온톨로지를 이용한 정보 추출

김인수, 이복주
단국대학교 전자 컴퓨터 공학과
insu@ai.dankook.ac.kr , blee@dankook.ac.kr

Information Extraction Using the Ontology

Insu Kim and Bogju Lee
Department of Computer Engineering, Dankook University, Korea

요 약

정보 추출은 텍스트로 되어 있는 비 정형화된 데이터로부터 정형화된 정보를 추출하는 분야이다. 기존의 정보 추출이 구문 중심의 방법인데 비해 본 논문에서는 시맨틱 웹과 온톨로지를 이용한 의미 기반의 정보 추출을 시도한다. 또한 본 논문에서는 기존의 정보 추출 모델을 분류해 보고 반자동 정보 추출이라는 새로운 모델을 제시한다. 이 모델에 기반하여 개인 정보를 자동으로 정형화 시켜주는 정보 추출 도구를 개발하고 이를 소개한다.

1. 서론

회원 가입이나 전자상거래시 배송지 정보와 같은 개인 신상 정보를 입력해야 하는 페이지 또는 이력서나 다른 일반 문서를 작성할 때 매번 똑같은 개인 신상 정보를 입력해야 하는 번거로움이 있다. 이러한 개인 신상 정보를 자동으로 채워지게 하여 간편하게 만들면 편하게 될 것이다. 여기에 정보 추출(Information Extraction)을 이용할 수 있다. 정보추출은 텍스트로 되어 있는 비 정형화된 데이터(unstructured data)로부터 정형화된 데이터(structured data)를 추출하는 분야이다. 기존의 정보 추출이 구문 중심의 방법인데 비해 본 논문에서는 시맨틱 웹과 온톨로지를 이용한 의미 기반의 정보 추출을 시도한다.

기존의 정보 추출 모델은 두 가지로 분류할 수 있다. 즉, 수동 정보 추출과 자동 정보 추출이 그것이다. 수동 정보 추출은 텍스트로부터 사람이 필드와 필드 값을 인지하여 추출하는 방식이고 자동 정보 추출은 그 것을 전부 컴퓨터 프로그램에 맡기는 것이다. 지금까지의 연구 성과로도 완전한 자동 정보 추출은 달성하지 못한다. 본 논문에서는 이 두 가지 정보 추출 사이에 존재하는 반자동 정보 추출이라는 새로운 모델을 제시한다. 즉 프로그램으로 하여금 일단 정보 추출을 하게 한 다음 사용자가 확인하고 수정하게 하는 방식이다. 현실적으로 이 방식이 완전 자동 정보 추출보다 훨씬 유용하다.

본 논문에서는 이 모델에 기반한 정보 추출 도구를 개발하고 이를 실험하였다. 정보 추출 도구의 개발 도메인은 웹 사이트에서 흔히 사용하는 사용자의 인적 정보에 관한 것이다. 상업적 웹 사이트 또는 일반 웹 사이트에 회원 가입할 때 사용자는 흔히 신상 정보를 웹에 입력하게 된다. 또는 이력서나 다른 일반 문서를 작성할 때도 테이블을 신상 정보로 채우게 된다. 이럴 때 텍스트 기반으로 되어 있는 만들어진 원래 신상 정보 파일이 있고 이로부터 자동적으로 정보 추출이 되어 해당 목표 페이지나 테이블에 " 로드" 된다면 매우 편리할 것이다. 로드 된 후 사용자는 혹시 잘못 옮겨진 필드를 수정하면 될 것이다.

2. 기존 연구

현재 인터넷에는 수많은 웹사이트가 존재하고 이들 중 대부분은 사용자 확인을 위해 회원가입을 요구한다. 그래서 사용자는 수많은 웹사이트를 일일이 가입하고 각 사이트의 계정 정보를 관리해야 하는 번거로움이 생긴다. 이러한 문제를 해결하기 위해 마이크로소프트에서는 닷넷 패스포트를 제안하였다.

닷넷 패스포트는 http://www.passport.net 에서 등록하여 사용할 수 있다. 닷넷 패스포트에 등록된 하나의 계정으로 MSN메신저와 닷넷 패스포트를 채택한 사이트의 로그인을 할 수 있다. 윈도우에서는 사용자 별로 닷넷 패스포트를 저장하여 사용 가능하다. 또 MSN메신저에 로그인한 상태라면 메신저의 계정 정보로 웹사이트의 로그인이 이루어 진다.

MSN메신저는 기본적으로 닷넷 패스워드를 계정으로 사용한다. 웹사이트의 경우, 로그인 방식을 닷넷 패스포트로 채택한 사이트에 한하여 간단히 로그인 버튼을 클릭하는 것 만으로 자동으로 로그인이 된다. 그러나 닷넷 패스포트를 채택한 웹사이트는 MSN홈페이지(http://www.msn.co.kr) 이외에는 찾아보기 힘들다. 데브피아(http://www.devpia.com) 는 닷넷 패스포트 로그인을 서비스하였지만 현재 중단한 상태이다.

이러한 방식은 회원관리부분을 닷넷 패스포트에 맡김으로써 편리하지만, 모든 웹사이트가 이 방식을 채택하기란 사실상 힘들다. 수많은 웹사이트에 독립적으로 사용자의 정보를 관리하여 로그인 폼이나 회원가입 폼을 채워주는 프로그램이 좀 더 현실적인 방법이 될 것이다.

사용자 정보를 관리해 주는 프로그램 중 하나로 '알프레드'가 있다. '알프레드'는 각 사이트 마다 로그인시 미리 저장된 아이디와 패스워드를 이용해 자동으로 로그인하는 기능과 회원 가입시 사용자 정보를 채워주는 기능을 한다. '알프레드'를 처음 시작하면 개인정보를 입력하게 되고, 이 정보를 회원 가입시 이용하게 된다. 웹브라우저에 회원 가입 폼이 나타나게 되면 '알프레드'는 이를 감지하여 저장된 사용자 정보를 입력할 것인지 묻는 팝업창을 띄우게 된다.

사용자 정보를 입력한 후에도 그림 1 과 같이 '휴대폰 번호' 필드는 채워지지 않은 것을 확인 할 수 있다. 이는 각 필드의 명칭이 정확히 매치가 되지 않아서, 요구하는 정보가 있음에도 불구하고 필요한 곳에 입력되지 않은 것이다.

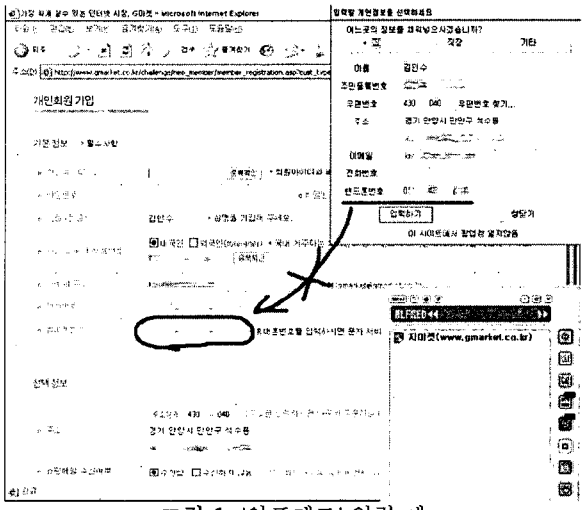

그림 1. '알프레드' 입력 예

이를 해결하기 위해서는 각 사이트가 요구하는 정보의 필드명과 '알프레드'에 저장되어 있는 정보의 필드명이 정확히 매치되도록 추가적인 정보가 필요하다. 각 사이트 마다 회원가입페이지의 폼이 다르게 작성되어 있기 때문에, 추가적인 정보의 양은 사이트의 수만큼 필요하게 된다. 또 새로운 사이트에 대해서도 적용될 수 있도록 지속적인 업데이트가 요구되어 해결방법으로는 적합하지 않을 수 있다.

이러한 상황에서 온톨로지를 사용하여 요구되는 필드명이 어떤 필드와 일치하는지 추론하여 매치 시킨다면 추가적인 정보 없이 매치 가능할 것이다.

3. 시스템 소개 및 구조

본 논문에서 소개할 시스템은 미리 사용자의 신상 정보를 가지고 있는 파일이 있고, 정보 입력 페이지의 field name에 따라 해당되는 정보를 파일로부터 가져와서 채워주게 된다. 이때 서로 다른 field name을 가지고 있지만 같은 정보를 요구하는 field의 경우, 서로 다른 field name 이 같은 field 라는 것을 알려주어 해당 정보를 채워야 한다. 여기서 Ontology를 사용하여 field가 같다는 것을 명시해준다. 예를 들어 '성명'과 '이름'은 같은 의미를 갖는 field 이므로 같이 취급해야 한다.

Ontology 를 사용하여 채워진 정보는 정확하지 않은, 관계없는 정보가 채워질 수 있기 때문에 사용자가 채워진 정보를 확인하고 수정한 후 Submit 하는 반자동 시스템으로 한다. 본 논문에서는 이 시스템을 Personal Information Manager라고 명칭 하기로 한다.

다음 그림 2 는 Personal Information Manager 의 전체적인 구조를 나타낸 그림이다.

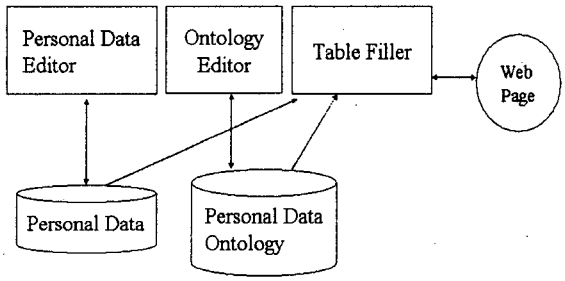

그림 2. Personal Information Manager 전체구조

각 구성요소를 살펴보면 크게 Personal Data Editor, Ontology Editor 그리고 Table Filler로 구성되어 있다.

Personal Data Editor 는 사용자의 신상정보를 입력 및 수정하여 Personal Data를 작성하는 역할을 한다. Personal Data 는 사용자의 컴퓨터 내에 저장되어있다가 필요에 따라 사용되게 된다.

Ontology Editor 는 Personal Data Ontology를 관리하기 위한 목적으로 사용된다. Personal Data Ontology 에는 사용자의 신상정보에 대한 각 필드의 유사어의 정보와 각 필드간의 관계 정보가 저장되어 있다. 그림 3 은 Ontology 내의 Field Name 간의 관계를 트리 구조로 보여주고 있다.

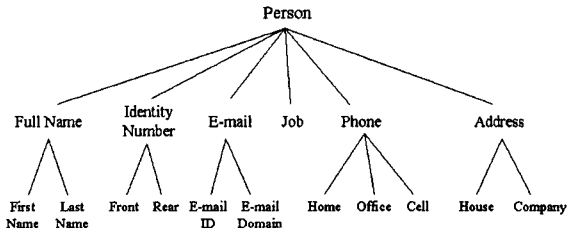

그림 3. Personal Data Ontology

이 정보를 이용하여 웹페이지에서 요구하는 사용자의 신상정보가 무엇인지 의미 기반으로 파악하게 된다.

Web Page 에서 사용자의 신상정보가 필요한 경우, Table Filler 는 해당 페이지에서 요구하는 정보가 무엇인지 추출해 온다. 그리고 Ontology를 사용하여 추출된 Field 정보가 어떤 Filed에 해당하는지 알아낸 후, 저장되어 있는 사용자 정보를 해당 Field 에 채워주게 된다.

4. Table Filler 의 내부 구조

전체 구성요소 중에서 Table Filler 는 다음 그림 4 와 같은 내부 구조로 이루어져 있다.

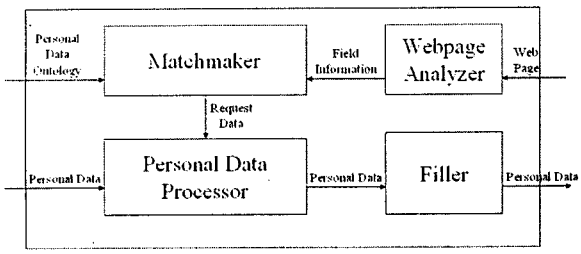

그림 4. Table Filler 내부 구조

웹페이지가 요구하는 정보를 분석하는 Webpage Analyzer, 웹페이지에서 추출된 Field Information을 온톨로지를 사용하여 필요한 정보가 무엇인지 파악하는 Matchmaker, 필요한 정보에 어떤 데이터를 넣을 것인지 선택하는 Personal Data Processor, 그리고 웹페이지로 해당 정보를 채워주는 Filler로 구성된다.

사용자가 웹페이지에서 요구하는 정보를 채우기 원하면, Webpage Analyzer는 입력 양식의 HTML문서를 분석한다. 다음은 입력 양식 HTML의 예이다.

```
<TD width=100><FONT color=#808080>?</FONT> 비밀번호 </TD>
<TD><INPUT type=password maxLength=12 size=12 value="" name=PWD><IMG
src="../images/required.gif">      비밀번호 확인 <INPUT
type=password maxLength=12 size=12 value="" name=PWD_tmp><IMG
src="../images/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 </TD>
<TD width=188><INPUT maxLength=20 name=Zip_phone></TD></TR>
<TR class=font9>
<TD width=100>직장 </TD>
<TD width=188><INPUT maxLength=20 name=Offi_phone></TD></TR>
<TR class=font9>
<TD width=100>이동통신 </TD>
<TD width=188><INPUT maxLength=20
name=Cell_phone></TD></TR></TBODY></TABLE></TD></TR>
```

그림 5. 입력 HTML의 예

정보를 입력 받는 곳은 <INPUT>태그로 되어 있고 name 속성으로 구분된다. 각각의 입력 폼이 요구하는 데이터는 name 속성과 <INPUT>태그의 전후 문맥에서 추출된 Field 정보를 이용하여 분석 가능하다. 그림 5 의 HTML 구문 중에서 직장 전화번호의 경우를 예로 들어보면, <INPUT>태그의 name속성값은 ' offi_phone' 이고 바로 앞의 텍스트가 ' 직장' 이라는 것을 알 수 있다. ' 직장' 이라는 정보로는 어떤 정보를 요구하는지 추론하기 어렵기 때문에 테이블 내의 다른 정보가 더 필요하다. Name속성값으로 추론할 수도 있지만 name속성값은 웹프로그래머가 임의로 정하기 때문에 추론에 부적합한 경우가 생길 수 있다. 따라서 테이블 전체를 분석하여 ' 전화 번호' 라는 정보를 추가로 추출해 낸다. Webpage Analyzer 에 의해 추출된 정보는 Matchmaker로 보내지게 되고, 온톨로지를 사용하여 필요한 해당 정보가 무엇인지 파악한다. ' offi_phone' ,' 직장' ,' 전화 번호' 라는 정보는 온톨로지의 유사어와 Field 관계를 사용하여 온톨리지 내의 Phone-Office에 해당하는 의미로 해석 가능하다. 해석된 정보는 Personal Data Processor로 보내지고 Personal Data로부터 해당 정보를 읽어 들이게 된다. 만약 웹페이지에서 ' 전화번호' 라는 의미의 정보만을 요구하였다면, Personal Data Processor는 온톨로지의 Phone에 해당하는 Home, Office, Cell 중에 하나의 정보를 선택하도록 사용자에게 질의하게 된다. 이와 같이 읽어온 정보는 각<INPUT>태그의 name속성값과 함께 Filler로 전송되어, 해당 입력란에 사용자의 신상정보를 채워주게 된다.

5. 결론 및 향후 과제

기존의 자동 입력툴이 Syntax 기반으로는 자동 입력되지 않는 문제를 개선하기 위해, 본 논문에서는 온톨로지를 사용한 의미 기반의 정보 추출을 제안 하였다.

향후 웹페이지의 HTML문서를 분석하여 의미를 추출하기 위해 XML문서로의 문서변환이 필요하다. 또 웹페이지마다 입력양식이 다양하기 때문에 드롭다운방식이나 검색을 통한 입력을 요구하는 경우에도 사용 가능하도록 방안이 필요할 것이다.

Personal Information Manager는 웹브라우저에 내장되는 형식으로 간단하고 편리한 사용자 인터페이스로 구현될 예정이다.

6. 참고 문헌

1. http://www.alfred.to/
2. http://www.cyber-card.co.kr
3. Alani, H., Sanghee Kim, Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H., Shadbolt, N.R. : Automatic ontology-based knowledge extraction from Web documents(2003)
4. Sheth, A., Bertram, C., Avant, D., Hammond, B.,Kochut, K., Warke, Y. : Managing semantic content for the Web(2002)
5. Visciola, M.A. : Search types and context of use in the semantic Web(2003)
6. Ning Zhang, Hong Chen, Yu Wang, Shi-Jun Cheng, Ming-Feng Xiong : Odaies: ontology-driven adaptive Web information extraction system(2003)
7. http://www.passport.net
8. 김재훈, " 정보추출의 기술 현황" , 정보과학회지. 22(4) : 35-46, 2004.