

대화형 에이전트의 적응적 웹 지식구조 구축을 위한 Semi-structure 기반 정보 추출

김경민*, 홍진혁, 조성배

*연세대학교 대학원 인지과학 협동과정

연세대학교 컴퓨터과학과

{*kminkim, hjinh}@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Semi-structure Information Extraction for the Construction of Adaptive Web Knowledge Base of Conversational Agents

Kyoung-Min Kim*, Jin-Hyuk Hong, Sung-Bae Cho

*Graduate Program in Cognitive Science, Yonsei University

Dept. of Computer Science, Yonsei University

{*kminkim, hjinh}@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

요 약

인터넷의 성장으로 웹 사이트를 통해 다양한 정보를 제공하고 있다. 현재의 웹은 사용자가 사이트에 접속하고 원하는 목적지에 이르기까지 링크를 따라 검색해야 한다. 따라서 웹 사이트의 구조를 제대로 이해하고 있지 않으면 정보 획득에 많은 어려움이 따른다. 특히 사이트의 구조가 복잡할 경우 정보 검색은 더욱 어려운 작업이 된다. 이런 한계를 극복하기 위해서 검색엔진 등의 다양한 기법이 적용되고 있으며, 사용자와의 대화를 통해 원하는 정보를 쉽고 친숙하게 제공하는 대화형 에이전트의 연구가 최근 활발히 진행되고 있다. 대화형 에이전트가 알맞은 답변을 제공하기 위해서는 지식구조를 사전에 구축해야 하는데, 대부분의 대화형 에이전트는 사용자 질의에 대해 수동으로 답변 지식구조를 구축해 왔다. 이런 수동 구축 방식은 비용이 많이 들고, 지속적으로 업데이트해야 하는 한계가 있다. 본 논문에서는 웹 사이트를 소개하는 대화형 에이전트의 지식구조 구축을 위해 랩퍼(Wrapper) 기법을 이용하여 사이트의 이질적인 구조 정보를 템플릿 형식으로 구성하고, 온톨로지(Ontology)를 적용하여 매칭시의 의미 정보를 활용함으로써 사용자가 원하는 정보를 웹으로부터 자동 추출하는 방법을 제안한다. 실제 상품 판매 사이트를 대상으로 테스트를 실시하여 제안한 방법의 가능성을 확인한다.

1. 서론

인터넷의 발달로 웹을 이용한 정보 제공 서비스가 확대되고 있다. 사용자가 원하는 정보를 보다 쉽고 빠르게 제공하기 위해 검색엔진이나 대화형 에이전트 등이 연구되고 있다[1]. 많은 연구에서 이미 정보를 교환하고 사용자 의도를 파악하는 수단으로 대화가 매우 효과적이라고 알려져 있다. 대화를 이용한 정보 제공 에이전트는 사용자 질의에 대해 답변 스크립트로 구성되는 지식구조를 사전에 구축해야 한다. 이는 매우 비용이 많이 드는 작업일 뿐만 아니라 지속적으로 갱신되는 웹 사이트의 정보를 효과적으로 처리하

기 어렵다.

따라서 본 논문에서는 빈번히 내용이 업데이트 되는 웹 사이트에 대해 지속적으로 지식구조가 갱신되어 적절한 답변이 제공되도록 답변 지식을 웹으로부터 자동으로 추출하는 방법을 제안한다. 일반적으로 웹으로부터의 정보 추출은 한 문서에서 그 문서의 중심적 의미를 나타내는 특정 구성 요소를 인식하여 자동으로 추출하는 작업을 말한다[2,3]. 대부분의 웹 문서는 이질적이고 복잡한 실세계 정보 소스들로 구성되어 있다.

본 논문에서는 비정형적 웹 문서에서 semi-structural 매칭 기술을 이용해 정보를 자동으로

추출하는 방법을 제안한다. 추출하고자 하는 정보에 대한 웹 페이지에서의 구조를 템플릿의 형태로 구성하고, 랩퍼를 적용하여 웹 페이지의 이질적인 구조 정보를 분석한다. 원하는 정보를 추출하여 XML 기반의 문서 정보를 생성하고, 이에 온톨로지를 고려하여 도메인 지식의 의미 정보를 부여한다. 따라서 지식구조를 설계함에 있어 사전에 수동으로 정의한 정보가 아닌 웹으로부터 자동으로 추출된 정보를 이용하기 때문에 사용자의 질의에 대한 동적 답변 생성이 가능하다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 기존 대화형 에이전트 시스템에 대한 소개 및 웹 정보 추출에 있어 랩퍼와 온톨로지의 역할에 대한 개략적인 내용을 살펴본다. 3장에서는 본 연구에서 제안하는 구조적, 의미적 정보 추출 및 매칭 방법에 대해 구체적으로 기술한다. 4장에서는 시스템의 성능 평가를 위한 실험 및 결과를 분석하고, 마지막으로 5장에서는 제안하는 방법의 결론 및 향후 연구 과제에 대해 언급한다.

2. 관련 연구

2.1 대화형 에이전트

대화형 에이전트는 사용자와 에이전트 사이에서 인간이 사용하는 자연어 대화를 통해 정보를 주고받는 시스템이다[4]. 대화형 에이전트 시스템은 사용자가 원하는 정보를 단순한 키워드나 메뉴 등의 방식으로 습득하는 것이 아니라 자연어로 된 문장을 입력하게 함으로써, 사용자에게 보다 친숙한 인터페이스를 제공한다. 표 1은 대부분의 대화형 에이전트에서 사용되는 스크립트 기반의 지식 구축 방법을 보여준다.

표 1. 대화형 에이전트의 지식표현 방법의 예

<pre> <alice>::=<category><category>* <category>::=<question><answer>* <answer>* <question>::="<question>"<질문패턴>"</question>" <answer>::="<answer>"<답변패턴>"</answer>" <질문패턴>::=<한글> * <한글> <답변패턴>::=<한글> * <한글> </pre>
<pre> <alice> <category> <question> 질문형_문장 단위로 입력 </question> <answer> 답변형_문장 단위로 입력 </answer> </category> </alice> </pre>

기존의 대화 시스템에서 사용된 방법들은 단순 스크립트 기반으로 미리 정의된 질의와 답변 쌍을 매칭시켜 출력으로 내보내기 때문에 답변 데이터베이스 구축 시 많은 노력이 소모된다. 또한 한번 구축된 답변 스크립트는 수정이 쉽지 않다. 자주 내용이 갱신되는 웹 사이트에서 이러한 방식의 대화형 에이전트는 동적인 지식구조를 가지지 않기 때문에 적용하기 어렵다.

2.2 랩퍼

인터넷의 웹 문서들은 HTML, CGI, 서블릿 등 각기 다른 형식으로 구성되어 있고, 비정형적이며 불완전한 데이터 구조를 갖는다. 따라서 웹 문서의 내용을 추출하는데 어려움이 있다. 이를 해결하기 위해 이질적인 웹 문서를 동일한 형식으로 변환시키는 랩퍼에 대한 연구가 활발히 진행되고 있다[5]. 랩퍼는 다른 형식의 문서에 대해 다른 시스템에서 동일하게 인식되도록 동형의 웹 문서로 변환하는 기법이다. 각 시스템별로 따로 처리해야 하는 과정을 하나의 데이터 처리로 묶는 장점이 있으며, 시스템에 포함되지 않은 서버 연결 시 랩퍼를 추가하여 시스템 전체의 처리 기능을 강화할 수 있다. 주로 HTML 문서를 구조 및 의미 정보를 포함하는 XML 문서로 변환하는 연구에서 많이 활용되고 있다.

랩퍼를 이용한 정보 추출 에이전트로 Florid 시스템, Lore 시스템, Strudel 시스템 등이 있으며[6], 이들은 주로 규칙 기반이나 정규화 방법을 이용하여 대상 문서의 지역 모델로부터 원하는 데이터를 추출한다. 이 때, 사용자가 랩퍼의 처리 규칙을 설계해야 할 필요가 있다.

2.3 온톨로지

온톨로지는 공유 가능한 개념화의 명세로서 [7], 급속히 발전하는 웹의 정보를 공유해 다양한 분야에서 활용하기 위해 많이 연구되고 있다. 다양한 환경에서 도메인의 지식을 공유하고 재사용하기 위해 온톨로지라는 개념이 도입되었으며, 개념과 관계로 모든 지식을 표현하는 방식으로 현재 정보검색, 의료정보와 바이오 정보, 전자상거래, 지능형 에이전트 등 다양한 분야에 적용되고 있다.

특히 웹을 통한 정보검색 및 추출 등의 문제에서 어휘적 분석만으로는 파악하기 어려운 문서의 의미를 보다 정확히 분석하기 위해 도메인 지식

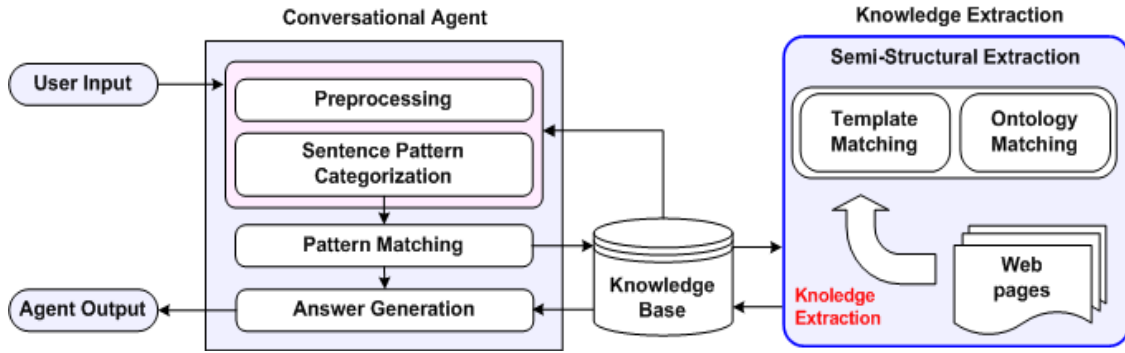


그림 1. 제안하는 대화형 에이전트 구조

의 표현 방법으로 온톨로지의 연구가 활발히 진행되고 있다[8]. HTML 형태의 문서들로 이루어진 현재의 웹은 사용자에게 정보를 제공하는 역할은 하고 있지만, 이들 정보가 의미적, 구조적으로 구성되어 있지 않기 때문에 시스템이 각 문서의 내용을 정확히 이해할 수 없다. 문서에서 특정 정보를 검색하기 위해서는 각 내용에 대한 의미와 구조 정보가 함께 명시되어 있어야 한다.

3. 제안하는 방법

본 논문에서 제안하는 대화형 에이전트 시스템의 전체적인 구조는 그림 1과 같이 대화형 에이전트[9,10]와 지식 추출 부분으로 구분된다.

3.1 대화형 에이전트 시스템

그림 1에서와 같이 대화형 에이전트는 상호작용을 통해 사용자가 원하는 정보를 제공한다[11]. 정적인 답변 스크립트를 기반으로 동작하는 기존의 대화형 에이전트와는 달리 지식구조가 동적으로 갱신되기 때문에, 답변도 동적으로 생성되어야 한다.

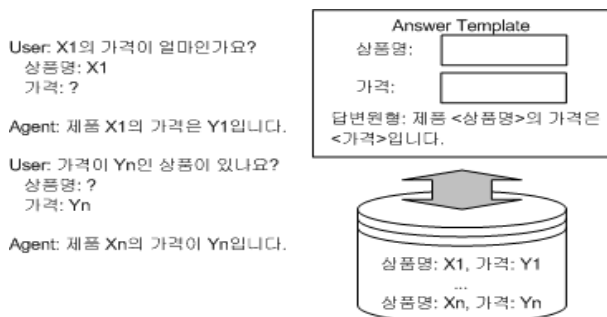


그림 2. 사용자 질의에 대한 동적 답변 생성

본 논문에서는 답변 템플릿을 구성하여 사용자가 요구하는 정보를 답변 템플릿에 삽입하여 사

용자에게 제공한다. 답변 템플릿에는 대상 영역에서 얻어진 온톨로지 정보와 답변 제공을 위한 답변 원형이 포함된다. 추출된 온톨로지 정보가 “제품명”과 “가격”이라면 템플릿은 제품명과 가격을 변수로 가지며, 답변원형으로 “제품 <제품명>의 가격은 <가격>입니다.”를 포함한다. 그림 2는 사용자 질의에 대해 답변이 동적으로 생성되는 모습을 보여준다.

3.2 의미 정보를 고려한 구조 매칭

웹 문서로 HTML 문서를 주로 사용하는데, 이는 HTML 문서가 쉽게 작성 가능하다는 장점 때문이다[12]. 그러나 HTML은 문서의 내용을 충분히 묘사하기 어렵고, 구조적으로 모호하기 때문에 태그를 추출하기 위한 파싱 작업을 수행하는데 어려움이 있다[13]. XML은 HTML과 비교해 구조적으로 명확하게 정의된 확장 가능한 웹 문서 정의 언어이다[14]. 또한 서로 다른 소스로부터 추출된 데이터를 비슷하게 통합시켜 구조화된 데이터를 사용자 인터페이스와 분리시킬 수 있다는 강점이 있다. 이질적이고 복잡한 웹 페이지로부터 원하는 정보를 정확히 추출하기 위해 본 논문에서는 HTML 문서에서 추출하고자 하는 정보의 개체 및 속성값에 대한 구조를 템플릿 형태로 설계하고, XML 형식의 지식 구조로 변환하는 랩퍼를 생성한다.

그림 3은 추출하고자 하는 정보의 의미와 구조를 표현한 랩퍼의 구조를 보여준다. 구조 매칭 과정은 먼저 대상 웹 페이지 내의 콘텐츠를 수집한다. 이 과정에서 불필요한 HTML 태그들을 제거하는 작업이 동시에 수행되며, 수집된 콘텐츠 중에서 대상 정보에 해당하는 콘텐츠의 패턴(태그의 헤더, 테일)을 분석하여 필요 데이터만을 추

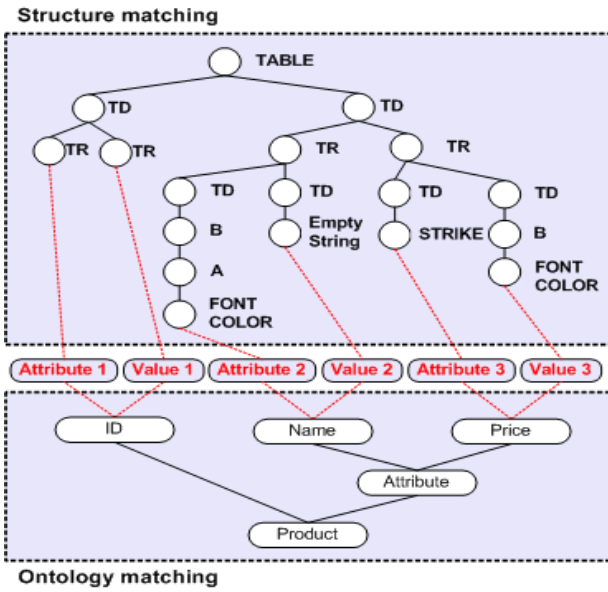


그림 3. 온톨로지와 구조 정보를 이용한 랩퍼 구조

출한다. 추출된 데이터는 XML과 같이 의미 정보를 갖는 문서 파일로 저장된다. 이때 저장되는 문서에 표현되는 XML 태그들의 정보는 설계자에 의해 대상 도메인에 따라 템플릿 형식으로 정의된 개체 및 속성 값을 적용한 것이다. 의미 정보를 고려하는 온톨로지 매칭은 얻어진 구조에 의미 정보를 포함하여 수행된다. 얻어진 변수와 값들에 각각 의미를 부여하여 매칭 시 정의된 의미에서 벗어나는 정보는 제거한다. 이렇게 의미와 구조를 동시에 고려하면 보다 정확한 정보를 추출할 수 있게 된다.

3.3 의미 정보를 이용한 지식 추출

본 논문에서는 semi-structural 웹 사이트로부터 구조 정보와 의미 정보의 매칭을 통해 원하는 지식 정보를 추출한다. 추출된 지식 정보는 XML 형식의 문서로 변환된다. 시스템의 처리 과정은 표 2와 같다.

우선 웹 페이지내의 구조 정보 분석을 통해 얻어진 개체와 의미 정보 표현의 온톨로지 개체를 매칭하는 작업을 수행한다. 다음으로 매칭된 개체에 해당 속성값을 적용하여 XML 형식의 스크립트 문서를 생성한다.

4. 실험 및 결과

4.1 구조와 의미 정보를 통한 답변 지식 생성

랩퍼를 이용한 구조 정보 추출과 온톨로지를

적용한 의미 정보 매칭으로 웹으로부터 자동적으로 필요 정보를 추출함으로써 사용자 입력 질의에 해당 답변 지식을 구축한다. 실제 제안한 방법의 성능을 평가하기 위해 상품 판매 사이트를 대상으로 실험을 수행하였다.

표 2. 지식 추출의 처리 과정

```

[Domain] Attribute:  $A = \{a_1, a_2, a_3, \dots, a_n\}$ 
Value :  $V = \{v_1, v_2, v_3, \dots, v_n\}$ 
Element :  $E = \{e_1, e_2, e_3, \dots, e_m\}$ 

While (no more  $e_i$ )
{
  if (structure_match( $e_i$ )==true)
  {
    V = ontology_match(A,  $e_i$ );
    generate_script( $a_r, v_r, 0$ );
  } //  $a_r, v_r$ : root node
}
generate_script( $a_i, v_i, depth$ )
{
  if ( $depth$ ==leaf_node depth)
  {
    print_script(< $a_i$ >);
    print_value( $v_i$ );
    print_script(</ $a_i$ >);
  }
  else
  {
    print_script(< $a_i$ >);
    print_value( $v_i$ );
     $depth++$ ;
    For all  $a_i$ 's child  $a_{is}$ 
      generate_script( $a_{is}, v_{is}, depth$ );
    print_script(</ $a_i$ >);
  }
}

```



Jenga Extreme
 제조사 : Hasbro / Milton Bradley
 시중가 : 39,000 원
 판매가 : 38,000 원

장바구니담기

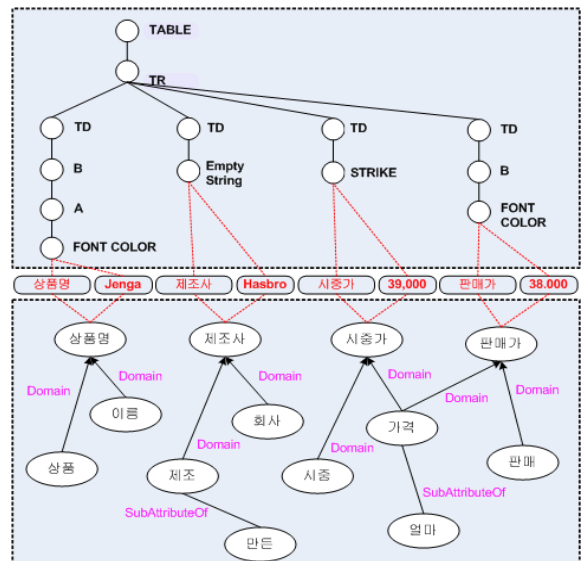


그림 4. 상품 정보 추출을 위한 랩퍼

그림 4는 상품 판매 사이트로부터 상품 정보를 추출하기 위해 설계된 랩퍼의 모습을 보여준다. 그림 5는 웹 페이지로부터 정보를 추출하여 XML 기반으로 구성된 결과를 보여준다. 구축된 지식 구조를 이용하여 대화형 에이전트에서 입력 질의와의 매칭을 수행하여 원하는 정보를 제공할 수 있다.

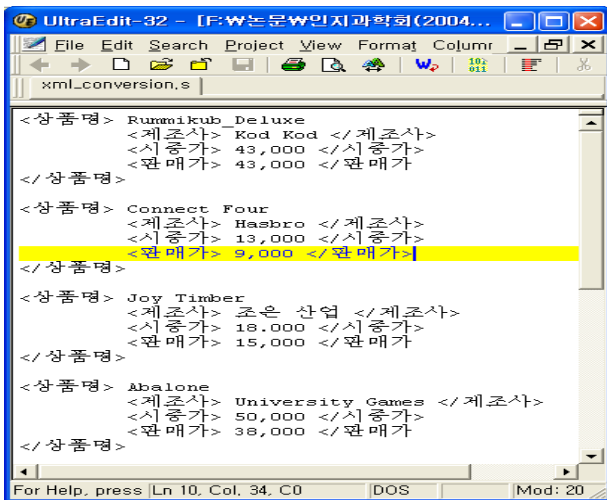
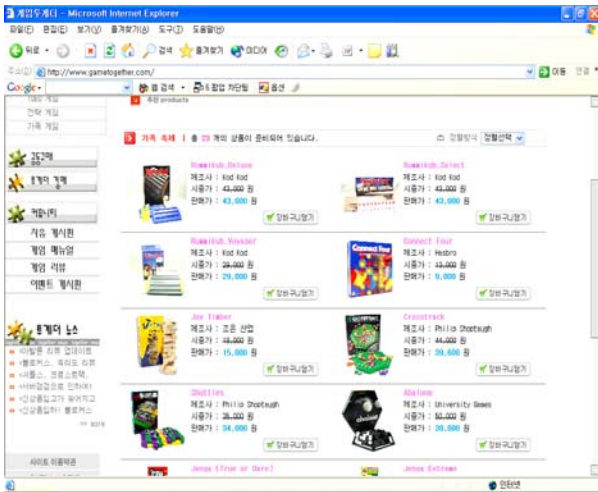


그림 5. 추출된 정보의 XML 문서로의 변환

4.2 대화형 에이전트에서의 적용

제안한 방법을 통해 구축된 지식구조를 바탕으로 대화형 에이전트의 답변 제공을 수행하였다. 앞서 언급한 과정으로 사용자 질의를 분석하여 지식구조로부터 동적으로 답변을 구성한다. 처리하는 사용자의 질의는 웹 페이지에 존재하는 제품에 대한 정보를 요구하는 것을 대상으로 하였

으며, 데이터베이스에서 정보를 검색하는 SQL 질의 형식으로 사용자 질의를 분석하고 정보를 검색하였다. 일반적인 내용에 대한 사용자 질의는 단순 패턴매칭 기술을 이용하여 답변을 선택하도록 하였다.

표 3은 사용자와 에이전트간의 대화 진행의 예를 보여주며, 그림 6은 본 논문에서 제안한 방법으로 대상 웹 페이지에서 동작하는 대화형 에이전트의 모습을 보여준다. 웹 정보와 함께 대화를 통해 정보를 제공하여 보다 나은 사용자 인터페이스를 제공한다.

표 3. 대화 수행의 예

구축된 온톨로지 변수: 상품종류, 상품명, 가격, 시종가, 제조사 등
<p>User : 보드 게임에는 어떤 것들이 있나요? 상품종류: 보드게임 상품명: ?</p> <p>Agent: Joy Timber, Crosstrack, Jenga 등이 있습니다.</p> <p>User : 2인용 게임은 어떤 것이 있나요? 상품종류: 2인용게임 상품명: ?</p> <p>Agent: Jenga, Abalon, Shuttles 등이 있습니다.</p> <p>User : Crosstrack은 얼마인가요? 상품명: Crosstrack 가격: ?</p> <p>Agent: Crosstrack의 가격은 39,600입니다.</p> <p>User : Abalon의 시종가는 얼마인가요? 상품명: Abalon 시종가: ?</p> <p>Agent: Abalon의 시종가격은 38,000입니다.</p> <p>User : Jenga는 어디에서 만들었나요? 상품명: Jenga 제조사: ?</p> <p>Agent: Jenga의 제조사는 Hasbro/Milton Bradley입니다.</p> <p>User : 여긴 판매가가 엄청 싸네요.</p> <p>Agent: 네..^^</p>

5. 결론

본 논문에서는 대화형 에이전트의 답변 지식을 웹 정보로부터 자동 추출하여 사용함으로써 기존 대부분의 대화형 에이전트에서 나타나는 지식구

조 구축의 노력을 감소하였다. 또한 온톨로지를 이용한 의미 정보 매칭으로 다양한 사용자 질의에 따른 중의성 문제 및 도메인 지식의 보다 정확한 분석을 가능하게 하였다. 의미 정보를 계층적으로 표현하기 때문에 이후 새로운 도메인을 대상으로 하더라도 확장성의 부분을 해결할 수 있다. 제안한 방법은 추출하고자 하는 정보의 기본 구조를 사람이 직접 설계하여야 하기 때문에 웹 정보 추출의 완전 자동화를 구현하기에 아직 어려움이 있다. 향후에는 웹 문서로부터 다양한 semi-structure 후보 집합을 생성하고, 이들에 대한 평가를 통해 정보를 추출하는 자동화 기법을 연구할 것이다.



그림 6. 대화형 에이전트에서의 응용 예

감사의 글

이 논문은 한국학술진흥재단(2002-005-H20002)의 연구과제에 의해 지원되었음.

참고문헌

[1] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, "Towards conversational human-computer interaction," *AI Magazine*, vol. 22, no. 4, pp. 27-38, 2001.

[2] N. Kushmerick, "Gleaning the web," *IEEE Intelligent Systems*, vol. 14, no. 2, pp. 20-22, 1999.

[3] 최중민, "인터넷 정보추출 에이전트," *정보과학회지*, vol. 18, no. 5, pp.48-53, 2000.

[4] S. Macskassy and S. Stevenson, "A conversational agent," *Master Essay, Rutgers University*, 1996.

[5] J. S. Deogun, "Structural abstractions of hypertext documents for Web-based retrieval," *Database and Expert Systems Applications*, pp. 385-390, 1998.

[6] T. Dao, "An indexing model for structured documents to support queries on content, structure, and attributes," *ADL 98. Proc. IEEE International Forum*, pp. 88-97, 1998.

[7] P. Borst, H. Akkermans, and J. Top, "Engineering ontologies," *Int. Journal of Human-Computer Studies*, vol. 46, pp. 365-406, 1997.

[8] M. Shamsfard and A. A. Barforoush, "Learning ontologies from natural language texts," *Int. Journal of Human-Computer Studies*, vol. 60, pp. 17-63, 2004.

[9] 이승익, 조성배, "A Web-based conversational agent," *정보과학회논문지*, vol. 9, no. 5, pp. 530-540, 2003.

[10] S. I. Lee and S. B. Cho, "An Intelligent agent with structured pattern matching for a virtual representative," *Intelligent Agent Technology*, pp. 305-309, Maebashi, Japan, Oct. 2001.

[11] 홍진혁, 조성배, "2단계 베이지안 네트워크를 이용한 대화형 에이전트의 문맥 관리," *정보과학회논문지*, vol. 10, no. 1, pp. 89-98, 2004.

[12] <http://www.w3.org/MarkUp>

[13] X. Gao and L. sterling, "Semi-structured data from heterogeneous sources," *Int. Workshop on Innovative Internet Information Systems(IIS'99)*, 1999.

[14] R. Khare and A. Rifkin, "XML: A door to automated web applications," *IEEE Internet Computing*, vol. 14, pp. 78-87, 1997.