2001 년 한국정보처리학회 추계 학술발표논문집 제 8 권 제 2 호

가상문서를 XTM 문서로 변환해 주는 자동변환시스템 설계 및 구현

* 윤여준, *조숙현, **김태현, **맹성현
충남대학교 컴퓨터과학과
*{mistone, shcho}@enya.cnu.ac.kr
**{heemang, shmyaeng}@cs.cnu.ac.kr

Design and Implementation of an Automatic Translator for Converting a Virtual Document to a XTM Document

Yeo-Jun Yun, Suck-Hyun Cho, Tae-Hyun Kim, Sung-Hyon Myaeng
*Dept. of Computer Science, Chungnam National University

요 약

가상문서 개념에 기반을 둔 디지털도서관에서는 기존의 멀티미디어 문서를 활용하여 새로운 view 를 생성하고 지식을 창출할 수 있는 기능을 제공한다. 이 접근 방법의 초점은 링크를 사용하여 기존의 자원을 연결하여 가상문서를 생성하고 필요한때 필요한 부분만 가져다가 통합하여 재현시킬 수 있는 기능과 이렇게 생성된 가상문서를 총체적으로 혹은 부분적으로 검색할 수 있다는 것이다. 반면에 W3C 에서는 Topic Maps 라는 개념을 표준안으로 하였는데, 이 개념은 근본 취지에 있어 가상문서의 개념과 매우 유사하나 문서를 끌어서 통합하는 기능이나 부분문서를 가져오는 기능 등이 없다. 여기서도 기존 문서에 링크를 걸어 사용자가 원하는 토픽이 연결될 수 있도록 하는 기능을 규정하고 있는데, 지식 표현 및 추론 등 다양한 응용에 관한 연구가 진행되어 오고 있다. 따라서 본 연구에서는 충남대학교에서 개발한 가상문서 기반 디지털도서관 개념에 이 Topic Maps 기능을 연결하여 호환성을 제공하는 것을 목표로 한다. 이러한 호환성은 디지털도서관 분야에서의 핵심 이슈이며, 실용적인 관점에서 가상문서 기반 디지털도서관의 가용성을 향상시켜 궁극적으로 외부 시스템과도 연계될 수 있는 기반을 제공할 것이다.

1. 개요

급격한 정보사회의 다변화가 이루지면서 다양한 형태의 디지털 정보가 인터넷이라는 거대 매체에 쏟아져 들어오고 있다. 이로 인해 이제는 사용자가 원하는 수많은 정보들을 인터넷에서 손쉽게 얻을 수 있게 되었다. 그러나 너무나 많은 양의 정보가 중복되고 정제되지 못한 채 뒤엉켜 있어, 디지털 정보 속에서 길을 잃고 헤매는 원인이 되었다.

이러한 문제를 해결해 줄 수 있는 방법으로 디지털 도서관의 중요성이 증대되고 있다. 체계적이고 정제된 정보들을 이용하기 때문에 디지털 도서관에서는 양질의 정보를 빠르고 쉽게 찾을 수 있다. 그러나 기존의 정보를 이용하여 새로운 문서를 만들 때 일반적인 방법을 이용할 경우 문서작성에 있어서 시간적,공간적 낭비를 가져오게 된다. 선행 연구에서는 이러한 문제를 해결하기 위해 가상문서 기술을 제안하였다. 그러나 이러한 가상문서 기술은 특정 디지털도서관에서만 사용될 수 있으며 가상문서 기술을 지원하지 않는 다른 디지털도서관과는 상호 호환이 어렵다는 문제점이 있었다. 이러한 문제점을 해결하기 위해 본 연구에서는 가상문서를 의미적 정보표현의 국제 표준 기술로 자리잡은 XTM 문서로 변환하여 가상문서 기반의 디지털 도서관의 호환성 을 극대화 하고자 가상문서를 XTM 문서로 변환해 주는 자동변환시스템을 개발하였다.[1][2]

본 논문은 과학기술부/한국과학재단 지정 소프트웨어연구센터의 지원을 받았음

2. 관련연구

2.1 가상문서

가상문서란, 기존에 존재하는 문서를 재사용할 수 있도록 XML 및 하이퍼링크의 특성을 이용하여 표현된 구조적인 문서를 의미한다. 이는 디지털 도서관 응용과 관련한 선행 연구에서 제안된 문서표현 기법으로, 이를 이용하여 정보를 표현 및 관리할 경우 새로운 문서의 작성과 관리에 드는 노력을 최소화할 수 있다.[2]

가상문서는 문서의 내용을 구성하는 역할을 하는 내포링크와 참조링크로 구성되며, 해당 문서 자체를 설명하기 위한 메타정보를 포함한다. 내포링크는 물

리적으로 존재하는 외부 문서의 내용을 링크기법을 이용하여 그대로 가상문서에 포함시킴으로써, 해당 링크의 내용이 문서를 브라우징할 때 화면에 바로 보여지게 된다. 참조링크는 하이퍼텍스트 문서에서 일반적으로 사용하는 링크와 유사하게, 가상문서가 브라우징될 때 앵커만이 표시되며 사용자의 요구에 따라 참조문서를 브라우징할 수 있게 한다.

가상문서의 이러한 구조는 원격지에 있는 멀티미디어 데이터를 복사하지 않고 링크만으로 이들을 포함시켜 이용할 수 있게 하므로 저장에 필요한 장소를 절약할 수 있으며, CD-ROM 과 같은 읽기 전용 문서에도 문서의 복사 없이 링크를 연결할 수 있다는 장점을 갖는다. 또한, 전체 문서가 아닌 링크가 된 일부분만을 다운 받아서 복합문서가 생성될 수 있으므로 문서작성에 필요한 시간을 절약할 수 있을 뿐 아니라, 가상문서를 활용하여 쉽게 자료를 생성하고 조직할 수 있으므로 개인 맞춤의 문서 구성이 가능하다는 특성을 갖는다.

2.2 XTM 문서의 구조

XTM 은 수많은 디지털 정보를 조직화하여 보다 빠르고 쉽게 원하는 정보를 찾을 수 있도록 도와주는 역할을 한다. 모든 디지털 정보(information resource)의 최소 구성 단위를 topic 으로 정의할 수 있으며 구조적이고 체계적인 구성을 위해 topic 들간의 의미적 연관성을 association 을 통해 정의할 수 있다. 즉, XTM 을 이용하여 정보를 표현한다면 물리적으로 인쇄된 책의 인덱스처럼 디지털 정보를 보다 쉽고 빠르게 이용할 수 있게 된다.[1][3]

XTM 은 디지털 정보에 포함된 지식 구조를 모델링하는 표준화된 방법을 제공하므로, 이러한 정보들을 네비게이션하고, 검색하고, 새롭게 구성하는데 있어서 새로운 수단을 제공할 수 있다. 따라서 " 지식표현" 과 " 정보관리" 사이에 존재하는 기술적·의미적인 격차를 줄이는데 기여할 수 있다.

2.3 가상문서와 XTM 문서의 비교

가상문서는 디지털 정보(information resource)와 그들 간에 있을 수 있는 관계를 링크를 이용하여 명시할 수 있기 때문에 구조적으로 새로운 문서를 작성하고 표현하는데 있어 공간적, 시간적인 노력을 최소화할 수 있다. 그러나, 이들 링크는 단순히 내포관계나 참조관계로서만 그 연관성이 표현될 뿐 의미적인 정보를 표현하기에는 다소 미흡한 특성을 갖는다.[2]

이에 반해, XTM 문서는 디지털 정보뿐만 아니라 추상적인 객체들의 의미까지도 구조적으로 명시할 수 있으며, 이들간의 관계 또한 구조적으로 표현할 수 있어, 의미표현을 목적으로 하는 구조적인 문서의 구성에 있어 뛰어난 표현력을 갖는다. 또한 XTM 문서는 XML Topic Maps(XTM)1.0 TopicMaps.Org Specification 으로 표준화되어 있어 일반성 면에서의 장점도 가지고 있다.[1]

가상문서와 XTM 문서는 모두 XML 을 기반으로 하고 있어, 문서를 구조적으로 기술할 수 있고, 이러한 구조적인 특성을 이용하여 손쉽게 문서를 가공할 수 있는 여지가 있다. 또한 하이퍼텍스트의 장점인 링크를 적극적으로 활용하여 문서를 구성하므로 다양한 정보를 복합적으로 구성할 수 있어 새로운 정보를 보다 편리하게 생성할 수 있다. 본 연구에서는 이러한 가상문서와 XTM 문서의 공통점을 최대한 이용하고, 가상문서를 보편화 시키기 위해 가상문서를 XTM 문서로 변환하는 자동 문서변환 시스템을 설계 및 구현하였다.

2.4 DTD(Document type Definition)및 스키마 개념

XML 을 기반으로 하는 문서들은 DTD 를 이용하여 그 구조적인 틀을 정형화할 수 있다는 장점을 갖는다. 가상문서와 XTM 문서는 이러한 XML 의 장점을 이용하여 구조적으로 추상화된 새로운 문서의 틀을 제공한다. 그러나, XTM 문서의 경우 정보간의 구조적인 형식뿐만 아니라 그 의미관계의 표현에 있어서의 일관성이 요구된다.

XML 문서에서는 일반적으로 의미표현의 일관성을 제공하기 위해 XML 스키마를 이용한다. XTM 문서에서도 이러한 목적으로 XML 스키마를 이용할 수 있다. XML 스키마를 이용하여 가상문서 및 XTM 문서 표현 자체의 의미를 체계적으로 정의할 수 있을 뿐만 아니라 특정 분야에서 사용되는 공통적인 의미들을 미리 정의해 두고 이를 참조하는 방식으로 사용함으로써 문서의 의미적 일관성을 이룰 수 있는 것이다.

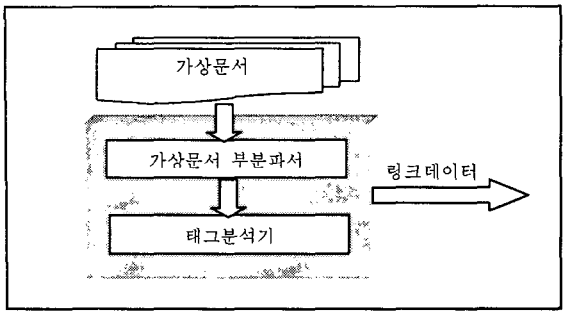
본 연구에서는 의미적 일관성을 제공하기 위해 XTM 을 위한 스키마와 가상문서를 위한 스키마를 정의하여 사용한다.

3. 변환기 설계 및 변환과정

변환기는 크게 문서처리기, 파서, 자동변환기로 분류될 수 있다. 문서처리기에서는 가상문서를 입력 받아 링크 태그를 분리하여 각 파서에 전달하는 역할을 하며, 파서에서는 링크태그의 세부 정보를 추출하여 자동변환기에 넘겨준다. 자동 변환기에서는 가상문서와 XTM 문서의 구조 정보를 이용하여 실제적인 문서 변환작업을 수행하게 된다. 각 모듈에 대한 설명은 아래에 기술한다.

3.1 문서처리기

문서처리기의 구조는 <그림 1>과 같으며 각 모듈에 대한 설명은 다음과 같다.



<그림 1> 문서처리기 구조

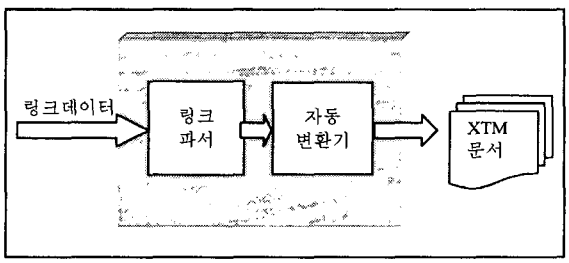
□ 가상문서 부분파서

입력 받은 가상문서를 순차적으로 처리해 나가면서, 가상문서 내에 존재하는 내포링크, 참조링크, 메타데이터를 구분하여 태그분석기에 해당 엘리먼트의 시작을 알리는 역할을 한다.

□ 태그분석기

태그분석기는 가상문서 부분파서에서 주어지는 입력에 따라, 찾아낸 엘리먼트 정보를 입력문서로부터 읽어들여 다음 단계의 파서에서 처리할 수 있는 단위의 데이터로 나눈 다음, 해당 파서에 이를 전달한다.

3.2 파서 및 자동변환기



<그림 2> 파서 및 자동변환기 구조

□ 링크 파서

태그 분석기에 의해 전달된 각 링크 정보를 파싱하여 링크 세부정보를 추출한다. 추출된 데이터는 구조체에 저장되며, 이는 자동변환기에 전달된다.

□ 자동변환기

구조체에 저장된 데이터는 가상문서 및 XTM 문서의 DTD 를 기본으로 한 매핑 관계에 따라 XTM 문서로 변환된다. 이에 대한 자세한 내용은 다음의 문서변환과정에서 기술한다.

3.3 문서변환과정

<그림 3>은 입력으로 주어지는 가상문서 예이다. 이와 같은 가상문서를 문서변환기의 입력으로 주면, 해당 문서는 문서처리기의 가상문서 부분파서에 의해 우선 <ELink>, <Rlink>, <Meta> 태그 부분이 추출된다. 추출된 태그는 태그 분석기에 의해 각각의 링크별로 나뉘어 처리되는데 여러 개의 링크가 나오기 때문에 한번에 처리할 수 있는 단위로 나뉘어 각각의 파서로 이동한다.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<Elink id="aa" href="http://168.188.128.31/k4.txt#text(124,255)"
       date="2001/02/13" actuatedefault="auto" autoDelete="NO"/>
<RLink id="bb" showdefault="new" actuatedefault="auto">
<Source id="cc" Is_generic="YES" href="#E1/text(66,4)"
        autoDelete="YES"/>
<Destination id="dd" href="http://enquest21.com" date="null"
        autoDelete="YES"/>
<MetaData>
       <DC_TITLE value="정보검색"/>
```

<그림 3> 입력 가상문서의 예

링크파서에서는 링크단위의 세부 정보(예를들어 Elink 에서 id 나 date 값)를 추출하여 정의된 구조체에 저장한다. 구조체에 저장된 데이터는 자동변환기에서 매핑 과정을 거치게 되는데 대략적인 의미 매핑 관계는 <표 1>과 같다.

	가상문서	XTM 문서
Elink	Elink id	Topic id
	href	resourceRef
	date	xlink:href=" vdate.xtm#date"
Rlink	Source id	Topic id
	Destination id	Topic id
		association
Meta-data	Metadata	topic
	title	xlink:href=" vbase.xtm#title"

<표 1> 의미적 매핑관계

여기에서 Elink, Rlink, Metadata 엘리먼트가 각각 topic 으로 매핑 되는데 이것은 가상문서를 구성하는 각 링크 엘리먼트 자체가 의미적인 단위가 될 수 있기 때문이다. Elink 부분에서 date 을 살펴보면 XTM 으로 변환될 때 스키마를 이용하여 의미적인 부분을 정확히 명시 할 수 있게 하였다. Rlink 에서는 source 와 destination 두 부분이 나오는데 각각이 topic 으로 매핑되며 그들의 관계(일대일, 일대다 등)가 다시 association 부분에서 정의된다. Metadata 부분의 title 역시 스키마를 이용하여 표현하였다.[1][2]

```
<?xml version="1.0" encoding="iso-8859-1"?>
<topicMap>
<topic id = "aa">
<subjectIdentity>
<resourceRef= "http://168.188.128.31/k4.txt#text(124,255)">
                ---             ---
<instanceOf>
<topicRef xlink:href="vdate.xtm#Date" />
                ---             ---
<topic id="cc">
<occurrence>
<instanceOf>
<topicRef xlink:href="vlink.xtm#generic"/>
                ---             ---
<association>
<topicRef xlink:href="vlink.xtm#OneToOneGeneric"/>
                ---             ---
<occurrence>
<topicRef xlink:href="vbase.xtm#Description" />
<resourceData>"정보검색"</resourceData>
```

<그림 4> 변환된 XTM 문서

<그림 4>는 입력예제를 문서변환기에 입력으로 주었을 경우에, 그 결과로 나오는 XTM 문서를 나타낸 것이다. XTM 은 구조적 문서구성에 대한 뛰어난 표현력을 가지고 있다. 따라서 가상문서가 본래 갖고 있는 구조나 의미를 변형시키지 않으면서 가상문서를 XTM 문서로 변환이 가능하다.

4. 결론 및 향후연구과제

본 논문에서는 디지털 도서관의 핵심 기술로 만들어진 가상문서를 XTM 문서로 자동변환 시키는 방법에 대하여 기술하였다. 가상문서의 구조적 특성과 의미를 그대로 살릴 수 있도록 가상문서와 XTM 간의 매핑 관계를 정의하여 XTM 문서로 변환함으로써 (가상문서의 장점과 XTM 의 장점을 모두 사용할 수 있도록 하여) 가상문서의 호환성을 극대화하여 궁극적으로 디지털 도서관에 표현된 가상문서 정보를 보다 표준화된 외부환경에서도 사용할 수 있는 기술적인 방법론을 제시하였다.

" XTM 문서→ 가상문서 변환연구 수행중" 향후 연구에서는 XTM 문서를 쉽게 작성할 수 있는 인터페이스를 제공하고 XTM 에서 표현된 주제들과 이들이 이루고 있는 관계구조를 가시화하며, 이 구조를 쉽게 네비게이션하고 구조검색을 가능하게 할 수 있도록 하는 것이 필요하다.

5. 참고문헌

[1] XML Topic Maps (XTM) 1.0 TopicMaps.Org Specification
" http://www.topicmaps.org/xtm/1.0/"
[2] 맹성현, 이만호, 강지훈, 외 6 인, " A Digital Library System for Easy Creation/Manipulation of New Documents over Existing Resources", RIAO 2000.
[3] Steve Pepper, " Navigating haystacks and discovering needles", Markup Languages: Theory & Practice 1.4 (1999): 41-68.