지능형 에이전트의 비구조화 Ontology를 이용한 정보의 분류와 획득

양성기*, 배상현**
*조선대학교 대학원 전산통계학과
**조선대학교 자연과학대학 전산통계학과

Classification and Acquisition of Information using Unstructured Ontology of Intelligent Agent

Sung-Ki Yang*, Sang-Hyun Bae**
*Dept. of Computer Science & Statistics, Graduate School, Chosun University
**Dept. of Computer Science & Statistics, Chosun University

요 약

광역 네트워크 정보원으로부터 정보량이 증가함에 따라 효율적인 정보검색 도구의 필요성이 강조되고 있다. 기존의 정보검색 도구는 내용기반 검색방법으로 대상영역에 관계되는 체계적인 지식이 결여되어 사용자의 요구에 정확한 정보의 제공이 어려웠다. 본 논문에서는 광역 네트워크 환경에서 시시각각으로 생성·소멸되는 정보 중 사용자가 원하는 정보를 정확한 시간에 정확하게 제공하기 위해 지능적인 처리가 가능한 Ontology를 이용하였다. 광역 네트워크에 산재하는 대량의 정보원에서 Ontology를 이용하여 사용자가 필요한 정보를 자동적으로 수집·분류하는 지능형 에이전트인 정보검색 시스템을 제안한다.

1. 서론

최근 네트워크의 발달과 Web등의 멀티미디어 기술 보급으로 인하여 사용자는 광역 네트워크상의 정보를 이용하고, 제공하는 일이 가능해졌다. 이에 따라 네트워크의 사용자는 급속하게 증가하고 있고, 광역 네트워크에서 제공되는 정보원도 다양화·대규모화되고 있다. 따라서 사용자의 요구에 맞는 정보의 선택과 제시라는 지능적인 처리가 가능하여야 한다. 기존의 정보검색 도구는 내용기반 검색방법으로 대상영역에 관한 체계적인 지식이 결여되어 있기 때문에 사용자가 필요로 하는 정보가 어느 분야에 속하는 것인지, 검색결과를 체계적으로 분석하여 정확한 정보의 제공이 어려웠다. 많은 검색도구가 네트워크 정보원에서 사용자가 원하는 정보를 찾는 것을 도와주는 지능형 에이전트를 이용하고 있다.

본 논문에서 제시한 특정의 대상영역에서 Ontology를 이용하여 광역네트워크에 산재하는 이질적인 정보를 자동적으로 수집하고, 분류하는 지능형 정보검색 시스템(이하 검색시스템)을 제안한다.

2. Ontology

2.1 Ontology의 역할

에이전트가 특정 도메인에 관한 질의를 하거나 통신하고자 할 때, 먼저 그 도메인에 대한 개념화가 요구되어지며 이를 Ontology라 한다. 즉 Ontology는 어휘와 이론으로 구성되는 개념화의 명세이다[1]. Ontology는 특정 도메인에 대한 공통적 어휘기반을 제공하며 이기종 시스템과의 운용과 Intelligent Agent 시스템의 개발에 필수적이다. 사용자가 원하는 서비스를 에이전트에게 정확하고 분명하게 기술하여 에이전트가 보다 지능적인 검색 결정을 하기 위해 Ontology가 필요하다.

Ontology는 사용자와 검색시스템과의 공유된 공통의 배경 지식을 명시하고, 사용자의 요청과 정보제공자 사이에서 정보를 조정해 주며, 사용자 개체 범주를 할당한다. 또한 정보검색과 여과에 관련된 Ontology를 이용한 검색시스템은 상태변환 네트워크 문법과 개념 구조를 사용하여 정보를 추출한다.

2.2 비구조화 Ontology

Ontology는 Ontolingua[1]로 대표되는 프레임형 언어와 일차 서술 논리에 기초한 지식표현 언어로 기술된다. 이러한 언어에 의해 탑 다운적으로 대규모의 *Ontology를 구축하기 위해서는 상당한 시간과 노력이* 필요하여 현실적이지 않다. 또한 실세계의 정보는 모순을 많이 포함하고 있어 미리 모든 것을 고려하여 체계적, 전체적으로 기술하는 것은 곤란하다.

본 논문에서는 동일한 관점으로부터 기존의 개념체계와 전문용어 시소리스를 기반으로 Ontology를 구축한다. 이러한 Ontology의 큰 장점은 비구조화[2] 즉 개념을 나타내는 어휘의 집합과 개념간의 연상적인 관계를 기술하는 것이다. [그림 1]은 인공지능에 대한 비구조화 Ontology의 일부분이다. 각 노드는 개념을, 각 아크는 개념간의 연상관계를 나타낸다. 아크에는 개념간의 연결강도를 나타내는 가중치가 부여되어 있으나 Class-Instance, 부분-전체와 같은 개념의 관계는 구별하지 않는다. 또한 개념자체의 정의도 기술되어 있지 않다.

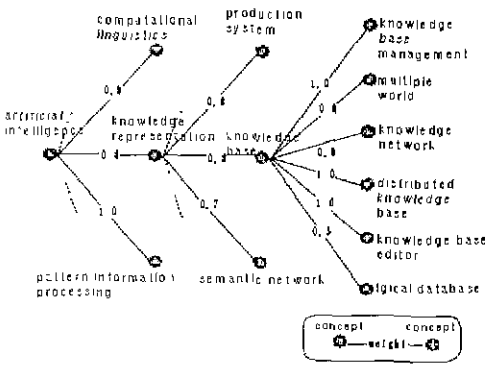

그림 1 비구조화 Ontology

3 지능형 정보검색 시스템

정보검색 시스템이란 시스템의 이용자가 필요로 하는 정보를 수집하여 내용을 분석한 뒤, 범주 형태로 조직하여 두었다가 정보에 대한 요구가 발생할 때 적합한 정보를 검색하여 제공하는 시스템이다. 정보검색 시스템이 지능적이기 위해서는 데이터나 정보 이외에 체계화된 지식을 이용할 수 있어야 하며, 자연언어의 이해능력과 문제해결을 위한 추론능력을 가져야 한다. 1983년 스파크존스는 지능형 정보검색시스템을 "정보요구와 문헌간의 관계를 결정하기 위한 추론능력과 지식베이스를 갖는 시스템"이라고 정의했다[3]. 본 논문에서 제시한 지능형 정보검색 시스템의 구조는 [그림2]와 같다.

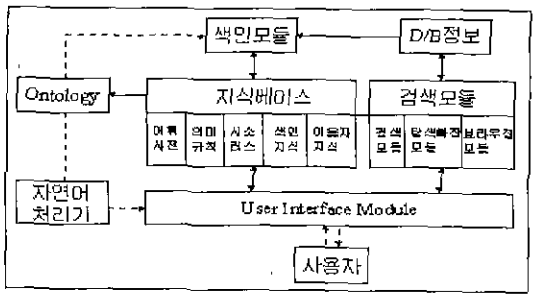

그림 2 지능형 정보검색 시스템

3.1 정보수집

최근에는 WWW상의 정보를 폭우선 탐색하는 Worm형 Agent[4]와 행동이력으로부터 사용자의 관심사항을 학습하는 능력을 갖는 Agent[5]의 연구가 행해지고 있다. 그러나 이러한 시스템은 대상영역에 관한 체계적인 지식이 결여되어 있기 때문에 사용자가 필요로 하는 정보가 어떤 분야에 속하는 것인가, 관련된 정보에는 어떤 것이 있는지 이해할 수 없다. 이에 Ontology를 이용한 정보검색 시스템에 체계적인 지식을 부여하여 지적인 정보수집 방법을 제안한다[그림3].

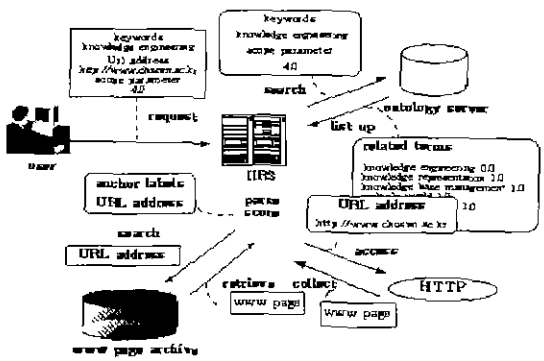

그림 3 WWW상의 정보수집

3.2 텍스트 분류

Text 분류에 있어서 Ontology의 각 개념을 각각 범주로 설정하고 수집된 Text를 이중 하나의 범주에 할당한다. Ontology 기반 텍스트 분류는 특징벡터와 분류벡터 사이의 유사도 계산과 분류벡터 사이의 유사도 계산의 일련의 처리과정에 의해 분류된다. 특징벡터는 그 문서의 특징을 나타내는 벡터에 있다. 계산방법은 다음 절에서 기술한다. 분류벡터는 카테고리의 특징을 나타내는 벡터로서 그 카테고리에 분류된 텍스트의 대표벡

터로부터 구해진다.

3.3 벡터 공간 모델

단어의 가중치 부여와 수집한 문서의 특징 벡터를 계산하기 위해서, 정보 탐색의 분야에서 폭넓게 이용되고 있는 벡터 공간 모델을[6] 사용하였다. 단어의 가중치 부여는 출현한 텍스트에서 그 단어의 상대 출현빈도 *tf*(term frequency)와 텍스트의 집합에 있어서 그 단어의 역 문헌빈도 *idf*(inverse term frequence)의 적(積)에 의해서 주어진다 즉,

$$W_{ik} = tf_{ik} \times idf_k$$

여기서, tf_{ik}는 문서 i에 있어서 단어 t_k의 출현빈도, idf_k는 문서 집합에 있어서 단어 t_k가 출현한 문서 수의 역수이다. 일반적으로 사용된 *idf*의 척도는 다음과 같이 주어진다.

$$idf_k = \log(N / n_k)$$

여기서, N은 문서의 총수이고, n_k는 키워드 t_k을 포함하는 텍스트의 수이다

4 분류실험

"Artificial Engineering"에 관해 네트워크 뉴스의 기사400건을 대상으로 분류 실험을 하였다. 뉴스 그룹은 "comp"를 대상으로 했다. 뉴스그룹 기사를 75개의 카테고리에 분류하여 Table 1에 그 결과를 제시하였다. 분류 결과를 평가하기 위해서 다음의 식을 이용하여 재현율과 정확률를 구했다

$$\text{재현율} = \frac{\text{검색된 적합 텍스트 수}}{\text{적합 텍스트 총 수}}$$

$$\text{정확률} = \frac{\text{검색된 적합 텍스트 수}}{\text{검색된 텍스트 총 수}}$$

상위10 카테고리	텍스트 수	하위10 카테고리	텍스트 수
program	48	VLSI	1
planning	31	statistics	1
artificial intelligence	25	SQL	1
prolog	17	signal	1
software	16	psycology	1
inference engine	14	PC	1
classification	13	Lisp	1
cognitive	12	interface	1
expert system	10	informatics	1
C	9	DOS	1

Table 1. News기사 분류 결과

계산결과는 Table 2에 나타냈다. 재현율과 정확률은 모두 전 카테고리의 평균치이다.

정확률(%)	재현율(%)	정확도(%)
77.0	76.2	76.0

Table 2 . 분류실험 평가

5. 결론

Ontology와 같은 도메인을 전문분야로 하는 연구자와 초보자에 있어서는 자동적으로 수집한 정보를 Ontology에 의해 체계적으로 분류할 수 있지만, 전문분야에 전혀 생소한 사람에게는 Ontology에 의한 분류가 어렵다.

향후 연구에서는 Ontology의 구조가 고정적이기 때문에, 사용자의 흥미와 새로운 정보에 유연하게 대응할 수 있도록 학습 방법을 향상시키고, 이용 가능한 Ontology의 수가 적기 때문에, 수집한 데이터로부터 새로운 링크를 수행하고 Ontology를 사용자에게 적응시키는 방법과 새로운 개념을 학습하는 방법을 연구한다.

참고문헌

[1]. T R. Gruber, J M.Tenenbaum, and J.C.Weber. "Toward a knowledge medium for collaborative product development". In Proc. 2nd Int. Conf on Artif. Intell. in Design, pp. 413-432, 1993

[2]. 化川賢治, "일상지식의 약구조화에 관한 연구", 나라첨단 과학기술대학원대학 석사논문, 1995.

[3]. K.Sparck Jones, "Intelligent Retrieval", in K.P.Jones, ed., *Intelligent Information Retrieval : Informatics 7*(London Aslib), 1983.

[4]. O. McBryan. Genvl and wwww.tools for taming the web *In Proceedings of 1st International WWW* Conference, 1994.

[5]. P.Maes and R Kozierok. "Learning interface agents". In Proceedings of AAAI, 1993

[6] G Salton and M.J. McGill. "Intoroduction to modern infomation retrieval", MacGraw-Hill, 1983.

[7]. 정영미, "정보검색론", 구미무역(주), 1993년 2월

[8]. 최충민, "인터넷 정보 가공을 위한 에이전트", 정보처리학회지, 1997년 9월

[9]. 맹성현, 주종철 "문서구조화와 정보검색", 정보과학회지, 1998년 8월