인터넷 질의/응답을 위한 지식베이스 구축

장문수° 장명길 김현진 오효정 이재성*
한국전자통신연구원 언어공학연구부
*충북대학교 컴퓨터교육과
{jms62804, mgjang, jini, ohj}@etri.re.kr, jasonl@cbucc.chungbuk.ac.kr

Construction of Knowledge Base for Question/Answering on Internet

Moon-Soo Chang° Myung-Gil Jang Hyun-Jin Kim Hyo-Jung Oh Jae-Sung Lee*
Dept. of Language Engineering, ETRI
*Dept. Of Computer Education, Chungbuk National University

요 약

차세대 검색 엔진의 모형으로 일컬어지는 질의/응답 시스템을 개발하는데 있어서, 보다 정확하고 유연한 검색 결과를 제공하기 위하여 개념망에 기반한 지식베이스 구축의 필요성이 대두되고 있다. 본 논문은 기존의 개념망에 속성 구조를 추가한 확장 개념망과 속성에 의해 분류되는 정답 문서 집합으로 구성되는 지식베이스를 이용한 질의/응답 시스템을 제안한다. 본 논문의 지식베이스에서 정의한 속성은 질의/응답에서 정답문서를 효과적으로 연계시켜 사용자에게 보다 유연한 정답을 제공할 수 있게 한다. 본 논문에서는 경제 분야의 지식베이스의 활용예를 설명한다.

1. 서론

기존의 검색 시스템은 빠른 시간 내에 얼마나 많은 문서를 처리하고 얼마나 많은 문서를 제공하는가 하는 것이 시스템 성능의 척도였다. 검색된 문서의 정확성이나 사용자의 요구에 대한 적응성은 상대적으로 간과되어, 사용자는 자신이 검색한 내용을 찾기 위해 검색 결과로 제공되어진 많은 문서를 또다시 읽어야 했다.

따라서, 좀더 편리하고 정확한 검색에 대한 사용자들의 요구는 인터넷의 사회 전반으로의 급속한 확산과 더불어 자연적으로 발생하는 초보자 층의 확대와 맞물려 검색 기술의 당면 과제로 떠오르게 되었다. 이러한 요구는 컴퓨터 시스템 성능의 빠른 향상과 자연어처리 기술과의 접목으로 질의어 확장, 자연어 질의, 질의/응답 구조 등과 같은 많은 기술들을 낳고 있다. 우리는 이러한 기술들을 이용하는 검색 기술을 인간의 지적 활동의 대표적인 도구인 언어를 매개로 한다는 뜻에서 지식 정보 검색이라고 한다.

지식 정보 검색 중에서도 특히 질의/응답은 단순히 사용자가 입력한 단어가 들어있는 문서를 나열해주던 기존의 방식에서 탈피하여, 사용자의 질의에 대해 직접적인 응답이 될 수 있는 정보를 제공하는 검색 방법으로 차세대 검색엔진의 모형으로 알려지고 있다.

본 논문에서는 이러한 질의/응답 시스템에 있어서 보다 정확하고 유연한 검색 결과를 제공하기 위해서 개념망에 기반한 지식베이스를 구축한다. 2장에서는 질의/응답의 기술과 검색 서비스의 현황을 살펴보고, 3장에서는 본 논문에서 구현하고자 하는 질의/응답 시스템의 개요를 설명한다. 그리고, 4장에서 제안하는 지식베이스의 구조와 구축과정에 관하여 설명하고, 5장에서 결론과 전망에 대해서 논한다.

2. 관련 연구

질의/응답은 디렉토리의 많은 정보의 층을 거쳐서 제공되는 기존의 horizontal 검색과는 달리, 한두번의 클릭으로 원하는 정보까지 수직으로 접근한다는 뜻에서 vertical 검색이라고도 한다. 질의/응답은 최근에 국내외적으로 많은 연구가 이루어지고 있으며, 상용 검색엔진도 나오고 있다. 상용 검색엔진의 대표적은 것

으로 AskJeeves(www.ask.com) 를 들 수 있는데, 이 엔진에서는 사용자로부터 자연어 문장 질의를 받아들여 분석하고, 그 결과로써 검색엔진이 새로운 정교한 질의어의 패턴을 생성하여 사용자로 하여금 선택하게 한다. 사용자가 하나의 패턴을 선택하면 검색엔진은 거기에 적합한 사이트를 정답문서로서 보여준다. 국내에서는 최근에 네이버(www.naver.com)에서 AskJeeves와 유사한 방식의 엔진을 선보이고 있고, 그 밖에 에스크존(www.askzone.come), 인포구루(www.infoguru.com), 엑스퍼트(www.xpert.co.kr) 등이 서비스를 시작하고 있지만, 대부분이 질의어 수집이나 전문가 조언, 즉 FAQ형식을 벗어나지 못하고 있어, 사용자의 요구에 접근할만한 수준까지는 많은 보완이 요구되고 있는 실정이다.

질의/응답에 있어서 지금까지 개발된 중요한 기술적 요소로는 사용자 질의 혹은 사용자 의도의 해석, 응답의 결정, 사용자 인터페이스 등을 들 수 있다. 사용자 질의의 해석은 보다 정확한 사용자의 의도를 파악하기 위하여 자연어 질의문 해석 기술이 많이 응용되고 있다. LASSO 시스템[1,2]은 다양한 질의 유형의 분석을 통하여 응답 유형을 결정하고 있다. 응답의 결정은 현재 대부분의 시스템이 기존의 정보검색 매칭 방법에 따르고 있는데, LASSO의 경우 paragraph 색인에 근거하고 있다. 응답의 결정에 있어서는 응답의 종류, 즉 정답문서, 단어, 정답 기술 문장 등에 따라 다양한 형태의 연구가 이루어지고 있다. 이러한 기술들은 TREC에서 QA track을 시작하여 활발한 연구가 진행되고 있다[3,4]. 사용자 인터페이스에 관한 기술은 주로 검색 엔진을 제공하는 벤치 기업에서 사용자 만족도를 위한 연구로 많이 이루어지고 있다.

지식 정보 검색에서 질의어 해석이나 질의에 적합한 정답 문서의 결정에 있어서 지식의 개념을 적용시키기 위해서는 지식을 표현하는 도구가 필요하다. 이러한 도구를 지식 처리 분야에서는 지식 베이스라고 한다. 정보 검색 분야에서는 방대한 정보량과 빠른 처리 속도를 요구하는 정보 검색의 특수성 때문에 제한적이고 변형적인 자연어 처리 기술만이 응용되고 있을뿐 체계적인 지식처리 기술이 논의된 경우는 많지 않다. 본 논문에서는 보다 정확한 질의어 분석과 체계적인 데이터베이스의 구축을 위해서 개념망을 기반으로 하는 지식베이스를 구축하고자 한다.

3. 개념망을 이용한 질의/응답 시스템

본 논문에서는 보다 지능적인 검색엔진을 실현하기 위하여 개념망을 이용한 질의/응답 시스템을 제안한다. 질의/응답은 사용자의 질의를 통해 사용자의 의도를 충분히 파악하여 그것에 가장 적합한 응답을 제공해야 한다. 따라서, 질의/응답 시스템은 사용자에게 세련게 질의문을 생성하여 선택을 하게 하거나, 자연어 질의문에 대해서 문법적, 의미론적 해석을 통해 유사한 내용이 있는 문서나 문장을 응답으로 제공하고 있다. 본 논문에서는 이러한 질의문 생성이나 해석을 통해 정보를 검색함에 있어서 개념적인 체계를 제공하기 위하여 개념망을 이용하는 질의/응답 시스템을 착안하였다.

개념망은 검색에 사용되는 단어들을 단순한 키워드의 매칭이 아닌 개념들의 연결을 통한 의미적으로 유사한 매칭으로 연결시켜준다. 따라서, 개념망을 이용한 질의/응답 시스템은 사용자 질의에 대한 적절한 해석이 이루어지고, 이에 대해 개념적으로 보다 유사한 내용이 응답으로 제공되어질 수 있다.

그림 1은 본 논문에서 제안하는 개념망을 기반으로 하는 지식베이스를 활용한 정보 검색 시스템의 개념도를 나타낸 것이다.

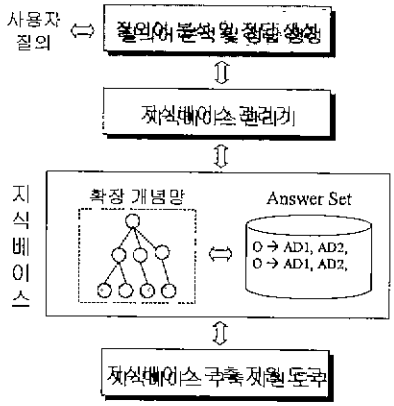

그림 1. 지식베이스를 활용한 질의/응답 시스템

4. 지식베이스의 구축과 활용

본 논문에서는 사용자에게 보다 정확한 검색 결과를 제공하기 위하여 지식베이스를 구축한다. 본 논문의 지식베이스는 확장 개념망과 정답문서 집합의 두 가지로 구성된다. 확장 개념망은 질의/응답 시스템에서 정답문서 집합과의 연계를 위하여 기존의 개념망[5]을 확장한 것이고, 정답문서 집합은 확장 개념망으로부터 유도된 정답문서들의 리스트이다.

4.1 확장 개념망

본 논문에서 도입한 개념망은 기존의 검색엔진에서 주로 사용하는 디렉토리와 상호보완적인 특징을 가지고 있다. 개념망과 디렉토리를 비교하면 표 1과 같다.

개념망은 모든 현상을 표현할 수 있는 반면 실제 정답문서와의 매칭은 이루어져 있지 않다. 디렉토리는 빠르고 정확한 검색이 가능하지만, 분류가 가능한 임의의 영역으로 제한되어 개념적으로 누락되는 부분이 존재한다. 그리고, 개념망은 개념적인 유추에 의해 색인어와 정확하게 일치하지 않더라도 검색이 가능

하며, 다의어를 말뜻에 따라서 분리하므로 큰 문제가 되지 않지만, 디렉토리에서는 단어의 모호성이나 유사어와 같은 언어적 지식에 기반하는 문제에 있어서는 처리가 쉽지 않다.

표 1. 개념망과 디렉토리의 특징 비교

개념망	디렉토리
하위계층	
is-a나 part-of와 같은 개념적인 연결	웹문서의 내용적 분류의 대중소 분류와 같은 세부항목으로의 연결
대상범위	
대표되는 단어에 의해 모든 현상과 사물, 원리를 나타낼 수 있음.	웹문서 분류에 의한 임의의 영역
장점	
개념적인 유추에 의한 접근이 가능 정답문서와 직접적으로 연결	
다의어의 구별이 가능하고, 유의어와 관련어를 지원할 수 있음 빠르고 정확한 검색	

이와 같이 개념망과 디렉토리는 서로 보완할 수 있는 장단점들을 지니고 있다. 본 논문에서는 질의/응답 시스템 구축에 있어서 사용자의 질의에 대해 보다 유연한 응답을 제공하기 위하여 개념망을 기본으로 하는 지식베이스를 이용하는 시스템을 제안하고 있다. 개념망에 있어서 문제가 되는 정답문서와의 연계를 보완하기 위해 속성 구조를 도입하고, 속성이 추가된 개념망을 확장 개념망이라 한다.

4.2 확장 개념망의 속성

질의에 대한 정답문서는 각각의 질의에 대해서 분류, 결정된다. 본 논문에서는 개념망의 각 개념 노드에 대해서 정답문서를 연결시킨다. 그리고, 개념 단어의 특징에 따라서 속성을 정의하고, 이 속성에 정답 문서를 분류하여 연결시킨다. 따라서, 실제 질의문에 대한 분석을 통해서 중심이 되는 개념 단어와 그 사용영역, 즉 속성을 결정하게 되면 정답문서로 바로 연결되게 된다.

따라서, 속성은 추상적인 개념들의 집합체인 개념망과 실제 웹문서를 연결시켜 주는 도구로서 개념망에 디렉토리의 장점을 살려주는 역할을 한다.

이러한 속성은 다음과 같은 분류 기준에 의해 정의된다.

- 개념에 대한 categorization : 특정 개념의 정답문서들에 대한 categorization 관점에서 속성을 부여한다.
- 문서집단의 성격을 살릴 수 있는 labeling : 분류되는 웹문서들의 내용을 포괄할 수 있는 단어를 속성으로 정의한다.
- 상위 개념으로 통합할 수 있는 속성의 일반화 : 특수한 용어를 지양하고 가능한 한 상위 개념으로 통합할 수 있는 단어를 선택한다
- 개념에 따른 특징을 살릴 수 있는 속성 : 개념의 특징에 따라 분류 가능한 속성을 추출하여 개념간의 차별화를 유도한다.

본 논문에서는 속성 분류의 예로써 경제 분야를 대상으로 속성을 분류하였다. 속성을 분류하기 위해서 본 논문에서는 다음과 같은 과정을 수행한다.

① 경제 용어 검색 및 분류
② 분류된 단어에 대한 웹 검색
③ 검색된 문서의 분류 : 정답문서
④ 속성의 결정
⑤ 속성의 정규화
⑥ 상위 개념 속성으로의 통합

여기서 사용된 경제 용어는 경제용어 사전에 수록된 표제어 중에서 일반적이고 시사적인 용어를 선정하였으며, 각 단어당 다섯개의 검색엔진을 통해 100개의 웹문서를 검색하였다. 검색된 문서를 분류, 즉 질의에 대한 정답문서의 추출은 검색엔진에서 제공하는 요약문과 실제 문서를 참조하여 결정하였다. 표 2는 속성 분류 실험의 결과를 나타낸 것이다.

표 2. 개념망의 속성 분류 결과

경제 용어 수	135
분류대상 웹 문서 수	13500
정답 문서 수	2605
정규화된 속성 수	17

표 3은 실험을 통해 정의된 속성과 분류를 위한 근거를 나타낸 것이다. 표 3에서의 속성은 정답문서의 분류를 목적으로 정의된 것이므로, 각각의 문서의 요약과는 다를 수 있다. 예를 들어, 동향은 동향, 전망, 현황으로 분류되는 문서를 묶어서 하나의 속성으로 나타내었다. 과다한 속성의 분류는 정답문서의 결정이

곤란하고, 상위 개념 노드로의 통합을 어렵게 하기 때문이다.

4.3 질의/응답 시스템에서의 지식베이스의 활용예

여기서는 실제 예제를 통해서 제안하는 확장 개념망이 질의/응답 시스템에 어떻게 활용되는가를 나타낸다.

표 3. 속성의 종류의 분류 근거

속성	분류 근거
정의	검색어의 의미(meaning)가 중심
정보	검색어를 이해하는 데 도움을 줌
뉴스	특정 시기에 일어난 사건,사고
사설	신문사설, 논문 등 개인적인 지식이 반영됨
규정	규칙으로 정하는 것
유형	공통의 성질, 특징이 있는 것끼리 묶은 틀
사례	어떤 행위나 일에 관하여 일어난 낱낱의 사건
표	어떤 사항을 보기쉽게 기록한 것
대책	사회에 악영향을 주거나 불건전한 경제 행위에 대한 대책이나 방책
동향	어떤 집단이나 현상에 대한 방향이나 흐름
학습자료	경제와 관련된 시험문제나 교과 내용
피해사례	사회에 악영향을 주거나 불건전한 경제 행위에 의해 발생한 결과
공고	광고, 게시 등으로 일반 공중에게 알리는 일
경력	인물이 중심이 된 검색어일 경우 어떤 사람의 학력이나 이력
자격증	자격증 취득에 관한 정보
판례	판결 사례를 말하며, 주로 법적인 내용을 담고 있음
상담사례	질의와 응답으로 이루어짐

본 논문에서는 경제 분야에 대해서 속성을 분류하였으므로, 경제 용어 중에 하나인 "불공정거래"에 관한 질의어를 가정하여 설명한다. 불공정거래에 대한 자연어 질의는 다음과 같은 것들이 있을 수 있다.

- 불공정거래란 무엇인가?
- 불공정거래의 사례를 보여주시오.
- 불공정거래에 관한 최근 뉴스를 보여주시오.
- 불공정거래에 관한 법률에 관해서 알고 싶다.
- 1999년에 일어난 불공정거래의 사례가 있는가?

불공정거래에 대해서 시스템은 다양한 형태의 질의 패턴을 가지고 있다. 이 패턴은 개념 단어와 속성의 조합으로 표현되며, 이러한 조합은 정답문서 집합으로 연결된다. 그림 2는 불공정거래를 중심으로 하는 확장 개념망과 정답문서 집합의 일부를 나타낸 것이다.

복합적인 질의는 질의 패턴들의 조합이나 키워드 검색과의 연계로 새로운 정답문서 집합을 만들어 낼 수 있다. 예를 들어 "1999년에 일어난 불공정거래의 사례가 있는가?"와 같은 질의문은 (불공정거래, 사례)와 연결된 정답문서 중에서 "1999년"과 관련된 문서만 추출해서 정답문서로 제공할 수 있다.

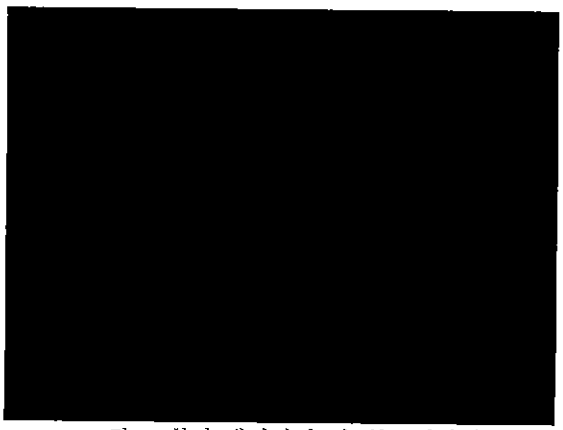

그림 2. 확장 개념망의 예: 불공정거래

5. 결론

속도와 문서 수에 중점을 두는 기존의 키워드 기반 검색 엔진의 문제점을 해결하기 위해 사용자 중심의 질의/응답 시스템에 관한 연구가 활발해 지고 있는 현 시점에서, 보다 정확하고 유연한 검색 결과를 유도하기 위해 본 논문에서는 개념망에 기반한 지식베이스를 질의/응답 시스템에 적용하였다. 본 논문에서 제안한 지식베이스는 기존 개념망에 디렉토리의 장점을 살린 속성을 부여하는 확장 개념망과 그에 따른 정답문서로 구성되어, 지식과 정보가 유기적으로 결합된 검색엔진의 모형을 제안하였다.

본 연구는 앞으로 자동 분류 기법을 적용하여 기존 검색 데이터베이스를 능가하는 지능적인 검색 데이터베이스를 구축하는 방향으로 나아갈 것이다. 이를 위하여 지식베이스에서 이미 구축된 확장 개념망의 속성과 정답문서를 학습데이터로 이용하여 확장 개념망의 확장과 정답문서의 자동 분류에 활용할 예정이다. 또한 본 논문의 지식베이스를 이용한 질의/응답 시스템은 질의 유형 분석 및 응답 유형 결정, AskJeeves와 유사한 시스템이 제공하는 질의 패턴 생성에 관한 연구를 진행할 계획이다.

6. 참고 문헌

[1] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju, and V. Rus, "LASSO: a tool for surfing the answer net", *Proc. of the Eighth Text REtrieval Conference (TREC-8)*, 1999.

[2] Sanda Harabagiu, Marius Pasca and Steven Maiorano, "experiments with open-domain textual question answering", *Proc. of COLING-2000*, August 2000.

[3] E. Voorhees and D.Tice, "The TREC-8 question answering track evaluation", *Proc. of the Eighth Text Retrieval Conference (TREC-8)*, 1999.

[4] K.Humphreys, R. Gaizauskas, M. Hepple, and M.Sanderson, "university of *Sheffield* TREC-8 Q & A system", *Proc. of the Eighth Text Retrieval Conference (TREC-8)*, 1999.

[5] 한국전자통신연구원, "내용기반 멀티미디어 정보검색 기술 개발", 최종연구보고서, 정보통신부, 1999.