온톨로지를 이용한 의미 기반 정보 채움 시스템

민 영 근[†] · 김 인 수[††] · 이 복 주[†††]

요 약

테이블 형태로 이루어진 이력서 양식이나 인터넷 회원 가입에서 개인의 신상 정보를 매번 입력하는 일은 매우 반복적이고 번거로운 일이다. 개인의 신상 정보를 컴퓨터에 저장하고 있다가 인터넷 회원 가입 페이지에 자동으로 채워 주는 몇 개의 시스템이 나와 있으나 필드와 필드 값이 잘못 매치되는 등 정확도가 떨어지는 면이 있다. 본 연구는 컴퓨터에 개인의 신상정보를 저장하고 있다가 개인 데이터 온톨로지를 이용하여 회원가입 페이지(목표 페이지)에서요구하는 사용자의 정보를 추론하고 자동으로 채워주는 시스템을 제안하였다. 추론의 과정에서 먼저 목표 페이지를 분석하여 요구하는 필드명을 추출하고, 유사어 온톨로지를 이용하여 요구 필드명을 표준 필드명으로 변환한다. 표준 필드명으로 변환된 요구 필드는 온톨로지 매치 메이킹을 이용하여 개인 데이터 온톨로지 상의 적절한 레벨을 찾아서 최종적인 필드값을 생성한다. 본 시스템은 목표 페이지와 유사한 필드를 가져올 뿐만 아니라 온톨로지 계층 상에 해당되는 필드를 추론하여 정확한 필드값을 가져오게 된다. 몇 개의 회원 가입 페이지를대상으로 실험한 결과 본 시스템이 기존의 시스템에 비해 정확도에서 우수함을 보였다. 본 시스템은 이력서 양식 등 반복적으로 동일한 정보를 채우는 경우에도 쉽게 적용 가능하다.

키워드 : 의미 기반 정보 추출, 온톨로지, 정보 채움

A Semantic-Based Information Filling System Using Ontology

Youngkun Min[†] · Insu Kim[††] · Bogju Lee[†††]

ABSTRACT

It is very iterative and complicated work to enter the personal information every time one fills the form-based resume or one joins the new membership page on the internet. Although there are some systems that have the personal information on the computer and fill the membership page automatically, their accuracies are not often satisfactory in that the fields and their values do not match exactly. The research proposes and implements a system that has user's information on the computer and reasons and fills the information automatically that a membership web page (target page) requests using the personal information ontology. During the reasoning process, the target page is analyzed to extract the requested fields. Then the requested field names are converted to the standard field names using synonym ontology. The converted requested fields find the appropriate level in the personal information ontology using ontology match making to generate the final field value. The system not only finds the similar fields but also generates the exact field values by reasoning on the information ontology hierarchy. By experimenting with several membership pages on the web, the system showed higher accuracy over the existing systems. The system can be easily applicable to the cases where one iteratively fills the same information such as resume form.

Key Words : Semantic-Based Information Extraction, Ontology, Information Filling

1. 서 론

상업적 웹 사이트 또는 일반 웹 사이트에 회원 가입할 때 사용자는 신상 정보를 웹에 입력하게 된다. 사용자는 한 사이트에만 국한하여 사용하지 않고 다양한 컨텐츠를 원하기 때문에 여러 사이트를 방문한다. 게다가 웹사이트는 계속해서 생겨나고 발전하여 사용자는 반복해서 자신의 정보를 입력해야 하는 상황을 벗어날 수 없다. 또한 사용자는 회사에 지원하거나 자신의 이력을 알릴 양식 형태로 이루어진 이력서를 무수히 만나게 된다. 이때에도 회원 가입과 유사하게 매번 같은 신상 정보를 입력해야 한다.

이때 컴퓨터에 저장되어 있는 개인 신상 정보 프로파일이 있고 이로부터 자동적으로 정보 추출이 되어 해당 목표 페이지나 양식에 올려진다면 매우 유용할 것이다. 올려진 후 사용자는 혹시 잘못 옮겨진 필드를 수정하면 될 것이다.

현재의 웹에서 사용자 정보 관리 시스템의 하나로 마이크

※ 본 연구는2005년도 단국대학교 교내 연구비로 수행되었음.
† 준 회 원 : 단국대학교 대학원 전자컴퓨터공학 석사과정
†† 정 회 원 : 서호전기 시스템연구부 연구원
††† 정 회 원 : 단국대학교 전자컴퓨터공학부 조교수
논문접수 : 2007년 1월 2일, 심사완료 : 2007년 6월 11일

(그림 1) 알프레드 입력 예

로 소프트의 닷넷 패스포트[1]가 있다. 닷넷 패스포트에 등록된 하나의 계정으로 MSN 메신저와 닷넷 패스포트를 채택한 사이트의 로그인을 할 수 있다. 이러한 방식은 회원관리 부분을 닷넷 패스포트에 맡김으로써 편리하지만, 모든 웹사이트가 이 방식을 채택하기란 사실상 힘들다. 독립적으로 구동되어 사용자의 정보를 관리하며 웹 사이트로의 로그인 폼이나 회원가입 폼을 채워주는 프로그램이 좀 더 현실적인 방법이 될 것이다.

이러한 시스템 중 하나로 알프레드[2]가 있다. 알프레드는 회원가입 시 사용자 정보를 자동으로 채워주는 기능을 한다. 알프레드를 처음 시작하면 개인 정보를 입력하게 되고, 이 정보를 사용자의 컴퓨터에 저장한다. (그림 1)에 나타난 바와 같이 일반 사이트에 회원 가입 폼이 나타나게 되면 (왼쪽 아래 화면) 알프레드는 이를 자동으로 감지하여 저장된 사용자 정보를 입력할 것인지 묻는 팝업창(오른쪽 위 화면)을 띄우게 된다. 이러한 과정은 때때로 완전하지 않다. 사용자 정보를 입력한 후에도 어떤 필드는 채워지지 않는 경우가 생긴다. 이는 각 필드의 명칭이 정확히 매치가 되지 않아서, 요구하는 정보가 있음에도 불구하고 필요한 곳에 입력되지 않은 것이다.

이러한 문제의 해결을 위하여 사용자 개인정보 필드명과 사이트의 필드명을 매치시키는 추가적인 정보가 필요하다. 그리고 각 사이트 마다 회원 가입 페이지의 폼이 다르기 때문에, 추가적인 정보는 계속 필요하게 된다. 새로운 사이트에 대해서도 적용될 수 있도록 지속적인 업데이트가 요구되어 영구적인 해결방법으로는 적합하지 않다. 실제로 알프레드에서는 자동 입력창이 자동으로 팝업되지 않거나 사용자의 정보가 제대로 입력되지 않는 사이트가 존재하고, 이를 사용자의 제보를 통해 지속적인 업데이트를 해야 한다. 또 다른 사용자 정보 관리 도구로서 eCARD[3]가 있다. eCARD는 사용자의 정보를 휴대 가능한 스마트카드에 저장하여 움직이는 사용자를 대상으로 하는 모바일 지향적이고 RFID 기술과 접합 가능한 시스템이다. 사용자는 eCARD 리더기를통해서만 카드 정보를 읽고 쓸 수 있기 때문에, 사용자 정보를 개인 컴퓨터에 저장해 놓고 쓰는 타 프로그램과의 차이는 개인 정보를 가지고 다닐 수 있다는 보안상의 이점 밖에 없다. eCARD도 알프레드와 유사한 사용자 인터페이스를 가지고 있다.

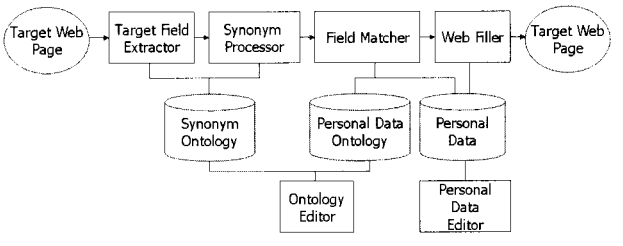
본 논문은 알프레드와 eCARD 같은 시스템의 결점을 해결하고 좀더 정확한 매치를위해 온톨로지를 사용한 의미 기반 필드 매치를 제안한다. 컴퓨터에 개인의 신상정보를 저장하고 있다가 개인 데이터 온톨로지를 이용하여 회원가입 페이지(목표 페이지)에서 요구하는 사용자의 정보를 추론하고 자동으로 채워준다. 개인의 신상정보는 개인 데이터 온톨로지의 "표준" 필드명에 따라 저장되어 있다. 추론의 과정은 다음과 같다. 먼저 목표 페이지를 분석하여 요구하는 필드명을 추출하고, 유사어 온톨로지를 이용하여 요구 필드명을 표준 필드명으로 변환한다. 표준 필드명으로 변환된 요구 필드는 온톨로지 매치 메이킹을 이용하여 개인 데이터 온톨로지 상의 적절한 레벨을 찾아서 최종적인 필드값을 생성한다. 본 시스템은 목표 페이지와 유사한 필드를 가져올 뿐만 아니라 온톨로지 계층 상에 해당되는 필드를 추론하여 정확한 필드값을 생성하게 된다. 몇 개의 회원 가입 페이지를 대상으로 실험한 결과 본 시스템이 기존의 시스템에 비해 정확도에서 더 나았다. 2장에서는 정보추출과 온톨로지를 소개하고 이를 이용한 제안된 시스템을 설명하였다. 3장은 시스템의 인터페이스 구현과 실험이며 4장은 결론이다.

2. 제안된 시스템 구조

정보 추출은 텍스트로 되어 있는 비 정형화된 데이터(unstructured data)로부터 정형화된 데이터(structured data)를 추출한다. 즉 신문 기사, 웹 문서, 전자우편 등과 같이 정형화되지 않은 문서를 입력으로 받아서 미리 정해놓은 (찾기를 원하는) 정보를 찾아내어 주어진 문서를 요약하는 시스템이다[4, 5, 6]. 비 구조화된 혹은 반 구조화된 문서로부터 구조화된 데이터베이스를 구축하는 시스템을 의미한다. 정보추출 시스템은 가격 비교 정보 사이트, 채용 정보 사이트 등에서 응용되고 있다. 정보 추출을 하기 위해서는 자연언어 이해 기술이 필요하다. 방대한 자료들 사이에서 원하는 정보를 추출하려면 문서의 내용을 이해할 수 있어야 한다. 현재 자연 언어 처리 기술로는 자연 언어로 되어 있는 문서를 완전히 이해 하기란 어렵다. 그러나 신문 기사로부터 사건 사고 일지, 구인 구직 광고로부터 채용 분야 및 연봉, 세미나 공고로부터 날짜와 주제 정보, 학술대회 공고로부터 날짜와 마감 시간, 웹 페이지로부터 주식 정보와 날씨 정보 그리고 가격 정보 등은 현재의 기술로도 충분히 추출이 가능하고 일부 기술은 상용화되어 사용되고 있다. 이와 같은 정보 추출은 완전한 자연 언어 이해를 필요로 하지 않는다. 유사한 규칙과 형식으로 이루어진 문서들 사이에서 정해진 정보를 추출하는 것은 부분적인 자연 언어 처리로 필요한 정보만을 이해하고 정확히 분석하면 된다.

최근 시맨틱 웹[7, 9, 10, 11]과 함께 온톨로지(ontology)에 대한 연구가 활발히 진행되고 있다. 온톨로지란 존재하는 것과 그것의 기본적인 범주를 연구하는 학문이라고 일반적으로 정의되며, 데이터베이스나 지식베이스를 기반으로 하는 시스템에서 온톨로지의 역할은 어플리케이션 영역에 존재하거나 존재할 수 있는 사물들의 카테고리를 결정하는 것이다[8]. 인공지능과 웹 연구자들은 온톨로지를 개념들 사이의 관계를 정의하는 문서를 위해 사용한다. 웹에서 온톨로지는 분류체계(taxonomy)와 더불어 일련의 추론과정을 포함하고 있다. 분류체계는 객체들의 클래스와 그 객체들 사이의 관계를 정의 한다. 온톨로지는 예를 들면 정보검색 시 에이전트(agent)가 이용자의 요청에 관련된 정보를 추론하는 지식을 제공하는 역할을 하며 개념(concept)-값(value), 클래스(class)-인스턴스(instance), 상위 클래스(super class)-하위 클래스(sub class)와 같은 개념적 관계를 제공한다. 온톨로지를 이용하여 웹에서 정보를 추출하는 시도도 있다[12]. 온톨로지를 유사어 정의에 사용한 사례로서는 WordNet[13,14] 또는 EuroWordNet[15]가 대표적이다. WordNet에는 일반적인 영어 온톨로지로서 약 95600개의 다른 단어를 가지고 있다. 각 단어의 의미가 유사어 집합과 어구 주석으로 표현된다. WordNet은 단어의 의미 범위가 너무 넓기 때문에, 의료 영역과 같은 특정 영역에서 온톨로지를 유사어를 정의하는데 사용한 사례[16, 17]가 많다.

본 논문에서는 정보추출과 온톨로지 기술을 이용하여 회원가입 페이지에서 요구하는 필드값을 생성하는 시스템을 목표로 하였다. (그림 2)는 본 시스템의 전체적인 구조를 나타내고 <표 1>은 본 시스템의 개략적인 알고리즘이다. 그림에서 왼쪽 상단의 목표 웹 페이지(Target Web Page)는 시스템에서 분석하여 자료를 채워 넣을 목표 페이지이다. 사각형으로 이루어진 일련의 처리기를 거쳐 목표 웹 페이지(오른쪽 상단에 표시된)에 정보가 채워지게 된다. 일련의 처리기는 목표 필드 추출기 (Target Field Extractor), 유사어 처리기(Synonym Processor), 필드 매치기 (Field Matcher), 웹 채움기 (Web Filler)로 이루어져 있다. 목표 필드 추출기(알고리즘 단계 1)는 목표 웹 페이지를 읽고 분석하여 페이지가 요구하는 필드 즉 정보 채움이 필요한 필드를 찾아내는 역할을 한다. 페이지를 분석할 때 목표 필드를 찾아내기 위한 태그 분석 과정을 수행하여 목표 필드명과 채워야 할 해당 필드 값의 위치를 찾아낸다. 목표 웹 페이지에 들어있는 용어는 매우 다양하므로 유사어 온톨로지를 필드명 사전으로 간주하여 온톨로지에 용어가 있으면 목표 필드로 간주하여 추출하게 된다. 유사어 처리기(알고리즘 단계 3)는 목표 필드 추출기가 페이지에서 찾아낸 필드에 대해 유사어 온톨로지를 사용하여 일종의 "표준" 필드 이름으로 변환하는 역할을 한다. 필드 매치기(알고리즘 단계 4)는 유사어 온톨로지가 변환한 표준 필드 이름을 개인 데이터 온톨로지(Personal Data Ontology)에 질의하여 매치시키고 더 세부적인 여러 필드로 변환한다. 이 세부적인 여러 필드를 기준으로 사용자 데이터(Personal Data)로부터 데이터를 가져와



(그림 2) 제안된 시스템의 구조

〈표 1〉 제안된 시스템의 개략적인 알고리즘

1. (Target Field Extractor) 목표 웹 페이지를 분석하여 채움이 필요한 목표 필드명을 추출한다.
2. 각 목표 필드명에 대해 {
3. (Synonym Processor) 유사어 온톨로지를 사용하여 유사어 처리를 통해 목표 필드명을 표준 필드명으로 변환한다.
4. (Field Matcher) 개인 데이터 온톨리지를 이용하여 온톨로지 추론을 수행한다. {
5. 변환된 목표 필드명으로 개인 데이터 온톨로지에서 Exact 관계로 노드 N을 찾는다.
6. N과 N의 자식 노드가 PART-OF 관계이면 N을 넘긴다.
7. N과 N의 자식 노드가 ISA 관계이면 자식 노드 중 대표 자식 노드 M을 넘긴다.
8. N 또는 M을 서브트리의 루트로 하여 그 아래 모든 단말 노드를 넘긴다.
9. 단말 노드에 해당하는 필드들에 대해 개인 데이터로부터 필드 값을 가져와서 접합한다.
   }
10. (Web Filler) 필드값을 목표 웹 페이지에 채운다.

}

"접합"하면 하나의 목표 필드에 대한 필드 값 생성 작업을 마치게 된다. 웹 채움기(알고리즘 단계 10)는 이 접합된 필드 값을 목표 웹 페이지의 필드 값 위치에 채워준다.

본 시스템은 두 개의 온톨로지를 사용하는데 개인 데이터 온톨로지(Personal Data Ontology)와 유사어 온톨로지(Synonym Ontology)이다. 개인 데이터 온톨로지에는 사용자의 신상 정보에 대한 각 필드간의 관계 정보가 트리 구조로 저장되어 있다. 트리의 각 노드 명칭은 대표적인 "표준" 명칭을 지정하여 사용하였다. 그리고 유사어 온톨로지에는 각 사용자 정보 필드의 유사어 정보가 저장되어 있다. 웹 페이지에서 추출해온 각 입력 필드에 대한 정보를 유사어 온톨로지의 유사어 정보에 맞추어, 개인 데이터 온톨로지에 사용된 각 표준 필드명으로 대체하는데 이용된다. 온톨로지 편집기(Ontology Editor)는 이 두 가지 온톨로지를 생성하거나 수정한다. 개인 데이터(Personal Data)는 사용자의 실질적인 정보를 가지고 있고, 여기서 사용되는 필드 이름은 개인 데이터 온톨로지의 용어와 일치한다. 개인 데이터 편집기(Personal Data Editor)는 사용자의 신상 정보를 입력 및 수정하여 개인 데이터를 작성하는 역할을 한다.

〈표 2〉 입력 HTML의 예

```
<TD width=100><FONT color=#808080>?</FONT> 비밀번호
</TD>
<TD><INPUT type=password maxLength=12 size=12 value=""
name=PWD><IMG src="../im-
ages/required.gif">     비밀번호 확인
<INPUT type=password maxLength=12 size=12 value=""
name=PWD_tmp><IMG src="../im-
ages/required.gif"></TD></TR>
<TR>
<TD vAlign=top width=100><FONT color=#808080>?</FONT>
전화 번호<BR><SPAN class=a3>(1개 이상)</SPAN></TD>
<TD>
<TABLE cellSpacing=0 cellPadding=3 width="95%" border=0>
<TBODY>
<TR>
<TD width=100>자택 전화번호</TD>
<TD width=188><INPUT maxLength=20
name=Home_phone></TD></TR>
<TR class=font9>
<TD width=100>직장 전화번호</TD>
<TD width=188><INPUT maxLength=20
name=Offi_phone></TD></TR>
<TR class=font9>
<TD width=100>휴대 전화번호</TD>
<TD width=188><INPUT maxLength=20
name=Mobile_phone></TD></TR>
```

각 처리기를 좀더 자세히 살펴보면 다음과 같다. 필드 추출기는 목표 웹 페이지의 HTML 문서를 분석한다. <표 2>는 목표 웹 페이지 HTML 문서의 예이다. 이를 살펴보면 목표 웹 페이지에서 정보를 입력 받는 부분은 <INPUT> 태그로 되어 있고 그 이름은 NAME 속성으로 명시되어 있다. 본 시스템에서는 <INPUT> 태그에서 나오는 NAME 속성을 중심으로 근처의 단어 토큰들을 추출하고 이를 사전에 질의하여 사전에 단어가 존재하면 필드명으로 간주하는 방식을 취하였다. 여기서 사전 역할을 하는 것은 유사어 온톨로지이다. 유사어 온톨로지에는 가능한 여러 후보 필드명이 정의되어 있으므로 이를 사전으로 사용한다. <표 2>의 HTML 구문 중에서 직장 전화 번호의 경우를 예로 들어보면, <INPUT> 태그의 NAME 속성(Offi_phone)이 있고 이를 중심으로 가까이 "직장 전화번호"가 있고 이것은 유사어 온톨로지에 들어 있으므로 목표필드명으로 추출된다.

유사어 처리기는 다음과 같다. 웹 페이지에 표현된 정보는 의미는 같으나 서로 다른 단어 또는 언어로 표현될 수 있다. '성명', '이름', '실명', 'Name', 'Full Name' 등은 모두 유사어 관계이다. 각 웹 페이지들은 통합된 단일 명칭을 사용하지 않는다. 이러한 상황에서 컴퓨터가 추출된 정보를 이해하고 정보간의 관계를 알지 못하면, 정확한 매칭은 불가능하다. 이를 위해 추출된 필드를 "표준" 필드명으로 변환하여 정보의 의미를 명확히 하는 과정이 필요하게 된다. 유사어 온톨로지는 추출된 정보의 의미를 정확히 하기 위한 표준명과그에 해당하는 유사어들을 가지고 있고 단어들 사이의 관계를 정의한다. <표 3>은 유사어 온톨로지를 보여주

〈표 3〉 유사어 온톨로지

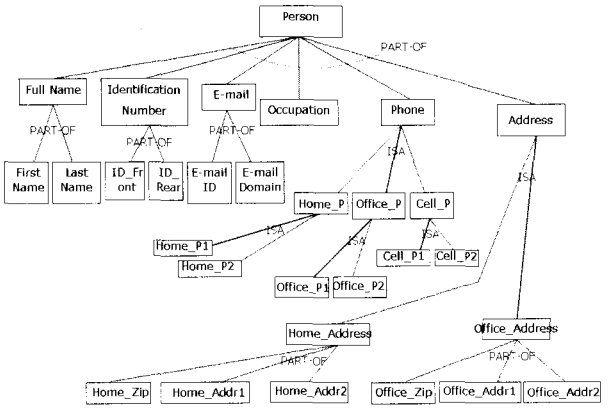
```
<Person rdf:ID="Full Name">
 <owl:sameAs>
  <Person rdf:ID="full_name_captin">
   <owl:sameAs>
    <Person rdf:ID="성명">
     <owl:sameAs rdf:resource="#full_name_captin"/>
    </Person>
   </owl:sameAs>
   <owl:sameAs>
    <Person rdf:ID="이름">
     <owl:sameAs rdf:resource="#full_name_captin"/>
    </Person>
   </owl:sameAs>
   <owl:sameAs rdf:resource="#full_name_captin"/>
   <captin
    rdf:datatype="http://www.w3.org/2001/
    XMLSchema#string"> true</captin>
    </Person>
  </owl:sameAs>
</Person>
<Person rdf:ID="phone">
 <owl:sameAs>
  <Person rdf:ID="phone_captin">
   <captin
   rdf:datatype="http://www.w3.org/2001/
   XMLSchema#string">true</captin>
```

고 있다. 유사어 온톨로지에는 Person 클래스가 하나 있고 클래스의 속성으로 captin을 가지고 있다. 이 captin 속성이 유사어들 중에대표어를 표시하도록 정의되어 있다. 각 유사 단어들은 Person 클래스의 인스턴스(instance)가 되어, 해당 단어가 들어올 경우 captin 속성에 참값을 갖도록 한다. 여기에서는 유사어 온톨로지 중 '이름', '성명'에 대한 유사어 관계를 정의한 것을 보이고 있다. 본 시스템의 유사어 온톨로지에서는 "Full Name"이 대표어로 지정되어 captin 속성 값을 참으로 설정하였다. 같은 의미를 갖는 모든 유사어들은 기본 속성인 sameAs 속성으로 연관되어 유사 관계를 정의한다.

목표 페이지에서 추출된 필드명을 유사어 온톨로지에 질의하여 표준 필드명을 얻어오는데 <표 4>와 같은 질의를 통해 sameAs 속성으로 연관된 동시에 captin 속성이 참값인 대표 유사어 중 표준 필드명을 찾아낸다.

〈표 4〉 유사어를 받아오는 질의

```
String queryString = "SELECT ?same" +
"WHERE (feat:"+inputname+" owl:sameAs ?same),"
+ "(?same feat:captin \"true\")"
+ USING feat FOR
http://ai.dankook.ac.kr/person_data_syninym.owl#"
```



(그림 3) 개인 데이터 온톨로지

다음은 필드 매치기로서 이는 개인 데이터 온톨로지를 사용한다. 개인 데이터 온톨로지는 사용자의 신상 데이터를 세부적으로 분류하여 각 정보들 간의 관계를 명시하는 온톨로지이다. (그림 3)과 같은 구조를 가진 개인 데이터 온톨로지는 하나의 클래스(Person)와 그에 속한 속성들로 구성된다. 각 속성들은 사용자의 신상정보를 상위 속성과 부 속성의 관계로 분류하여 트리 구조를 가진다.

사용자의 신상 정보는 이름, 주민등록번호, 이메일, 직업, 전화번호, 주소 등으로 구성된다. 여기에서 전화번호는 집전화, 회사전화, 휴대폰으로, 주소는 집주소와 회사주소로 세부적으로 분류된다. 이 온톨로지의 노드에 들어가는 용어는 유사어 온톨로지의 표준 필드명을 사용한다는 것을 볼 수 있다. 노드와 자식 노드의 관계, 즉 속성과 부 속성의 관계는 PART-OF 또는 ISA 관계를 가진다. PART-OF 관계는 First Name과 Full Name 관계처럼 부 속성이 모여서 속성을 만드는 관계이다. 반면 ISA 관계는 Home_P과 Phone 관계처럼 부 속성이 속성의 특별한 경우임을 나타낸다. ISA 관계일 경우 여러 자식 노드 중에서 하나가대표 자식 노드로 지정되어 있다(그림에서 굵은 실선으로 표시되어 있다). 예를 들면 Phone의 자식 노드Home_P, Office_P, Cell_P 중에서 Office_P이 대표 자식 노드로 지정되어 있다. 개인 데이터 온톨로지는 그림에서 보는 것처럼 사용자 신상 정보의 성명을 성과 이름, 주민등록번호의 앞과 뒤, 이메일의 ID와 도메인 등으로 더 이상 나눌 수 없는 수준까지 세부적으로 나누어 가지고 있다. 이는 목표 웹 페이지가 어떤 수준의 필드를 원하는지 미리 알 수 없고, 특정 데이터를 나누는 것 보다는 이미 분리된 데이터를 접합시키는 작업이 더 간단하기 때문이다. <표 5>는 개인 데이터 온톨로지를 RDF로 표현한 것이다. 개인 데이터 온톨로지의 구조 중 주소 정보에 대한 관계의 일부를 명시하고 있다. 주소를 의미하는 Address 속성은 집주소와 회사주소를 의미하는 Home_Address와 Office_Address를 부 속성으로 갖는다. 각 Property들은 서로 관계된Property와 subPropertyOf 속성으로 연결되어 전체적으로 (그림 4)에서 본 바와 같은 트리 구조를 형성한다.

이러한 상황에서 필드 매치기는 변환된 표준 목표 필드명으로 개인 데이터 온톨로지에서 Exact 관계로 목표 노드를

<표 5> 개인 데이터 온톨로지의 RDF 표현

```
<owl:DatatypeProperty rdfabout="#Home_Address">
   <rdfs:range
   rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
   <rdfs:subPropertyOf>
        <owl:DatatypeProperty rdf:ID="Home"/>
   </rdfs:subProperty>
   <rdfs:domain rdf:resource="#Person"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdfabout="#Home_Address1">
   <rdfs:range rdf:re
   source="http://www.w3.org/2001/XMLSchema#string"/>
   <rdfs:subPropertyOf>
        <owl:DatatypeProperty rdf:ID="Home_Address"/>
   </rdfs:subProperty>
   <rdfs:domain rdf:resource="#Person"/>
</owl:DatatypeProperty>
```

<표 6> 개인 데이터의 저장된 예

First Name: 길동
Last Name: 홍
ID_Front: 790116
ID_Rear: 1234567
E-mail ID: minyk
E-mail Domain: ai.hankook.ac.kr
Occupation: 대학원생
Home_P1: 02-380-3478
Home_P2: 02-381-3894
Office_P1: 02-394-1930
Office_P2: 02-383-9334
Cell_P1: 010-3840-3944
Home_Zip: 430-283
Home_Addr1: 강남구 압구정동
Home_Addr2: 394-382
Office_Zip: 394-293
Office_Addr1: 서울시 용산구 한남동
Office_Addr2: 한국대학교

찾는다. 이 목표 노드를 N 이라 하자. 목표 노드 N과 N의 자식 노드의 관계를 파악하여 그 관계가 PART-OF 인 경우는 N을 서브트리의 루트로 하는 서브트리의 모든 단말 노드를 가져온다. 만약 그 관계가 ISA인 경우는 자식 노드 중 대표 노드를 찾는다. 이를 M이라 하자. 이 경우는 M을 서브트리의 루트로 하는 서브트리의 모든 단말 노드를 가져온다. 예를 들어 목표 웹 페이지의 필드가 '주소'를 요구하였고 유사어 처리에서 표준 필드명인 'Address'를 얻었다면 필드 매치기는 온톨로지로부터 Exact 매치를 수행하여 'Address'를 찾고 'Address'와 자식 노드의 관계가 ISA이므로 그 자식 노드 중 대표 노드인 Office_Address 를 얻고 이를 서브트리로 하여 모든 단말 노드인 Office_Zip, Office_Addr1, Office_Addr2를 얻게 된다.

다음은 실제 개인 데이터로부터 필드값을 가져와 접합하

는 단계이다. 개인 데이터는 <표 6>와 같이 개인 데이터 온톨로지의 단말 노드에 해당하는 필드와 필드값으로 구성되어 있다. 개인 데이터는 개인 데이터 온톨로지에서처럼 사용자 신상 정보의 성명을 성과 이름, 주민등록번호의 앞과 뒤, 이메일의 ID와 도메인 등으로 더 이상 나눌 수 없는 수준까지 세부적으로 나누어 가지고 있다. 이는 특정 데이터를 나누는 것 보다는 이미 분리된 데이터를 접합시키는 작업이 더 간단하기 때문이다. 이전 단계에서 얻은 필드명들을 이용하여 필드값을 차례로 읽어서 접합한 후 최종 필드값을 생성한다. 예를 들어 이전 단계에서 Office_Zip, Office_Addr1, Office_Addr2을 얻었다면 "394-293 서울시 용산구 한남동 한국대학교"가 생성된다.

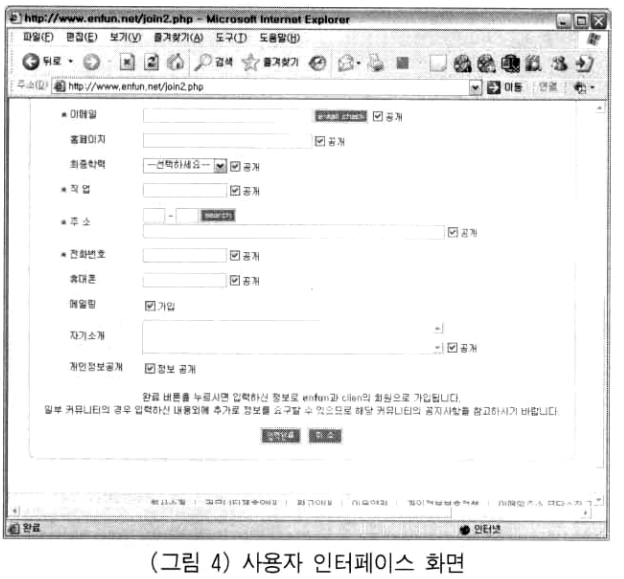
웹 페이지 채움기는 이전 단계에서 접합하여 생성된 필드값을 목표 웹 페이지의 해당 필드값 위치에 채움으로써 하나의 필드에 대한 과정을 마치게 된다. 이 과정은 필드 추출기가 찾아온 모든 필드에 대해 반복된다.

3. 실험 및 결과

본 시스템을 구현하기 위하여 Protégé-2000 온톨로지 편집기, Jene 2.2 RDF 파서, 그리고 Java와 C++가 사용되었다. 본 시스템의 사용자 인터페이스는 (그림 4)과 같이 익스플로러의 도구 모음(오른쪽 상단 +표시 아이콘)에 등록되어, 사용자가 개인 신상 정보를 입력해야 하는 상황에서 아이콘을 클릭하는 것으로 자동 입력을 수행하게 된다.

본 시스템의 성능 테스트를 위해 다음과 같이 웹 사이트 회원가입 페이지를 임의로 15개 선정하였다.

- 사이트 1: http://myjoins.joins.com/register/member_reg.asp
- 사이트 2: http://www.10x10.co.kr/member/joinpage.asp
- 사이트 3: http://member.auction.co.kr/membership/signup/SignupPerson1.aspx?ssn=6311291558812&afterregisturl=
- 사이트 4: http://www.bluemonkey.co.kr/shop/main/main.htm
- 사이트 5: http://login.danawa.com/join_member_step3.php
- 사이트 6: https://login.empas.com/register/register_user_input.html
- 사이트 7: http://www.gundamshop.co.kr/Member/GD_MemR.html
- 사이트 8: https://cymember.cyworld.nate.com/main/regist/regist_foreigner.jsp
- 사이트 9: http://www.cdpkorea.com/
- 사이트 10: http://www.lotteimall.com/regist/NameCheck.jsp



(그림 4) 사용자 인터페이스 화면

- 사이트 11: https://member.shinsegaepoint.com/ccom/servlet/ccom001.do? action=ccom001_reg
- 사이트 12: https://www.aladdin.co.kr/account/waccount_makenew.aspx
- 사이트 13: http://register.freechal.com/ApplyEntry/FcNwJoinForm.asp
- 사이트 14: http://kangcom.com/common/member/form.asp
- 사이트 15: http://www.e-himart.co.kr/member/memberAddInfo.jsp

<표 7>은 15개 사이트에 대하여 기존의 시스템인 알프레드와본 시스템의 정확도에 관한 성능을 비교한 결과이다. 각 사이트 별로 정상적으로 채워진 필드와 오류 필드를 나열하였다. 정상적인 필드의 비율을 정확도로 계산하였다. 정상적/오류 필드 구분은 알프레드의 필드 목록에 기준하여 테스트하였다. 알프레드의 필드 목록은 알프레드가 처음 설치될 입력 받는 필드명, 아이디, 비밀번호, 이름, 주민등록번호, 우편번호, 주소, E-mail, 전화번호, 휴대폰번호를 이른다. (알프레드는 이중 우편번호와 주소는 집과 직장을 구분하여 받는다.)

기존 시스템과 비교해 보았을 때 본 시스템은 정확도에서 높은 수치를 보여 (67% 대 89%) 더 나은 성능을 보였다. 정상적으로 채워진 필드와 오류 필드를 기준으로 알프레드와 본 시스템을 비교해 보면 일단 알프레드가 찾는 필드를 본 시스템에서 찾지 못하는 경우는 발생하지 않았다. 알프레드가 찾지 못한 필드를 본 시스템이 찾은 경우는 표에서 밑줄로 표시하였다. 이에 해당하는 필드는 "휴대폰/호출", "연락처(자택)", "이메일주소", "전화번호", "핸드폰번호",

〈표 7〉 기존의 시스템과 본 시스템의 정확도 비교

사이트	알프레드			본 시스템		
	정상적으로 채워진 필드	오류 필드	정확도	정상적으로 채워진 필드	오류 필드	정확도
사이트 1	성명, 주민등록번호		100%	성명, 주민등록번호		100%
사이트 2	이름, 주민번호, 이메일, 전화번호, 우편번호, 주소	휴대전화	86%	이름, 주민번호, 이메일, 전화번호, 우편번호, 주소, 휴대전화		100%
사이트 3	주민등록번호, 주소지	성명, email주소, 휴대폰번호, 전화번호,	33%	주민등록번호, 주소지, 성명, 전화번호	휴대폰번호, email주소	66%
사이트 4	이름, 주민번호, 전자우편, 우편번호, 주소(자택)	휴대폰/호출, 연락처(자택)	57%	이름, 주민번호, 전자우편, 우편번호, 주소(자택), 휴대폰/호출, 연락처(자택)		100%
사이트 5	이메일, 우편번호, 주소	전화번호, 휴대폰번호	60%	이메일, 우편번호, 주소	전화번호, 휴대폰번호	60%
사이트 6	휴대폰번호, 우편번호, 주소, 전화번호,	이메일주소	80%	휴대폰번호, 우편번호, 주소, 전화번호, 이메일주소		100%
사이트 7	주민번호, 우편번호, 주소, E-mail	전화번호, 핸드폰번호	67%	주민번호, 우편번호, 주소, E-mail, 전화번호, 핸드폰번호		100%
사이트 8	이름	이메일주소, 연락처 자택, 연락처 직장, 핸드폰	20%	이름, 이메일주소, 핸드폰	연락처 자택, 연락처 직장	60%
사이트 9	Name, E-mail, 주민등록번호, Home Address	Home Phone, Cellular	67%	Name, E-mail, 주민등록번호, Home Address, Home Phone, Cellular		100%
사이트 10	우편번호, 주소, 전화번호, 휴대폰번호, 이메일주소		100%	우편번호, 주소, 전화번호, 이메일주소, 휴대폰번호		100%
사이트 11	이름, 주민등록번호, 자택주소, 이메일주소	자택전화번호, 이동전화번호, 직장명, 직장전화번호, 직장주소	44%	이름, 주민등록번호, 자택주소, 이메일주소, 이동전화번호	자택전화번호, 직장전화번호, 직장명, 직장주소	56%
사이트 12	실명, 주소	E-Mail, 주민등록번호, 전화번호, 휴대전화	33%	실명, 주소, E-Mail, 전화번호, 휴대전화, 주민등록번호		100%
사이트 13	이름, 주민등록번호, 핸드폰, 메일		100%	이름, 주민등록번호, 핸드폰, 메일		100%
사이트 14	성명, E-mail, 우편번호, 주소, 전화번호	주민등록번호, 이동전화번호	71%	성명, E-mail, 우편번호, 주소, 전화번호, 주민등록번호, 이동전화번호		100%
사이트 15	아이디, 전화번호, 우편번호, 주소, E-mail 주소	휴대전화	83%	아이디, 전화번호, 우편번호, 주소, E-mail 주소, 휴대전화		100%
평균			67%			89%

"Home Phone", "Cellular", "이동전화번호" 등 유사어 온톨로지와 개인 데이터 온톨로지의 사용에 기인한 것으로 분석된다. 목표 필드 추출기가 유사어 온톨로지를 사전으로 사용하여 "휴대폰", "연락처" 등을 정확히 추출할 수 있었고 표준 필드명으로 변환된 후 개인 데이터 온토로지를 사용하여 접합하여 정확한 값을 생성하였다. 본 시스템도 찾지 못한 필드명은 "email주소", "연락처 자택", "연락처 직장" "자택전화번호" 등 단어 토큰의 분리에 어려움이 있거나("email주소"), 필드명이 웹에계층적으로 배치 되어 있는 경우 ("연락처" 아래 "자택", "직장"), 유사어 온톨로지에 등록되어 있지 않은 경우 ("자택전화번호" 등)이다.

전체적인 처리 시간 비교에서는 본 시스템이 알프레드에 비해 시간이 더 걸렸다. 이는 본 시스템이 두 개의 온톨로지를 사용함에 따라 부가적인 처리를 하는 시간에 기인한 것이다.

4. 결 론

기존의 웹 페이지 자동 입력 도구는 구문 기반으로 동작하기 때문에 웹 페이지의 해당 필드에 요구되는 정보를 정확하게 매치하지 못하는 경우가 발생한다. 본 논문에서는 이를 개선하여 의미 기반의 자동 입력 도구를 설계하였다. 의미 기반의 필드명 추출을 위해, 유사어 온톨로지를 사용하여 추출된 정보의 유사어 관계와 상관 관계를 정의하였

다. 구문 중심의 매칭은 저장된 정보에 한하여 수행되기 때문에 다른 웹 페이지를 사용하게 되면 그에 따라 추가적인 정보가 필요하다. 이에 반해 의미 중심의 매칭을 사용하면 지정된 정보 이외에도 그와 의미가 같은 정보에 대해서도 추가적인 정보 없이 처리 가능하게 된다. 의미 기반의 필드명 추출은 유사어 온톨로지와 개인 데이터 온톨로지를 사용함으로써 구현되었다. 본 논문에서는 온톨로지가 단순히 유사어를 매치시키는데 사용될 뿐만 아니라 추론을 통하여 적절한 필드 값을 "생성"하는데 사용됨을 보였다. 개인 데이터 온톨로지의 계층 관계와 필드와 필드의 관계를 이용하여 목표 웹 페이지가 원하는 필드 정보를 정확하게 생성할 수있었다. 실험 결과 본 시스템이 기존의 사용자 정보 채움 도구에 비해 우수한 정확도를 보였다. 테스트 웹 페이지에 대해 정상적으로 채워진 필드와 오류 필드를 분석한 결과 온톨로지의 사용이 매우 효과적임을 확인하였다.

본 시스템은 또한 회원 가입 웹 페이지뿐만 아니라 이력서 정보를채워야 하는 웹 페이지나 문서 편집기에서 흔히 발견되는 이력서 양식 등 반복적으로 동일한 정보를 채우는 경우로 쉽게 확장될 수 있다.

참 고 문 헌

[1] 마이크로스프트 패스포트, http://www. passport.net

[2] 알프레드, http://www.alfred.to/

[3] eCARD 솔루션, http://www.cyber-card.co.kr.

[4] Claire Cardie, "Empirical Methods in Information Extraction," AI Magazine, Vol.18, No.4, 1997.

[5] 엄재홍, "은닉 마르코프 모델을 이용한 정보추출", 제5회 한국 과학기술 정보인프라 워크샵 학술발표 논문집, pp. 132-146, 2000.

[6] 김재훈, "정보추출의 기술 현황", 정보과학회지, 제 22권 제 4호, pp. 35-46, 2004.

[7] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web", Scientific American, 2001.

[8] D. Fensel, F. van Hamelen, I. Horrocks, D. L. Mcguinness, and P. F. Patel-Schneider, "An Ontology Infrastructure for the Semantic Web", 2001.

[9] A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, and Y. Warke, "Managing Semantic Content for the Web", 2002.

[10] 최중민, "시맨틱 웹의 개요와 연구동향", 정보과학회지, 제 21권, 제3호, pp.4-10, 2003.

[11] M.A. Visciola, "Search types and context of use in the semantic Web," 2003.

[12] Ning Zhang, Hong Chen, Yu Wang, Shi-Jun Cheng, and Ming-Feng Xiong, "ODAIES: ontology-driven adaptive Web information extraction system", 2003.

[13] C. Fellbaum, "WordNet: An Electronic Lexical Databasc", MIT Press, 1998.

[14] http://www.cogsci.princeton.edu/~wn

[15] H. Rodriguez, S. Climent, P. Vossen, L. Bloksma, W. Peters, A. Alonge, F. Bertagna, and A. Roventint, "The top down strategy for building EuroWordNet: Vocabulary coverage, base concepts and top ontology," Comput. Humanities, vol. 32, pp. 117-159, 1998.

[16] G. Leroy and H. Chen, "Meeting medical terminology needs-the ontology-enhanced Medical Concept Mapper," IEEE Transactions on Information Technology in Biomdeicine, vol. 5, Issue 4, pp. 261-270, 2001.

[17] A. T. McCray, S. Srinivasan, and A. C. Browne, "Lexical methods for managing variation in biomedical terminologies," in Proc. Annual Comput. Applicat. Med. Care Symp., pp. 235-239, 1994.

민 영 근

2005년 단국대학교 전기전자컴퓨터 공학과 (학사)

2005년~현재 단국대학교 대학원 전자컴퓨터공학 석사과정

관심분야 : 시맨틱 웹, 기계 학습

김 인 수

2004년 단국대학교 전기전자컴퓨터 공학과 (학사)

2006년 단국대학교 대학원 전자컴퓨터 공학 (석사)

2006년~현재 서호전기 시스템연구부 연구원

관심분야 : 시맨틱 웹, 센서제어, 지능형로봇

이 복 주

1986년 서울대학교 컴퓨터공학과 (학사)

1992년 University of South Carolina 컴퓨터학과 (석사)

1996년 Texas A&M University 컴퓨터학과 (박사)

1997년~1999년 AT&T

2000년~2001년 한국정보통신대학교 (ICU) 조교수

2001년~현재 단국대학교 전자컴퓨터공학부 조교수

관심분야 : 기계 학습, 데이터마이닝, 시맨틱 웹