의 유효성을 증명한 후에 위와 같은 연구 사항들을 사이트 카탈로그에 반영하고자 한다.

참고문헌

[1] Daniela Florescu, Alon Levy and Alberto Mendelzon, " Database Techniques for the World Wide Web: A Survey"ACM SIGMOD Record 27:3, September 1998, 59-74.

[2] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. "The TSIMMIS Project: Integration of Heterogeneous Information Sources". In Proceedings of IPSJ Conference, pp. 7-18, Tokyo, Japan, October 1994.

[3] J. Hammer, M. Breunig, H. Garcia-Molina, S. Nestorov, V. Vassalos, R. Yerneni. "Template-Based Wrappers in the TSIMMIS System". In Proceedings of the Twenty-Sixth SIGMOD International Conference on Management of Data, Tucson, Arizona, May 12-15, 1997.

[4] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo. "Extracting Semistructured Information from the Web". In Proceedings of the Workshop on Management of Semistructured Data. Tucson, Arizona, May 1997.

[5] Naveen Ashish and Craig Knoblock. "Wrapper Generation for Semi-structured Internet Sources." ACM SIGMOD Workshop on Management of Semi-structured Data, Tucson, Arizona, 1997

[6] Nicholas Kushmerick, Daniel S. Weld, Robert Doorenbos. "Wrapper induction for information extraction" In International Joint Conference on Artificial Intelligence, Nagoya, Japan, 1997

[7] Jean-Robert Gruser, Louiqa Raschid, M. E. Vidal, Laura Bright: "Wrapper Generation for Web Accessible Data Sources" 14-23 Proceedings of the 3rd IFCIS International Conference on Cooperative Information Systems, New York, USA, August 20-22, 1998

[8] Arnaud Sahuguet, Fabien Azavant. "Building Intelligent Web Applications Using Lightweight Wrappers" Data and Knowledge Engineering (to appear 2000)

[9] http://www.w3c.org/TR/1999/REC-html401-19991224

[10] "A Report on the Applicability of Mediation in ALP" Global InfoTek,, DARPA.