한국지능시스템학회 논문지 2008, Vol. 18, No. 1, pp. 86-93

온톨로지 인스턴스 구축을 위한 주제 중심 웹문서 수집에 관한 연구

A Study on Focused Crawling of Web Document for Building of Ontology Instances

장 문 수*
Moon-Soo Chang

* 서경대학교 소프트웨어학과

요 약

복잡한 의미관계를 정의하는 온톨로지를 구축하는 일은 매우 정밀하고 전문적인 작업이다. 잘 구축된 온톨로지를 응용 시스템에 활용하기 위해서는 온톨로지 클래스에 대한 많은 인스턴스 정보를 구축해야 한다. 본 논문은 온톨로지 인스턴스 정보 추출을 위하여 방대한 양의 웹 문서로부터 주어진 주제에 적합한 문서만을 추출하는 주제 중심 웹 문서 수집 알고리즘을 제안하고, 이 알고리즘을 바탕으로 문서 수집 시스템을 개발한다. 제안하는 문서 수집 알고리즘은 URL의 패턴을 이용하여 주제에 적합한 링크만을 추출함으로써 빠른 속도의 문서 수집을 가능하게 한다. 또한 링크 블록 텍스트에 대한 퍼지 집합으로 표현된 주제 적합도는 문서의 주제 관련성을 지능적으로 판단하여 주제 중심 문서 수집의 정확도를 향상시킨다.

키워드 : 주제 중심 문서 수집, 온톨로지, 웹 문서, URL 패턴, 링크 추출, 주제 적합도

Abstract

The construction of ontology defines as complicated semantic relations needs precise and expert skills. For the well defined ontology in real applications, plenty of information of instances for ontology classes is very critical. In this study, crawling algorithm which extracts the fittest topic from the Web overflowing over by a great number of documents has been focused and developed. Proposed crawling algorithm made a progress to gather documents at high speed by extracting topic-specific Link using URL patterns. And topic fitness of Link block text has been represented by fuzzy sets which will improve a precision of the focused crawler.

Key Words : Focused Crawling, Ontology, Web Document, URL Pattern, Link Filtering, Topic Fitness

1. 서 론

2003년 기준으로 한국의 웹문서는 약 7천만 문서가 존재한다[1]. 해마다 두배 정도 늘어난다고 가정하면 현재 한국어 웹문서만 수억 페이지가 웹 상에 존재한다. 최근에는 블로그나 신문 기사 등의 다이나믹 페이지들을 누구나 손쉽게 작성할 수 있는 환경이 갖추어져 웹 문서의 증가를 가속화시키고 있다.

웹 문서가 기하급수적으로 늘어나는 원인은 여러 가지가 있지만 최근 몇 년간의 웹 문서 증가에는 블로그의 발달, 게시판의 활성화, 오프라인 매체의 온라인화 등이 큰 역할을 하고 있다. 블로그는 정보 이용자로만 역할을 하던 많은 사람들을 정보 생성자로 만들면서 새로운 웹 영역을 탄생시켰다. 그리고, 웹 데이터베이스가 일반화되면서 대부분의 사이트에 게시판 기능이 추가되어 웹 환경이 단방향의 정보 제공에서 양방향 정보 교류 환경으로 전환되는데 큰 역할을 하고 있다. 그 결과로 한 사이트의 문서 중에 게시판 문서가 차지하는 비중이 점점 높아지고 있다. 2000년대 이후, 대표적인 언론 기관을 비롯하여 대부분의 언론 매체들이 웹사이트를 운영하면서 실시간으로 수많은 뉴스와 정보 페이지를 만들어 내고 있다. 또한, 전자상거래가 활성화되면서 많은 인터넷 쇼핑몰과 가격비교 사이트 등 쇼핑 관련 사이트들이 상품에 관한 많은 문서들을 만들어 내면서 전체 웹에서 이들 상업용 웹사이트들이 상당 부분을 차지하고 있다.

상업용 웹사이트에서 제공하는 정보 페이지는 일반인이 만드는 블로그 문서나 게시판 문서보다 신뢰성이 높은 문서로서 그 정보의 가치가 상대적으로 높다. 정보 제공업체의 정보는 기존 오프라인 시절부터 각종 통계나 정보 시스템을 위해 활용되어 왔는데, 온라인화가 이루어진 이후에 이들 사이트에서 제공하는 수많은 구조화된 정보가 연구자들의 기초 데이터로 많이 요구되고 있으며 이를 활용한 연구나 시스템도 다수 발표되어 왔다[2].

본 논문에서는 이러한 상업용 웹사이트의 정보를 수집하기 위하여 이들 사이트의 문서 중에서 특정 주제에 맞는 문서만을 수집하는 주제 중심 문서 수집 방법(focused crawl-

접수일자 : 2007년 11월 13일
완료일자 : 2007년 12월 25일
본 논문은 정통부 및 정보통신연구진흥원의 정보통신 선도기반 기술개발사업의 연구결과로 수행되었습니다.

ing)을 제안하고, 이 방법으로 웹 문서 수집 시스템을 구현하고자 한다.

한편, 처리할 수 없을 정도로 범람하는 정보로 혼란스러운 현재의 인터넷 환경을 개선하기 위하여 최근 시맨틱 웹(Semantic Web)이 주목받고 있다. 이 시맨틱 웹은 무의미한 HTML의 태그 대신에 의미를 가지는 태그를 사용함으로써 기계가 스스로 문서의 의미를 파악하여 지능적으로 처리가 가능하도록 하고 있다. 여기서 태그에 부여하는 의미는 의미망의 발전된 형태인 온톨로지(Ontology)로 구현되고 있다. 이 온톨로지를 구축하기 위해서는 많은 정보가 필요하며 그 정보의 신뢰성도 매우 높아야 한다. 특히, 온톨로지를 특정 시스템에서 활용하기 위해서는 다양한 온톨로지의 인스턴스 정보들이 요구되는데, 이러한 정보를 수작업으로 입력하는 것은 불가능한 일이다. 본 논문에서는 제안하는 주제 중심 웹 문서 수집 방법을 이용하여 IT 분야 온톨로지 인스턴스 구축을 위한 IT 분야 웹 문서를 수집하고자 한다.

2장에서는 기존 웹문서 수집기의 동향과 시맨틱 웹 및 온톨로지에 관한 관련 연구를 살펴보고, 3장에서는 상업용 웹 사이트의 특징과 본 논문에서 제안하는 수집 방법에 사용될 URL 패턴에 대해서 분석한다. 4장에서는 본 논문에서 제안하는 문서 수집 방법을 IT분야 문서 수집을 위한 문서 수집기 구조에 따라 설명한다. 5장에서는 제안된 방법으로 구성한 문서 수집기로 웹 문서를 수집한 실험 결과를 분석하고, 6장에서 결론 및 향후 연구방향을 기술한다.

2. 관련 연구

2.1 웹 문서 수집기

웹 문서 수집기(Web Crawler) 분야는 인터넷 정보 검색에서 색인에 필요한 웹 문서를 수집하는 기술로서 과중한 네트웍 트래픽을 유발하지 않으면서 빠르게 대량의 웹 문서를 수집하는 것을 목표로 한다. 초기의 웹 크롤러는 인터넷 상의 모든 문서를 다운로드하는 것이 목표였기 때문에 한 문서에 나타나는 모든 URL 링크를 대등한 위치에서 탐색하여 수집을 시도하였다. 1990년대 후반 이후 웹문서의 양이 폭발적으로 늘어나면서 수집되는 문서의 수도 같이 늘어나고 이를 이용한 색인의 수도 엄청나게 늘어났다. 그 결과 사용자가 원하는 문서에 비하여 색인을 통해 검색되는 문서의 수가 너무 많아져서 검색의 정확도(precision)와 재현율(recall)이 좋아짐에도 불구하고 사용자의 검색 만족도는 떨어지는 결과를 가져오게 되었다.

2000년대 이후 이러한 문제점을 해결하기 위한 시도들이 연구되었다. 구글의 페이지랭크(PageRank)가 대표적인 것으로 백링크[1](back link)의 양을 문서의 중요도로 파악하여 이를 토대로 한 랭킹 정보를 검색 결과에 반영함으로써 검색 만족도를 높이고 있다[3].

한편, Chakrabarti는 웹 상에 존재하는 문서를 주제별로 분류하여 수집하는 주제 중심 웹 문서 수집(focused crawling) 방법을 제안하였다[4]. 웹 문서의 양이 일반적인 웹 크롤러로 수집하고 유지하는 한계를 넘어섰기 때문에 정보를 요구하는 쪽의 관점에서 관련이 있는 문서를 분류하여 따로 수집함으로써 검색 만족도를 높이는 방법이 제시되었다. 이 방법은 버티컬 검색[2](vertical search)이 보편화되면서 필요한 문서만 구별하여 수집하는 목적으로 최근 주목을 받고 있다[4-7]. 또한 최근에는 검색 분야를 넘어서 다른 연구 분야의 정보 수집 목적으로도 활용되고 있다[8].

2.2 시맨틱 웹과 온톨로지

구글의 페이지랭크를 이용한 검색이나 버티컬 검색과 같은 지능적인 검색 방법들이 나오고 있지만 기하급수적으로 성장하고 있는 웹 환경을 통제하기는 현실적으로 어려워지고 있다. 그 원인의 하나로 웹이 가진 구조적인 문제를 들 수 있다. 웹 문서를 기술하는 HTML은 웹 문서의 시각적인 표현을 위한 태그들이 대부분이고 문서의 내용(contents)을 관리하거나 기술하는 태그는 거의 없다. 웹 검색과 같은 웹 서비스들은 컴퓨터에 의해서 수행되는데 정작 컴퓨터는 자신이 관리하고 있는 문서가 가리키는 정보의 내용을 알지 못한다.

인터넷의 창시자인 팀 버너스 리는 이러한 문제를 해결하기 위하여 차세대 웹의 표준으로 시맨틱 웹(semantic web)을 주창하고 있다[9]. 시맨틱 웹은 텍스트의 의미를 XML 기반 태그를 이용하여 표현함으로써 컴퓨터가 자연어 처리를 통하지 않고 문서의 내용을 파악할 수 있다. 이것을 가능하게 하는 것은 시맨틱 웹의 지식 구조인 온톨로지이다. 온톨로지는 개념들의 상호관계를 정의하는 의미망의 확장으로서 특정 분야의 사물이나 개념들의 다양한 관계를 표현한다.

온톨로지를 구축하기 위해서는 대상이 되는 분야의 전문 용어를 알아야 할 뿐만 아니라 그 용어들 간의 관계를 정확하게 파악할 수 있어야 하기 때문에 그 분야의 전문적인 지식이 필수적이다. 그리고, 온톨로지의 활용을 위해서는 개념 관계를 정의하는 온톨로지 클래스뿐만 아니라 실제로 사용되는 사례인 온톨로지의 인스턴스의 구축도 필요하다. 즉, '고급 승용차'가 개념이라면 '에쿠스'나 '체어맨'과 같은 상용차의 이름이 인스턴스가 된다.

온톨로지의 활용은 시맨틱 웹에서 뿐만 아니라 의미관계에 대한 지식베이스가 필요한 많은 응용 시스템에서 이루어지고 있으며, 각 분야에 대한 온톨로지가 정보 인프라라는 관점에서 국가적인 지원으로 구축이 이루어어지고 있다. 그 한 예로서 우리나라 산업의 큰 핵심인 IT분야의 온톨로지 구축이 진행되고 있다. 특히 IT분야는 웹 정보와 밀접한 관련이 있기 때문에 대부분의 IT관련 정보가 웹 상에 존재한다. 따라서, IT 분야 온톨로지 구축에는 웹 정보의 역할이 필수적이며, 이러한 웹 정보의 수집 또한 매우 필요한 과정이다.

3. 상업용 웹문서 분석

본 논문에서는 온톨로지 인스턴스를 구축하는데 필요한 웹 정보를 수집하고자 한다. 온톨로지는 신뢰성이 매우 중요한 리소스이기 때문에 여기에 사용되는 웹정보는 상당한 신뢰성이 보장되어야 한다. 본 논문에서는 이를 위하여 어느 정도 신뢰성이 있는 상업용 웹사이트를 선정하여 해당 사이트 내에서 원하는 분야의 정보를 가진 페이지를 수집하고자

1) 한 문서에서 다른 문서로 연결해서 나가는 링크를 포워드링크(forward link)라고 하고 다른 문서가 현재의 문서를 참조하고자 링크하는 것을 백링크라고 한다. 백링크가 많은 문서일수록 유용한 문서일 가능성이 높다.

2) 일반적인 웹 검색을 수평적 검색(horizontal search)라고 하고 특정 목적에 따라 맞춤 정보를 제공하는 검색을 버티컬 검색이라고 한다. 네이버나 구글에서 제공하는 이미지검색, 지도검색, 지식검색 등이 여기에 속한다.

한다. 본 장에서는 상업용 웹사이트의 문서 구조를 분석하고 그 특징에 맞는 URL 패턴 기반 문서 수집 방안을 제시한다.

3.1 상업용 웹사이트의 문서

인터넷이 발달하면서 웹 문서를 기술하는 기본 언어인 HTML의 버전이 4.0을 넘어서고 그밖에 DHTML, XML, javascript 등 웹 문서를 표현하는 방법이 다양해지고 있다. 현재 인터넷을 구성하는 웹문서는 문서의 소스를 보기 전에는 어떻게 기술되어 있는지 판단하기 어렵다. 또한, 우리나라는 외국에 비하여 네트워크 하드웨어 인프라 구축이 매우 잘 되어 있어 용량이 큰 웹 문서에 대한 거부감이 적다. 따라서 국내 웹 문서는 문서의 장식을 위해 이미지와 태그를 많이 사용한다. 특히, 각종 포털이나 언론, 쇼핑관련 사이트 등의 유명 정보 제공 사이트들에 있어서 이러한 현상이 더욱 두드러져, 이들 사이트의 한 문서의 크기가 평균적인 웹 문서 크기인 20kByte의 10배 이상인 200kByte에서 300kByte에 이르고 있다. 여기에 문서에 포함된 이미지의 크기를 고려하면 훨씬 더 커지게 된다.

상업용 웹사이트의 문서의 소스를 분석해 보면 몇 가지 특징이 있다. 첫째, 페이지가 구조화되어 있다. 한국 웹 환경은 일찍부터 웹 저작도구가 발달되어 있었기 때문에 웬만한 웹사이트는 저작도구에 의해서 만들어져 <table> 태그를 이용한 구조화된 페이지 형태를 가지고 있다. 둘째, 멀티미디어 데이터를 많이 사용한다. 한국의 빠른 인터넷 환경은 많은 트래픽을 유발시키는 멀티미디어 데이터를 웹 문서에 삽입하는 것을 쉽게 허용하고 있다. 셋째, 대부분의 상업용 웹사이트는 데이터베이스와 연동하여 ASP나 자바스크립트를 이용한 서버 사이드 페이지(SSP: server side page)로 운영되고 있다. 이러한 특징으로 인하여 이 사이트들의 문서의 HTML 소스 파일은 일반적인 웹 문서와 달리 문서의 텍스트에 비하여 태그나 URL이 차지하는 비율이 월등히 높다. 그리고 이러한 추세는 상업용 웹 문서를 넘어서 점차 일반 웹 문서로 확대되고 있다. 또 다른 특징으로 문서의 컨텍스트 측면에서는 문서의 중심 내용의 양에 비하여 부가적인 내용이 차지하는 비율이 매우 높다. 부가적인 내용은 사이트 내 다른 영역으로의 링크나 광고 링크 등이다. 그림 1은 상업용 웹사이트의 일반적인 문서의 소스 중 한 부분을 나타내고 있다. 이들 페이지는 이와 같이 대부분 태그와 링크로 이루어져 있다.

```
<tr>
<td>
<div align="center">
<table width="890" border="0" cellspacing="0" cellpadding="0">
<tr>
<td height="3" colspan="3"></td>
<td colspan="15" bgcolor="b6cee5"></td>
</tr>

<tr><!-- 즐겨찾기버튼 및 무료서비스등 START -->
<td width="384" rowspan="2" align="left" valign="top" style=
"padding-left:9;padding-top:6;"><a href="http://www.saramin.co.kr"
onfocus="this.blur();" target="_top"><img src=
"http://images.saramin.co.kr/index/top/logo_01.gif" width="149"
height="46" border="0" alt=
"&raquo;&ccedil;&para;&divide;&Agrave;&Icirc;"></a></td>
<td width="109" height="19"><a href="javascript:bookmark();"
onfocus="this.blur();"><img src=
"http://images.saramin.co.kr/index/top/bookmark.gif" width="109"
height="19" border="0" alt=
"&Aacute;&ntilde;&deg;&Uuml;&Atilde;&pound;&plusmn;&acirc;&Atilde;&sz
</a></td>
<td width="7"></td>
<td width="39"><img src=
"http://images.saramin.co.kr/index/top/free_01.gif" width="39"
```

그림 1. 상업용 웹사이트의 문서 예.
Fig. 1. An example document of commercial Web site.

상업용 웹사이트 문서의 이러한 특징은 기존 주제 중심 문서 수집기에서 많이 사용하는 문서 분류기의 성능에 영향을 줄 수 있다. 기존 수집기는 웹 문서를 분류할 때 문서의 내용을 비교하기 위하여 내용과 관련이 없는 부분인 태그를 제외하기 때문에 전체 문서 중 일부만의 정보로 분류하게 된다. 그리고 중심 내용 이외 부가적인 내용이 많은 것도 분류를 어렵게 하는 원인이 된다.

따라서, 상업용 웹사이트의 문서 수집은 체계적으로 잘 정리되어 있는 URL을 이용하는 것이 효율적이다. 본 논문에서는 상업용 웹사이트 문서 수집을 위하여 URL로 구성된 링크의 구조를 분석하여 이 링크 구조와 링크 주변 텍스트를 이용하여 수집 대상 문서를 찾고자 한다.

3.2 링크 구조 분석

3.2.1 기본 URL 패턴

국내 대부분의 대용량 데이터베이스를 가진 웹사이트는 카테고리별로 컨텐츠를 분류한다. 정보 컨텐츠는 웹서버 측의 서버 사이드 페이지(ASP, ASP.NET, JSP, PHP 등)를 이용하여 보여준다. 서버 사이드 페이지는 문서를 동적으로 생성하기 때문에 정적인 웹페이지와 다른 URL 패턴을 가지고 있다[10].

http://www.web.com/detail.aspx?prod_id=100320&view=contents http://웹사이트주소/서버사이드페이지?변수전달용 쿼리

그림 3. 서버 사이드 페이지의 URL 패턴.
Fig. 2. URL pattern of server side page.

그림 2는 웹사이트에서 사용하는 일반적인 서버 사이드 페이지 URL의 형식과 한 예를 나타낸 것이다. 이 URL의 뒷부분은 변수 전달 쿼리로 구성되는데 여기서 데이터베이스와 연결되는 ID값으로 데이터베이스에 저장된 각각의 컨텐츠와 연결된다. 따라서 이 ID값을 이용하면 사이트 내의 문서에 접근할 수 있다.

3.2.2 상업용 웹 문서의 유형

본 논문에서는 100여개의 국내 정보제공 사이트를 조사하여 사이트의 구성과 페이지 형태, 접근 방법 등을 분석하였다. 그림 3은 분석 결과를 도표로 나타낸 것이다.

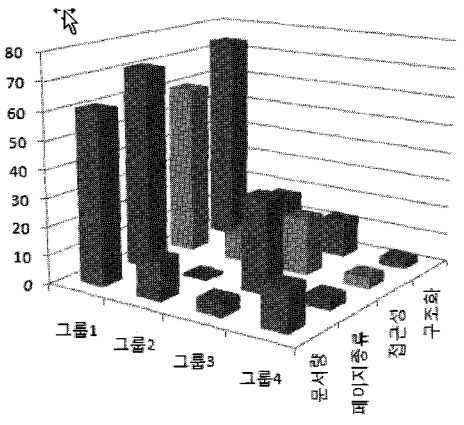


그림 3. 웹사이트 분석 자료
Fig. 3. A Chart of Web sites analysis.

상품 정보나 뉴스와 같이 많은 양의 데이터베이스화된 페이지를 가진 사이트들은 몇 가지 페이지 유형으로 분류할 수 있는데, 일반 페이지, 리스트 페이지, 상세내용 페이지로 구분할 수 있다. 일반 페이지는 웹사이트 내에서 다른 메뉴로 가기 위한 중간 단계, 즉 정적인 자료를 서비스하기 위한 페이지들이다. 여기에는 사이트의 시작 페이지를 비롯하여 특정 주제의 카테고리를 가리키는 링크를 많이 포함하고 있다. 문서 수집에서는 이러한 페이지로부터 주제에 적합한 카테고리 링크를 추출할 수 있다. 따라서, 일반 페이지는 문서 수집기 측면에서는 카테고리 페이지라고 볼 수 있다. 그림 4는 상품비교 사이트의 시작 페이지의 카테고리 부분을 나타내고 있다. 리스트 페이지는 특정 카테고리의 상세내용 페이지들의 링크와 상세내용 페이지의 내용을 요약하여 보여준다. 그림 5는 리스트 페이지의 예를 나타낸 것으로 특정 상품 정보의 요약된 내용을 나열하고 있는데 이 텍스트에 상세내용 페이지의 링크가 연결되어 있다. 상세내용 페이지는 웹사이트에서 제공하는 주 컨텐츠로, 상품 정보의 경우 표를 이용한 구조문서와 특정한 문자열 패턴으로 표시되는 반 구조 문서로 되어 있다[11]. 그림 6은 상세내용 페이지의 예를 보여주고 있다.

그림 4. 일반 페이지의 예.
Fig. 4. An example of general page.

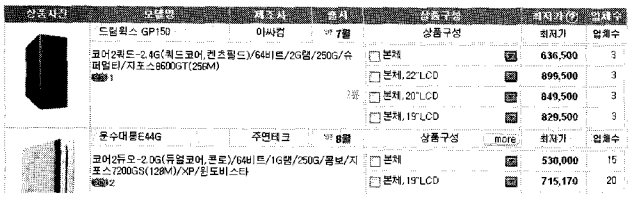


그림 5. 리스트 페이지의 예.
Fig. 5. An example of list page.

제조회사	COMPAQ	출시년월	2005년
케이스 타입	미들타워	운영체제	
LCD 크기		LCD 해상도	
프로세서			
CPU 분류		탑재 CPU	

그림 6. 상세내용 페이지의 예.
Fig. 6. An example of detail page.

본 논문에서는 위에서 분석한 것과 같이 상업용 웹 문서를 카테고리 페이지[3], 리스트 페이지, 상세정보 페이지로 나누어 분류하여, 각 문서 유형의 특징을 이용한 문서 수집 알고리즘을 개발한다.

3) 일반 페이지는 본 논문에서 주목하는 카테고리 부분뿐만 아니라 다른 정보를 포함하는 경우가 많다. 또한 카테고리 정보가 없는 일반 페이지도 존재한다. 본 논문에서는 이러한 일반 페이지에서 문서 수집에 필요한 카테고리 정보만 추출하고 나머지 정보는 제외하는 것을 원칙으로 한다.

4. URL 패턴과 주변 텍스트를 이용한 주제 중심 문서 수집기

4.1 시스템의 구성

본 논문에서는 3장에서 분석한 상업용 웹 문서의 URL 패턴과 링크 주변의 텍스트를 이용하여 목표 주제에 대한 문서를 수집하는 시스템을 제안한다. 그림 7은 제안하는 시스템의 구조를 나타내고 있다. 시스템은 문서 수집기, Link 추출기, 주제 분류기의 세 가지 모듈과 수집 대상 문서의 URL 리스트, URL 분석에 사용될 패턴 스크립트, 주제 분류기에 적용되는 주제 지식 등의 리소스로 구성된다.

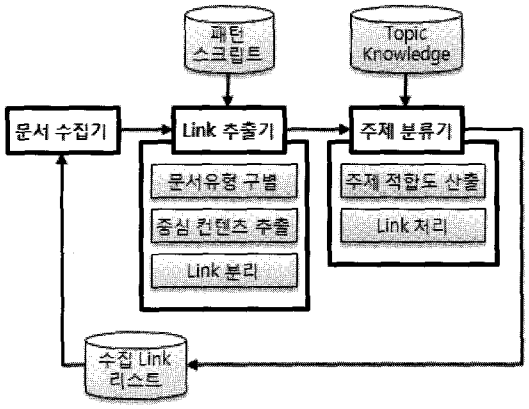


그림 7. 시스템 구성도
Fig. 7. Diagram of system architecture.

문서 수집기는 수집 링크 리스트로부터 하나의 URL을 가져와서 웹에서 문서를 수집해온다. 수집된 문서는 분석용 문서와 수집용 문서로 나뉘는데, 분석용 문서는 카테고리 페이지나 리스트 페이지처럼 실제 수집 대상인 상세정보 페이지로 가기 위한 링크 문서로서, 수집 후 링크 추출기의 입력으로 사용된다. 수집용 문서는 상세정보 페이지로서, 주제 분류기를 통하여 주제에 적합한 링크로 판단된 URL로부터 문서를 수집하여 적절한 문서번호를 붙여서 로컬디스크에 저장한다. 이때 문서 수집에 대한 정보를 XML 형태로 저장하여 수집된 문서를 이용하는 시스템에 부가 정보로 제공한다.

링크 추출기는 한 문서에 나오는 모든 링크 중에서 해당 문서의 중심 영역에 존재하는 링크만을 추출함으로써 주제 분류에 대한 부하를 줄여준다. 주제 분류기는 1차적으로 주제 범위를 줄여서 입력되는 링크 블록의 텍스트를 주제 지식을 참조하여 수집 대상 링크로 분류해낸다. 본 논문에서는 이 두 단계의 모듈을 연동함으로써 빠르고 정확한 문서 수집을 가능하게 한다.

4.2 링크 추출기

링크 구조와 URL 패턴을 이용하여 필요없는 링크는 제거하고 관심 영역의 링크만 필터링하는 링크 추출기는 세 가지 처리 모듈로 나뉘는데, 문서유형 구별과 중심 컨텐츠 추출, 그리고 링크 분리 모듈로 구성된다.

4.2.1 문서유형 구별

3장에서 대용량 데이터베이스를 사용하는 상업용 웹사이트의 문서들의 유형을 나누었다. 카테고리 페이지는 특정 주제

의 문서들을 목록으로 제시하는 리스트 페이지로 가는 링크들을 모아놓았거나 혹은 또 다른 세부 카테고리 페이지로 이동하는 링크들의 나열일 수도 있다. 리스트 페이지는 웹사이트에서 제공하는 정보를 나타내는 상세정보 페이지의 링크들의 목록으로 이루어져 있다. 그리고, 리스트 페이지의 경우 많은 상세정보 페이지를 한 페이지에 나타낼 수 없기 때문에 리스트 하단에 다음 리스트 페이지의 링크들이 나열되어 있다.

각 문서 유형들은 페이지의 특징이 다르고 그 속에 포함된 링크들의 역할이 다르게 때문에 그 문서로부터 추출한 링크의 처리 방법이 다르다. 따라서 문서로부터 링크를 분리하기 전에 문서의 유형을 먼저 구별해야 한다. 문서 유형에서 상세정보 페이지는 다른 유형과 페이지 구성이 다르기 때문에 쉽게 구분되지만 카테고리 페이지와 리스트 페이지는 URL 목록의 나열이라는 점에서 유사하기 때문에 소스 파일에서 두 유형을 쉽게 구별하기 어렵다. 본 논문에서는 몇 가지 휴리스틱을 이용하여 두 유형의 문서를 구별한다.

- 리스트 페이지는 URL 패턴에서 "category=?" 부분이 동일한 경우가 많다.
- 카테고리 페이지에 비하여 리스트 페이지는 한 항목의 텍스트 양이 훨씬 많다.
- 리스트 페이지의 링크 텍스트에는 같은 문자열이 반복적으로 나오는 경향이 많다.
- 리스트 페이지가 카테고리 페이지보다 한 항목 내의 링크 수가 많다.

4.2.2 중심 컨텐츠 추출

중심 컨텐츠 추출과정은 문서 내 링크의 분포를 분석하여 부가 정보를 제거하고 해당 문서의 중심 내용이 있는 영역을 구별한다. 이때 URL 패턴 정보를 이용하여 링크의 분포를 분석한다. 링크 분포를 분석하기 위하여 본 논문에서는 하나의 웹 문서를 링크를 중심으로 하는 블록으로 간주하고, 그림 8과 같이 링크 블록을 정의한다. 하나의 링크 블록은 하나의 "<a> ... " 태그와 이어서 나오는 링크 주변 텍스트로 이루어진다. 링크 주변 텍스트의 경계는 웹 문서의 형태에 따라서 달라진다.

 [<img ...>] 링크 텍스트 주변 텍스트

그림 8. 일반적인 링크 블록 구조.
Fig. 8. General structure of Link block.

카테고리 페이지나 리스트 페이지는 그림 8과 같은 링크 블록이 같은 패턴으로 연속해서 나타난다. 이 패턴을 찾기 위하여 본 논문에서는 링크 근접도와 URL 패턴의 유사도를 이용한다. 링크 근접도는 HTML 문서의 DOM(Document Object Model) 트리를 만들어 트리에서의 계층 간 거리를 이용하여 나타낸다. 식 (1)은 링크 근접도를 구하는 식을 나타낸다.

$$close(a,b) = d(a,b) + |depth(a) - depth(b)| \times w \quad (1)$$

$d(a,b)$는 그림 9와 같은 DOM 트리에서 트리를 추적하여 목적한 노드까지 도달하는 데까지의 노드 간 거리를 나타내는 것으로, 그림 9에서 A노드와 B노드 사이의 거리는 6이 된다. $depth(a)$는 트리 구조에서의 깊이를 나타내는 것으로 그림 9에서 노드 A의 깊이는 3이 된다. w는 가중치를 나타낸다. 일반적으로, 웹 문서에서 리스트의 항목들은 DOM 트리에서 같은 깊이를 가진다(그림 9에서 점선 블록 참조). 그러나 드물게는 상위 노드나 하위 노드에 있는 경우도 있기 때문에 깊이 차에 대해 가중치를 부여하여 링크 근접도를 나타낸다.

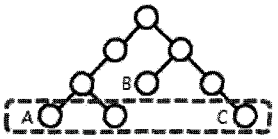


그림 9. DOM 트리에서의 URL간 거리.
Fig. 9. Distance between URLs in the DOM tree.

URL 패턴의 유사도는 하나의 링크와 주변 링크들을 비교하여 동일한 URL 패턴의 길이를 측정하여 나타낸다. 같은 카테고리 그룹이나 리스트 그룹에 속하는 링크들은 문서 내의 물리적인 위치도 근접해 있지만 두 항목의 링크 사이에는 같은 위치를 가리키는 다른 링크 태그들이 존재한다. 그림 10에서 리스트 페이지의 일부를 나타낸 것으로 리스트의 한 항목에 많은 링크가 존재하는 문서의 예를 나타내고 있다. 리스트의 첫 번째 항목에서 링크A와 같은 링크가 휴대폰 이미지와 휴대폰 설명 텍스트에도 존재한다. 이 문서의 경우, 링크A와 링크B 사이에 이들 링크와 같거나 보조 기능의 링크들이 20개 존재한다. 본 논문에서는 문서마다 링크 수가 다른 점을 고려하여 한 문서의 링크 수에 비례하여 가변적으로 비교 대상 링크의 개수를 정하여, 현재 링크 전후의 최대 30개의 링크와 URL 패턴 비교를 하도록 한다.

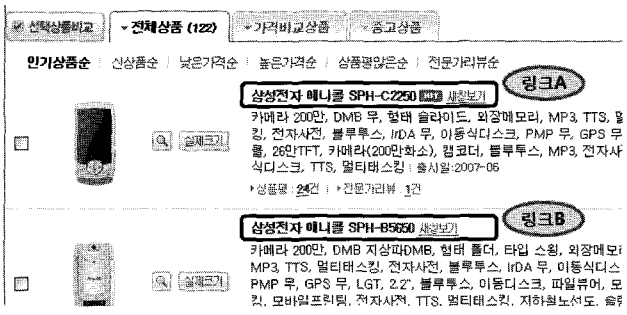


그림 10. 리스트 페이지에서의 링크 배열.
Fig. 10. Link array in the list page.

일반적으로 같은 주제의 리스트 URL은 유사한 패턴을 가진다. 그러나, 카테고리의 경우에는 각 카테고리의 주제에 따라 다른 URL 패턴을 가지는 경우도 많이 존재한다. 그래서 문서 유형에 따라서 URL 패턴의 유사도의 영향은 다르다. 본 논문에서는 DOM 트리에서의 링크 근접도와 물리적 거리에 기반한 URL 패턴 유사도를 함께 고려함으로써 오류를 최소화시키고 있다.

4.2.3 링크 분리

중심 컨텐츠 추출과 문서 유형의 분석 결과로부터 링크를 추출한다. 링크 추출은 문서 유형에 따라 달라진다. 카테고리 페이지는 링크를 추출하여 분석용 링크 리스트에 저장한다. 리스트 페이지는 항목 리스트와 다음 리스트 페이지들의 링크로 구성된다. 항목 리스트는 수집용 링크 리스트에 저장하

여 문서 수집기에서 수집이 이루어지도록 한다. 다음 페이지 링크들은 분석용 링크 리스트로 보내어 다음 페이지를 가져오도록 한다. 본 논문에서는 빠른 문서 수집을 위하여 상세정보 페이지에 대해서 페이지 분석을 하지 않는다. 그러므로 상세정보 페이지에서는 링크 추출이 이루어지지 않는다. 단, 상위 페이지의 주제 판별을 위해 수집되는 경우가 있는데, 이 경우에는 링크 추출기를 통과하여 주제 분류기로 바로 넘어가게 된다.

4.3 주제 분류기

링크 추출기에 의해 주변 정보가 제거되고 문서 수집에 필요한 링크 목록을 구할 수 있다. 본 논문에서는 주제 분류기를 이용하여 이 링크 목록에서 목적한 주제에 맞는 문서를 분류해낸다.

본 논문에서는 IT분야 온톨로지 구축을 위해 IT분야 문서를 수집하는 것을 목표로 하고 있다. 따라서, 주제 분류 항목은 관련 문서와 비관련 문서, 두 가지로 제한되지만, 정확한 분류 성능이 요구된다. 나이브 베이즈 분류(naive bayesian classification)등을 이용한 일반적인 문서 분류기는 여러 그룹의 문서를 효과적으로 분류할 수 있으며 많은 주제 중심 문서 수집기는 기계학습에 의한 기존 분류기를 사용한다 [4,6,7]. 그러나 학습 문서의 정확도에 따라 분류기의 성능이 좌우되며 대상 문서의 텍스트의 양이 적을 경우 성능이 떨어지게 된다. 온톨로지는 개념의 종류가 많고 개념 간의 관계도 복잡하다. 잘못된 개념 관계는 그것을 이용하는 시스템에 심각한 오류를 유발할 수 있다. 본 논문에서는 정밀한 온톨로지 구축에 사용될 정보의 신뢰성 향상을 위하여 수작업으로 IT 관련 주제어들을 수집하여 지식으로 사용하는 분류 알고리즘을 개발한다. 본 논문에서 제안하는 주제 분류기는 퍼지 추론을 이용한 적합도 산출 모듈과 페이지 유형에 따른 링크 처리기로 구성된다.

4.3.1 주제 적합도 산출

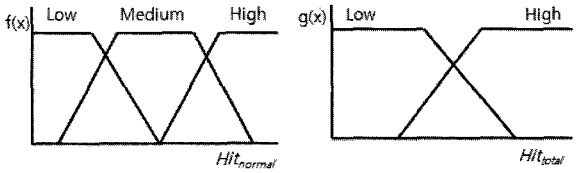
선택된 링크가 주제에 적합한지 판별하기 위하여 본 논문에서는 퍼지 추론을 이용하여 주제 적합도를 산출한다. 대상이 되는 텍스트는 4.2.2절의 그림 8에서 나타낸 링크 블록으로 링크 텍스트뿐만 아니라 URL과 이미지 태그의 문자열, 링크 텍스트에 따라오는 주변 텍스트 등을 포함한다. 이것은 링크 텍스트의 양이 극히 적거나 없는 경우 URL이나 이미지의 파일명 등에 출현하는 주제 관련 단어를 이용하기 위해서이다.

주제 적합도는 주제 관련어 지식에 포함된 단어의 Hit율을 나타내는 Hit_{total}과 주제 관련어의 가중치 계산값을 정규화한 값인 Hit_{normal}을 사용한다. 이 두 값은 식 (2)와 식 (3)으로 표현된다. w_i는 주제 관련어의 가중치로서 중요도에 따라 다른 값을 가지며, 주제 비관련어에는 음수값을 할당한다. Hit_{normal}은 주제 관련어가 많을수록 큰 값을 가지며, 주제 비관련어가 많으면 작은 값을 가지게 된다. Hit_{total}은 관련어의 빈도가 높으면 큰 값을 가지게 된다. 두 가지 값을 보는 것은 주제 관련어와 비관련어가 비슷하게 출현하여 Hit_{normal} 값이 큰 값을 가지지 못하더라도 관련어의 빈도수가 높게 나올 경우 보다 정밀한 판단을 위하여 링크의 문서를 수집하여 하위 문서에서 주제 관련성을 판단하게 하기 위함이다.

$$Hit_{total} = \sum_{i=1}^{n} |x_i w_i| \tag{2}$$

$$Hit_{normal} = \frac{1}{2}\left(\frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} |x_i w_i|} + 1 \right) \tag{3}$$

그림 11은 식 (2)와 식 (3)으로 표현되는 두 가지의 단어 Hit값 대한 퍼지집합을 나타내고 있다. 그리고, 그림 12는 주제 적합도를 구하기 위한 퍼지 추론 규칙을 나타낸다. Hit_N은 Hit_{normal}을 나타내고, Hit_T는 Hit_{total}를 나타낸다.



(a) *Hitnormal*의 퍼지집합. (b) *Hittotal*의 퍼지집합.
(a) Fuzzy set of *Hitnormal*. (b) Fuzzy set of *Hittotal*.

그림 11. 주제 관련어의 퍼지집합.
Fig. 11. fuzzy sets of topic words.

IF Hit_N is HIGH and Hit_T is HIGH then Fitness is GOOD
IF Hit_N is HIGH and Hit_T is LOW then Fitness is GOOD
IF Hit_N is MEDIUM and Hit_T is HIGH then Fitness is INTER
IF Hit_N is MEDIUM and Hit_T is LOW then Fitness is BAD
IF Hit_N is LOW and Hit_T is HIGH then Fitness is BAD
IF Hit_N is LOW and Hit_T is LOW then Fitness is BAD

그림 12. 주제 적합도를 위한 퍼지 추론 규칙.
Fig. 12. Fuzzy inference rule for fitness of topic.

주제 적합도(Fitness)는 주제 관련어 출현 정도에 따라서 적합(GOOD), 부적합(BAD), 관심영역(INTEREST)으로 나뉜다. 주제 적합도의 퍼지집합은 Hit_{normal}의 퍼지집합과 유사한 형태를 가진다. 여기서 적합 판정을 받은 링크는 수집 대상이 되고 부적합 판정을 받은 링크는 리스트에서 제외시킨다. 관심 영역으로 판정 받은 링크는 하위 페이지에서 주제 관련성을 다시 조사한다.

4.3.2 페이지 유형에 따른 링크 처리

페이지 유형에 따라 링크에 대한 처리 방법이 다르다. 카테고리 페이지는 일반적으로 세부 카테고리나 주제 문서 모음인 리스트 페이지로 연결된다. 리스트 페이지는 웹사이트의 주 컨텐츠 페이지인 상세정보 페이지들의 링크 리스트를 가지고 있다. 따라서 리스트 페이지의 링크가 주제 관련 링크로 판단되면 해당 링크의 문서를 수집하여 저장한다. 그림 13은 문서 유형에 따른 링크 처리 방안에 관한 알고리즘을 나타내고 있다.

```
if(page_type is CAT) {  // 카테고리 페이지
    if(Fitness is Bad)
        Link 제외;
    else if(Fitness is Good or Interest)
        Link 문서에서 판단;
}
else if(page_type is LIST) {  // 리스트 페이지
    if(Fitness is Bad)
        Link 제외;
    else if(Fitness is Good)
        Link 문서 수집;  // 상세 페이지 수집
    else if(Fitness is Interest)
        Link 문서에서 판단;
}
else if(page_type is DETAIL) {  // 상세정보 페이지
    if(Fitness is Good)
        해당 문서 수집;
    else if(Fitness is Interest or Bad)
        해당 문서 제외;
}
```

그림 13. 링크 처리 알고리즘
Fig. 13. Link processing algorithm

5. 실험 및 분석

5.1 링크 추출 실험

본 논문에서 제안하는 문서 수집 알고리즘은 문서의 중심 영역을 링크 분포를 이용하여 찾아내어 중심 영역의 링크만 추출함으로서 전체적인 수집 속도를 향상시킨다. 실험을 위하여 9개 사이트의 약 2만 문서를 수집하여 각 문서에서 전체 링크 수에 대한 추출된 링크 수를 조사하였다. 표 1은 링크 추출 실험 결과를 나타내고 있다. 평균적으로 92% 이상의 링크가 제거되고 7.8% 의 링크만 주제 분류기로 넘어가고 있음을 확인하였다. 이 결과는 본 논문에서 대상으로 하는 사이트가 상업용 사이트이기 때문에 광고나 기타 주변 링크들이 일반 웹문서보다 많아서 필터링 효과가 크게 나타났다. 최근에는 일반 웹 문서도 상업용 사이트의 문서와 비슷해지는 경향이 있기 때문에 중심 영역의 링크 추출 효과는 클 것으로 기대한다.

표 1. 링크 추출 실험 결과.
Table 1. The result of Link filtering.

	전체 링크수	추출 링크수	비 율
A사이트	696	111	15.9 %
B사이트	1492	103	6.9 %
C사이트	1144	106	9.3 %
D사이트	1238	59	4.8 %
E사이트	711	210	29.5 %
F사이트	1293	292	22.6 %
G사이트	2539	167	6.6 %
H사이트	6821	142	2.1 %
I사이트	2336	239	10.2 %
전 체	18270	1429	7.8 %

5.2 문서 수집 실험

제안하는 알고리즘으로 구현한 시스템에서 문서를 수집하는 실험을 실시하여 주어진 주제에 적합한 문서의 비율을 조사함으로써 문서 수집 정확도를 측정한다.

수집을 위한 주제는 IT분야 문서이고, 수집 대상 사이트로는 상품가격 비교 사이트, 인물 정보 사이트, 도서 판매 사이트, 기업 정보 사이트 등 다양한 영역의 정보 제공 사이트를 선정한다. 각 사이트는 IT 관련 정보와 비 IT 정보가 함께 제공되는 사이트이며 IT 분야 문서만 존재하는 사이트는 수집 실험 대상에서 제외한다. 표 2는 위 조건을 만족하는 8개 사이트에서 약 7000개 문서를 수집하여 무작위로 10%정도의 문서를 수작업으로 주제 적합 여부를 확인한 결과를 나타내고 있다. 774개 문서를 확인한 결과 715개 문서가 주제에 적합한 문서로 판명되어 92%의 정확률을 나타내었다. 기존 주제 중심 문서 수집 알고리즘의 정확도가 80%~90% 정도 [6,7,13]이므로 만족할만한 수준의 결과로 볼 수 있다.

그러나, 기존 알고리즘의 경우 사이트 제한을 두지 않고 수집을 한 경우이고, 본 논문에서는 온톨로지 구축을 목적으로 하기 때문에 사전에 수집 대상 사이트를 제한하고 있다는 점에서 정확한 성능 비교를 할 수 없다. 기존 알고리즘이 넓은 영역의 문서 수집에서 우수한 반면, 본 논문의 알고리즘은 지정된 사이트 내에서 우수한 성능을 나타낸다. 그 예로 본 논문의 실험에서 제외한 정제된 IT관련 사이트의 문서를 포함할 경우 제안하는 알고리즘의 정확률은 98%이상이 된다.

표 2. 주제 중심 문서 수집 실험 결과
Table 2. The result of focused-crawling

	샘플링 문서수	IT관련 문서수	정확률
A사이트	200	197	99 %
B사이트	77	64	83 %
C사이트	50	39	78 %
D사이트	42	39	93 %
E사이트	50	46	92 %
F사이트	200	189	95 %
G사이트	105	97	92 %
H사이트	50	44	88 %
전체	774	715	92 %

6. 결 론

현재의 방대한 웹 환경에서 관심있는 문서만 찾아서 수집하는 주제 중심 웹 문서 수집기는 정보 검색을 비롯하여 많은 양의 정보나 지식이 필요한 응용 분야에 매우 유용한 기술로서 활용 범위가 점차 확대되고 있다. 본 논문은 온톨로지 인스턴스 구축을 위한 IT 분야 웹 문서를 수집하기 위하여 링크 추출 기법과 링크 및 주변 텍스트의 주제 적합도를 이용하여 주제 중심 웹 문서 수집 알고리즘을 제안하고, 문서 수집 시스템을 개발하였다. 온톨로지는 매우 정밀한 정보 리소스로서 이를 구축하기 위해서는 높은 신뢰도의 정보가 필요하다. 기존 주제 중심 문서 수집기는 기계학습에 의한 문서 분류기를 이용하여 필요한 문서를 수집하지만, 본 논문에서는 보다 정확한 주제 관련성을 확보하기 위하여 수작업으로 주제 관련 지식을 수집하여 문서 분류에 활용하였다. 그 결과 비교적 만족스러운 92%의 정확률을 얻을 수 있었다. 그리고 링크 추출을 통하여 불필요한 많은 링크를 제외시킴으로써 전체적인 문서 수집 속도를 향상시켰다.

제안한 웹 문서 수집기는 IT 분야라는 특정 주제에 대한 문서만 수집하였지만, 문서 분류 기능을 강화하면 여러 분야에 대한 문서 수집을 동시에 진행할 수 있다. 그리고, 본 논문에서 제한시킨 외부 사이트로의 링크를 링크 분석 범위 안으로 포함시키면 제안한 문서 수집 알고리즘을 일반 문서 수집 알고리듬을 확대할 수 있을 것으로 기대한다. 그 밖에 문서 수집 시스템 구현 상에 있어서 일부 수작업 및 스크립트 처리 부분에 대한 자동화 연구가 보충되어야 한다.

참 고 문 헌

[1] 김성진, 이상호, "웹 로봇 구현 및 한국 웹 통계보고," *한국정보처리학회논문지C*, 제10권, 4호, pp.509-518, 2003.

[2] 정한민, 성원경, "과학기술 용어에 대한 용어 생명주기 고찰," *한국콘텐츠학회 종합학술대회 논문집*, 제4권 2호, pp.84-89, 2006.

[3] Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *Stanford InfoLab Publication Server*, 1999.

[4] Soumen Chakrabarti, Martin van den Berg and Byron Dom, "Focused crawling: a new approach to topic-specific Web resource discovery," *Computer Networks*, Vol.31, No.11-16, pp.1623-1640, 1999.

[4] G. Almpanidis, C. Kotropoulos and I. Pitas, "Combine text and link analysis for focused crawling - An application for vertical search engines," *Information Systems*, Vol.32, No.6, pp.886-908, 2007.

[5] 하은용, 최선완, "정확도 높은 검색 엔진을 위한 문서 수집 방법," *한국정보과학회 학술발표논문집*, 제26권 2호(III), pp.471-473, 1999.

[6] 조창희, 이남용, 강진범, 양재영, 최중민, "주변정보 분할을 이용한 주제 중심 웹 문서 수집기," *정보처리학회논문지B*, 제12권 6호, pp.697-702, 2005.

[7] 이정훈, 전서현, 김선희, "웹 문서 수집을 위한 효율적인 문서 분류," *한국정보과학회 학술발표논문집*, 제33권 2호(B), pp.397-401, 2006.

[8] 김기주, 최영식, "포커스드 크롤러를 이용한 웹 검색 및 모니터링 개인화 시스템," *한국인터넷정보학회 춘계학술발표대회 논문집*, 제5권 1호, pp.297-300, 2004.

[9] 김중태, *시맨틱 웹*, 디지털미디어리서치, 2006.

[10] 정준영, 장문수, "URL 패턴을 이용한 웹문서의 선택적 자동 수집 방안," *퍼지 및 지능 시스템학회 추계학술대회*, 제17권 2호, pp41-44, 2007.

[11] 장문수, 강선미, "도메인지식의 계층화를 통한 온톨로지 인스턴스의 속성정보 추출", *퍼지 및 지능시스템학회 논문지*, 제17권 3호, pp.291-296, 2007.

[12] 김원우, 변영태, "Link와 Clustering을 이용한 적극적 문서 수집 기법," *한국지능정보시스템학회 학술대회논문집*, 제1권, pp.393-398, 2001.

[13] 조광제, 김준태, "하이퍼링크 정보를 이용한 HTML 문서의 자동 분류," *한국정보과학회 학술발표논문집*, 제24권 2호(II), pp.277-280, 1997.

저 자 소 개

장문수(Moon-soo Chang)

1992년 : 고려대학교 전자전산공학과 학사.
1994년 : 동 대학원 전자공학과 석사
2001년 : 동경공업대학 지능시스템전공 박사
2000년~2003년 : 한국전자통신연구원 선임연구원
2003년~현재 : 서경대학교 소프트웨어학과 전임강사

관심분야 : 언어이해, 대화처리, 지능시스템, 정보검색, 온톨로지

E-mail : cosmos@skuniv.ac.kr