XML 기반의 이질 정보의 통합 방법론 *

이경하 °, 이강찬, 이규철
충남대학교 컴퓨터공학과 데이터베이스 연구실

A XML-Based Approach to Integrate Heterogeneous Information

Kyongha Lee °, Kangchan Lee, Kyuchul Lee
Department of Computer Engineering, Chungnam National University

요 약

현재의 인터넷은 HTML 뿐만 아니라 다양한 멀티미디어 문서 포맷 및 질의 가능한 정보를 제공할 수 있도록 발전함에 따라 단순한 정보 전달만이 아닌 하나의 통신 수단으로써 활용되는 양상을 띄고 있다. 또한, XML의 등장으로 인하여 구조적 문서 정보를 전달할 수 있도록 발전하고 있다. 인터넷의 비약적인 발전에 따라 기존의 정보 시스템들은 인터넷을 통하여 기존에 존재하던 데이터들을 서비스할 수 있도록 새로 작성되거나 재구축되어 왔다. 이런 경우 기존의 정보시스템들이 제공하는 데이터들은 질의 형식 및 데이터 모델, 스키마 구조, 사용하는 시스템에서 이질적인 특성을 가지고 있으며, 서로 자치적인 시스템으로써 분산되어 존재한다는 특성을 지니고 있다.

본 논문에서는 이런 이질적으로 분산되어 있는 인터넷 데이터들을 XML을 공통 데이터 모델로 이용하는 미디에이터 방식을 이용하여 통합하는 방법(XMF: XML-Based Mediation Framework)을 제시한다.

1. 서론

많은 정보 자원들이 인터넷을 통하여 서비스됨으로써 인터넷은 하나의 정보 교환 및 검색 수단으로써의 양상을 띄게 되었다. 하지만, 인터넷 상의 정보 자원들은 정보 자원의 타입에 따라 많은 이질성을 띄게 된다. 예를 들어, 어느 한 정보 자원에서 사용하는 질의는 단순한 text가 되고 text 문서를 결과로 취하는 대신에 다른 정보 자원에서 사용하는 질의는 SQL 형식을 취하면서 관계형 데이터모델을 보여줄 수가 있다. 인터넷 상에 존재하는 정보 자원들은 다음과 같은 특징을 지니고 있다.

분산성(Distribution): 인터넷 상의 정보자원들은 지역적으로 분산되어 존재한다.

자치성(Autonomy): 인터넷과 연결되기 전부터 정보 자원들은 자치적으로 구축, 운영되어왔다. 이는 여러 정보 자원들이 통합되더라도, 서로 독립적으로 운영될 수 있는 자치성을 포함하는 것을 의미한다.

이질성(Heterogeneity): 인터넷 정보 자원들의 이질성은 사용하는 시스템 및 데이터 모델, 질의어, 스키마 등 다양한 부분에서 서로 다른 이질성을 가지게 된다.

* 본 연구는 소프트웨어연구센터의 핵심응용기술과제인 XML 저장/검색 및 분산 문서 시스템의 설계 및 구현(과제호:99-11-02-01-A-2) 으로 수행되는 과제임

이외에도, 모든 데이터 모델을 WWW 상에서 표현하기 위하여 HTML로 변환하면서 발생하는 문제가 존재하는데 그 대표적인 경우가, HTML 자체에는 구조적 정보를 표현할 수 있지 못하므로 정보 자원이 가지고 있던 스키마 정보 등을 내포하지 못하게 된다는 것이다. 따라서 본 논문에서는 이러한 정보자원들의 이질성, 분산성을 극복하면서도 각 정보 자원들의 자치성을 보장할 수 있도록 XML[1]을 이용한 미디에이터 시스템으로의 통합 방법(XMF: XML-based Mediation Framework)을 제시한다.

2. XMF의 특징

XMF는 기존의 통합 방법과는 다르게 다음과 같은 특징을 가지고 있다.

① XML을 데이터 교환/중재/표현 모델로 사용한다.

각기 다른 정보 자원들이 제공하는 데이터는 XML로 변환되어 표현된다. XML은 HTML과 달리 구조적인 데이터를 표현할 수 있으므로 그 자체가 데이터 모델이 될 수 있다. 또한 서로 다른 XML로 표현된 데이터들은 미디에이션 언어(Mediation Language)로 작성된 미디에이션 규칙(Mediation Rule)에 따라, 하나의 XML 문서로 통합되어 사용자에게 보여지게 된다.

② 미디에이터 구조를 사용한다.

미디에이터-래퍼 구조를 사용함으로써, 다른 통합 방법보다도 정보 자원에 자치성 및 확장성을 보장하면서 통합을 가능케 하며, 동적인 통합이 가능하다[2][3].

③ 미디에이션 언어 또한 XML로 작성되므로 일관된 시스템을 구성할 수 있다.

이 외에 XML을 이용함으로써 정보 자원의 인터넷 서비스 구축이 곧바로 가능하다는 점, 여러 미디어 타입의 지원이 가능하다는 점 등을 가지고 있다.

3. XMF 구조

XMF의 전체 구조는 다음과 같다.

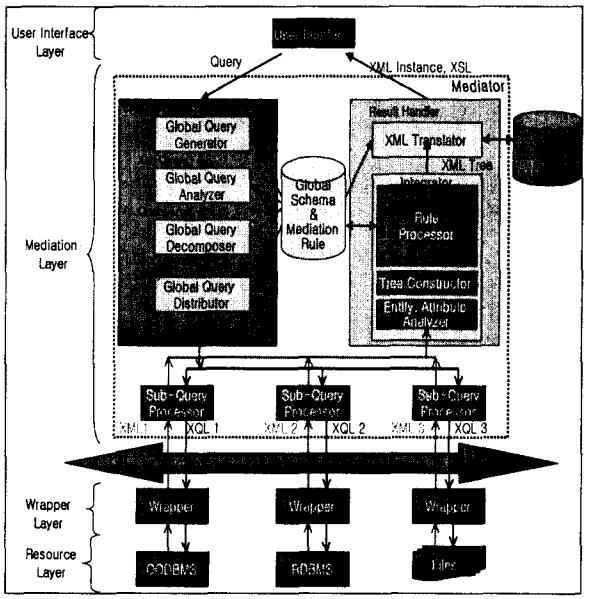

그림 1. XMF 전체 구조

XMF는 4계층 구조(User Interface 계층, 미디에이션 계층, 래퍼 계층, 정보 자원 계층)로 이루어진다. 정보 자원들은 자원 계층에 존재하게 되며, 래퍼들을 통해 인터넷 상에 연결된다. 새로운 정보 자원이 추가될 경우, 해당 정보 자원에 대해 래퍼를 추가시킴으로써 확장이 용이하다. 래퍼는 XQL 형식으로 넘어온 질의를 해당 정보 자원에서 사용하는 로컬 질의로 변환하여 데이터를 추출한 후 이를 XML 인스턴스로 변환하여 미디에이션 계층에 전달하게 된다.

미디에이션 계층은 사용자 질의를 글로벌 질의로 변환한 후 글로벌 스키마와 미디에이션 규칙을 참조하여 질의 재작성 과정을 거쳐 해당 래퍼들에 전달할 서브-질의들로 분해된 후, 서브-질의 처리기를 통하여 해당 래퍼들에게 전달하게 된다. 글로벌 스키마와 미디에이션 규칙(mediation rule)은 미디에이션 언어를 통하여 작성되며, 글로벌 질의 처리기와 결과 통합기 모듈에서는 질의 분해와, 결과 통합에 이를 참조하여 해당 연산을 수행하게 된다.

사용자 인터페이스 계층은 웹 브라우저를 이용하거나, 별도의 클라이언트로써 사용자 질의와 XML 문서의 디스플레이를 책임지게 된다.

4. XMF 미디에이션 언어

XMF에서의 미디에이션 언어(Mediation Language)는 실제 사용자에게 보여지는 글로벌 스키마의 정의, 글로벌 스키마와 각 실제 정보 자원들의 스키마 간의 변환(mapping) 규칙을 기술하는 미디에이션 규칙(Mediation Rule)을 기술한다. 또한 XMF의 중재 언어는 그 자체로도 XML 응용이며, XML 관련 표준 또는 최신 동향 기술인 XMLnamespace[4]와 XPath[5]를 따른다.

실제 중재 언어를 이용하여 두 정보 자원의 데이터들이 통합되는 예를 보이도록 한다. 정보 자원 1,2가 다음과 같은 형식의 XML 인스턴스로 래퍼를 통해 미디에이터에 export 시켰다 가정한다. 아래는 wrapper 1과 wrapper 2로부터 미디에이터에 넘어온 XML 인스턴스의 pseudo-code 표현 예이다.

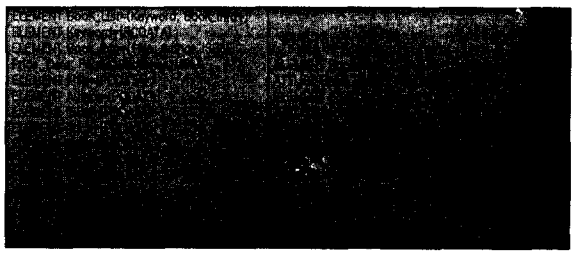

그림 2. 정보 자원 1,2에서 반환한 XML 문서 구조

미디에이션 언어는 전처리 부분 글로벌 스키마 정의 부분, 그리고 규칙 처리 부분으로 이루어진다. 전처리 부분은 각 정보 자원의 위치 정보와 메타 데이터들을 포함하며, 미디에이션시 필요한 계산을 수행하기 위한 외부 함수 이름과 변수들을 선언 할 수 있다. 정보 자원의 위치 정보는 XML namespace를 사용한다. 다음은 미디에이션 언어의 전처리 부분이다.

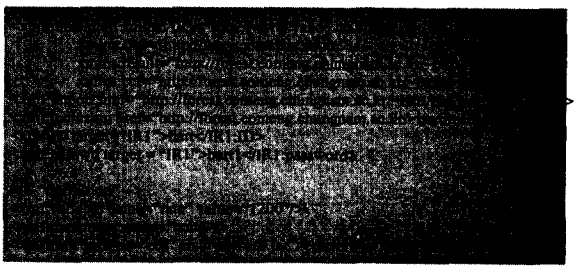

그림 3. 미디에이션 언어의 전처리 부분

위의 예에서 정보 자원 IR1, IR2는 해당 URI로 표기되며, 이 URI는 실제 정보 자원의 위치 정보이다. <xmf:ID>와

<xmf:password> 태그는 ID와 password가 필요할지 모르는 사이트에 접근 시에 필요한 ID, password를 표현하기 위한 메타 데이터 태그이다.

XMF의 미디에이션 언어는 기존의 미디에이터 시스템이 새로운 정보 자원을 통합 시에 기존의 미디에이터 위에 새 미디에이터를 놓는 계층적 구조시 발생하는 성능 문제를 극복하기 위하여 XSLT[6]에서처럼 <xmf:import>, <xmf:inclusion>을 이용하여 기존의 미디에이션 규칙을 이용, 확장하여 새 정보 자원에 대한 통합을 수행하는 구조를 지원한다.

다음은 미디에이션 언어의 글로벌 스키마 정의 부분의 예이다. 표 1에서 보인 IR1, IR2의 데이터는 다음과 같은 구조를 갖는 XML 문서로 최종 통합된다.

```
<xmf:view>
  <Integration>
      <Meta_Info>
          <Register_Info> $A</Register_Info>
          <keyword>       $B    </keyword>
          <Result_count> $C </Result_count>
          <Around_trip>  $D</Around_trip>
        <Wrapper_count> $E  </Wrapper_count>
      </Meta_Info>
      <Result no="$F">
          <Title>            $G </Title>
          <Price no="$O">$H </Price>
          <Author>           $I  </Author>
          <Publisher>        $J </Publisher>
          <Date>             $K </Date>
          <URL>              $L </URL>
          <Pages>            $M</Pages>
          <ISBN>             $N</ISBN>
      </Result>
  </Integration>
</xmf:view>
```

그림 4. 미디에이션 언어의 글로벌 스키마 정의 부분

$A-$O는 실제 미디에이션 규칙이 처리할 목표를 지칭하는 식별자이자, 미디에이터가 처리한 해당 엘리먼트와 어트리뷰트의 모든 값을 가질 수 있는 변수이다. <Meta_Info>는 미디에이터가 데이터 통합 연산을 수행하면서 얻은 여러 메타 데이터들을 포함하는 엘리먼트이다.

```
<xmf:processing>
    . . . . . .
  <xmf:rule target= "$B">
    <xmf:source>IR1:Book_List/Keyword</xmf:source>
  </xmf:rule>
  <xmf:rules integrate= "UNION">
    <xmf:rule target= "$G" operator= "$EQ">
        <xmf:source>IR1:Book_List/Book_Info/TITLE</xmf:source>
        <xmf:source>IR2:Book_Search_Result/Book_Item/Book_Title</xmf:sou  rce>
    </xmf:rule>
    <xmf:rule target= "$H" operator=$EQ ">
        <xmf:source>IR1:Book_List/Book_Info/Price</xmf:source>
        <xmf:source invoe= "ccl:convert() ">
          IR2:Book_Search_Result/Book_Item/Info/Book_Pri  ce</xmf:source>
    </xmf:rule>
    . . . . . .

  </xmf:rules>
</xmf:processing>
```

그림 5. 미디에이터 언어의 Rule processing 부분

그림 5.는 각 엘리먼트와 어트리뷰트에 따른 미디에이션 언어의 규칙 처리 부분이다. 규칙 처리 부분은 <xmf:processing> 태그로 시작이 되며, 하위 엘리먼트로 <xmf:rules> 또는 <xmf:rule> 엘리먼트를 가질 수 있다. 여기서 각 엘리먼트 또는 어트리뷰트에 해당하는 미디에이션 규칙들은 <xmf:rule> 엘리먼트의 "target" 어트리뷰트로 글로벌 스키마의 엘리먼트 또는 어트리뷰트를 지칭하고, <xmf:source> 엘리먼트로 해당 글로벌 스키마의 엘리먼트 또는 어트리뷰트가 실제 정보 자원 1,2 의 어느 부분에 해당함을 기술하게 된다. 전처리 부분은 글로벌 질의를 분해하여 각 정보 자원에 해당하는 질의를 추출하는데 필요하며, 이 후 결과 통합에서도 해당 룰을 이용하게 된다. $H에 해당하는 <xmf:source> 가 함수를 호출하는 것은 정보 자원 2의 가격 표시가 Dollar일 경우 ccf:convert() 함수로써 이를 Won으로 변환시키기 위해 외부 함수를 호출하는 예이다. 해당 엘리먼트 또는 어트리뷰트를 찾는 구문 비교(Pattern Matching)은 XPath[5]의 구문을 사용한다.

5. 결론

XMF는 많은 양의 인터넷 정보 자원들을 XML을 통하여 동일한 뷰(view)를 사용자에게 보이도록 하는 통합 모델링 방법이다. 이를 이용하면 서로 다른 정보자원들 각각의 질의를 통일시킬 수 있으며 결과의 일관적으로 통일된 뷰를 사용자에게 제공할 수 있을 것이다. 더 나아가 구조적, 의미적 이질성을 가진 데이터들이 인터넷 서비스되면서 발생하는 해당 구조 및 의미적 정보의 손실을 겪지 않으면서 통합할 수 있는 훨씬 나은 데이터 통합 방법이 될 수 있을 거라 생각되며, 인터넷 상의 정보 자원 통합 뿐만 아니라, 일반적 목적으로서의 미디에이터[2]가 사용될 수 있던 개발 환경에 모두 적용이 가능할 것이라 기대된다.

6.참고 문헌

[1]W3Consortium, "XML, eXtensible Markup Language", http://www.w3.org/TR/1998/REC-xml-19980210

[2]H. Garcia-Molina, Y.Papakonstantinou, D. Quass, A Rajaraman, Y.Sagiv, J. Ullman, V.Vassalos, J. Widom. "The TSIMMIS approach to mediation: Data models and Languages", *In Journal of Intelligent Information Systems*, vol.8, p117-132, 1997.

[3]Kangchan Lee, Kyuchul Lee, "Hybrid Database Integration(HyDIM) for Product Data Management". Proceedings of the 4th International Conference on Information Systems, 1998

[4]W3Consortium, "Namespaces in XML",http://www.w3.org/TR/REC-xml-names

[5]W3Consortium, "XPath, XML Path Language",http://www.w3.org/TR/xpath

[6]W3Consortium, "XSLT, "XSL Transformation", http://www.w3.org/TR/WD-xslt