에이전트 기반의 인터넷쇼핑몰 사업자신원정보 조사

Agent-based Investigation of Business Information from Internet Shopping Malls

성 낙 현 (Nahk Hyun Sung) 용인대학교 컴퓨터정보학부

요 약

국내에서는 2002년 7월부터 '전자상거래 등에서의 소비자보호에 관한 법률'에 의해, 인터넷쇼핑몰 운영자가 자신의 신원정보를 표시하는 것을 의무화하고 있다. 이러한 사업자신원정보의 속성(Attribute)에는 상호, 대표자성명, 사업자등록번호, 사업장주소, 전화번호, 팩스번호, 전자메일주소 등이 있다. 신원정보를 밝히지 않는 사이트를 조사하여 기재를 권유하는 일은 전자상거래의 신뢰도를 높이는 데에 기여할 것이다. 본 연구에서는 인터넷쇼핑몰 사업자신원정보의 기재여부를 조사하기 위해, 조사대상으로서의 인터넷쇼핑몰의 URL을 수집하는 방법, 신원정보 속성을 추출하는 방법, 그리고 신원정보 조사에이전트시스템의 구조를 제시한다. 본 연구에서 정보의 추출은 속성명의 동의어와 속성값의 지시어를 이용한다. 연구의 유용성을 보이기 위해 신원정보 추출의 정확도가 89.3%인 조사에이전트의 실험결과를 제시한다.

키워드 : 전자상거래, 인터넷쇼핑몰, 사업자신원정보, 에이전트, 정보추출

Ⅰ. 서 론

소비자가 인식하는 인터넷쇼핑몰의 신뢰도에 영향을 미치는 요인으로는 쇼핑몰사업자의 인지도, 사업자신원정보, 개인정보보호대책, 지불방법의 보안수준 등이 있다. 이 중에서도 사업자의 신원정보는 소비자가 사업자의 정보를 파악하는 기본적인 정보이다.

OECD에서는 1999년에 전자상거래 소비자보호를 위한 가이드라인을 이사회 권고사항으로 발표하였으며, OECD 소비자보호 가이드라인과 이에 따른 각국의 전자상거래 소비자보호 가이드라인들은 사업자가 자신의 신원에 관한 정보를 소비자에게 온라인으로 제공할 것을 명시하고 있다(강성진, 1999). OECD 가이드라인은 B2C 인터넷쇼핑몰은 최소한의 사업자신원정보(Business Information)를 제공해야 한다고 제안한다. 이러한 사업자신원 정보의 속성 (Attribute)에는 상호, 대표자성명, 사업자등록번호, 사업장주소, 전화번호, 팩스번호, 전자메일주소 등이 있다.

한편, 국내에서는 OECD 권고에 따라 2002년 7월부터 시행되는 '전자상거래 등에서의 소비자 보호에 관한 법률'에 의해, 소비자가 사업자의 신원정보를 쉽게 알 수 있도록, 인터넷쇼핑몰 운영자가 자신의 신원정보를 표시하는 것을 의무화하고 있다. 이에 따라 사업자신원정보는 국내에서 인터넷쇼핑몰의 필수기재사항이 되었다. 따라서 어떤 인터넷쇼핑몰이 고의적으로 사업자신원정보의 전체 또는 일부분을 기재하지 않는다면, 이 경우에는 소비자를 속이는 전자상거래에서의 부당행위를 할 의도를 가진 쇼핑몰로 간주할 수 있

다. 사업자의 신원을 제대로 밝히지 않는 인터넷 쇼핑몰을 조기에 발견하여 사업자신원정보의 기재를 권유하고 이를 소비자에게 알리는 일은 전자상거래의 신뢰도를 높이는 데에 크게 기여할 것이다.

통계청에 의하면 2002년 11월 사이버쇼핑몰조사 결과 전년동월 대비 사업체수는 2,874개로 34.6% 증가, 거래액은 5,526억원으로 71.2% 증가하여 지속적인 증가추세를 보이고 있다(통계청, 2003). 2002년 12월 현재, 국내 인터넷 이용률은 전체인구의 59.4%로 추정되며, 최근 6개월 이내 인터넷 이용자의 31.0%가 인터넷쇼핑을 이용한 것으로 추정되어 인터넷 이용자의 인터넷 활용도가 점차 증대되고 있는 것으로 나타났다(한국인터넷정보센터, 2003).

이러한 전자상거래와 인터넷이용의 증가에 따라 인터넷쇼핑몰의 부당행위에 의한 소비자의 피해도 증가하고 있어 전자상거래에서의 소비자보호가 상당히 큰 문제로 대두되고 있다. 2002년 상반기 중 한국소비자보호원에서 처리한 전자상거래관련 소비자상담실적은 4,631 건으로 전년 상반기에 비해 110.1% 증가하였다. 이는 전체 상담 건 중 2.9%로 작년의 1.2%에 비해 전자상거래 관련 상담비율이 증가하고 있음을 보여주고 있다(한국소비자보호원, 2002).

전자상거래 사이트는 매일 새롭게 나타나고 사라지며, 그 숫자도 매우 많기 때문에, 사람이 인터넷쇼핑몰 홈페이지를 일일이 검색하여 사업자신원정보 기재여부를 조사하는 것은 한계가 있다. 사이버소비자협의회에서는 2000년 8월과 9월의 2개월에 걸쳐 인터넷쇼핑몰 380개(종합쇼핑몰 165개, 전문쇼핑몰 215개)에 대하여 6개의 소비자보호단체에서 33명의 감시요원이 인터넷쇼핑몰의 사업자신원정보 기재여부를 조사한 바 있다(공정거래위원회, 2000). 그러나 위와 같은 조사를 자주 시행하기는 현실적으로 어렵기 때문에 조사의 효율성을 위해서는 소프트웨어 에이전트를 이용하는 것이 바람직하다. 또한 사업의 진출과 퇴장이 빠른 전자상거래의 특성상 에이전트를 이용하여 24시간 감시하는 상시적 감시체계를 수립할 필요가 있다. 인터넷쇼핑몰의 사업자신원 정보조사에이전트시스템은 인터넷쇼핑몰의 URL 수집방법, 홈페이지의 변경 및 폐쇄사이트의 모니터링, 사업자신원정보의 추출 및 기재권유 등의 기능을 고려하여 구축되어야 한다.

위와 같은 문제를 해결하기 위하여 본 논문을 다음과 같이 구성하였다. 제Ⅱ장에서는 관련연구를 살펴보고, 제Ⅲ장에서는 인터넷쇼핑몰의 신원정보 조사체계로서 에이전트를 이용한 조사대상 인터넷쇼핑몰 URL의 수집 및 변경추적, 속성명의 동의어 및 속성값의 지시어 체계를 이용한 사업자신원정보 조사방법을 제시한다. 제 Ⅳ장에서는 사업자신원정보 조사실험을 통하여 조사에이전트 시스템의 유용성을 보이고, 제 Ⅴ장에서는 결론으로 사업자신원정보 조사 에이전트의 활용방안을 제시한다.

Ⅱ. 문헌 연구

OECD 가이드라인은 인터넷쇼핑몰이 사업자신원정보를 제공해야 한다고 제안하지만, 이를 실현하여 법제화하는 방향은 국가마다 차이가 있다. 전자상거래의 선진국인 미국을 비롯한 대부분의 국가에서는 사업자신원정보의 제공이 법제화되어 있지 않다. 따라서 미국의 인터넷쇼핑몰들은 사업자신원정보를 제공하는 것이 필수적이 아니므로, 신원정보를 제공하지 않는 경우도 있고, 신원정보를 제공하더라도 대개 'Contact us' 또는 'About us' 등의 웹페이지에서 신원정보를 자유로운 형식의 비구조적(Un-structured) 문서로서 제공하고 있다. 반면에 국내에서는 법률에 의해, 사업자신원정보의 각 항목이 인터넷쇼핑몰의 필수기재사항이 되었다. 그러나 사업자신원정보의 표현항목은 통일되었지만, 각 항목의 순서 등 표현방법은 통일되어 있지 않는 반구조적(Semi-structured) 특징을 가지고 있다.

구조적(Structured), 비구조적(Un-structured), 반구조적(Semi-structured) 문서의 학문적 정의는 다음과 같다. 구조적 문서란 한 튜플 내의 각 항목이 구분자 또는 항목의 순서 등과 같이 일정한 문법적 어구나 배열의 단서에 근거하여 정확하게 추출될 수 있는 문서를 말한다. 비구조적 문서는 항목을 정확하게 추출하기 위해서는 언어학적인 지식이 요구되는 문서를 지칭한다. 반구조적 문서는 비구조적이 아닌 문서이지만, 항목이 누락되어 있거나, 한 항목이 여러 개의 값을 갖거나 항목 표현의 일관성이 일부 결여된 문서를 지칭한다(Hsu and Dung, 1998). 따라서 앞서 언급한 법제화의 정도에 따른 구조적 문서의 정의와 학문적 정의는 내용상으로 일치한다고 볼 수 있다. 즉, 사업자신원정보를 제공하는 것이 전혀 법제화되어 있지 않고, 이를 표현하는 형식도 제시되어 있지 않다면, 사업자신원정보는 비구조적 문서로서 표현될 것이다. 반면 사업자신원정보를 제공하는 것이 법제화되어 있고, 이를 표현하는 형식도 통일되어 있다면, 이를 구조적 문서라 할 수 있을 것이다. 또한 사업자신원정보를 제공하는 것이 법제화되어 있지만, 이를 표현하는 형식이 통일되어 있지 않다면, 이를 반구조적 문서라고 할 수 있을 것이다. 즉 우리나라의 경우처럼 웹페이지에 사업자신원정보를 제공하는 것이 법제화는 되어 있지만, 표현형식이 대통령령 등으로 정해지지 않은 경우에 이를 반구조적 문서라고 할 수 있을 것이다.

반구조적 웹페이지로부터 정보를 추출하기 위해서는 Wrapper의 이용이 필수적이다 Wrapper는 주어진 웹페이지로부터 정보를 추출하여 구조적 데이터 튜플로 결과를 보여주는 소프트웨어를 말한다(Hsu and Dung, 1998). Wrapper는 초기에 각 사이트에 맞게 수작업으로 프로그램이 작성되었으나, 이러한 방법은 다양한 웹문서의 변화에 대응하기가 힘이 든다. 이를 극복하고자 Wrapper의 빠른 구축 방법에 대한 연구가 진행되어 왔다. 첫째, Wrapper 구축용 프로그램을 이용하는 연구(Atzeni and Mecca, 1997; Hammer et al., 1997; Smith and Lopez, 1997)가 있으나, 이는 전문가가 아닌 경우 사용이 어려우며 수작업으로 대응하기 힘든 단점이 있다. 다음으로는 이러한 문제점을 해결하기 위하여 다음과 같은 Wrapper 자동구축에 관한 여러 연구들이 있다. 1) 상품정보 속성과 동의어 기반의 휴리스틱검색, 패턴매칭, 상품정보 표현형태의 귀납적 학습을 이용하여 웹페이지로부터 상품정보를 추출하는 비교쇼핑 에이전트(ShopBots)에 관한 연구(Doorenbos, 1997), 2) 추출속성 주변에 HTML 태그를 넣어 정보를 추출하는 연구(Ashish and Knoblock, 1997), 3) HTML 문서의 FAQ 정보추출에 관한 템플릿기반 접근(Hsu and Yih, 1997), 4) HTML 태그를 구분자로 이용하는 연구(softbots)(Kushmeick, 1997), 5) 속성 간에 분리자(separator)를 이용한 연구(Hsu and Dung, 1998), 6) 추출규칙에 기반을 둔 Wrapper induction 알고리즘에 관한 연구(Stalker)(Muslea et al., 1998) 등이 있다.

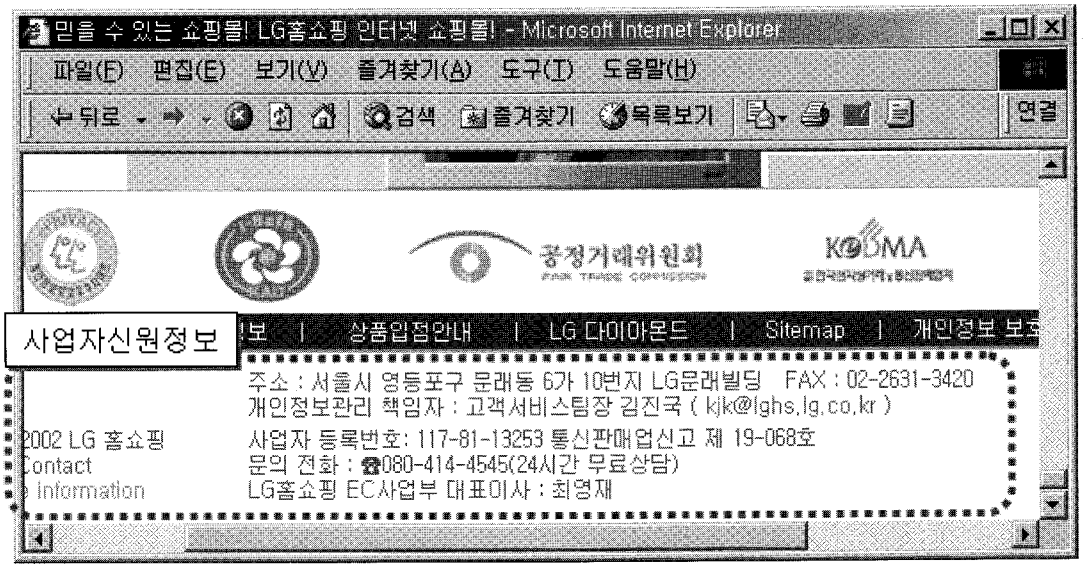
인터넷쇼핑몰의 사업자신원정보는 다수의 속성들로 구성되어 있다. 본 연구는 기존의 연구를 확장하여 반구조적 형식으로 기재된 사업자신원 정보에 대하여 속성명의 동의어와 속성값의 지시어를 이용한 표현구조에 기반을 둔 사업자 신원정보 추출방법을 제안하고 있다.

향후 시멘틱웹(Semantic Web)에 대한 연구가 활발히 이루어져서 현재의 웹환경이 바뀌어 웹페이지의 정보를 소프트웨어 에이전트가 읽을 수 있도록 표현하는 시대가 온다면, 새로운 사업자 신원정보 추출방법이 필요할 것이다.

Ⅲ. 사업자신원정보

3.1 사업자신원정보 표현구조

<그림 1>에서 (a)는 LG홈쇼핑 인터넷쇼핑몰의 홈



(a) LG홈쇼핑 인터넷쇼핑몰의 홈페이지 (굵은 점선 내부 영역은 사업자신원정보를 나타냄)

```
<table width="750" border="0" cellspacing="0" cellpadding="0">
<tr>
<td align="left" valign="top">
<img src="http://menu.lgeshop.com/top_2002/top_bottom/MainBottom-bi.gif" width="111" height="35">
</td>
<td>주소 : 서울시 영등포구 문래동 6가 10번지 LG문래빌딩   FAX : 02-2631- 3420 <br>  개인정보관리
책임자 : 고객서비스팀장 김진국 (
<a href="mailto:kjk@lghs.lg.co.kr;kwlee@lghs.lg.co.kr">kjk@lghs.lg.co.kr</a> )
</td>
</tr>
<tr>
<td>Copyright 1998 - 2002 LG 홈쇼핑<br>All rights reserved Contact</font><br>
    <a href="/jsp/jseig_customerMail.jsp"><font color="#999999">webmaster</font></a>
    <font color="#999999"> for more information</font>
</td>
<td valign="bottom">사업자 등록번호: 117-81-13253 통신판매업신고 제 19-068호<br>
문의 전화 : 080-414-4545(24시간 무료상담)<br>  LG홈쇼핑 EC사업부 대표이사 : 최영재
</td>
</tr>
</table>
```

(b) 사업자신원정보를 표현하는 HTML 문서의 일부

```
사업자신원정보(lgeshop.com)
  {(상호, 'LG홈쇼핑 EC사업부')
   (대표자성명, '최영재')
   (사업자등록번호, '117-81-13253')
   (사업장주소, '서울시 영등포구 문래동 6가 10번지 LG문래빌딩')
   (전화번호, '080-414-4545')
   (팩스번호, '02-2631- 3420')
   (전자메일주소, 'kjk@lghs.lg.co.kr')}
```

(c) 사업자신원정보

〈그림 1〉 LG홈쇼핑 인터넷쇼핑몰(www.lgeshop.com)의 사업자신원정보

페이지를 나타낸 것이고, (b)는 해당 인터넷 쇼핑몰의 HTML 문서, 그리고 (c)는 해당 인터넷 쇼핑몰의 사업자신원정보를 나타낸 것이다. 인터넷 쇼핑몰의 사업자신원정보는 사업자신원정보의 각 속성을 나타내는 단어들과 각 속성의 값들로 구성되어 있다. LG홈쇼핑 인터넷쇼핑몰의 경우 대표자성명이라는 사업자신원정보 속성은 '대표이사'라는 속성명과, '최영재'라는 속성값으로 표현되어 있다. 상호의 경우는 속성명이 없고 'LG홈쇼핑 EC사업부'라는 속성값 만으로 표현되어 있다. 사업장주소의 경우 속성명은 '주소'라는 동의어를 사용하고 있고, 속성값은 '서울시 영등포구 문래동 6가 10번지 LG문래빌딩'라는 값으로 표현되어 있다. 이 경우에는 '○○도 ○○시 ○○동'과 같은 속성값의 표현구조 자체가 사업장주소를 의미한다. 일반적으로 사업자신원 정보는 다음과 같은 특징이 있다.

- 사업자신원정보는 대부분 반구조적 형식으로 표현되어 있다.
- 사업자신원정보 속성들 간에는 기재순서가 없다.
- 일부 사업자신원정보 속성들은 누락되기도 한다.
- 사업자신원정보 속성명은 다양한 표현들, 즉 동의어들로 표현된다 (전화번호의 예: '전화', 'Tel' 등).
- 일부 사업자신원정보 속성값은 문자 자체가 속성값의 일부임을 의미하는 지시어와, 지시어를 포함한 표현구조를 갖는다(전화번호의 예: 지역(숫자)-국(숫자)-번호 (숫자)).
- 어떤 사업자신원정보 속성값은 두 개 이상의 값으로 표현되기도 한다(전화번호의 예: 02-555-1234, 1235).

3.2 사업자신원정보 추출을 위한 지식

사업자신원정보 속성들은 속성명과 속성값으로 구성되어 있다. 각 속성명은 다수의 동의어로 표현되기도 하며, 어떤 속성값은 지시어를 이용한 표현구조를 가진 특정한 표현형식이 있다. 동의어는 사업자신원정보 속성명과 동일한 뜻으로 사용되는 단어를 의미하며, 상호의 경우는 '상호명', '회사명' 등이, 대표자성명은 '대표', '대표이사', '대표자' 등이 동의어로 사용된다. 한편, 지시어는 속성값을 표현하기 위해 사용되는 문자로서, 상호의 경우 '주식회사', '㈜' 등의 지시어가 있으며, '○○주식회사', '○○㈜'와 같은 형식으로 표현된다. 전화번호는 '-'이 지시어로 사용되며, '지역-국-번호' 의 표현구조를 가지고 있다.

<표 1>에 사업자신원정보의 각 속성명의 동의어와 속성값의 지시어와 그 표현구조를 기술하였다. 상호의 경우는 속성명의 동의어가 매우 다양하며 또한 이러한 다양한 동의어를 생략하는 경우가 많다. 상호 속성값의 지시어와 표현구조 역시 매우 다양하며 <표 1>에 정리한 것은 가장 빈도가 높은 것들 만을 열거하였다. 상호의 경우에는 주로 이 지시어와 표현구조를 이용하여 속성값을 추출한다. 대표자성명의 경우에도 속성명의 동의어가 매우 다양하다. 대표자성명 속성값의 지시어와 표현구조는 대개 세 음절로 이루어진다는 것 이외에 특정한 규칙을 발견하기가 어려웠다. 따라서 대표자성명의 경우에는 주로 동의어를 이용하여 추출한다. 상호와 대표자성명을 제외한 다른 속성들의 경우에는 속성명의 동의어를 대개 사용하고 있고, 속성값을 나타내기 위해서는 지시어를 이용하고 있다. 이 경우 속성값이 지니는 특정 표현구조에 따라 표현하는 것이 유효한 표현방법이며, 또한 소비자가 알아보기 쉽다.

〈표 1〉 사업자신원정보 추출을 위한 지식

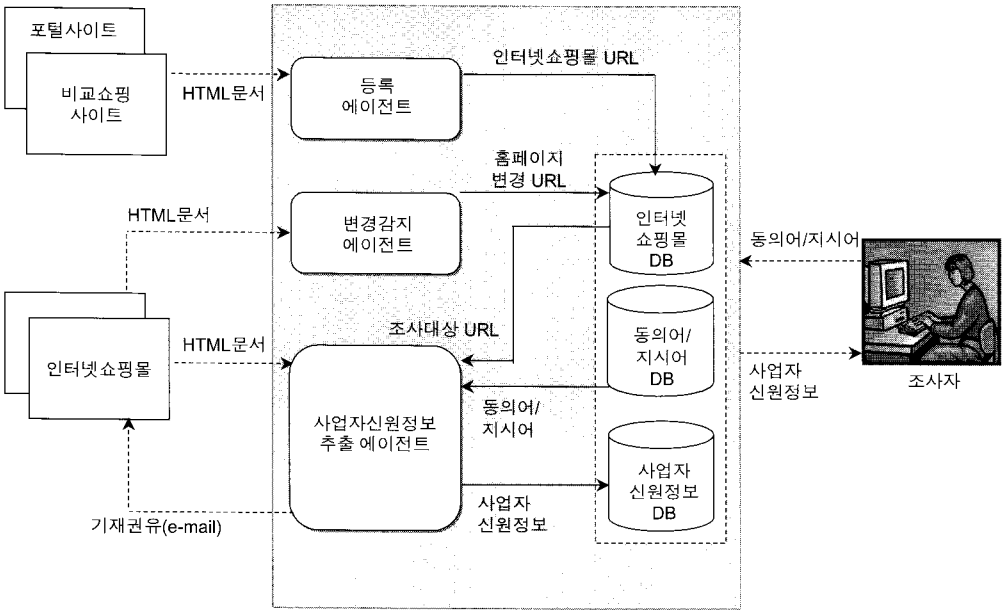
사업자신원정보 속성	사업자신원정보 속성명의 동의어	사업자신원정보 속성값	
		지 시 어	표 현 구 조
상 호	상호, 회사명, 상호명, 상점명	㈜, 주식회사, Co. Ltd, 쇼핑몰, 인터넷쇼핑몰	○○㈜, 주식회사○○, ○○쇼핑몰, ○○인터넷쇼핑몰
대표자 성명	대표자, 대표, 대표이사, 사장, CEO, 대표이사, 대표자		
사업자등록번호	사업자등록번호, 사업자등록번호 안내, 사업자 등록 번호, 사업자번호, 사업자등록, 사업자 등록, 사업자등록번호 안내, 사업자등록번호, 사업자등록증, 사업자 번호, 사업자등록번호안내	–	지역번호(3자리 숫자) – 업종(2자리 숫자) – 일련번호(5자리 숫자)
사업장주소	주소, 사업장주소, 사업장, 사업장소재지	특별시, 광역시, 도, 시, 구, 군, 동, 읍, 면, 리, 동, 가	○○특별시|광역시 ○○구 ○○동|가, ○○광역시 ○○군 ○○읍|면 ○○리, ○○도 ○○시 ○○구 ○○읍|면 ○○리, ○○도 ○○시 ○○구 ○○동, ○○도 ○○시|군 ○○읍|면 ○○리, ○○도 ○○시 ○○동
전화번호	전화번호, Tel, 대표전화, 문의 전화, 문의전화, 고객문의, 고객문의, 전 화, 고객센터, 전화/팩스, 고객상담, 고객지원센터, □, 전화번호, 문의, 연락처, 고객센타, T., T E L	–	지역(숫자)-국(숫자)-번호(숫자)
팩스번호	팩스, Fax, 팩스번호, 팩 스, 전화/팩스, F., FAX	–	지역(숫자)-국(숫자)-번호(숫자)
전자우편주소	전자우편, E-mail, Contact Us, 이메일, 운영자메일, Email, Contact, Mail	@	사용자ID@전자메일서버명

주) "|"는 선택적 항목을 의미함.

Ⅳ. 사업자신원정보 조사에이전트

4.1 사업자신원정보 조사에이전트

주기적으로 그리고 자동으로 인터넷쇼핑몰의 사업자신원정보를 조사하기 위해서는 <그림 2>와 같이 인터넷쇼핑몰의 URL 등록 및 홈페이지 변경감지 에이전트, 사업자신원정보 조사 및 기재권유를 담당하는 에이전트들로 구성된 사업자신원정보 조사에이전트시스템이 필요하다.



〈그림 2〉 전화번호 표현을 위한 HTML 문서의 작성 예

이하에서는 위의 에이전트에 대해 각각 설명한다.

4.2 인터넷쇼핑몰 등록 및 변경감지 에이전트

이 에이전트는 조사 대상이 되는 인터넷쇼핑몰의 URL을 수집하여 데이터베이스에 관리하는 에이전트이다. 현재 '전자상거래 등에서의 소비자보호에 관한 법률'에 의하면 인터넷쇼핑몰은 지방자치단체 또는 한국전자상거래 및 통신판매협회(www.kedma.or.kr)에 신고를 하게 되어 있다. 그러나, 신고를 하지 않고 영업을 하고 있는 인터넷쇼핑몰의 숫자가 신고를 하고 영업을 하는 쇼핑몰의 숫자보다 훨씬 많다. 이렇게 공식적인 통신판매업자 등록을 하지 않는 사이트들도, 광고를 위해 포탈사이트에는 등록을 하는 경향이 있다. 이러한 사이트들은 공식적인 판매업자로 등록한 사이트보다 부당행위를 할 가능성이 더 많다. 인터넷쇼핑몰들은 광고를 위하여 포털사이트에 등록하거나, 비교쇼핑사이트에 등록하여 가격경쟁을 통한 광고를 한다. 따라서 방문자가 많은 포털사이트와 비교쇼핑사이트를 에이전트가 방문하여 신규 인터넷쇼핑몰에 관한 정보를 수집하는 것이 조사대상 인터넷쇼핑몰의 URL을 수집하는 효과적인 방법이 될 수 있다. 본 연구에서는 이 방법을 이용하여 인터넷쇼핑몰의 URL을 데이터베이스화 하였다.

포탈사이트 및 비교쇼핑사이트로부터 URL을 수집하는 과정에서의 주요 고려사항은 다음과 같다. 첫째, 포탈사이트 및 비교쇼핑사이트에 등록된 사이트가 모두 조사대상이 되는 인터넷쇼핑몰은 아니다. 일반적으로 ".co.kr" 또는 ".com"으로 끝나는 도메인네임을 가진 사이트를 인터넷쇼핑몰로 간주할 수 있다. 둘째, 인터넷쇼핑몰의 정보를 수집하는 과정에서 포탈 및 비교쇼핑사이트들은 사이트마다 나름의 디렉토리 분류체계를 가지고 있다. 따라서 조사대상이 되는 인터넷쇼핑몰 사이트들은 통합된 분류체계에 따라 저장해야 할 필요가 있다. 이 경우 통합된 분류체계는 수집대상이 되는 사이트 모두의 분류체계를 수용할 수 있어야 한다. 셋째, 이렇게 등록된 인터넷쇼핑몰에 대한 에이전트를 이용한 방문조사를 주기적

으로 모두 시행한다는 것은 많은 시간이 소요된다. 인터넷쇼핑몰은 개폐가 매우 빠르게 일어나고 홈페이지 변경이 자주 일어나기 때문에 인터넷쇼핑몰의 활동성과 홈페이지 변경 여부를 주기적으로 조사하여 현실에 맞게 유지 관리함으로써, 이전에 조사한 사이트의 홈페이지의 변경 여부(Douglis and Ball, 1996)를 감지하여 선별적으로 재조사함으로써 조사의 효율성을 높일 수 있다.

4.3 사업자신원정보 추출

다음으로는 조사 대상 인터넷쇼핑몰 사이트를 방문하여 홈페이지로부터 사업자신원정보를 추출하여, 사업자의 신원공개수준을 판별하는 방법이 필요하다. 다양한 동의어로 표현된 사업자신원정보 속성명을 확인하고 그 다음에 나오는 속성값이 정확히 표현되어 있는지 판별이 가능해야 한다. 이를 위해서는 <표 1>에 정리되어 있는 속성명의 동의어와 속성값의 지시어와 표현구조에 관한 사업자신원정보 추출지식을 이용한다.

사업자신원정보는 웹페이지에서 다양한 웹프로그래밍 언어에 의한 표현방법으로 구현 가능하다. 이러한 경우에는 다양한 문서형태를 반영하여 사업자신원정보 속성명과 속성값을 판별해야 한다. 예를 들어 전화번호를 표시하는 HTML 문서의 작성 방법은 <그림 3>과 같이 여러 가지가 있을 수 있다. 이러한 경우에는 사업자신원정보 속성명의 동의어와 속성값 사이에 존재하는 " ", ":", "< ..... >" 등을 제외하고 해석해야 한다.

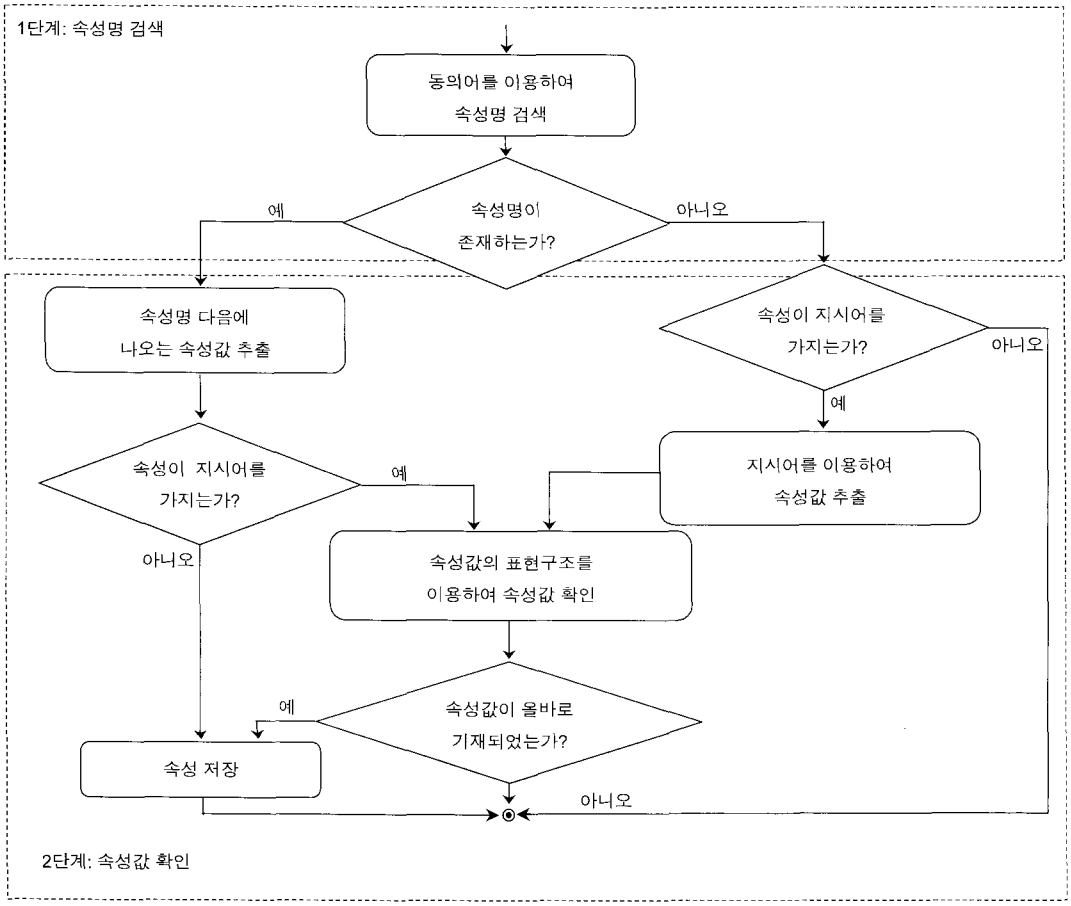
예1) 전화번호: 02-555-7777
예2) 전화번호 02-555-7777
예3) 전화번호 02-555-7777

〈그림 3〉 전화번호 표현을 위한 HTML 문서의 작성 예

사업자신원정보를 추출하는 기본 알고리즘은 <그림 4>와 같다. 먼저 사업자신원정보의 속성명의 동의어를 이용하여 속성값을 추출한 후, 속성값의 지시어를 포함하는 표현구조에 적합한지를 확인하는 절차를 거친다. 속성명이 존재하면 속성명의 뒤에 있는 속성값을 추출하되 상호 또는 대표자성명이 아닌 속성은 속성값이 지시어를 포함하는 표현구조에 맞는지를 확인하여 표현의 적합성 여부를 판별한다. 속성이 지시어를 가지는 경우에는 속성값이 지시어를 포함하는 표현구조에 적합하다고 판정하면 그 속성값을 저장한다. 그러나, 대표자성명은 속성이 지시어를 가지지 않으므로 속성값을 확인할 수 있는 방법이 없기 때문에 바로 저장한다. 상호의 경우에는 속성명의 동의어가 생략되는 경우가 많다. 상호의 지시어와 표현구조는 매우 다양하기 때문에, 지시어를 포함하고 있으면 이를 상호로 판별하여 바로 저장한다.

속성명이나 그 동의어가 없고, 속성이 지시어를 가지는 경우에는 지시어를 포함하는 표현구조에 맞는 속성값을 웹문서에서 발견하여, 그 값이 표현구조에 적합하다고 판정하면, 그 속성값을 저장한다. 그러나 이 경우에는 사업자신원정보에 대한 정확성을 보장할 수 없다. 왜냐하면 사업자신원정보가 속성값이 아닌 어휘를 지시어를 포함한다는 이유로 속성값이라고 판단할 수 있기 때문이다. 속성명이나 그 동의어가 없고, 지시어를 가지지 않는 경우에는 속성값이 없는 것으로 판정하고 종료한다.

조사 후에는 인터넷쇼핑몰의 사업자신원정보 공개수준에 따라 기재를 권유하는 등의 후속조치를 취한다. 사업자신원정보 조사에이전트시스템은 사업자신원정보의 공개수준에 따라 인터넷쇼핑몰 사업자에게 사업자신원정보의 공개를 요구하는 시정메일 발송 등의 사후조치를 할 수 있어야 한다. 이 경우에 포털사이트 또는 비교쇼핑사이트로 메일을 발송하는 것이 아니라, 쇼핑몰로 직접 메일을 발송하여야 한다.



〈그림 4〉 사업자신원정보 추출 알고리즘

V. 사업자신원정보 조사 실험

앞에서 제시한 사업자신원정보 조사에이전트 시스템의 유용성을 검증하기 위하여, Pentium Ⅲ PC, 운영체제로서는 Windows-2000 Advanced Server, DBMS로서는 MS-SQL Server, 그리고 JAVA 언어를 이용하여 사업자신원정보 조사에이전트시스템을 개발하였다.

조사대상 인터넷쇼핑몰을 등록하기 위하여 우리는 포털사이트인 야후(www.yahoo.co.kr), 라이코스(www.lycos.co.kr), 다음(www.daum.net), 엠파스(www.empas.com)와 가격비교사이트인 오미(www.omi.co.kr), 에누리(www.enuri.com), 야비스(www.yavis.com), 베스트바이어(www.bestbuyer.co.kr) 등에서 인터넷쇼핑몰의 URL을 추출하였다. 이들 사이트로부터 인터넷쇼핑몰 등록에이전트가 등록한 전자상거래 사이트의 수는 2002년 3월 25일 현재 총 14,610 사이트이었다.

이 중에는 종합쇼핑몰이 1,140개, 가전/컴퓨터 부문이 3,348개, 생활/문화부문이 8,052개, 레저/오락/기타 부문이 1,299개, 서비스 부문이 771개였다.

통계청이 2002년 11월 발표한 인터넷쇼핑몰의 숫자는 2,874개이다. 본 연구의 인터넷쇼핑몰 등록에이전트가 등록한 숫자와는 큰 차이를 보이고 있다. 이러한 큰 차이는 4.2절에서 언급한대로 많은 인터넷쇼핑몰들이 등록을 하지 않고 영업을 하고 있음을 나타내고 있다.

에이전트의 사업자신원정보 추출의 정확성을 측정하기 위하여, 야후(www.yahoo.co.kr)에 등록된 213개

인터넷쇼핑몰(종합쇼핑몰 66개, 전문쇼핑몰 147개)의 홈페이지를 실험표본으로 설정하였다. 이 과정에서 에이전트시스템이 그래픽 정보를 조사하지 못하는 한계 때문에, 사업자신원정보가 문자로 표현된 웹사이트 만을 실험표본으로 하였다.

실험표본으로부터 사업자신원정보 속성명의 표현에 많이 사용되고 있는 동의어를 조사하여 실제 검색 시 표현순위가 높은 동의어를 우선하여 조사함으로써 조사효율성을 높이도록 하였다. 전화번호 속성의 경우 'Tel', '대표전화', '전화 번호', '전화주문', '고객상담', '문의', '문의전화', '연락처' 등의 동의어 들이 순서대로 많이 나타나고 있다.

실제로 사람이 웹사이트를 일일이 방문하여 사업자신원정보를 발견하여 기재한 결과와 사업자신원정보 추출에이전트를 이용하여 추출한 사업자신원정보를 비교한 조사결과는 <표 2>에 나타나 있다. 실제 속성값이 기재된 빈도수와 에이전트시스템이 조사한 빈도수를 비교하여 조사정확도를 나타내고 있다. <표 2>에 사용된 표기에 대한 설명은 다음과 같다.

- $f(Y)$: 실제 웹페이지에 사업자신원정보 속성이 기재된 빈도수
- $f(N)$: 실제 웹페이지에 사업자신원정보 속성이 기재되지 않은 빈도수
- $f(Y|Y)$: 실제 웹페이지에 기재된 속성을 정확하게 추출한 빈도수
- $f(E|Y)$: 실제 웹페이지에 기재된 속성값을 틀리게 추출한 빈도수
- $f(N|Y)$: 실제 웹페이지에 기재된 속성값을 기재하지 않은 것으로 조사한 빈도수
- $f(N|N)$: 실제 웹페이지에 기재되지 않은 속성값을 기재되지 않았다고 정확하게 추출한 빈도수
- $f(E|N)$: 실제 웹페이지에 기재되지 않은 속성을 기재된 것으로 부정확하게 추출한 빈도수

조사정확도(%): $[\{f(Y|Y)+f(N|N)\}/\{f(Y)+f(N)\}]\times 100$.

〈표 2〉 에이전트에 의한 사업자신원정보 추출의 정확도 (2003년 1월 15일 현재)

(조사 사이트 수 = 213사이트)

사업자신원 정보 속성	실제 기재수	에이전트에 의한 추출 빈도수			조사 정확도
	$f(Y)$(비율)	$f(Y\|Y)$	$f(E\|Y)$	$f(N\|Y)$	
	$f(N)$(비율)	$f(N\|N)$	$f(E\|N)$		
상 호	176(83%)	143	12	21	80.8%
	37(17%)	29	8		
대표자성명	129(61%)	122	1	6	93.0%
	84(39%)	76	8		
사업자등록 번 호	163(77%)	156		7	96.2%
	50(23%)	49	1		
사 업 장 주 소	186(87%)	157		29	85.9%
	27(13%)	26	1		
전화번호	188(88%)	145	1	42	79.8%
	25(12%)	25			
팩스번호	139(65%)	124		15	93.0%
	74(35%)	74			
전자우편 주 소	148(69%)	141	1	6	96.2%
	65(31%)	64	1		
평균 조사정확도					89.3%

실험표본에 대한 조사결과 사업자신원정보의 평균 조사정확도는 89.3%로서 만족할 만한 수준을 보이고 있다. 향후 사업자신원정보 속성명의 동의어를 모두 발견하여 이용하고, 지시어를 이용한 표현구조를 더욱 정확하게 표현하는 등, 사업자신원정보 추출을 위한 지식이 더욱 정교해지면, 정확도는 더욱 높아질 것이다.

Ⅵ. 결 론

6.1 연구의 의의

인터넷쇼핑몰 홈페이지에 사업자신원정보를 기재하는 것은 소비자보호를 위해 전자상거래의 신뢰도를

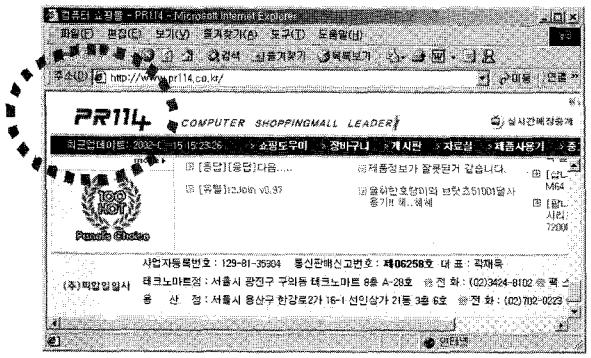
제고하기 위한 첫걸음이 될 것이다. 인터넷쇼핑몰들이 사업자신원정보 공개권고를 잘 지키고 있는지를 조사할 때 에이전트의 활용은 매우 효율성이 있다. 동의어 및 지시어가 조사결과에 많은 영향을 미치기 때문에 이들 동의어와 지시어에 기반을 두고 지식을 관리하는 것이 중요하다. 에이전트에 의한 사업자신원정보 추출의 정확도 실험을 통하여 이를 확인할 수 있었다.

본 연구는 소비자보호와 건전한 전자상거래 발전을 위해 정부기관 및 민간 소비자보호단체에서 다음과 같은 영역에 유용하게 이용될 수 있다. 첫째, 전자상거래업체 중 미신고업체를 검색할 수 있다. 둘째, 신고사항과 웹의 기재사항을 비교하여 부당행위 가능성이 높은 사이트를 파악할 수 있다.

6.2 연구의 한계와 과제

현재 에이전트를 이용한 인터넷쇼핑몰의 신원정보 조사는 짧은 시간에 조사가 가능하여 조사 효율성은 좋은 편이나, 판단 능력에 있어서는 아직 한계가 있다. 그 이유는 홈페이지의 사업자신원정보 기재방법이 정형화되어 있지 않고, 다양한 형태로 기재되는 속성값을 정확하게 판단하기 어렵기 때문이다. 구체적 한계는 다음과 같다.

- 이미지로 표현된 정보: <그림 5>는 이미지 형태로 상호명을 기재한 예로서, 에이전트가 정보를 추출하기 어려운 경우를 보여준다.



〈그림 5〉 사업자신원정보가 이미지 형태로 표현된 인터넷쇼핑몰의 예

범례) 굵은 점선의 원은 이미지 형태로 표현된 상호를 나타냄

텍스트로 표현된 사업자신원정보의 비율을 알아보기 위해 2003년12월 현재 야후코리아의 종합쇼핑몰 리스트의 첫 웹페이지에 등록된 291개 사이트를 대상으로 사업자신원정보의 기재방식을 조사한 결과 텍스트로 표시한 사이트가 181개 사이트였고, 이미지로 사업자신원정보를 표시한 사이트가 37개 사이트였다. 그리고 서비스가 중지되거나 사업자신원정보가 없는 경우가 73개 사이트였다. 291개 사이트 중 이미지로 사업자신원정보를 표현한 사이트의 비율이 12.7% 이었다. 향후 사업자신원정보를 텍스트로 표현하도록 정하지 않는다면, 이미지로 표현한 사업자신원정보를 파싱하여 추출하는 방법론에 대한 연구도 필요할 것이다.

- 상호와 대표자성명과 같이 사업자신원정보 속성명이나 속성명의 동의어가 기재되어 있지 않고, 속성값이 지시어를 갖지 않는 경우: 속성명이 기재되어 있지 않으면서 속성값의 지시어가 없는 경우는 속성에 대한 판단이 어렵다.
- 사업자신원정보 속성값의 애매한 구분: 사업자신원정보 속성값의 시작과 끝이 애매하거나, 또는 속성값 내에 HTML 태그 등이 섞여 있을 때 속성값의 정확한 추출이 어렵다.

현재의 웹 환경에서 앞서 언급한 문제들을 해결하기 위한 두 가지 관점은 다음과 같다.

첫째, 기술의 발전에 따라 인터넷쇼핑몰 홈페이지의 사업자신원정보 표현 방법이 갈수록 다양해지고 복잡해지기 때문에 일일이 에이전트시스템을 수정하여 그에 대응하기에는 쉽지 않다. 실용적인 시스템으로서 유용성을 가지기 위해서는 보다 지능적인 Wrapper에 대한 연구가 필요하다.

둘째, 웹 환경의 표준화 문제이다. 소비자를 보호하고 에이전트의 조사 효과를 높이기 위해, 향후 사업자신원정보 기재를 표준화하도록 권고할 필요가 있다. 현재의 기술수준에서는 다음과 같은 사항이 표준화 권고사항이 될 수 있다. 1) 모든 사업자신원정보는 초

기화면인 홈페이지에 기재하도록 한다. 2)모든 사업자 신원정보는 홈페이지 내 사업자신원정보 영역에 기재하도록 한다. 3) 모든 사업자신원정보는 문자 형태로 기재한다.

참 고 문 헌

강성진, "OECD 전자상거래 소비자보호 가이드라인의 내용과 과제", 한국소비자보호원, 1999. 12.

공정거래위원회, "인터넷쇼핑몰 현황조사 및 부당한 광고행위에 대한 직권조사 실시결과", 2000. 10.

통계청, "2002년 11월 사이버쇼핑몰 통계조사 결과", 2003. 1.

한국소비자보호원, "전자상거래 소비자피해 백서", 2002. 12.

한국인터넷정보센터, "인터넷 이용자수 및 이용행태 조사", 2003. 1.

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, pp. 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, pp.39-48.

Douglis, F. and Ball, T., "Tracking and Viewing Changes on the Web," *In Proceedings of the USENIX Technical Conference*, San Diego, Jan. 1996, pp.165-176.

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

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, pp.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, pp.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.

Smith, D. and Lopez, M., "Information Extracting for Semistructured Document," *In Proceedings of the Workshop on Management of Semi-Structured Data*, Tucson, AZ, 1997.

Information System Review
Volume 6 Number 1
June 2004

Agent-based Investigation of Business Information from Internet Shopping Malls

Nahk Hyun Sung*

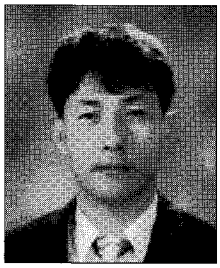
Abstract

In Korea, The Consumer Protection Law in Electronic Commerce, which came into effect in July 2002, forced internet shopping malls to provide a minimum of 7 forms of business information, including the name of the business, the name of the representative, geographical address, telephone number, fax number, e-mail address, and business license number, so that consumers could easily identify them. To investigate the shopping malls which do not provide their business information, can lead to confidence in the electronic commerce environment. To investigate the completeness of the business information with the internet shopping malls, this paper proposes the methods of gathering URLs of internet shopping malls, of extracting business information attributes, and an architecture of the agent system. Information extraction in our research is based on synonyms and indicator words of the attributes. With the experiment we showed that the accuracy of our agent system to find out the right business information is 89.3%.

Keywords: Electronic Commerce, Internet Shopping Mall, Business Information, Agent, Information Extraction

* Department of Computer & Information, Yongin University

◐ 저 자 소 개 ◑



성 낙 현 (nhsung@yongin.ac.kr)

1981년에 서울대학교 경영학과를 졸업하고, 1990년에 한국과학기술원에서 경영정보시스템으로 석사학위를 취득하고, 2000년에 동 대학원에서 박사학위를 취득하였다. 1983년부터 1996년까지 ㈜한화 및 ㈜한화유통의 정보시스템팀에서 근무하였으며, 1997년부터 용인대학교 컴퓨터정보학부 교수로 재직 중이다. 주요 관심분야는 전자상거래에서의 지능형시스템의 응용 및 표준화, 유통 및 물류정보시스템 등이다.