분산된 웹 정보의 효과적 통합·추출을 위한 동적 Wrapper 조합

백주흠[0], 홍진혁, 조성배
연세대학교 대학원 인지과학 협동과정[0], 연세대학교 컴퓨터과학과
{bjh[0], hjinh}@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Dynamic Wrapper Composition for Integrative Extraction of Distributed Web Information

Joo Huem Baek[0], Jin-Hyuk Hong, Sung-Bae Cho
[0]Graduate Program in Cognitive Science, Yonsei University,
Dept. of Computer Science, Yonsei University

요 약

웹 정보 통합은 사용자 질의에 적합한 정보를 분산된 웹에서 추출하여 제공하는 방법으로 질의응답 속도의 향상을 위해 질의처리 방식을 주로 사용한다. 질의 처리는 Wrapper를 이용해 웹으로부터 제약조건을 만족하는 정보를 추출하고 사용자가 원하는 형태로 결합하는 방식인데, 통합과정에서 제거될 정보까지 미리 추출하는 문제가 있다. 본 논문에서는 이를 해결하기 위해 튜플 단위 웹 정보 추출 방법을 제안한다. 제안하는 방법은 F-Logic으로 표현된 도메인 모델과 CHR(Constraint Handling Rule)로 정의한 규칙을 이용해 질의를 확장하고 적절한 Wrapper들을 선택한 뒤 추출에 필요한 Wrapper를 동적으로 조합한다. 쇼핑몰 사이트에 분산된 웹 정보 획득에 제안하는 방법을 적용하여 유용성을 확인하였다.

1. 서론

웹 상의 정보가 증가함에 따라 이를 통합하려는 시도가 활발히 이루어지고 있다. 웹 정보는 HTML로 표현되며 데이터베이스 상의 정보, 혹은 XML로 표현되는 정보처럼 완전히 구조화된 형태로 표현되어 있지 않기 때문에 정확한 추출이 쉽지 않다. 따라서 웹 정보 통합을 위해 보통 HTML상의 데이터를 구조화된 형태로 추출해주는 Wrapper를 이용한다. 여러 사이트에 분산된 정보를 획득하고자 하는 경우에는 다수의 Wrapper가 필요하다.

웹 정보 통합 시스템은 사용자 질의에 적합한 정보를 분산된 웹에서 추출하여 제공하는 시스템으로 추출 과정에서 Wrapper를 이용한다. 보통 질의에 대한 결과는 분산된 정보의 결합된 형태로 제공된다. 기존 연구의 경우 추출 및 결합에서 Wrapper별로 추출한 후 결합하지만 외래키(Foreign Key) 기준의 분산 정보 결합 질의 등에 대해, 결합에서 제거될 정보까지 미리 추출하는 문제점이 있다.

이를 해결하기 위해 본 논문에서는 튜플 단위로 추출한 다음 결합하는 동적 Wrapper조합방법을 제안한다. 제안하는 방법은 F-Logic[1]으로 표현된 도메인 모델과 CHR(Constraint Handling Rule)[2]로 정의된 규칙을 이용해 질의를 확장하고 Wrapper 메타 정보를 이용해 질의에 적합한 기본 Wrapper를 선택하고 동적으로 조합하여 질의에 최적화된 Wrapper를 생성해낸다. 실험을 통해 분산된 웹 정보 획득에 Wrapper조합방법이 유용함을 확인한다.

2. 배 경

2.1. 웹 정보 통합

최근 많은 관심을 받는 정보 통합의 한 분야로 Ariadne[3]과 같은 시스템이 제안되고 있다. 사용자 질의에 따라 통합된 정보 제공을 위해 대부분 질의 처리 방식(Query Processing)을 이용한다. 이는 사용자 질의를 확장하고 각 사이트에서 질의될 수 있는 형태로 분리, 변형한 후 실행 계획을 통해 질의를 최적화하여 수행하고 추출된 출처별 정보를 사용자 요구에 부합하도록 통합한다.

분산된 웹 정보를 제공하는 방식은 정보를 추출하는 시점과 사용자 질의를 수행하는 시점의 선후관계에 따라, 미리 추출된 정보를 기반으로 질의를 수행하는 방식과 질의에 따라 동적으로 추출작업을 수행하는 방식으로 구분된다[4]. 전자의 경우는 질의-응답작업의 신속성 측면에서 유리하지만 사전에 방대한 데이터를 미리 추출해두어야 한다. 후자의 경우에는 다양한 질의에 반응적으로 충분한 정보를 제공하지만 질의에 따라 동적으로 추출작업을 수행하기 때문에 질의-응답시간이 증가하는 한계가 있다. 따라서 질의-응답 시간을 최소화하기 위해, 가장 많은 시간이 소요되는 웹 정보 추출작업 이전에 질의에 적합한 추출 대상만을 선택하는 등의 질의 처리계획을 추론하는 방법[5] 등이 제안되고 있다.

- 규칙 1. 단순화 규칙(Simplification)
  - Head <=> Guard | Body
  - Guard가 참일때 Head라는 제약을 Body라는 제약으로 대체할 수 있다.
- 규칙 2. 확장 규칙(Propagation)
  - Head==> Guard | Body
  - Guard가 참일때 Head라는 제약에 Body 제약을 더한다.
- 규칙 3. 단순화+ 확장 규칙(Simpagation)
  - Head1 | Head2 <=> Guard | Body
  - Guard가 참일 때 Head1은 남겨두고 Head2를 Body로 대체한다.

그림 1. CHR의 3가지 기본 규칙

2.2 F-Logic과 CHR

F-Logic은 논리기반 언어로 명확할 뿐 아니라 추론을 지원한다. Prolog와 같은 전통적인 논리기반 언어와 달리 고정된 수의 인자를 가지는 predicates를 사용하지 않으며 위치기반 인자 접근방식을 사용하지 않는다. 따라서 동적으로 변화하는 웹 상의 잘 구조화되어 있지 않은 데이터를 다루는 웹 통합분야에서 도메인 모델를 표현하는 논리 언어로 많이 사용되고 있다.

CHR은 사용자 정의 제약 문제를 풀기 위해 사용되며 제약규칙의 정의가 비교적 자유로워 정보통합 분야의 질의 계획 추론에 주로 이용된다. CHR의 기본적인 규칙은 그림 1과 같이 제약을 대체하는 "단순화 규칙," 제약을 확장시키는 "확장 규칙," 앞의 두 가지 규칙을 합친 "단순+확장 규칙"으로 구성된다.

3. 웹 정보 통합 추출을 위한 동적 Wrapper 조합방법

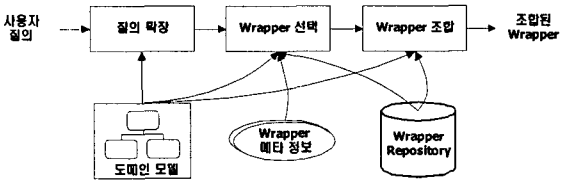


그림 2. 동적 Wrapper 조합과정

제안하는 Wrapper 조합방법은 그림 2와 같이 질의확장, Wrapper 선택, Wrapper 조합의 3 단계로 구성된다. 질의 확장 단계는 더 많은 Wrapper들을 선택하기 위해 입력된 사용자 질의를 도메인 모델 상에서 상하위 관계로 확장한다. Wrapper 선택 단계는 Wrapper 메타 정보에 기술된 제약정보를 바탕으로 질의에 적합한 Wrapper들을 선택하고, 마지막 단계인 Wrapper 조합에서는 선택된 Wrapper들을 정의된 규칙을 바탕으로 조합하여 최종적으로 조합된 Wrapper를 생성한다.

질의 확장은 그림 3과 같은 도메인 모델을 기반으로 진행된다. 그림에서 개념은 원으로, 속성과 함수관계 및 ISA관계는 선으로 표시되는데 상품이라는 개념은 하위 관계로 전자제품과 노트북을 가지며 상품명, 분류, 설명 등의 속성을 가진다. 또한 회사 개념과는 제조사 관계로 연결되어 있다.

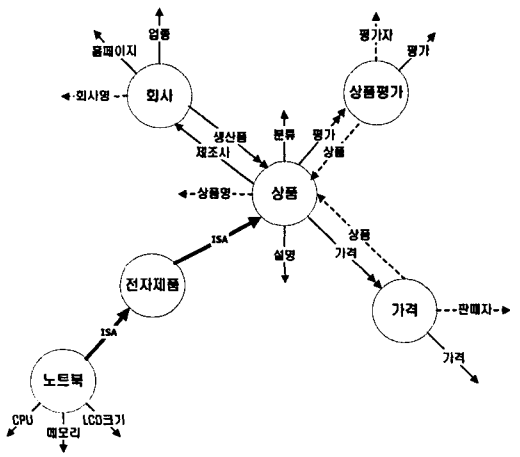


그림 3. 도메인 모델의 예 (온라인 쇼핑분야)

이 단계에서는 두 가지 규칙(그림4)에 따라 도메인 모델에서 질의 대상 개념과 상하위 관계를 가지는 개념들로 질의를 확장한다. 예를 들어 그림 3의 도메인 모델에서 "상품"에 대한 질의가 주어졌을 때 개념 상세화 규칙를 통해 "전자제품"에 대한 질의로 확장된다.

Wrapper를 이용해 웹 정보를 추출하는 작업은 시간이 매우 많이 걸리기 때문에 사전에 추출될 정보의 일관성을 고려해 적절한 Wrapper들을 선택한 뒤 조합되어야 한다. 따라서 Wrapper선택 단계에서는 질의대상이 되는 개념에 속한 Wrapper들을 모두 선택한 후, Wrapper 메타정보(그림 5)에 포함된 제약조건(Constraint 속성)들을 CHR로 확장하여 추출 이전에 일관성을 검사한다. 검사를 통과한 Wrapper들은 Wrapper 조합 단계에서 다음 2가지 기본 조합 방법들을 반복 사용하여 최종적으로 조합된 Wrapper를 생성한다.

- 규칙1. 개념 일반화
  - cClass ==> cClass::pClass | pClass ; Constraint
  - 하위 개념(cClass)에 관련된 질의를 상위 개념(pClass)으로 확장하기 위해 상위 개념에 제약을 가하는 규칙
- 규칙2. 개념 상세화
  - pClass ; Constraint ==> cClass:: pClass | cClass
  - 위의 "개념 일반화"의 반대 과정으로 상위 개념의 질의를 하위개념으로 확장하는 과정에서 포함 중인 관련 제약이 제거되는 규칙

그림4. 질의 확장규칙

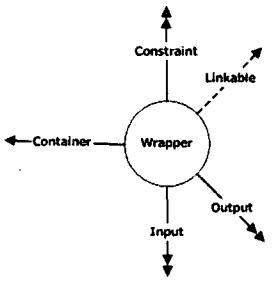
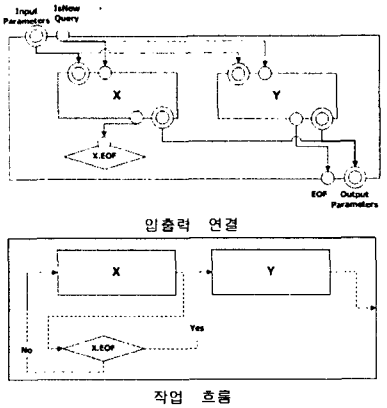


그림 5. Wrapper 메타 정보 구조

● 조합방법 1: "UNION 조합"

UNION조합은 이전 단계에서 선택된 기본 Wrapper들 중 도메인 모델 상에서 형제관계, 부모-자식관계의 개념이나 동일개념에 속한 것들을 조합하여 각각의 Wrapper들이 추출해내는 결과를 합하도록 한다. 그림 3의 모델에서 상품에 관한 질의가 입력될 경우 상품, 전자제품과 노트북 정보를 Container속성으로 가지는 Wrapper들이 "UNION 조합"으로 조합된다. 조합시 Wrapper의 입출력 연결과 작업 흐름은 그림 6과 같이 구성된다.



입출력 연결 작업 흐름

그림 6. X × Y (UNION조합)

● 조합방법 2: "INNER JOIN 조합"

INNER JOIN 조합은 도메인 모델 상에서 함수관계로 연결된 서로 다른 개념에 속해 있는 Wrapper들을 조합하여 각각의 추출결과를 튜플 단위에서 결합되도록 한다. 그림 3의 모델에서 상품정보와 그와 결합된 상품평가정보에 관한 질의를 입력할 경우 상품정보 추출 Wrapper와 상품평가정보 추출 Wrapper가 이 방법으로 조합되어 튜플 단위로 결합된 Wrapper가 생성된다. 조합시

Wrapper의 입출력 연결과 작업 흐름은 그림 7과 같이 구성된다.

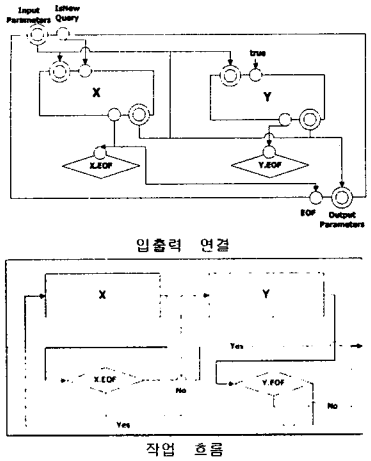


그림 7. X ∞ Y (INNER JOIN조합)

4. 실험 결과

여기에서는 제안하는 방법을 전자 상거래 분야에 적용하여 유용성을 살펴보고자 한다. 제안하는 방법에서 사용하는 도메인 모델 정보와 Wrapper 메타정보는 그림 8, 9와 같이 정의한다.

```
전자제품::상품
노트북::전자제품

상품[상품명->string,제조사->회사,분류->string,
평가->>평가,가격->> 가격]
노트북[CPU->string,메모리->integer, LCD크기->string]
가격[상품->상품, 판매자->string, 가격->integer]
평가[상품->상품, 평가자->string, 평가->string]
```

그림 8. F-Logic으로 표현된 도메인 모델

```
쇼핑몰A:Wrapper
쇼핑몰A[input->>{상품명}output->>{상품명,분류, 설명}]
쇼핑몰A[container-->전자제품]

쇼핑몰B:Wrapper
쇼핑몰B[input->>{상품명}output->>{상품명,분류,설명,메
모리}]
쇼핑몰B[container-->노트북]
쇼핑몰B[constraint -->" 메모리 > 128"]

쇼핑몰C:Wrapper
쇼핑몰C[input->>{상품명}output->>{상품명,분류,설명}]
쇼핑몰C[container-->노트북]

가격비교웹A:Wrapper
가격비교웹A [input->>{상품명}output->>{가격,판매자}]
가격비교웹A [container-->가격]
가격비교웹A [constraint -->" 상품.분류-->노트북"]

가격비교웹B:Wrapper
가격비교웹B [input->>{상품명}output->>{가격,판매자}]
가격비교웹B [container-->가격]
가격비교웹B [constraint -->" 상품.분류-->MP3"]
```

그림 9. F-Logic으로 표현된 Wrapper의 메타정보

가격이 200만원 미만이고 메모리가 128MB 이상인 노트북에 대한 상품과 판매자를 찾고자 하는 사용자 질의는 다음과 같이 조합된다.

사용자 질의 : *?- 상품[상품명->X, 분류->"전자제품>>노트북", 가격->> 가격[가격<200, 판매자->Y]]*

● 질의 확장

입력된 사용자 질의는 도메인 모델정보와 그림 1의 Propagation 규칙을 통해 아래와 같이 확장된다.

?-상품[상품명->X, 분류->"전자제품>>노트북",가격->> 가격[가격<200, 판매자->Y]] --> 전자제품::상품 / ?-전자제품[상품명->X, 분류->"노트북",가격->> 가격[가격<200, 판매자->Y]]

?-전자제품 [상품명->X, 분류->"노트북",가격->> 가격[가격<200판매자->Y]] --> ?-노트북::전자제품 / 노트북 [상품명->X, 가격->> 가격[가격<200, 판매자->Y]]

● Wrapper 선택

먼저 질의에 표현된 개념인 상품, 전자제품, 노트북, 가격을 Container로 가지는 Wrapper들을 선택(쇼핑몰A, B, C, 가격비교웹A, B, C)한 뒤 제약조건 확장을 거쳐 일관성 검사를 수행하면 쇼핑몰A, B, 가격비교웹A가 최종적으로 선택된다.

● Wrapper 조합

선택된 3개의 Wrapper들은 두가지 Wrapper 조합 방법에 의해 *"(쇼핑몰A X 쇼핑몰B) ∞ 가격비교웹A"*의 형태로 조합된다.

사용자는 조합된 Wrapper를 이용해 질의 결과를 추출한다.

5. 결론

웹 통합 분야에서 분산된 웹 정보를 동적으로 추출하는 방법에 관한 기존 연구들은 정보 추출을 위해 단일 형식의 웹 사이트에 하나의 Wrapper를 이용한다. 통합된 정보에 대한 사용자 질의가 주어질 경우 다수의 Wrapper들이 Record Set 단위로 정보를 추출한 다음 통합 작업이 수행되는데, 본 논문에서 제안하는 방법은 사용자 질의에 가장 적합한 Wrapper들을 조합하여 튜플 단위로 통합된 형태로 정보를 추출한다. 따라서 분산된 웹 정보 환경에서 기존 방식에 비해 더욱 최적화된 통합 추출이 가능하며, 그 결과 사용자 질의 후 수행되는 동적 추출 작업에서 추출 작업 시간을 최소화한다.

감사의 글

이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2004-005-H00005).

참고문헌

[1] G. Wiederhold, *Intelligent Integration of Information,* Kluwer, 1996.

[2] Y. Arens, C. Knoblock, H. Chun-Nan, "Query Processing in the SIMS Information Mediator," *Advanced Planning Technology, A.Tate (ed), AAAI Press*, 1996.

[3] C. Knoblock et al. "The ARIADNE Approach to Web-Based Information Integration," *Int. Journal of Cooperative Information Systems,* vol. 10, no. 1-2, pp. 145-169, 2001.

[4] M. Kifer, G. Lausen, J. Wu, "Logical Foundations of Object-oriented and Frame-based Languages," *Journal of the ACM,* vol. 42 , no. 4, pp. 741-843, 1995.

[5] T. Fruewirth, "Theory and Practice of Constraint Handling Rules," *JLP*: pp.95-138, 1998.