특정 영역 정보 에이전트의 지식베이스 확장을 위한 웹 정보추출

조은휘, 변영태
홍익대학교 컴퓨터공학과

Web Information Extraction for Expanding Knowledge Base of a Specific Domain Information Agent

Eun-Hwi Cho , Young-Tae Byun
Dept. of Computer Science, Hongik University

요 약

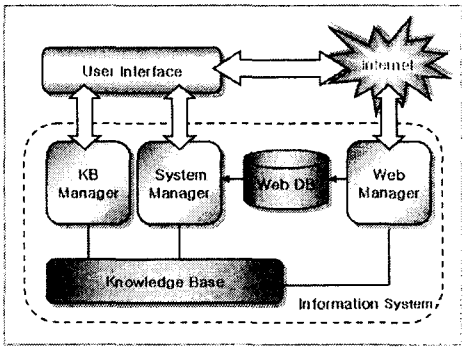
현재 연구개발 중인 웹 정보 에이전트는 Agent Manager와 KB Manager, Web Manager로 구성되어 있다. 이 시스템은 동물영역에 관련된 정보를 영어로 서비스하고 있어 국내 접근보다는 외국에서의 접근이 더 많았다. 그러므로 국내 사용을 높이기 위해 애완용 동물을 위주로 한 정보추출(IE)을 수행하여 지식베이스(KB)의 확장을 시도하고 있다. 이를 위하여 태그(tag) 및 심볼(symbol)의 패턴(pattern) 유사성 정보를 찾아내고, 기존 KB와 연계하여 KB의 확장 및 수정에 이용하기 위한 유효 정보 패턴 결정에 활용함으로써 정보 추출의 새로운 방법을 고찰하고 그 가능성을 제시하고자 한다.

1. 서론

정보 에이전트는 특정 영역에 대한 문서를 Database에 저장해 놓았다가, 사용자가 원하는 정보를 담고 있는 관련 문서를 제공해 주는 서비스를 수행한다[1].

개발 중인 시스템은 특정 영역으로 동물 영역을 선정하여 이 분야에 관련된 양질의 정보를 제공하기 위해 전문적인 정보를 지식베이스(KB; KnowledgeBase)로 구축해두고 이를 통해 사용자의 질의를 추론, 확장시켜 보다 적합한 정보를 제공해 주고 있다. 이는 Agent Manager, Web Manager, KB Manager로 구성되어 있고, 각각의 기능은 다음과 같다.

Agent Manager는 사용자들로부터 질의를 받아서 질의에 대한 정보를 제공해주는 역할을 한다. 제공하고 있는 정보는 질의의 추론에 의해 사용자가 원하는 지식을 제공하고 있고 또한 질의의 확장을 통해 웹 문서를 제공하고 있다. 질의어에 대한 확장은 사용자들이 질의를 하는 경우 추론, 확장 과정을 거쳐 문서 검색의 정확성을 높이는데 이용되어진다.



<그림1> HIIA 시스템 구성도

Web Manager는 웹을 기반한 적절한 웹 문서를 제공하기 위한 Web DB를 구축하고 유지, 관리하는 역할을 한다. 세부적인 과정들은 문서 필터링, 색인 작업, 모니터링으로 이루어진다.

본 연구는 뇌과학 연구 사업의 지원으로 진행 되었음.

KB Manager는 영역 전문가 혹은 선택된 사용자에 의해 동물 영역의 지식인 지식베이스의 구축과 유지, 관리하는 역할을 한다. 즉 사용자에게 원하는 지식이나 정보를 제공해주기 위해서 에이전트는 지식베이스를 기반으로 한다.

위에서 살펴본 정보 에이전트 시스템은 현재 영문 사이트로 운영되어지고 있다. 그러므로 국내 사용자들의 접근을 높이고 교육적 활용을 위해 한글화 작업을 하려고 한다. 이를 위해서 단순한 번역 작업을 통한 지식베이스의 구축은 그 한계가 있으므로, 기존 동물관련 사이트를 통한 지식베이스의 구축 및 확장 작업을 위해 애완동물 사이트를 위주로 하여 정보추출을 위한 방법을 고찰하고 그 가능성을 살펴보았다.

2. 관련연구

2-1. 정보추출

정보추출(IE)은 한 문서에서 그 문서의 중심적 의미를 나타내는 특정 구성요소를 인식하여 추출하는 작업을 말한다[2].

대개의 IE 모델들은 특정 영역에 관련된 정보들에서 자연어 프로세싱(NLP; Natural Language Processing)의 과정을 거쳐 유용한 값들을 추출해 내었다[3]. 관련 정보들을 인터넷의 웹 문서에서 찾아 IE 작업을 시도한 모델로는 WHISK와 SRV가 있다[4]. WHISK는 웹 문서에 Rule-based 기법을 적용한 모델이고, SRV는 Dictionary-based 기법을 사용한 정보추출 모델이다.

이 중 WHISK에 대해 살펴보겠다.

2-2. WHISK

WHISK는 Sentence 단위로 텍스트에 대한 Rule을 생성해 내는 Supervised Learning 기법을 사용한 모델이다. 기존 Free text 타입 및 Structured & Semi-structured 타입에 대해서도 적용 가능하도록 구현되었으며, 한 단어의 정보추출이 아닌 여러 단어로 이루어진 데이터도 추출 가능한 Multi-slot을 지원해준다[5].

WHISK의 전체적인 진행 순서는 다음과 같다.

- Creating hand-tagged training instances
- Creating a rule from a seed instance
- Hill climbing and horizon effects
- Pre-pruning and post-pruning

즉, WHISK는 정보를 통해 생성 가능한 의미있는(semantic) 데이터에 대한 정보를 사용자가 직접 가려내어 다수의 training instance를 생성해내고, 이들을 통해 Top-down 방식으로 가능한 많은 instrance를 cover할 수 있는 일반적인 Rule을 찾아내고, 확장시켜 나간다.

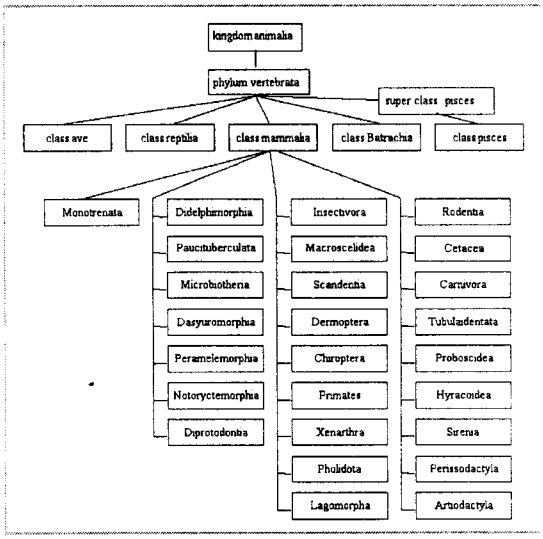
이러한 WHISK 방법을 통한 Rule의 생성은 Structured & Semi-structured 타입의 문서에 있어서는 좋은 결과를 얻어낼 수 있었다.

2-3. HIIA-1 KB

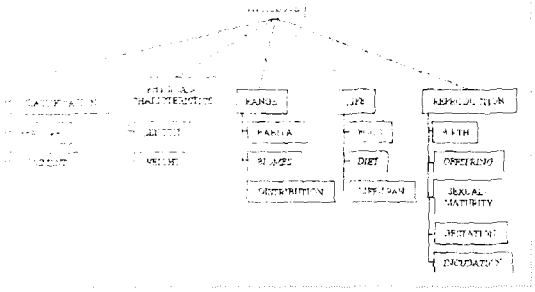
기존 HIIA 시스템의 KB는 동물에 대한 전문적인 지식으로서 동물의 계층 구조와 각 동물에 대한 속성과 속성값을 등을 가지고 있다[6].

HIIA-1은 총 1114개의 계층정보를 갖고 있으며, 포유류 데이터가 110개, 양서류 데이터가 15개, 파충류 데이터가 30개, 어류 데이터가 25개, 조류 데이터가 50개, 총 230개의 데이터가 존재한다. (<그림2-1>)

또, 객체 정보 외에 속성에 해당하는 것들에 대해서도 <그림2-2>과 같은 계층 구조를 가지고 있다. KB의 이 속성들을 이용해 문서에서의 속성을 찾아오게 된다.



<그림2-1> HIIA-1의 계층구조

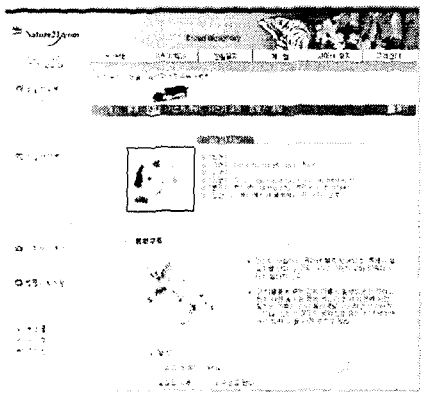


<그림2-2> 속성에 관한 계층 정보

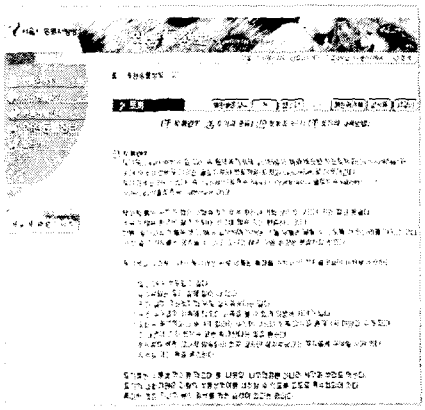
3. 실험 방법

실험을 위해 국내 사이트 중 동물과 관련한 정보를 가지고 있는 다음 두개의 사이트를 선택하였다.

[네이처21] : http://www.nature21.com/
[서울시 동물사랑방] : http://animals.seoul.go.kr/



<그림3-1> 네이처21



<그림3-2> 서울시 동물사랑방

이들 사이트들은 애완용 동물을 위주로 한 정보를 제공하고 있으며, 각 사이트마다 동물 정보를 '속성-속성값'의 연속한 형식으로 특정 형태를 갖춰 제공해주고 있다 (<그림3-1>, <그림3-2>).

'네이처21'에서는 동물관련 정보 중에서 포유류에 관한 문서 87개를 '서울시 동물사랑방'에서는 동물 정보에 관한 문서 27개를 정보추출을 위한 대상으로 선택하였다.

실험과정은 다음과 같다.

<Top-level 알고리즘>

```
AnilE(Web Site) {
    // step1:
      Pre-Processing(twp);
    // step2:
      ptn = PatternSelect(twp);
    // step3:
      InfoExtract(ptn, twp);
}
( twp; total web page in site, ptn; pattern )
```

동물관련 영역 추출 과정은 사이트 내의 모든 문서들을 가지고 이루어진다. Step1은 문서 내의 Tag와 Term을 분리하고, Term의 경우 단일 Word(영문자, 한글)나 연속된 Word 모두에 대해 '#'으로 전환되는 과정을 거친다. Word를 제외한 단어는 Symbol로 처리되며 패턴 추출 시에 패턴은 Tag, Word(#), Symbol로 구성되어진다.

```
Pre-Processing (total_web_page) {
    (1) Tag & Term을 나눈다
    (2) Term의 경우 Word와 Symbol을 구분한다.
}
```

(2)의 예)
" ... ♣ 귀를 긁거나 머리를 한 쪽으로 기울일 때 ♣ 토하거나 구역질 할 때 ... "

› " ... ♣ # ♣ # ..."

Step2는 전처리 과정을 거친 웹 문서 데이터를 가지고 패턴을 추출해 내는 단계이다.

여기서 패턴은 속성으로 유효한 Word를 찾아낼 수 있는 것을 말한다. 속성이 될 수 있는 것은 slot 5개 이하의 Word를 기준으로 하며 패턴에 의해 둘러싸여진 Word로 한정한다.

이때 기존 KB의 속성 데이터를 이용하여 반복적인 패턴과는 다르게 구분한다.

```
PatternSelect (total_web_page) {
    (1) slot 5개 이하, Word를 찾아낸다.
    (2) Word 앞뒤로 패턴을 추출한다.
        (가능한 모든 패턴을 선택하되, prev
        pattern(앞 패턴)은 닫히는 태그 </tag>
        까지 next pattern(뒤 패턴)은 열리는 태
        그 <tag>까지를 한정 범위로 한다.)
    (3) - Word가 기존 KB에 속하면 바로 패턴
        으로 선택
        - 아니면, 반복되는 패턴을 선택
    (4) return Pattern;
}
```

예)

Web Page

" ... </td> </tr> <tr> <td> 수 명 </td> <td> 5 ~ 10년 </td> </tr> <tr> <td> 먹이 섭취량 </td> <td> ... "

Pre-Processing

" ... </td> </tr> <tr> <td> # </td> <td> 5 ~ 10# </td> </tr> <tr> <td> # </td> <td> ... "

Pattern Select

Ptn1: "<tr> <td> # </td>"

Ptn2: "<tr> <td> # </td>"

› 위와 같은 문서는 하나의 pattern을 생성해내고. 이것은 수명이라는 기존 KB가 가지고 있는 속성이므로 바로 패턴으로 선택된다.

Step3은 선택된 속성 패턴을 가지고 속성-속성값의 쌍(set)을 추출해내는 단계이다.

동물 관련 웹 사이트의 구조가 '속성-속성값'이 연속적으로 제시한다는 것을 가정하고. Tag 중 구별자 역할을 하는 <table>~</table>, <hr> 태그에 주목하여 속성에 대한 속성값을 추출해 낸다.

```
InfoExtract (pattern, total_web_page) {
    (1) 웹 문서 내, 패턴에 의한 속성이 있는
        곳의 위치를 찾아낸다.
    (2) 속성이 있는 위치를 기준하여, 이후에
        나타나는 Term을 속성값으로 추출한다.
        - 이 경우 속성값은 Tag에 의해서만 식
        별되는데, <table>의 경우 </table>까지
        의 모든 Term을 속성값으로 선택하고
        <hr>의 경우 속성값의 추출을 끝낸다.
        - 이때, Symbol도 속성값이 될 수 있음
        에 주의!
}
```

4. 실험 결과 및 평가

3절에서 제시한 방법을 통해 실험한 결과는 다음과 같다. <표1-1>, <표1-2>는 각 사이트에서 추출된 패턴 정보이고, <표2-1>과 <표2-2>는 이들 패턴을 통해 얻어진 '동물명-속성-속성값'의 정보 중 토끼와 관련된 일부이다.

```
| <a> # </a> |
[ # ] : </td> </tr>
<td> <a> # </a> </td>
<tr> <td> <ul> <li> # </li>
| <font> # </font> |
<b> ◆ # </b>
[ # ] :
<ul> <li> # .
<li> # .
<tr> <td> <font> <ul> <li> # .
<b> ( # </b>
<font> <b> - # : </b> </font>
<u> # </u>
<a> <font> # </font> </a>
<li> # . </li>
```

<표1-1> '네이처21' 패턴 정보

('#'은 단어를 나타낸다. 즉, 각 패턴들은 단어를 중심으로 추출되어져 온다.)

<tr> <td> <tt> <img> <a> # </tt> </td>

</tr>

<a> 9 #

<tr> <td> # </td>

] [<img> <a> #] [

] [<img> <a> #] [

<img> #

] [<img> <a> #] </td> </tr>

<tr> <td> <img> #

 ♣ #

] [<img> <a> #] </td> </tr>

 ♣ # ? -

 ♣ # ?

 ♣ # :

<tr> <td>   <img> #

 ♣ #

<tr> <td> # </td>

 1 . #

  ▶ # >

<표1-2> '서울시 동물사랑방' 패턴 정보

<1> 토끼
<2> +--A:학명
<3> oryctolagus cuniculus (var. domesticus)

<1> 토끼
<2> +--A:도워프종
<3> 도워프종에 대해(dqarf rabbits)
원산지:오스트레일리아
체 중: 1.8kg 전후
특 징:소형, 귀가 짧고, 얼굴이 몰려 있다.

.
.

<1> 토끼
<2> +--A:멕시코 토끼
<3> - 원산지 : 멕시코 - 몸길이 : 27~35.7cm - 귀길이 : 4~4.4cm - 분 포 : 멕시코 근교의 2개의 화산지역(아우스코스 산지와 이즈타코 시라코르, 코코카페르 산지)의 표준높이 2800 ~ 4000m의 넓게 트인 소나무 산지 에서 번식 아마미노크로 토끼와 함께 오래된 타입의 토끼로서 토끼과 중에 쿠까끼 토끼"亞과에 "분류

<1> 토끼
<2> +--A:눈
<3> : 빨갛게 보이는 것은 홍채의 멜라닌 색소가 없기 때문에 망막의 안쪽에 있는 혈관의 혈액이 그대로 보여서 눈이 빨갛게 보이는 것이다.

.
.

<1> 토끼

<2> +--A:급수기
<3> 물을 마시기 위해서 먹이 그릇처럼 바닥에 놓는 것과 병처럼 생겨 토끼가 입으로 빨아먹을 수 장착된 것이 있다.
물 그릇보다는 급수기를 사용하는 것이 더 위생적이다.

.
.
.

<표2-1> '네이처21' 패턴을 통한 추출 정보
(추출 정보의 순서는 <1>동물명, <2>속성, <3>속성값 순이다. 그리고 <2>속성에서 '+--A:'는 KB의 속성 데이터를 통해 추출된 패턴의 정보이고, '+--P:'는 웹 페이지 내의 반복 패턴을 통해 추출된 속성을 표시한다.)

<1> 토끼
<2> +--A:토끼의 먹이
<3> 양질의 펠렛, 건초(알팔파, 큰조아재비, 귀리), 물, 신선한 야채로 먹이를 구성해야 한다. 그밖에는 간식으로 주고 그 양을 제한해서 줘야 한다. 토끼의 평균 먹이섭취량은 체중의 약 4%이다. 성숙한 뉴질랜드화이트는 하루에 120 ~ 180g(30 ~ 60g/kg/일)의 먹이를 먹는다. 1일 2회, 아침 저녁으로 먹이를 준다. 토끼는 새벽과 저녁, 야간에 먹이를 먹으며 익숙한 것 외에는 먹지 않는다. 그러나 애완으로 키우는 토끼는 낮에도 먹는 경우가 있다.

.
.

<1> 토끼
<2> +--P:잎이 많이 붙어 있다
<3> 잡초, 풀뿌리, 나무토막 같은 이물이 적다.

<1> 토끼
<2> +--A:체 온
<3> 섭씨 37.2 ~ 39.4도

<1> 토끼
<2> +--A:심박수
<3> 130 ~ 325회/분

<1> 토끼
<2> +--A:호흡수
<3> 32 ~ 60회/분

<1> 토끼
<2> +--A:수 명
<3> 5 ~ 10년

.
.
.

<표2-1> '서울시 동물사랑방' 추출 정보

결과를 보면, 각 사이트의 문서에서 제시하고 있는 다양한 형태의 패턴에 대한 속성과 속성값에 대한 정보들을 추출해 낼 수 있었다.

그러나 이들 정보들을 그대로 KB로 끌어오기는 어렵다. 기존 KB가 가지고 있는 속성에 대한 표현도 각 사이트마다 다르고, 다양한 속성들이 새롭게 나타나고 있는데 이를 기존 속성 계층의 어느 위치에 포함시켜야 할지도 결정되어야 하는 문제가 있다.

5. 결론 및 향후과제

특정 영역 정보 에이전트의 서비스를 향상시키기 위해서는 KB의 확장 및 갱신은 중요한 요소이다. 이번 논문에서는 기존 동물 영역 시스템의 한글 서비스를 위한 KB의 한글화 작업을 인터넷 웹 사이트에서의 정보추출을 통해 자동화하고자 하였다.

기존의 정보추출 모델들이 자연어 처리에 의존하고 있고, 특정 정보 패턴에 대한 사전 지식이 필요했던 것에 비하여 실험에 사용되어진 방법은 웹 HTML 문서의 Tag에 주목하였고, 특정 정보 패턴 보다는 반복적으로 나타나는 Tag와 Symbol을 통해 속성과 속성값을 구하고자 하였다.

이는 동물 관련 사이트에서 동물 정보에 대한 표현이 일정한 형태를 가지고 있고, 그 기술이 일반적으로 속성-속성값으로 이루어져 있다는 점에서 매우 유용하였다. 또한 기존 KB의 속성을 이용하여 보다 명확한 정보를 끌어오고자 노력하였다.

하지만, 기존에 가지고 있는 특정 속성에 대한 값 뿐만 아니라 다양한 속성들에 대한 정보가 각 사이트마다 여러 표현으로 나타나므로 이를 KB 확장에 그대로 사용하는 것은 힘들다.

그러므로 이후 속성과 속성값에 대한 평가를 위한 방안을 더 생각해보고, 각 사이트에 공통적으로 가지고 있는 속성에 대한 정보들을 통합하기 위한 정보통합 기법에 대한 연구가 이루어져야 할 것이다.

6. 참고문헌

[1] **이용현, "정보통신망에서 지능형 정보 에이전트와 특정 영역에서의 구현", 홍익대학교 박사학위 논문**, 1999

[2] N. Kushmerick, "Gleaning the Web", IEEE Intelligent Systems, 1999

[3] Fuchun Peng, "Models for Information Extraction", 1999

[4] Ion Muslea, "Extraction Patterns for Information Extraction Tasks: A Survey", 1999

[5] S. Soderland, "Learning Information Extraction Rules for Semi-structured and Free text", Machine Learning, 1999

[6] **오정민, "웹 기반의 지능형 정보 에이전트의 개발: 동물 영역 지식 베이스의 구축을 중심으로", 홍익대학교 석사학위 논문**, 1998