한국지능정보시스템학회논문지 제10권 제1호 2004년 6월(pp.27~45)

인터넷쇼핑몰의 사업자신원정보 구조화 방안

장용식
한신대학교 경상대학 e-비즈니스학과
(yschang@hs.ac.kr)

온라인 쇼핑이 증가하고 있는 가운데, 우리나라는 "전자상거래 등에서의 소비자보호에 관한 법률"로 사업자신원정보의 기재를 의무화하고 있다. 인터넷쇼핑몰들은 대부분 홈페이지 하단에 반구조적인 형태로 사업자신원정보를 기재하고 있으나, 기재항목과 표현형식이 구조화되어 있지 않아 사업자의 신원파악이 어렵기 때문에 소비자의 신뢰도에 나쁜 영향을 미칠 수 있다. 이에 본 연구는 사업자신원정보를 정확하게 표현하는 세 가지 구조화 방안 - HTML기반 구조, XML기반 구조, XML data island기반 구조 - 을 제시하고 비교하였으며, HTML기반구조와 XML data island 기반구조의 추출성능을 실험으로 비교하였다. 60개의 인터넷쇼핑몰 표본에 대해 실험결과, XML data island 기반구조는 사업자신원정보 추출시간이 웹문서의 크기와는 관계가 없으나, HTML기반구조는 웹문서의 크기에 비례하였다. 또한, 평균 추출시간을 비교한 결과 XML data island 기반구조가 HTML기반구조보다 정보추출면에서 더 효율적이며 효과적임을 검정하였다.

논문접수일 : 2003년 12월 게재확정일 : 2004년 3월 교신저자 : 장용식

1. 서론

온라인 쇼핑이 계속 증가하고 있는 가운데, OECD(Organization for Economic Co-operation and Development)에서는 1999년에 "전자상거래 소비자보호를 위한 가이드라인"을 이사회 권고사항으로 발표하였으며, 소비자보호 가이드라인과 이에 따른 각국의 전자상거래 소비자보호 가이드라인들은 온라인 사업자가 자신의 신원에 관한 정보를 소비자에게 온라인으로 제공할 것을 명시하고 있다(성낙현 & 장용식, 2002). OECD 가이드라인에서 인터넷쇼핑몰이 사업자신원정보(Business Information)로서 공개하도록 제안하는 정보로는 상호, 대표자성명, 사업자등록번호, 주사업장주소, 전화번호, 팩스, 전자우편주소 등이 있다(OECD, 1999).

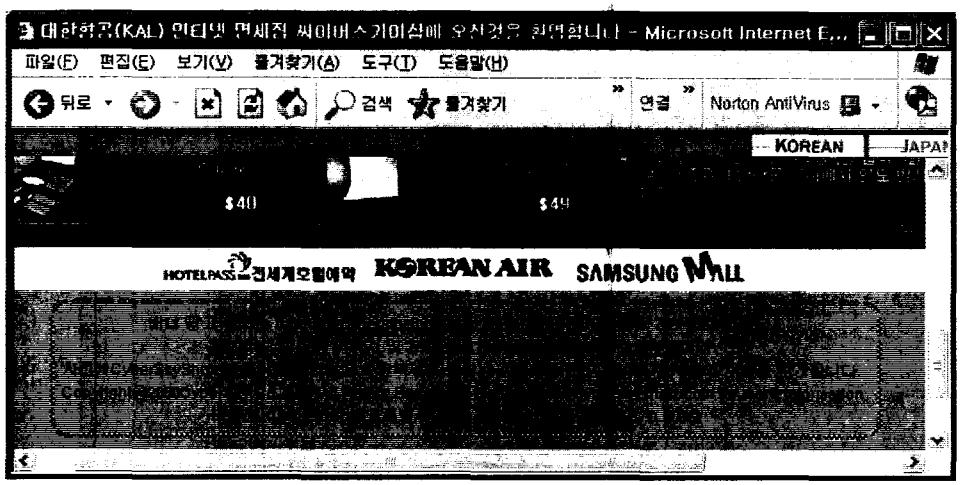
만일, 어떤 인터넷쇼핑몰이 고의적으로 사업자신원정보의 전체 또는 일부를 기재하지 않는다면 전자상거래 부당행위를 할 의도를 가진 쇼핑몰로 간주할 수 있기 때문에, 문제발생 이전에 사전 예방을 위한 조치로서 "전자상거래 등에서의 소비자보호에 관한 법률"에 의거하여 사업자신원정보를 제대로 기재하지 않은 인터넷쇼핑몰을 조기에 발견하고 기재를 권유하는 일은 전자상거래의 소비자 신뢰도 제고에 기여할 것이다(성낙현 & 장용식, 2002).

2003년 2/4분기 국내의 인터넷쇼핑몰 사업체 수는 3,284개로서 계속 증가하고 있으며(통계청, 2003), 2003년 12월 현재, 우리나라의 인터넷종합쇼핑몰 중 231개 사이트의 홈페이지를 조사한 결과, 사업자신원정보를 문자형태로 제공하고 있는 사이트는 186개, 이미지 형태로 표현하고 있는 사이트는 41개이며, 4개 사이트는 기재하지 않고 있다. 기재율은 상당히 높은 편이나, 기재항목과 표현내용이 구조화되어 있지 않아 각 인터넷쇼핑몰마다 다른 형식으로 기재하고 있기 때문에 사업자의 신원 파악과 정보추출이 쉽지 않다.

웹문서는 구조적(Structured), 반구조적(Semi-structured), 그리고 비구조적인(Un-structured) 형태로 작성되고 있다. 구조적 문서란 한 튜플내의 각 항목이 구분자 또는 항목의 순서 등과 같이 일정한 문법적 어구 배열의 단서에 근거하여 정확하게 추출될 수 있는 문서인데 비해, 비구조적 문서란 항목을 정확하게 추출하기 위해서는 언어학적인 지식이 요구되는 문서이다. 그리고, 반구조적인 문서란 비구조적이 아닌 문서로서 항목이 누락되어 있거나, 한 항목이 여러 개의 값을 갖거나, 항목 표현의 일관성이 일부 결여된 문서이다 (Hsu and Dung, 1998). <그림 1>은 사업자신원정보를 비구조적인 방법으로 표현한 웹문서의 예이다.

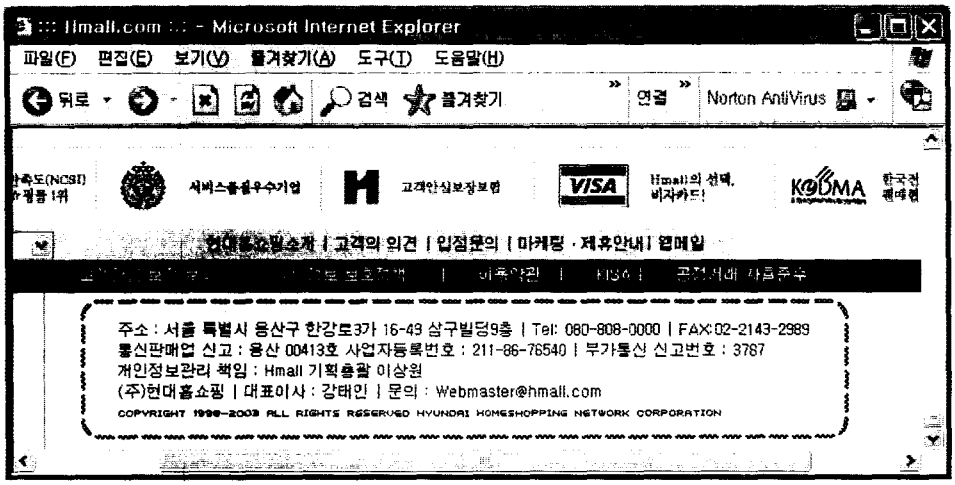
한편, <그림 2>는 ㈜현대홈쇼핑에서 운영하는 Hmall.com(www.hmall.com)의 홈페이지이며, <그림 3>은 HTML문서로 작성된 반구조적인 형태의 사업자신원정보 예이다. 상호의 경우에 속성명을 기재하지 않았으며, 속성 간의 구분이 각기 다르고, 각 속성명과 속성값의 표현에 대한 일관성이 결여되어 있다.

현재, 미국을 포함한 세계 대부분의 나라들은 사업자신원정보의 일부를 비구조적인 방법으로 작성하기 때문에 사업자의 신원을 정확하게 파악하기가 쉽지 않다. 이에 비해, 한국의 경우에는 OECD 권고에 따라 2002년 7월부터 발효된 "전자상거래 등에서의 소비자보호에 관한 법률"에 의



<그림 1> 사업자신원정보를 비구조적인 방법으로 표현한 웹문서의 예

(대한항공 면세점, http://www.cyberskyshop.com, 2003년 12월 현재, 점선 안은 사업자신원정보 일부를 표현하고 있음)



<그림 2> 사업자신원정보를 반구조적인 방법으로 표현한 웹문서의 예

((주)현대홈쇼핑, http://www.hmall.com, 2003년 12월 현재, 점선 안은 사업자신원정보를 표현하고 있음)

```
.....
<td width=587 valign=top style=padding-top:22px;>
주소 : 서울 특별시 용산구 한강로3가 16-49 삼구빌딩9층 | Tel: 080-808-0000 | FAX:02-2143-2989 <br>
통신판매업 신고 : 용산 00413호 사업자등록번호 : 211-86-76540 | 부가통신 신고번호 : 3787<br>
개인정보관리 책임 : Hmall 기획총괄 이상원<br>
(주)현대홈쇼핑 | 대표이사 : 강태인 |
문의 : <a href=mailto:Webmaster@hmall.com>Webmaster@hmall.com</a><br>
</td>
.....
```

<그림 3> Hmall.com의 HTML문서: 홈페이지의 사업자신원정보 부분

해, 온라인 사업자는 소비자가 사업자의 신원정보를 쉽게 알 수 있도록 표시하는 것을 의무화하고 있으며, 대부분의 인터넷쇼핑몰이 HTML(Hyper Text Markup Language)을 이용하여 홈페이지 하단부에 사업자신원정보를 반구조적인 형태로 제공하고 있다.

비구조적으로 작성된 사업자신원정보 또는 기재항목과 표현형식이 구조화되어 있지 않은 반구조적 사업자신원정보는 소비자의 사업자신원 파악에 어려움을 주고 신뢰도에 나쁜 영향을 줄 수 있다. 따라서, 사업자가 "전자상거래 등에서의 소비자보호에 관한 법률"에 의거 신원정보를 작성 시 준거할 만한 사업자신원정보의 구조적인 틀을 제공하는 것은 건전한 전자상거래 발전에 기여할 것이다. 이것은 소비자, 사업자, 공공기관 관점에서 다음과 같은 잇점이 있다.

- 소비자: 사업자 신원 파악이 쉽고 사업자를 신뢰할 수 있다.
- 사업자: 사업자 신원정보 작성이 용이하며, 사업자 신원을 정확하게 밝힘으로써 소비자의 신

뢰를 얻을 수 있다.

- 공공기관: 효율적이며 효과적인 방법으로 부당행위의 소지가 될 수 있는 사이트를 조기에 발견함으로써, 소비자 보호와 국가 전자상거래 발전에 기여할 수 있다. 최근 웹문서 표현의 복잡성과 조사대상사이트의 증가 및 웹페이지 변화에 따라 사람이 사업자신원정보를 조사하기보다는, 일차적으로는 자동화된 소프트웨어 에이전트를 이용하는 것이 효율적이다. 이 경우 HTML과 같은 반구조적 또는 비구조적 문서형태보다는, XML(extensible Markup Language)과 같은 구조화된 형태의 구조적 또는 반구조적 사업자신원정보는 표현이 명확하고 효율적이며 효과적인 정보검색이 가능하여 불필요한 자원과 시간을 줄일 수 있다.

본 연구는 목적은 소비자보호, 사업자의 신뢰도 제고, 사업자신원정보 표기에 대한 조사관점에서, 현재의 사업자신원정보 표현체계를 보완하는 구조화 방안을 제시하고 추출알고리즘의 성능을 검정하는데 있다. 이를 위하여 제 2장에서 관련문헌을 살펴보고, 제 3장에서 사업자신원정보구조의 틀을 제안하며, 제 4장에서는 사업자신원정보 표현을 위한 세가지 구조화 방안을 제안하고, 제 5장에서는 이 들을 서로 비교하였다. 제 6장에서는 현재 사업자신원정보 표현체계와 XML data island 구조에 대하여 실험을 통하여 효율성 및 효과성 관점에서 비교 검정하였으며, 마지막으로 결론을 제시하였다.

2. 관련연구

전자상거래의 발전과 더불어 인터넷쇼핑몰이 계속 증가하는 가운데 부당행위를 하는 사업자로부터의 소비자보호가 문제시 되고 있었다. 1999년에 이르러서야 OECD에서 "전자상거래 소비자보호를 위한 가이드라인"을 이사회 권고사항으로 발표하였으며, 소비자보호 가이드라인과 이에 따른 각국의 전자상거래 소비자보호 가이드라인 들은 온라인 사업자가 자신의 신원에 관한 정보를 소비자에게 온라인으로 제공할 것을 명시하고 있다. OECD가 권고하는 사업자신원정보(Business Information)는 상호, 대표자성명, 사업자등록번호, 주사업장주소, 전화번호, 팩스, 전자우편주소 등이 있다(OECD, 1999).

이에, 우리나라에서는 세계에서 가장 앞선 2002년 7월에 "전자상거래 등에서의 소비자보호에 관한 법률"을 공표하였으며, 사업자신원정보를 제대로 기재하지 않은 인터넷쇼핑몰을 조기에 발견하고 기재를 권유하여 전자상거래의 소비자 신뢰도 제고에 기여할 수 있는 기틀을 마련하였다. 이러한 가운데, 성낙현과 장용식(2002)은 처음으로 우리나라 인터넷쇼핑몰에 대한 사업자신원정보 평균기재율을 조사한 바 있다.

사업자신원정보의 기재권고도 중요하지만, 이를 표현하기 위해 준거할 구조적 표현체계를 제시하는 것도 상당히 중요하다. 아직까지 이러한 연구가 없기 때문에 소비자, 사업자, 공공기관 측면에서 제안하는 본 연구의 구조적 표현방안 들은 구체적인 의미를 갖게 될 것이다.

한편, 웹문서로부터 자동적인 정보추출을 위한 연구로는, 상품정보 속성과 동의어 기반의 휴리스틱 검색, 패턴매칭, 그리고 상품정보 표현형태의 귀납적 학습을 이용하여 상품정보를 추출하는 비교쇼핑 에이전트에 관한 연구(Doorenbos, 1997), 추출 대상의 속성 주변에 HTML 태그를 넣어 정보를 추출하는 연구(Ashish and Knoblock, 1997), HTML 문서의 FAQ 정보를 추출하기 위한 템플

릿 기반의 접근(Hsu and Yih, 1997), HTML 태그를 구분자로 이용하는 연구(Kushmeick, 1997), 속성간의 분리자를 이용한 연구(Hsu and Dung, 1998), 추출규칙에 기반을 둔 알고리즘 연구(Muslea et al., 1998)가 있다. 최근에 웹문서의 반구조적인 사업자신원정보 추출을 위해 동의어 및 지시어 기반의 사업자신원정보조사에이전트시스템의 구조와 추출알고리즘을 제안하고, 이를 이용하여 추출 효율성과 효과성을 검정한 연구(성낙현 & 장용식, 2002; 성낙현, 2004)가 있다. 본 연구에서는 사업자신원정보 추출의 효율성과 효과성을 검정하기 위하여 사업자신원정보의 속성에 기반을 둔 HTML 태그와 XML 태그를 이용하여 사업자신원정보를 추출하였다.

3. 사업자신원정보 구조

웹페이지에 기재하는 사업자신원정보의 각 속성은 (속성명, 속성값) 으로 구성되어 있다. <그림 2>의 Hmall.com이 제공하는 사업자신원정보를 프레임(Frame) 형태로 표현하면 <그림 4>와 같다.

사업자신원정보는 각 사업체마다 기재항목과 표현형식이 다양하다. 어떤 속성은 한 개 이상의 속성값으로 표기되기도 한다. 속성명의 경우, 다양한 동의어로 기재되어 있고, 어떤 경우는 동일한 용어가 서로 다른 의미로 사용되기도 한다. 예를 들면, Hmall.com 경우에서, "문의"는 "전자우편주소"라는 속성명을 의미하고 있으나, "Webmaster@hmall.com"이라는 속성값을 보지 않으면 "전화번호"라는 속성명으로 인식될 수도 있다. 이런 경우는 속성값의 의미로부터 속성명을 추론하여야 한다. 속성값 역시 다양한 형식으로 표현되어 있고, 정확하게 기재되어 있지 않은 경우도 있다. 예를 들면, 많은 쇼핑몰이 상호를 웹사이트명으로 기재하기도 하며, 주사업장 주소의 경우는 주소표현체계에 따르지 않고 기재하기도 한다.

"상호"와 "대표자성명"을 제외한 사업자신원정보 속성값은 표현구조를 가지고 있다. 즉, "사업자등록번호"는 "청(서구분(3자리 숫자)", "개인(법인구분코드(2자리 숫자)", "일련번호(4자리 숫자)", "검증번호(1자리 숫자)"의 조합으로 되어 있다. "주사업장주소"는 우리나라의 경우 다양한 형태의 주소표현체계가 있으며, "OO시 OO구 OO동(또는 로) OO가 OO번지" 또는 "OO도 OO시 OO읍(또는 면) OO리 OO번지" 등이 그 예이다. "전

{{ 사업자신원정보
  도메인명: www.hmall.com
  속성: (상호, (NULL, "(주)현대홈쇼핑"))
        (대표자성명, ("대표이사", "강태인"))
        (사업자등록번호, ("사업자등록번호", "211-86-76540"))
        (주사업장주소, ("주소", "서울특별시 용산구 한강로 3가 16-49 삼구빌딩 9층"))
        (전화번호, ("Tel", "080-808-0800"))
        (팩스번호, ("FAX", "02-2143-2989"))
        (전자우편주소, ("문의", "Webmaster@hmall.com")) }}

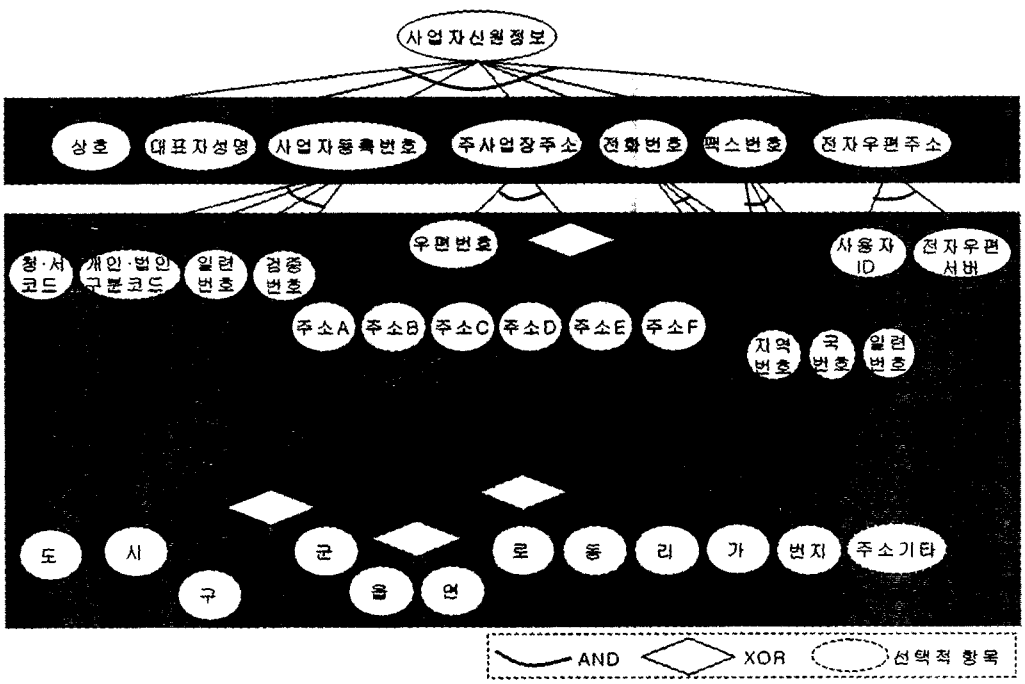
<그림 4> Hmall.com의 사업자신원정보

화번호"와 "팩스번호"는 "지역번호(숫자)", "국번호(숫자)", "일련번호(숫자)"로 되어 있으며, "전자우편주소"는 "사용자ID@전자우편서버"로 표현되고 있다. 사업자등록번호, 주사업장주소, 전화번호, 팩스번호는 각 나라마다 속성을 구성하는 요소들이 다를 수 있다.

<그림 5>는 사업자신원정보의 구조를 그림으로 나타낸 것이다. 사업자신원정보의 구조는 사업자신원정보를 구성하는 상호, 대표자성명, 사업자등록번호, 주사업장주소, 전화번호, 팩스번호, 전자우편주소 등을 나타내는 속성층과 각 속성값의 구성요소를 나타내는 표현구조층의 2계층 구조로 구분할 수 있다. 표현구조층은 각 나라마다 다르게 표현될 수 있다. <그림 5>의 사업자등록번호, 주사업장주소, 전화번호, 팩스번호는 우리나라의 경우에 해당하는 구조이다. 전화번호, 팩스번호, 전자우편주소 등은 여러 개의 값으로 기재되기도 하고, 주사업장주소의 경우 주소체계 특성상 주소A, 주소B 등과 같은 다양한 형태의 주소표현 중 한 가지로 표현하기 때문에 웹문서의 사업자신원정보는 반구조적인 형태를 갖게 된다.

4. 사업자신원정보 표현 구조화 방안

사업자신원정보의 기재를 위한 구조화 방안은 소비자, 사업자, 그리고 공공기관 관점에서 고려되어야 한다. 본 연구에서는 현 시점에서 적용이 가능한 구조로서, 사업자신원정보의 표현방식에 따라 HTML기반 구조, XML기반 구조, 혼합형인



<그림 5> 사업자신원정보 구조

XML data island기반 구조의 세가지 사업자신원정보 구조화 방안을 제시하였다.

4.1 HTML기반 구조

이 구조는 HTML을 이용하여 사업자신원정보를 표현하는 가장 간단한 방법이다. 인터넷쇼핑몰 사업자는 공공기관에서 권유하는 사업자신원정보의 구조화된 기재 방식에 따라 그들의 신원정보를 작성하여 홈페이지에 끼워 넣는다.

<그림 6>은 사업자신원정보를 <그림 5>의 속성층에 맞추어 HTML의 표 형식으로 작성한 예로서, 인터넷쇼핑몰 홈페이지의 하단부에 삽입하면 된다.

현재, 대부분의 인터넷쇼핑몰 홈페이지가 HTML기반으로 작성되어 있기 때문에 손쉽게 적용할 수 있다. 하지만, 이 구조는 사업자가 사업자신원정보의 구조와 속성명을 임의대로 변경하기가 쉬워서 정해진 속성값 표현체계를 따르지 않을 수도 있으며, HTML을 이용하여 표현구조층을 표현하는 것은 정보의 표현 및 이용측면에서 바람직하지 않다. 또한, 공공기관이 소프트웨어 에이전트를 이용한 조사 시에 추출 정확도를 높이기 위해서는 웹사이트의 변경 및 신규 진출에 따라 다양하게 표현되는 속성명을 시스템에 등록 관리하여야 하며(성낙현 & 장용식, 2002; 성낙현, 2004), 각 사업자별 홈페이지의 표현구조 변화에 따른 유지보수가 어렵고 비용이 많이 들며, 정보 추출 및 확인을 위한 노력이 많이 든다.

4.2 XML기반 구조

이 방법은 사업자신원정보의 구조와 표현구조를 XML 문서로 구조화하는 방법이다. 공공기관에서 제공하는 사업자신원정보 구조를 정의한 DTD(Document Type Definition) 문서와 사업자신원정보 표현형식에 관한 XSL(eXtensible Stylesheet Language) 문서를 기반으로 인터넷쇼핑몰사업자는 홈페이지를 XML로 작성하고 사업자신원정보를 나타낸다. <그림 5>의 사업자신원정보 구조에서 속성층과 표현구조층에 대하여 DTD로 정의하면 <그림 7>과 같다. 주사업장주소의 경우 다양한 주소체계를 반영하며, 전화번호, 팩스번호, 전자우편주소는 여러 개의 값을 가질 수 있도록 정의하였다. 이는 반구조적인 문서의 표현으로 사업자의 신원을 정확하게 표현 가능하다. <그림 8>은 DTD를 참조하여 XML 문서로 작성한 홈페이지의 예이다.

```
<!-- 사업자신원정보 -->
<table>
<tr><td>상호</td><td>(주)현대홈쇼핑</td></tr>
<tr><td>대표자성명</td><td>강태인</td></tr>
<tr><td>사업자등록번호</td><td>211-86-76540</td></tr>
<tr><td>주소</td><td>서울특별시 용산구 한강로3가 16-49 삼구빌딩9층</td></tr>
<tr><td>전화번호</td><td>080-808-0000, 02-500-0114</td></tr>
<tr><td>팩스번호</td><td>02-2143-2989</td></tr>
<tr><td>전자메일주소</td><td>Webmaster@hmall.com</td></tr>
</table>
```

<그림 6> HTML 형태의 사업자신원정보 문서의 예

```
<?xml version="1.0" encoding="euc-kr" ?>

.....

<!ELEMENT 사업자신원정보 (상호, 대표자성명, 사업자등록번호, 주사업장주소, 전화번호+,
          팩스번호+, 전자우편주소+)>
<!ELEMENT 상호 (#PCDATA)>
<!ELEMENT 대표자성명 (#PCDATA)>
<!ELEMENT 사업자등록번호 (청서코드, 개인법인구분코드, 일련번호, 검증코드)>
<!ELEMENT 주사업장주소 (주소A|주소B|주소C|주소D|주소E|주소F)>
<!ELEMENT 전화번호 (지역번호, 국번호, 일련번호)>
<!ELEMENT 팩스번호 (지역번호, 국번호, 일련번호)>
<!ELEMENT 전자우편주소 (사용자ID, 전자메일서버)>
<!ELEMENT 청서코드 (#PCDATA)>
<!ELEMENT 개인법인구분코드 (#PCDATA)>
<!ELEMENT 일련번호 (#PCDATA)>
<!ELEMENT 검증코드 (#PCDATA)>
<!ELEMENT 주소A(시, 구, (동|로), 가, 번지, 주소기타?)>
<!ELEMENT 주소B(시, 군, (읍|면), 리, 번지, 주소기타?)>
<!ELEMENT 주소C(도, 시, 구, (읍|면), 리, 번지, 주소기타?)>
<!ELEMENT 주소D(도, 시, 구, 동, 번지, 주소기타?)>
<!ELEMENT 주소E(도, (시|군), (읍|면), 리, 번지, 주소기타?)>
<!ELEMENT 주소F(도, 시, 동, 번지, 주소기타?)>
<!ELEMENT 시 (#PCDATA)>
<!ELEMENT 도 (#PCDATA)>
<!ELEMENT 구 (#PCDATA)>
<!ELEMENT 군 (#PCDATA)>
<!ELEMENT 읍 (#PCDATA)>
<!ELEMENT 면 (#PCDATA)>
<!ELEMENT 동 (#PCDATA)>
<!ELEMENT 로 (#PCDATA)>
<!ELEMENT 리 (#PCDATA)>
<!ELEMENT 가 (#PCDATA)>
<!ELEMENT 번지 (#PCDATA)>
<!ELEMENT 주소기타 (#PCDATA)>
<!ELEMENT 지역번호 (#PCDATA)>
<!ELEMENT 국번호 (#PCDATA)>
<!ELEMENT 사용자ID (#PCDATA)>
<!ELEMENT 전자메일서버 (#PCDATA)>
```

<그림 7> 사업자신원정보구조를 정의한 DTD 문서(Home.dtd)

```
<?xml version="1.0" encoding="euc-kr" ?>
<?xml:stylesheet type="text/xsl" href="Home.xsl" ?>
<!DOCTYPE 홈페이지 SYSTEM "Home.dtd">

        .....

        <사업자신원정보>
                <상호>(주)현대홈쇼핑</상호>
                <대표자성명>강태인</대표자성명>
                <사업자등록번호>
                        <청서코드>211</청서코드>
                        <개인법인구분코드>86</개인법인구분코드>
                        <일련번호>7654</일련번호>
                        <검증코드>0</검증코드>
                 </사업자등록번호>
                 <주소>
                        <주소A>
                            <시>서울특별시</시>
                            <구>용산구</구>
                            <로>한강로</로>
                            <가>3가</가>
                            <번지>16-49번지</번지>
                            <주소기타>삼구빌딩</주소기타>
                        </주소A>
      </주소>
                <전화번호>
                        <지역번호>080</지역번호>
                        <국번호>808</국번호>
                        <일련번호>0000</일련번호>
                </전화번호>
                <전화번호>
                        <지역번호>02</지역번호>
                        <국번호>500</국번호>
                        <일련번호>0114</일련번호>
                </전화번호>
                <팩스번호>
                        <지역번호>02</지역번호>
                        <국번호>2143</국번호>
                        <일련번호>2989</일련번호>
                </팩스번호>
                <전자메일주소>
          <사용자ID>Webmaster</사용자ID>
          <전자메일서버>hmall.com</전자메일서버>
                </전자메일주소>
        </사업자신원정보>
.....
```

<그림 8> 사업자신원정보를 XML 문서로 작성한 홈페이지의 예

XML체계는 사업자신원정보의 구조가 명확하고 정보의 추출이 쉬우나, HTML 보다 상대적으로 작성이 어렵다. 현재 대부분의 웹문서들이 HTML 또는 HTML에 기반을 둔 스크립트 언어(ASP, PHP, JSP 등)로 작성되어 있어, 이 구조의 적용은 웹페이지에 많은 변화가 요구되기 때문에 현실적으로 도입은 이른 편이다.

4.3 XML data island기반 구조

XML data island기반 구조는 HTML기반 구조와 XML기반 구조의 장점들을 이용하는 혼합형 구조이다. 이 구조는 공공기관이 제시하는 사업자신원정보 구조를 정의한 DTD 문서를 적용하여 사업자신원정보를 기술한 XML data island를 HTML 홈페이지에 데이터 바인딩(Data binding)하는 방법이다.

사업자신원정보를 <그림 5>의 속성층과 표현구조층을 고려하여 XML data island로 작성하면 데이터 바인딩이 복잡해지기 때문에, 구조를 단순화하여 속성층으로만 표현한 구조화된 DTD 문서는 <그림 9>와 같다.

<그림 10>은 Hmall.com의 홈페이지는 <그림 9>의 DTD를 참조하여 작성한 XML 형태의 사업자신원정보로서 XML data island이며, 하단부에 XML형태의 사업자신원정보를 HTML 테이블에 데이터 바인딩하였다(<그림 11> 참조). 사업자는 데이터 바인딩 형식에 따라 웹브라우저에서 다양한 형태로 표현할 수 있다.

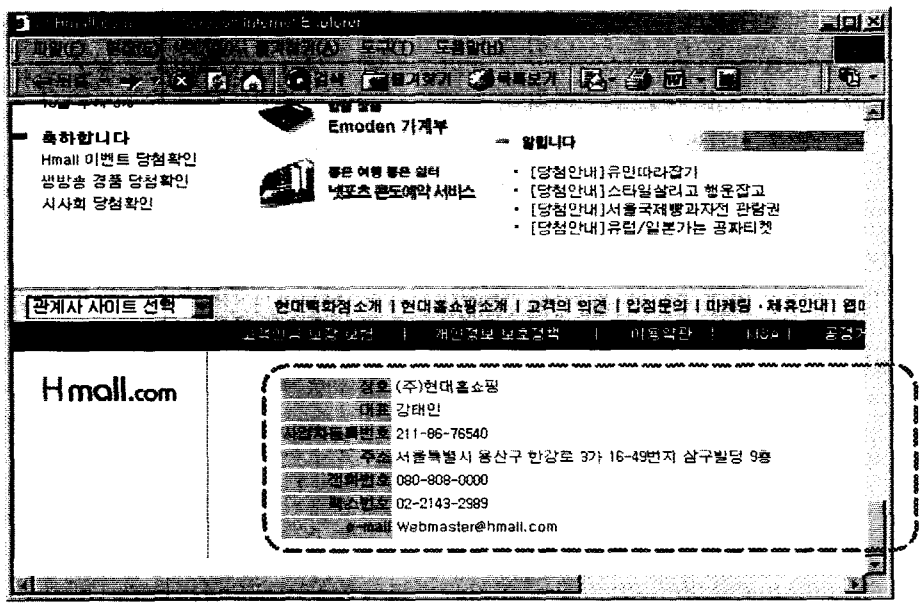
```
<?xml version="1.0" encoding="euc-kr" ?>
<!ELEMENT 사업자신원정보 (상호, 대표자성명, 사업자등록번호, 사업장주소, 전화번호,
          팩스번호, 전자우편주소)>
<!ELEMENT 상호 (#PCDATA)>
<!ELEMENT 대표자성명 (#PCDATA)>
<!ELEMENT 사업자등록번호 (#PCDATA)>
<!ELEMENT 사업장주소 (#PCDATA)>
<!ELEMENT 전화번호 (#PCDATA)>
<!ELEMENT 팩스번호 (#PCDATA)>
<!ELEMENT 전자우편주소 (#PCDATA)>
```

<그림 9> 사업자신원정보구조를 정의한 DTD 문서(BusinessInfo.dtd)

```
<html>
<body>

<!-- [1] 홈페이지 정보 -->
.....

<!-- [2] 사업자신원정보 -->
<!-- [2-1] XML Data Island -->
<!DOCTYPE 사업자신원정보 SYSTEM "BusinessInfo.dtd">
<xml id = "BI">
    <사업자신원정보>
        <상호>(주)현대홈쇼핑</상호>
        <대표자성명>강태인</대표자성명>
        <사업자등록번호>211-86-76540</사업자등록번호>
        <주사업장주소>서울특별시 용산구 한강로 3가 16-49번지 삼구빌딩 9층</주사업장주소>
```

```
        <전화번호>080-808-0000</전화번호>
        <팩스번호>02-2143-2989</팩스번호>
        <전자우편주소>Webmaster@hmall.com</전자우편주소>
    </사업자신원정보>
</xml>

<!-- [2-2] 데이터 바인딩 -->
<table datasrc = "#BI">
  <tr> <td align=right bgcolor=lightgrey>상호</td>
    <td><span datasrc="#BI" datafld="상호"></span></td> </tr>
  <tr> <td align=right bgcolor=lightgrey>대표</td>
    <td><span datasrc="#BI" datafld="대표자성명"></span></td> </tr>
  <tr> <td align=right bgcolor=lightgrey>사업자등록번호</td>
    <td><span datasrc="#BI" datafld="사업자등록번호"></span></td> </tr>
  <tr> <td align=right bgcolor=lightgrey>주소</td>
    <td><span datasrc="#BI" datafld="주사업장주소"></span></td> </tr>
  <tr> <td align=right bgcolor=lightgrey>전화번호</td>
    <td><span datasrc="#BI" datafld="전화번호"></span></td> </tr>
  <tr> <td align=right bgcolor=lightgrey>팩스번호</td>
    <td><span datasrc="#BI" datafld="팩스번호"></span></td> </tr>
  <tr> <td align=right bgcolor=lightgrey>e-mail</td>
    <td><span datasrc="#BI" datafld="전자우편주소"></span></td> </tr>
</table>
</center>
</body>
</html>
```

<그림 10> HTML의 홈페이지 문서(Home.html)



<그림 11> 사업자신원정보 홈페이지의 예: Hmall.com

현재 운영되고 있는 웹사이트들이 HTML을 이용하여 홈페이지를 작성하고 있기 때문에, XML data island와 데이터 바인딩 부분을 작성하여 홈페이지에 끼워 넣음으로써 구조화된 사업자 신원정보 문서 작성이 가능하다. 즉, 초기 적용과정에서 시스템의 변화가 적고, 사용이 간단하기 때문에 이 구조는 적용하기가 쉽다. 하지만, 공공기관이 자동으로 사업자신원정보를 추출하기 위해서는 다양한 데이터 바인딩의 형태를 반영하여야 하기 때문에 조사효율성과 효과성을 높이기 위해 적절한 노력이 요구된다.

5. 사업자신원정보 표현 구조화 방안의 비교

앞에서 제시한 소비자, 사업자, 공공기관 관점에서의 효과적인 사업자신원정보 표현을 위한 구조화 방안을 서로 비교한다. 현재의 표현체계인 HTML 비구조적 방식과 HTML 반구조적 방식을 구조화 방안으로 제시하는 HTML 기반구조, XML 기반구조, XML data island 기반 구조와 비교하면 <표 1>과 같다. <표 1>은 비교요인들을 구조화, 소비자, 사업자, 공공기관 차원으로 분류

<표 1> 사업자신원정보 표현체계의 상대적 비교

구분		현재 방식		구조화 방안		
		HTML 비구조적 방식	HTML 반구조적 방식	HTML 기반 구조	XML data island 기반 구조	XML 기반 구조
구조화	사업자신원정보의 구조화 형태	비구조적	반구조적	구조적, 반구조적	구조적, 반구조적	구조적, 반구조적
	사업자신원정보 표현수준	-	-	속성층	속성층, 표현구조층	속성층, 표현구조층
소비자	소비자의 이해도	매우 낮음	낮음	높음	높음	높음
사업자	적용 시 웹페이지의 변화	-	-	없음	매우 적음	많음
	사업자신원정보 표현의 일관성	매우 낮음	낮음	높음	높음	높음
	사업자의 작성 용이도	높음	높음	높음	조금 높음	낮음
	사업자의 도입용이성	-	-	매우 높음	매우 높음	조금 낮음
공공기관	조사에이전트의 추출정확도	매우 낮음	낮음	높음	조금 높음	매우 높음
	조사에이전트의 추출시간	매우 김	조금 김	짧음	짧음	짧음
	조사에이전트의 유지보수비	매우 높음	높음	조금 높음	조금 높음	조금 높음
비고		사업자신원정보의 구조화 방안에 따른 표현형태로의 전환이 필요함.		사업자편의에 따라 신원정보 구조와 속성값 표현체계가 변경될 수 있음.	현재 대부분 웹문서들이 HTML 기반이므로 적용이 용이함.	현재 대부분 웹문서들이 HTML 기반이므로 적용시기가 이름.

하여 비교한 내용이다. 구조화 차원에서 사업자신원정보의 표현 수준은 사업자신원정보의 구조화 관점에서 <그림 5>의 속성층 또는 표현구조층으로의 표현수준을 의미한다. XML data island 기반구조의 경우에는 속성층과 표현구조층의 사용이 다 가능하나 표현구조층을 이용한 표현 시 데이터바인딩이 속성층만으로 표현했을 때에 비해 복잡하다. 소비자 차원에서 소비자의 이해도는 구조적 표현체계의 경우가 높다. 사업자 차원에서는 HTML 기반구조와 XML data island 기반구조가 용이한 편이다. 공공기관 차원에서의 조사관점에서는 XML 기반구조, XML data island 기반구조, HTML 기반구조, HTML 반구조적 방식, HTML 비구조적 방식의 순으로 선호되는 편이다.

현재 사용되고 있는 HTML 비구조적 및 반구조적 방식은 구조화 방안에 따른 표현형태로의 전환이 필요하다. HTML 기반구조는 사업자 편의에 따라 신원정보 구조와 속성값 표현체계가 변경될 수 있으며, 현재 대부분의 사업자 웹문서들이 HTML 기반이므로 XML data island 기반구조가 적용이 용이하며, XML 기반구조의 적용은 시기가 이르다고 볼 수 있다.

6. 사업자신원정보 추출 검정

HTML 반구조적 방식으로 기재된 사업자신원정보와 XML data island 기반 구조에 의한 사업자신원정보의 추출에 관하여 효과성과 효율성 관점에서 비교실험하였다. 실험을 위한 표본으로는 야후! 코리아(www.yahoo.co.kr)에 온라인 종합쇼핑몰의 첫 목록에 있는 291개 사이트중에서 정상적으로 운영이 되고 사업자신원정보가 문자로 표현된 181개 사이트에서 임의로 60개의 사이트를 추출하였다. 추출한 표본들은 HTML 반구조적 방식으로 기재된 사업자신원정보를 XML data island 기반 구조로 변환하고 변환 전후의 사업자신원정보 추출실험을 비교하였다. 실험을 위해 사용한 시스템의 사양은 CPU가 인텔 펜티엄 4 프로세서 2.2GHz-M, 주기억장치가 256MB이며, 운영체제는 MS Windows XP Home Edition이다.

실험 1: 효과성(Effectiveness) 분석을 위한 추출정확도의 비교

성낙현(2004)은 동의어 및 지시어를 이용한 사업자신원정보 조사에이전트를 이용하여 현재 대부분의 사업자들이 취하고 있는 구조화되지 않은 사업자신원정보를 추출한 결과 89.3%의 정확도를 보였다. 향후, 이 방법에서는 동의어의 개수를 늘리고 다양하게 표현된 속성값의 추출을 위한 방법을 알고리즘에 반영한다면 더 높은 정확도로 추출이 가능할 것이다. 그러나, 이를 위한 노력과 비용이 많이 들며, 추출알고리즘이 복잡하여 사업자들이 작성하는 사업자신원정보 표현방식의 변화에 대응하기가 쉽지 않다.

<그림 12>는 4.3 절에서 제안한 XML data island 기반 구조에 대한 사업자신원정보 추출알고리즘이다. 먼저, 사업자의 웹문서를 읽어 와서, XML data island 부분과 데이터 바인딩 부분을 추출한 후, 각 사업자신원정보 속성별로 데이터 바인딩되어 있는지 검색하고 그에 대한 속성값을 XML data island로부터 추출한다.

이 알고리즘을 이용하여 표본에 대한 사업자신원정보를 추출한 결과 100%의 정확도로 추출이 가능하다. 그러나, 실제 사업자들은 사업자신원정보를 다양한 데이터 바인딩 형식으로 표현하기 때문에 이를 모두 반영하지 못한다면 추출 정확

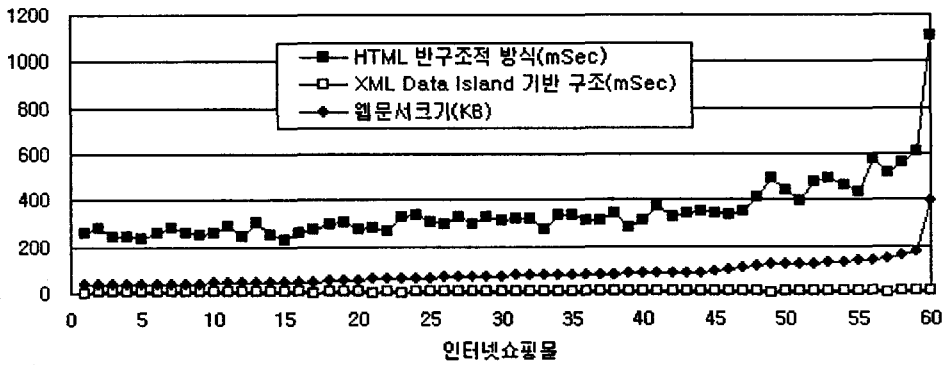
```
1. BI = {"상호", "대표자성명", "사업자등록번호", "주사업장주소", "전화번호", "팩스번호",
        "전자우편주소"}
2. strWebpage = ReadWebPage(URL)                    // 사업자의 URL에 해당하는 웹문서 검색
3. strIsland = ExtractDataIsland(strWebpage)        // 웹문서로부터 Data island 추출
4. strBinding = ExtractDataBinding(strWebpage)      // 웹문서로부터 Data binding 추출
5. FOR each item IN BI
     IF (Search(strBinding, item) != NULL)          // 데이터바인딩된 item 검색
          strItemStart = "<" + item + ">"
          strItemEnd = "</" + item " ">"
          strAttrValue = ExtractAttr(strIsland, strItemStart, strItemEnd)
                                                    // Data island에 표현된 속성값 추출
     END IF
   NEXT
```

<그림 12> XML data island 기반 구조의 사업자신원정보 추출알고리즘의 예

성은 낮아지게 된다. 그러나 이전의 연구(성낙현, 2004)에 비하면 추출 알고리즘이 단순하고 다양한 데이터 바인딩 형식에 대응하기가 쉽기 때문에 상대적으로 추출 정확도를 높이기가 쉬울 것이다.

실험 2: 효율성(Efficiency) 분석을 위한 추출속도의 비교

<그림 13>의 각 추출시간은 신뢰도를 높이기 위해 5회씩 측정한 평균치로 표시하였다. 추출시간은 인터넷을 통한 웹문서의 검색을 제외한 실제 정보추출시간으로 비교하였다(XML data island 기반 구조의 경우, <그림 12>의 3~5 부분에 해당함). 표본들의 웹문서 크기는 36~395KB 사이이며, 웹문서의 크기에 따른 추출시간의 변화를 보기 위하여 웹 문서의 크기 순으로 정렬하였다. XML data island 기반 구조는 웹 문서의 크기에 관계없이 거의 일정한 추출시간을 갖는데 비해, HTML 반구조적 방식은 웹문서의 크기에 영향을 받으며 크기가 클수록 추출시간이 역시 증가하는 현상을 볼 수 있다. 이는 XML data island



<그림 13> 인터넷쇼핑몰의 사업자신원정보 추출시간

기반 구조는 사업자신원정보의 기재 범위가 명확하고 표현이 구조화되어 있어서 추출방법이 단순하고 명확한데 비해서, HTML 반구조적 방식은 사업자신원정보의 시작과 끝을 구분하기 힘들며 추출알고리즘이 복잡하고 다양한 추출방법을 이용해야 하기 때문이다.

실험결과 <표 2>와 같이 HTML 반구조적 방식의 평균추출속도는 μ_H = 348 (msec), 표준편차는 s_H = 134 (msec)이며, XML data island 기반 구조는 μ_D = 6 (msec), s_D = 2 (msec)이다. 귀무가설 $\mu_D \leq \mu_H$ 에 대한 통계추정결과, Z = -19.76 < $-z_{0.01}$ = -2.325 이므로 HTML 반구조적 방식보다 XML data island 기반 구조의 평균추출속도가 낮음을 1% 유의수준에서 검정할 수 있다. 시간차의 단위가 작기는 하나, 계속 증가하는 많은 수의 사이트 조사 또는 HTML 반구조적 방식의 경우 웹문서의 복잡도가 증가하는 추세에서는 의미가 있다. HTML 반구조적 방식의 경우, 표준편차가 큰 것은 웹문서의 크기와 사업자신원정보 표현의 복잡도에 따라 달라지기 때문이다.

한편, 웹문서의 크기에 따른 HTML 반구조적 방식의 추출시간에 대해

HTML 반구조적 방식의 추출시간(추정치)
$= \alpha + \beta \cdot$ 웹문서의 크기

라는 단순회귀모형을 가정한다면, α = 148.1, β = 2.4가 되며, 귀무가설 $\beta \leq 0$를 검정하면 T= 32.26 $\geq t(58, 0.01)$ = 2.39 이므로 1%의 유의수준에서 귀무가설을 기각한다. 즉, 회귀직선은 유의하며, 웹문서의 크기가 커질수록 HTML 반구조적 방식의 경우 추출시간이 증가하게 되어 XML data island 기반 구조에 비해 그 차이가 점점 커질 것이다.

7. 결론

OECD의 "전자상거래 소비자보호를 위한 가이드라인"에서 인터넷쇼핑몰의 사업자신원정보 기재를 권유하고 있다. 다른 나라들이 아직 이런 권고를 따르지 않고 있으며, 그들 나라의 인터넷쇼핑몰은 비구조적인 형태로 사업자신원정보의 일부를 표현하고 있는 실정인데 비하여, 우리나라는 세계에서 가장 앞서 "전자상거래 등에서의 소비자보호에 관한 법률"을 제정하여 시행하고 있으며, 인터넷쇼핑몰 업체들은 거의 대부분 반구조적인 형태로 사업자신원정보를 제공하고 있다.

그러나, 우리나라의 인터넷쇼핑몰 업체들이 제공하는 사업자신원정보의 반구조적인 표현은 현재 기재항목과 표현방법에 있어서 구조화되어 있지 않아 소비자가 사업자의 신원을 정확하게 파악하기 어렵기 때문에 소비자의 신뢰도에 나쁜 영향을 줄 수 있으며, 정보를 자동으로 추출하는 알고리즘이 복잡하고 다양한 사업자신원정보 표현에 대응하기 위하여 많은 노력과 비용이

<표 2> 인터넷쇼핑몰의 사업자신원정보 추출시간에 대한 평균과 표준편차

표본의 수 = 60개 사이트

구 분	HTML 반구조적 방식	XML data island 기반 구조	웹문서 크기
평 균	μ_H = 348 mSec	μ_D = 6 mSec	82 KB
표준편차	s_H = 134 mSec	s_D = 2 mSec	54 KB

들기 때문에 사업자신원정보 조사는 아직 비효율적이다.

이에 본 연구는 소비자, 사업자, 공공기관 관점에서 건전한 전자상거래 발전을 위하여 사업자신원정보 기재를 위한 표현 구조화 방안으로 HTML 기반 구조, XML 기반 구조, XML data island 기반 구조를 제시하고 XML data island 기반 구조의 예에 대하여 추출알고리즘을 제시하여 유용성을 검정하였다. 현재, 대부분의 인터넷상거래 업체들이 HTML 반구조적 방식으로 웹페이지를 작성하고 있기 때문에 <표 1>의 세가지 구조 비교에서 보듯이 XML data island 기반 구조를 이용하는 것이 손 쉬운 접근 방안이 될 것이다. 한편, 구조화 방안이 새로운 규제로 작용하여 전자상거래 발전을 저해해서는 안 될 것이다.

본 연구는 사업자신원정보의 구조화 관점에서 정보의 표현에 초점이 맞추어져 있으나, 향후에 사업자는 사업자신원정보와 상품정보를 포함한 사업자정보를 소비자에게 알릴 필요가 있으며, RDF(Resource Description Framework) 등을 이용한 시맨틱웹(Semantic web) 기반구조의 의미론적인 사업자신원정보 체계에 대한 연구가 필요할 것이다.

Acknowledge

이 논문은 2003년도 한신대학교 교내 연구비 지원에 의하여 연구되었습니다.

참고문헌

성낙현, 에이전트 기반의 인터넷쇼핑몰 사업자신원정보 조사, *Information Systems Review*, (Forthcoming) 2004.

성낙현, 장용식, 에이전트 기술을 이용한 전자쇼핑몰 필수기재사항 조사, 2002 한국경영정보학회 춘계학술대회, 2002. 6.20. pp. 858-864.

통계청, 2003년 6월 및 2-4분기 사이버쇼핑몰통계조사 결과, 2003. 8.

Ashish, N. and Knoblock, C.A., "Semi-automatic Wrapper Generation for Internet Information Sources," *In Proceedings of the International Conference on Cooperative Information Systems (Coopis-97)*, Charleston, South Carolina, (1997).

Atzeni, P. and Mecca, G., "Cut and Paste," *In Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems (PODS-97)*, Tucson, Arizona, (1997), 114~153.

Doorenbos, R.B., Etzioni, O. and Weld, D.S., "A Scalable Comparison-Shopping Agent for the World-Wide Web," *In Proceedings of the First International Conference on Autonomous Agents*, ACM Press, New York, NY, (1997), 39~48.

Hsu, C.-N. and Dung, M.-T., "Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web," *Information Systems*, Vol. 23, No. 8 (1998), 521~538.

Hsu, J.Y.-J. and Yih, W.-T., "Template-based Information Mining from HTML Documents," *In Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, AAAI Press, Menlo Park, CA, (1997), 256~262.

Kushmerick, N., *Wrapper Induction for Information Extraction*, Ph.D. Thesis, Department of Computer Science and Engineering, University of Washington, Seatle, WA, 1997.

Muslea, I., Minton, S. and Knoblock, C.A., "STALKER: Learning Extraction Rules for Semistructured, Web-based Information Sources," *In Proceedings of AAAI-98 Workshop on AI and Information Integration*, Technical Report WS-98-01, AAAI Press, Menlo Park, CA, (1998).

OECD, Recommendation of OECD Council Concerning Guidelines for Consumer Protection in the Context of Electronic Commerce, 1999.

Abstract

An Approach to Structuralizing Business Information for Internet Shopping Malls

Yong Sik Chang*

While on-line shopping is increasing, the "Consumer Protection Law in Electronic Commerce" obliges each internet shopping mall to provide its business information. Although most internet shopping malls provide their business information in the semi-structured format on the bottom of their homepages, the attributes and expression forms of business information are different each other. It makes consumers difficult to identify their business information and lowers public confidence. Hence this study proposes three approaches - HTML-based structure, XML-based structure, and XML data island-based structure - to structuralizing business information for correct expression. The experiment results showed that the business information extraction time by XML data island-based structure is independent of the size of the web document, while the time by HTML-based structure is dependent on the size. By comparing the business information extraction times, we show that XML data island-based structure is more efficient and effective than HTML-based structure.

Key words : Internet shopping mall, Business information

* Department of e-Business, College of Management & Trade, Hanshin University