정보에이전트를 위한 지식 기반(동물) 질의 처리 시스템

• 오 정 옥, 변 영 태
홍익대학교 전자계산학과

A Knowledge-Based Query Processing System for an Information Agent

Jung-Ok Oh, Young-Tae Byun
Dept. of Computer Science, Honglk University

요 약

본 시스템은 현재 연구 개발 중인 정보에이전트 시스템의 일부로서 특정분야에 대한 사용자의 관심 주제에 관련된 정보와 함께 적절한 문서를 제공하는 지식 기반 시스템이다. 이러한 목적을 위해서 본 시스템의 지식베이스는 구조적인 방식으로 표현된 BKB(Biology Knowledge Base)와 DIC(DICtionary)로 구성된다. DIC는 특정분야에서 일반적으로 사람들이 사용하는 용어와 학명을 기준으로 하는 시스템에서 사용하는 용어들의 관계와 그러한 용어들간의 동의어 관계를 갖고 있다. 또한 BKB는 동물에 관련된 지식베이스로써 상위·하위 개념과 함께 사용자가 관심있는 정보를 제공하기 위해 객체의 속성과 이에 관계된 값들을 포함한다. 본 시스템은 문서를 검색할 때 사용자 초기 질의를 상위·하위 개념 그리고 동의어로 확장할 뿐만 아니라 사용자 의도의 정확한 표현을 위해서 제공하는 다양한 질의 형식에 따른 질의 처리 결과로도 확장하므로 효과적인 문서 검색 결과를 보인다.

1. 서론

웹(World Wide Web) 상의 정보 제공자가 폭발적으로 증가함에 따라 필요한 정보를 웹에서 찾아내야 하는 정보 사용자는 망망대해의 정보 바다 앞에 서 있는 상황이 되었고 필요한 정보를 신속하고 정확하게 획득하는 것이 더욱 어려워졌다. 따라서 이러한 어려운 일을 대신해 주는 지능적인 정보 검색 대리자, 즉 정보에이전트가 필요하게 되었다. 정보에이전트는 본 시스템의 지식베이스를 바탕으로 웹 상의 정보를 지능적으로 접근해서 사용자에게 효율적인 정보를 제공할 뿐만 아니라 정보를 별도로 웹 DB에 저장하고 이러한 정보의 변화를 지속적으로 감시하며 관련 지식을 추가 및 갱신할 수 있도록 현재 개발 중인 시스템이다. 본 연구는 정보에이전트의 일부로서 전처리(preprocessing)된 질의를 입력받아 질의를 처리·확장하고 정보에이전트의 웹 DB 검색에 사용하도록 확장된 질의를 정보에이전트에게 넘겨준다.

본 시스템의 지식베이스는 특정분야에서 일반적으로 사람들이 사용하는 용어와 학명을 기준으로 하는 시스템에서 사용하는 용어들의 관계와 그러한 용어들간의 동의어 관계를 갖고 있는 DIC과 구조적으로 표현된 특정 도메인에 관련된 전문 지식을 갖는 BKB로 구성된다. BKB는 단순한 객체들간의 관계만을 나타내지 않고[4] 계층 구조 정보와 계층 구조상의 모체에 대한 속성과 속성값을 갖고 있어서 사용자의 질의에 대해서 적절한 정보를 제공할 수가 있다.

지능적인 검색을 위해서 시소러스를 사용하는 연구들이 있다[1-3,5,6]. 이러한 연구들은 질의 상의 용어를 개별적으로 고려하여 상위·하위 개념 또는 관련어(related term)로 확장한다.

본 시스템은 특정 도메인의 지식베이스를 기반으로 사용자 질의에 대한 정보와 사용자 질의에 개념의 확장, 개념의 구체화 작업을 거쳐서 새롭게 획득한 질의를 갖고 문서들을 검색한다. 여기서 개념의 확장은 질의 상의 해당 객체에 대한 상위·하위 개념과 동의어로의 확장이고 개념의 구체화란 사용자 질의를 처리해서 얻은 결과를 입력된 질의에 확장시키는 것을 의미한다. 이러한 개념의 구체화를 통한 효율적인 검색 방법을 제시한다.

그리고 일반적으로 사용되는 불리언(boolean) 연산자 외에 OF, MAX, MIN 등의 별도의 연산자를 제공해서 불리언 형태만으로는 표현하기 힘든 사용자의 의도를 보다 명확하게 표현할 수 있도록 한다. 사용 OF 연산자는 객체의 속성값을 알기 위해서, MAX 연산자는 특정 속성값이 가장 큰 객체를 알고 싶을 때 사용한다.

본 연구는 과학재단 (과제번호 973010301)의 지원을 받았음.

2. 지식베이스의 표현 방법

본 시스템의 지식베이스는 DIC과 BKB로 구성되어 있다. DIC은 동의어와 시스템어 정보를 갖고 BKB는 동물에 대한 전문적인 지식을 갖고 있다. 시스템어란 한 객체(object)를 의미하는 여러 용어들 중 하나를 채택해서 시스템에서 그 객체를 표현하기 위해서 사용하는 용어로서 학명(scientific name)을 기본으로 하고 구(phrase)의 경우에는 처리의 편의를 위해서 단어들 사이를 '-'로 연결해서 만든다. 이러한 시스템어의 사용은 시스템 처리상의 편의를 위해서 이기도 하지만 여러 가지 용어를 하나로 표준화하고자 하는 의미가 있다. BKB의 지식은 시스템어를 사용해서 구축되었다.

정보에이전트는 특정 도메인에 대한 지식베이스를 갖는다. 서로 다른 지식베이스를 갖는 다중 정보에이전트 시스템을 위해서 지식베이스에 이름을 부여함으로써 각각의 지식베이스를 구분한다. 본 시스템의 지식베이스는 생물 분야이므로 그 이름을 BIOLOGY라고 한다.

2.1 DIC

DIC은 동의어와 일반 용어에 관련된 시스템어 정보를 갖고 있다. 이러한 지식을 이용해서 질의의 일반 용어를 시스템어로 바꿔서 BKB에게 건네주고 또 BKB의 결과에서 시스템어를 일반 용어로 바꿔주는 BKB의 front-end 역할을 한다.

(synonym (domain <도메인>)
(system-word <시스템어>) (general-word <일반어>))

DIC의 지식은 위와 같이 표현이 되며 <일반어>는 사용자가 입력할 수 있는 임의의 용어이고 <시스템어>는 <일반어>에 해당되는 시스템어이다. 예를 들면 학명이 ailuropoda melanoleuca인 giant panda는 다음과 같은 형태로 저장된다.

(synonym (domain BIOLOGY)
(system-word ailuropoda-melanoleuca) (general-word giant panda))

동의어 관계는 별도의 형식을 갖지 않고 같은 <시스템어>를 갖는 <일반어>는 동의어가 된다. 따라서 ailuropoda-melanoleuca를 <시스템어>

로 갖는 giant panda, great panda, bamboo bear, ailuropoda melanoleuca 들은 서로 동의어 관계가 된다.

2.2 BKB

2.2.1 계층 정보

분류 생물학에서는 동·식물들을 특성별로 분류해 놓은 계>문>강>목>과>속>종의 계층 구조가 있다. 학명이 ailuropoda melanoleuca인 giant panda는 animalia > chordata > mammalia > carnivora > ursidae > ailuropoda > ailuropoda melanoleuca 의 계층 구조를 갖게 된다. 본 시스템은 이러한 계층 구조에서 과까지의 계층 구조를 BKB의 계층 구조로 구축하였다. 계층 구조상의 모든 객체는 한 단계 상위 객체와 is-a 관계를 갖는다. is-a 관계의 표현은 다음과 같다

(is-a 〈객체〉 〈상위객체〉)

<객체>는 객체의 이름이고 <상위 객체>는 <객체>의 상위 객체의 이름이다 따라서 giant panda의 시스템어인 ailuropoda-melanoleucas와 ursidae와의 관계는 (is-a ailuropoda-melanoleucas ursidae)로 표현된다

2.2.2 객체의 표현

계층 구조상의 객체는 속성을 나타내는 slot의 집합이다 객체에 속한 slot은 다음과 같이 표현된다

(slot-of 〈속성〉 〈객체〉)

<객체>는 객체의 이름이고 <속성>은 <객체>에 속한 속성 이름이다 [표 1]은 mammalia를 표현하는 속성들이다.

(slot-of length mammalia)	(slot-of weight mammalia)
(slot-of habitat mammalia)	(slot-of biome mammalia)
(slot-of distribution mammalia)	(slot-of food mammalia)
(slot-of diet mammalia)	(slot-of sexual-maturity mammalia)
(slot-of offspring mammalia)	(slot-of life-span mammalia)
(slot-of birth mammalia)	

[표 1] mammalia의 표현하는 속성들

계층 구조는 상속성을 갖기 때문에 상위 객체의 속성은 하위 객체로 상속된다. 따라서 mammalia의 하위 객체들은 자신만의 고유한 속성을 가질 수 있을 뿐만 아니라 mammalia의 속성들 또한 갖게 된다.

2.2.3 인스턴스의 표현

인스턴스는 해당 객체에 속한 속성들에 해당하는 실제값의 집합으로 표현된다. 이러한 값들을 다음과 같은 형태로 표현이 된다

(value-of 〈속성〉 〈객체〉 〈속성값〉+)

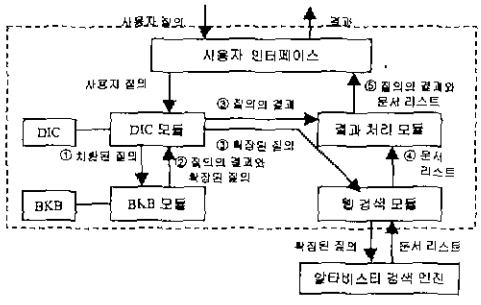
<객체>는 객체의 이름이고 <속성>은 <객체>에 속한 속성의 이름이고 <속성값>은 <객체>의 <속성>의 값이 된다 <속성값>은 하나 이상 존재한다. [표 2]은 giant panda를 의미하는 ailuropoda-melanoleuca를 표현하고 있다

```
(value-of birth ursidae vivipara)
(value-of length ailuropoda-melanoleuca 1 2 1 5)   range
(value-of weight ailuropoda-melanoleuca 75 160)
(value-of habitat ailuropoda-melanoleuca bamboo-forest)
(value-of biome ailuropoda-melanoleuca
   mountain temperate-forest-and-rainforest temperate-grassland)
(value-of distribution ailuropoda-melanoleuca oriental`
(value-of food ailuropoda-melanoleuca
                        bamboo insect bulb-grass rodent)
(value-of diet ailuropoda-melanoleuca herbivore)
(value-of sexual-maturity ailuropoda-melanoleuca 1825 2555)
(value-of offspring ailuropoda-melanoleuca 1 2)
(value-of life-span ailuropoda-melanoleuca 10950)
(value-of gestation ailuropoda-melanoleuca 97 163)
(value-of comment ailuropoda-melanoleuca "  ") . string
```

[표 2] ailuropoda-melanoleuca 의 표현

[표 2]에서 gestation은 carnivora로부터, comment는 animalia로부터, 나머지 속성들은 mammalia로부터 상속받은 속성이다. 객체는 상위 객체로부터 속성 외에 속성값도 상속받는다. 따라서 ailuropoda-melanoleuca의 birth의 속성값은 ursidae의 birth의 vivipara라는 값을 상속받는다 이러한 상속성(inheritance)을 이용하면 객체들을 효율적으로 특성별로 그룹화(grouping)할 수 있다. 값으로 갖는 속성들은 각각 기본 단위를 갖고 있다. weight는 kg(kilogram)으로 length는 m(meter)로 gestation, sexual-maturity, life-span은 day를 기본 단위로 갖는다. 정보로 제공될 때는 적합한 단위로 변환된다.

3. 확장된 연산자와 질의 확장



[그림 1] 시스템의 구성

본 시스템은 [그림 1]과 같이 구성되어 있다. 다양한 연산자에 따른 처리는 BKB 모듈에서 다루어지고 다른 모듈들에서는 모든 질의에 동일하게 동작한다 이러한 다양한 연산자는 사용자에 의해서가 아니라 정보에이전트의 사용자 질의 전처리(preprocessing) 과정에 의해 사용된다 즉 사용자의 초기 질의는 사용자 질의 전처리 과정에서 다양한 연산자를 사용한 질의로 변형되어 본 시스템의 입력 질의로 보내지는 것이다. 그러나 현재 정보에이전트가 개발 중인 관계로 본 시스템의 입력은 이러한 전처리과정을 거친 변형된 질의라고 가정하고 또한 질의 처리 결과의 효율성을 검사하기 위해서 정보에이전트 DB 검색을 웹 검색으로 대체하였다

3.1 BKB 모듈에서의 연산자에 따른 질의 확장

본 시스템에서는 불리언 연산자를 사용하지만 사용자 의도를 보다 명확하게 하기 위해서 다양한 연산자를 갖는 질의 형식을 제공한다

3.1.1 연산자가 없는 질의문 · <객체>

연산자가 없는 경우 질의문 전체가 하나의 <객체>를 의미하고 해당 객체에 관련된 모든 지식의 제공을 원한다고 간주한다 따라서 해당 객체의 계층 구조상의 상위·하위 개념으로 한 단계씩 확장하고 해당 객체에 대해서 알고 싶은 것들을 나타내기 위해서 속성의 이름만을 질의에 추가한다 해나하면 속성값을 나타내는 표현은 다양하지만 시스템에서는 가능한 표현을 모두 제공할 수가 없기 때문이다. 하지만 사용자에게 결과를 제공할 때는 객체의 속성과 속성값을 전부 제공한다. 자세한 예제는 다음 절에서 다룬다

3.1.2 OF 연산자 <속성> OF <객체>

<객체>의 <속성>의 값을 알고 싶을 때 사용한다 이런 경우 사용자에게는 속성값이 제공되지만 문서를 검색하기 위해 질의에 추가되지 않는다 질의는 <객체>에 대한 상위·하위 개념으로 한 단계씩 확장된다. 가령 cub OF giant panda라는 질의는 giant panda의 평균 산아수가 몇인지를 묻는 것이다 이러한 질의는 'cub OF giant panda [BT ursidae]'로 확장된다.

3.1.3 AND 연산자 . <객체1> AND <객체2>

<객체1> 과 <객체2> 둘 다 알고 싶은 경우에 사용하는 질의 구조이다 순수한 질의 외에 두 객체의 공통된 요소를 추론해서 질의를 확장한다 공통된 요소는 속성에만 국한되지 않고 상위·하위 개념에서의 공통점도 찾아본다

3.1.4 OR 연산자 <객체1> OR <객체2>

<객체1> 과 <객체2> 둘 중 하나만이라도 알고 싶은 경우에 사용되는 질의 구조이다. 따라서 질의는 <객체1>과 <객체2>의 속성이름, 상위 개념·하위 개념으로 확장된다. 이 때 확장은 중복을 허용하지 않는다

3.1.5 한정사들

▶ MAX 〈속성〉 animal [AMONG 〈객체집합〉] [IN 〈지역〉]
▶ MIN 〈속성〉 animal [AMONG 〈객체집합〉] [IN 〈지역〉]
▶ LT 〈속성값〉 〈속성〉 animal [AMONG 〈객체집합〉] [IN 〈지역〉]
▶ LE 〈속성값〉 〈속성〉 anmial [AMONG 〈객체집합〉] [IN 〈지역〉]
▶ GT 〈속성값〉 〈속성〉 anmial [AMONG 〈객체집합〉] [IN 〈지역〉]
▶ GE 〈속성값〉 〈속성〉 anmial [AMONG 〈객체집합〉] [IN 〈지역〉]

[표 3] 한정사를 사용하는 질의 구조

[표 3]은 한정사를 사용하는 질의 형식을 보여주고 있다. 위와 같은 구조는 단순한 불리언 연산만으로는 표현하기 힘든 개념을 나타내고 있다. 여기서 <속성>은 속성의 이름을 뜻하며 <객체집합>은 계층 구조상에서 하위 객체를 갖고 있는 객체에 해당된다 ursidae 혹은 mammal 등이 해당된다. <지역>은 지리적인 영역을 의미하는데 육지, 수중, 바다, 6개의 동물지구이름 등이 들어가게 된다 <속성값>은 조건이 되는 값을 의미한다 <객체집합>과 <지역>은 반드시 입력될 필요가 없으며 각각 animal과 world가 기본값이 된다.

<지역>에 서식하는 <객체집합>중에서 MAX 구문은 <속성>값이 가장 큰 animal을, MIN 구문은 <속성>값이 가장 작은 animal을 가리킨다. <지역>에 서식하는 <객체집합> 중에서 LT 구문은 <속성>의 값이 <속성값> 보다 작은 animal을, LE 구문은 <속성>의 값이 <속성값>과 같거나 작은 animal을, GT 구문은 <속성>의 값이 <속성값> 보다 큰 animal을, GE 구문은 <속성>의 값이 <속성값>과 같거나 큰 animal을 의미한다 예를 들자면 [MAX size animal AMONG mammal IN land]는 육지에 사는 포유류 중에서 크기가 가장 큰 동물을 알고 싶다는 의미로 추론과정을 통해서 african elephant라는 결과를 얻는다 따라서 이러한 결과는 웹 문서 검색을 위해서 상위·하위 개념으로 확장되어 질의에 추가된다

3.2 질의 확장 단계

본 시스템의 검색 절차를 'giant panda'라는 예제와 함께 살펴본다 각 단계는 [그림 1]의 번호에 해당된다

① 도메인 적합성 판단 및 용어의 치환 DIC 모듈에서는 질의의 도메인 적합성 검사와 입력된 질의의 일반 용어를 시스템어로의 치환하는 작업을 한다 질의가 도메인에 적합하지 않다고 판단되면 더 이상 처리되지 않는다 'giant panda'는 도메인이 BIOLOGY이므로 도메인이 적합하다고 판단돼서 시스템어로 치환된다 'ailuropoda-melanoleuca'은 치환된 질의어다

② 질의 처리 및 개념의 확장 KB 모듈은 치환된 질의를 상위·하위 개념으로 한 단계씩으로 확장하고 연산자에 따른 질의 처리를 하게 된다. 상위 개념은 BT로 하위 개념은 NT로 표현한다 치환된 질의는 연산자가 없으므로 ailuropoda-melanoleuca에 대한 전반적인 정보를 요청한다고 간주해서 ailuropoda-melanoleuca에 속한 속성의 이름이 질의에 추가된다 또한 질의는 ailuropoda-melanoleuca에 대한 상위 개념과 하위 개념으로 한 단계씩 확장을 한다 따라서 ailuropoda-melanoleuca는 개층 구조에서 가장 하위 객체에 해당되므로 상위 개념으로만 확장돼서 사용자 질의는 ailuropoda-melanoleuca [BT ursidae] length weight habitat, biome distribution food diet offspring sexual-maturity life-span birth feature'로 확장된다 또한 사용자에게 정보를 제공하기 위해서 각 속성과 속성값은 질의의 결과로서 별도로 저장된다 확장된 질의와 질의 처리 결과는 *DIC* 모듈로 전달된다

③ 동의어 확장 및 용어의 치환 확장된 질의와 질의의 결과를 넘겨받은 DIC 모듈은 각각에 대해서 시스템어의 일반 용어로의 치환과 동의어 확장을 거친 다음 확장된 질의는 웹 검색 모듈로 질의의 결과는 결과처리 모듈로 보낸다 동의어 확장은 SYN으로 표현한다. 따라서 확장된 질의는 'giant panda [SYN ailuropoda melanoleuca, great panda, bamboo bear] [BT ursidae [SYN bear]], length, weight [SYN size], habitat, biome, distribution [SYN geographical range, geographical distribution], food, diet, offspring [SYN litter, cub], sexual maturity, *life span, birth, feature [SYN characteristic]*'로 바뀐다

④ 웹 문서 검색 · 확장된 질의는 웹 검색 모듈에서 방대한 문서를 갖고 있는 알타비스타(altavista) 검색 엔진의 질의 형태에 맞게 질의를 변형시킨다 사용자 초기 질의는 반드시 포함하기 위해서 '+'를 붙이고 구(phase)는 " "로 묶은 다음 [,], SYN, BT, NT, ,(쉼표) 등은 제거된다 이렇게 변형된 질의에 대한 알타비스타의 검색 결과인 문서 리스트는 결과 처리 모듈로 보내진다

⑤ 결과 만들기 마지막으로 결과 처리 모듈에서는 질의 처리 결과와 문서의 리스트를 갖고 html 형식을 갖춘 결과 형태를 만들어서 사용자에게 제공한다

4. 구현 및 비교분석

본 시스템은 Windows95 운영체제를 바탕으로 Visual C++ 5.0와 CLIPS을 사용해서 구현하였다. DIC 모듈과 BKB 모듈은 CLIPS로 구현되었으며 사용자 인터페이스는 ISAPI를 사용해서 구현되었기 때문에 웹 브라우저를 통해서 접할 수 있다

본 시스템에 의해서 확장된 질의와 순수한 질의를 갖고 검색한 문서들 중 상위 30개를 테스트해 보았다 [표 2], [표 3], [표 4]의 내용을 살펴보면 본 시스템의 검색 결과가 단순 방식의 검색 결과보다 효과적임을 알 수 있으며 본 시스템의 검색 결과에서 대부분 적합한 문서가 상위권에 존재한다. 특히 [표 4]의 경우에는 연결에 실패한 한 문서를 제외하고 상위 10위 내의 모든 문서가 적합한 문서였다

시스템	질의	적합	비적합	연결실패
알타비스타	+"giant panda"	8/30	21/30	1/30
본 시스템	giant panda	25/30	5/30	0/30

[표 3] 연산자가 없는 질의

시스템	질의	적합	비적합	연결실패
알타비스타	+cub +"polar bear"	0/30	29/30	1/30
본 시스템	cub OF polar bear	9/30	20/30	1/30

[표 4] OF 연산자를 사용한 질의

시스템	질의	적합	비적합	연결실패
알타비스타	*+largest +land +mammal*	*5/30*	*21/30*	*4/30*
본 시스템	MAX size animal AMONG mammal IN land	16/30	9/30	5/30

[표 5] MAX 연산자를 사용한 질의

5. 결론 및 향후과제

본 논문은 동물에 관한 지식베이스를 기반으로 사용자 질의를 처리하여 사용자가 원하는 정보를 제공하고 또한 적당한 웹 문서를 추천하는 시스템에 대해서 기술하였다 도메인에 적합한 구조적인 방식으로 지식베이스를 구축해서 사용자 질의에 대해서 답할 수 있도록 만들었으며 사용자의 의도를 보다 정확하게 표현할 수 있도록 다양한 질의 구조를 제공한다 또한 사용자 질의의 단순한 개념 확장과 동의어 확장 이외에 질의 처리 결과를 질의에 확장시키는 방법을 사용하여 보다 효과적인 웹 문서의 검색 효과를 보였다

앞으로는 본 시스템을 웹 검색 엔진이 아닌 정보에이전트의 웹 DB에 결합시키려고 한다

참고문헌

[1] 신정훈, 안윤애, 류근호, 박현주, "문헌검색을 위한 지식기반 질의처리기 구현", 1997

[2] 신동욱, 임형묵, 윤용운, 최기선, "도메인 독립 및 종속자식을 이용한 효율적 정보검색", 1994

[3] 박영분, 김민구 마정태, "지식 기반의 정보 검색 시스템", 1994

[4] 최재훈, 한종진 박종진, 양재동, "구조적인 시소러스 구축을 지원하는 개체 기반 정보 검색 모델", 1997

[5] B P McCune, R M Tong, J S Dean and D G Shapiro, "RUBRIC A System for Rule-Based Information Retrieval," IEEE Transactions on Software Engineering, Vol. SE-11, No 9, pp 940-945, 1985

[6] Y Chiaramella and B Defude "A Prototype of an Intelligent System for Information Retrieval IOTA" Information Processing & Management, Vol 23, No 4, pp 285-303, 1987