질문구조의 전형을 이용한 정보요구의 모형화에 관한 연구

A Study on Modeling of Information Need Using Stereotype of Question Structures

김기영, 정영미
연세대학교 문헌정보학과

Gi Yeong Kim, Young Mee Chung
Dept. of Library and Information Science, Yonsei University

본 연구는 질의어 확장 및 단기 이용자 모형 구축에 응용할 수 있는 하나의 기법으로서 이용자 질문구조의 전형을 통한 정보요구의 모형화를 실험을 통해 제시한다. 실험방법은 이용자의 질문을 시소러스를 통해 분석, 구조화 하고 그 질문구조에서 전형을 추출한 후 전형에 따라 요구하는 정보가 질문구조내에 일정하게 위치하는지를 알아보았다. 이러한 실험을 통해 6가지 질문구조 전형을 추출할 수 있었으며 질문구조의 전형을 이용한 정보요구 모형의 구축이 타당성이 있음을 입증하였고 지능형 정보검색 시스템에의 적용가능성을 논의하였다.

1. 서 론

최근 정보검색시스템은 탐색전문가에 의한 중개탐색에서 이용자가 시스템에 직접 접근하는 방향으로 나아가고 있으며, 이에 따라 시스템이 개별 이용자의 정보요구를 수용하기 위하여 이용자 모형을 구축하는 방법이 논의되고 있다. 이용자 모형은 시스템이 가지는 이용자에 대한 지식으로 이용자의 특성과 탐색행태 간의 관계를 설정하여 이용자의 질문에 표현되지 않은 정보요구를 예견하도록 한다.

이러한 이용자 모형의 요소들은 기본적으로 정적인 성격을 지녀 정보요구 발생 시점에 대한 상황을 고려하기 힘들며, 이용자가 시스템에 처음 접근할 경우 시스템은 이용자에 대한 정보 대부분을 탐색전 사전 질문을 통해 수집하므로 시스템을 처음 이용하거나 가끔씩 이용하는 이용자에게 많은 대화를 해야 하는 부담을 가져오며, 또한 이용자 모형 요소들은 대체로 비주제적인 요소로 이루어져 있어 정보요구 주제에 대한 직접적인 분석 및 확장을 하기 힘든 단점이 있다.

따라서 본 연구는 이용자의 정보요구 발생 시점에서 시스템과 이용자간의 최소한의 대화 즉, 이용자가 시스템에 입력하는 질문에 의해 이용자의 정보요구를 추론할 수 있는지 그 가능성을 살펴 보고자 한다. 본 연구에서 정보요구는 이용자가 자신의 지식의 이상상태(ASK : Anomalous State of Knowledge)를 인지하고 이를 해결하기 위해서 일어나는 것이며,[1] 질문은 이용자가 이러한 자신의 정보요구를 자연언어로 표현한 것을 가정한다. 따라서 공식적인 지식구조인 시소러스를 이용하여 이용자의 질문구조의 전형을 구하고 이를 통해 정보요구를 파악할 수 있는 지를 실험을 통해 알아보고자 한다.

2. 질문구조 전형 추출 실험

2.1 실험 설계

본 실험의 목적은 질문을 통해 이용자의 정보요구를 추론함으로써 정보검색시스템이 이용자에 대한 더 많은 정보를 입수하여, 결과적으로는 시스템의 성능을 향상시키기 위한 것이며, 이를 위해 구축된 주제영역 지식을 이용하여 자연언어 질문과 이용자가 요구한 정보인 응답을 분석함으로써 질문구조의 전형을 추출하고, 이

를 통해 이용자의 정보요구에 대한 모형을 구축하는 것이 가능한지를 알아보는 것이다.

본 실험은 ㉠ 이용자의 정보요구는 이용자의 지식상태가 불완전하고 이를 이용자가 인지하기 때문에 일어나고, ㉡ 이용자의 질문은 자신의 정보요구를 표현하기 위하여 작성되지만 이 질문은 정보요구를 구체적으로 나타내지 않으며, ㉢ 실험자료로 사용된 컴퓨터통신 서비스, 동호회코너의 '질문/응답'에서 응답은 해당하는 질문이 표현하는 구체적 정보요구를 반영한다는 가정하에서 실시한다.

위의 가정중 첫번째, 두번째 가정은 ASK개념을 기초로 한 것이며, 세번째의 가정은 다음과 같은 배경에 의한다. 즉, 컴퓨터통신 서비스의 질문/응답에서 응답자는 해당 질문에 즉각 응답하는 경향이 있으며, 만약 질문자가 제공된 응답에 만족하지 않을 경우, 이용자는 재질문을 통해 자신의 정보요구를 다시 표현하므로 질문자가 응답에 대해 재질문할 경우 이 응답은 해당질문이 표현하는 정보요구가 아닌 것으로 간주되어 본 실험 자료에서 탈락된다. 다시 말해 응답은 재질문이 없는 한 이용자의 정보요구에 적합한 응답, 즉 질문자의 구체적인 정보요구로 간주한다.

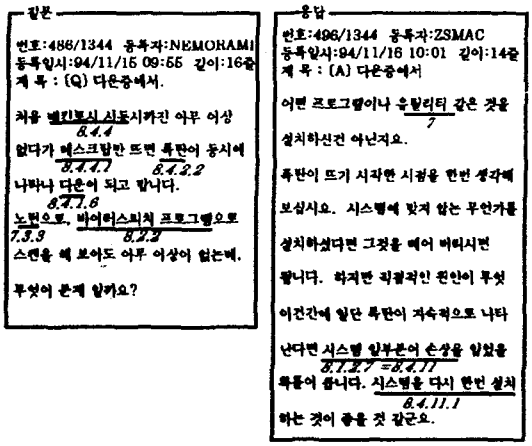
이러한 가정하에 이용자가 질문을 통해 정보요구를 표현할 때 사용하는 질문구조에 어떠한 전형이 있다는 가설을 세우고 이를 실험을 통해 증명하고자 하는 것이다.

본 실험에 사용될 자료의 주제영역은 PC중 한 기종인 매킨토시 기종에 대한 하드웨어 및 소프트웨어에 관한 분야로 제한하는데, 그 이유는 매킨토시의 환경이 독립적이고 호환성이 없으므로 관련 문헌이 일관적이기 때문에 질문/응답 분석의 틀인 시소러스의 구축과 질문/응답의 수집 및 분석을 용이하게 하였다.

실험자료인 질문과 응답은 컴퓨터통신 서비스의 해당 동호회에 있는 질문응답코너를 이용하여 67건의 질문/응답을 실험자료로수집하였다.

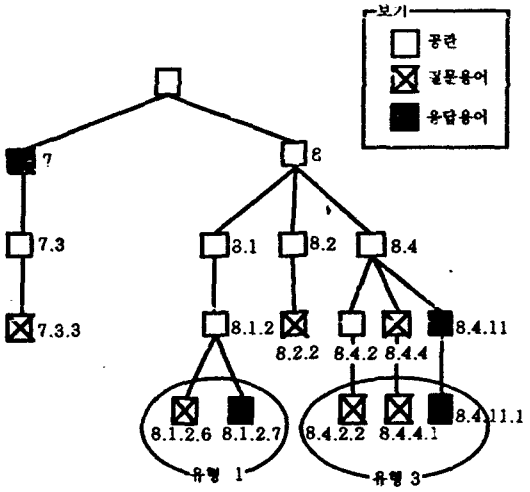
본 실험에서 시소러스의 구조 및 개념의 추출을 위해 가장 기본적으로 사용된 문헌은 『매킨토시 바이블』 3판이며, 이 문헌은 전체적인 조망에 중점을 두었기 때문에 세부적인 내용은 부차적인 문헌을 사용하여 확장하였다. 시소러스의 디스크립터는 기초문헌의 목차 및 색인에 나타난 개념을 추출하고 복합어가 위주가 되었다. 디스크립터의 관계 또한 기초문헌의 목차에 따라 BT와 NT관계만을 설정하였다.

질문/응답 분석방법은 우선 수집된 질문/응답에 존재하는 개념을 추출하여 이를 시소러스와 비교한 후 이를 기초로 질문/응답구조를 생성한다. 〈그림 1〉은 수집된 질문/응답에서 개념을 추출하여 시소러스와 비교한 후 질문/응답 구조상에 위치를 부여한 예이다.



〈그림 1〉 질문/응답 개념의 추출

이렇게 처리된 자료는 〈그림 2〉와 같이 질문응답구조도로 표현하며, 여기에서 전형적인 질문구조를 추출하여 응답이 질문구조에서 일정하게 위치하는 비율을 계산한다.



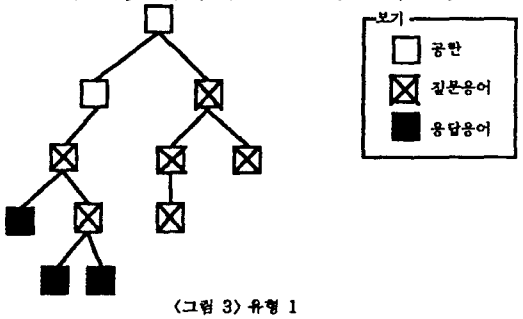
〈그림 2〉 질문응답 구조도 및 전형의 추출

2.2 질문구조 전형의 추출

실험결과 추출된 질문의 전형과 이에 따른 응답이 일정하게 위치하는 비율은 다음과 같다. 괄호안의 수치는 '해당 유형의 질문구조에서 응

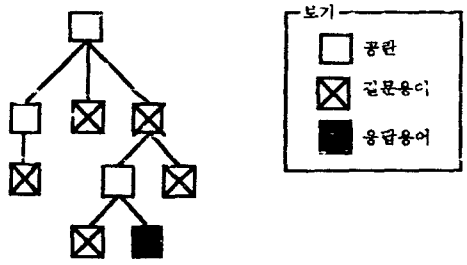
답의 위치가 일정한 질문구조의 수/전체 질문에서 해당 유형의 발생 횟수'와 '두 수치의 비율'을 각각 나타낸다.

㉠ 유형 1 : 질문개념이 두 가지 주제에 나타나고, 한쪽에 속한 개념의 수가 2개 이상 많을 때 ⇒ 응답의 위치는 개념의 수가 적은 주제의 질문 개념중 최하위 개념과 같은 수준 및 하위 수준의 개념(12/14, 86%)



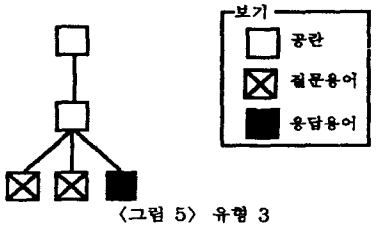
〈그림 3〉 유형 1

㉡ 유형 2 : 질문의 개념이 세 가지 이상 주제에 나타나고 한 주제에 압도적으로 많은(2개 이상) 수의 개념이 속해 있을 경우 ⇒ 응답개념은 질문개념이 많은 주제분야의 최하위 개념과 동일한 수준에 위치(4/5, 80%)



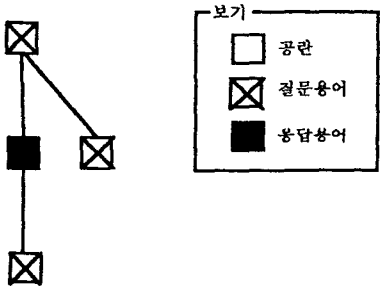
〈그림 4〉 유형 2

㉢ 유형 3 : 질문의 개념이 동일주제내 동일수준에서 2개이상 나타날 경우
⇒ 응답개념은 2개 이상의 개념이 나타난 수준에 위치 (17/25, 68%)



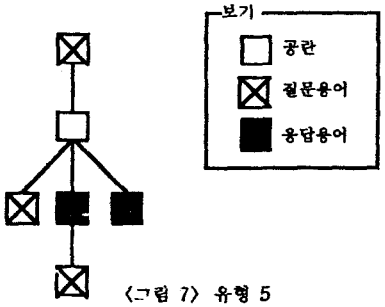
〈그림 5〉 유형 3

㉣ 유형.4 : 질문개념 중 두 개념이 할아버지-손자 관계인 경우 ⇒ 두 개념 사이에 응답 개념이 위치 (12/15, 80%)



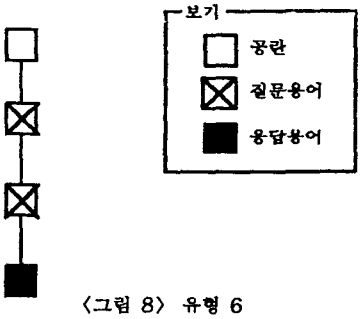
〈그림 6〉 유형 4

㉤ 유형 5 : 질문내의 개념이 특정주제군에 밀집해 있으며, 개념간의 관계가 상하로 2단계 이상 떨어져 있을 경우 ⇒ 응답용어는 두 개념 사이 단계에 위치 (5/7, 71%)



〈그림 7〉 유형 5

㉥ 유형 6 : 질문의 개념이 특정주제군에 밀집해 있으며, 특정용어와 그 아들단계의 용어가 동시에 발견되고, 아들단계의 용어가 최하위어가 아닌 경우 ⇒ 아들단계 용어의 하위수준에 응답 개념이 위치 (7/10, 70%)



〈그림 8〉 유형 6

한편 조사된 67건의 자료에서 76개의 전형적인 질문구조가 나타나 질문 1개당 평균 1.1개의 전형을 가지고 있었으며, 전형이 있는 질

문은 67건 중 57건으로 85%의 질문에서 질문 구조 전형이 나타났다. 또한 질문에 있는 전형적인 질문구조 76개 중 응답이 질문구조내에 일정하게 위치한 전형은 57개로 약 75%의 일치율을 나타내었다. 한편 질문에 나타난 주제영역에서 의미있는 총 개념의 수는 253개로 1개의 질문에서 평균 3.77개의 개념이 이용자가 자신의 정보요구를 표현하는데 쓰여졌다.

2.3 실험 결과에 대한 평가

실험 결과, 어떤 유형도 발견되지 않은 질문/응답은 몇가지 경우로 구분할 수 있는데 첫번째는 질문이 단순히 '예-아니오' 만을 묻는 경우이다. 이것은 본 실험의 질문구조 분석 방법이 갖는 한계로 질문분석시 언어적 방법을 병행하여 구문이 가지는 문제해결의 실마리를 이용해야 할 것으로 보인다. 두번째는 질문의 표현이 부족하거나, 언어적 처리가 미숙하여 질문내의 개념추출에 실패하는 경우이다. 세번째는 질문내에 추출한 개념의 수가 너무 적은 경우이다. 실제로 개념이 두개 이하인 경우 첫번째처럼 '예-아니오"식의 질문이거나, 단순히 개념의 의미를 묻는 경우가 많았다.

또한 질문에서 질문구조 전형은 발견되나 응답의 위치가 일정하지 않은 질문/응답 또한 몇가지의 경우로 구분할 수 있다. 첫번째는 질문이 내용상 두 가지를 묻고 있으나 응답이 그중 한 가지 질문에만 해당하는 경우이다. 두번째는 자료의 특성과 연관된 것으로 프로그래밍 관련 질문에서 질문자가 프로그램 코드의 일부분을 기재하였을 경우이다. 이 때, 프로그램에 있는 모든 함수 및 루틴명들이 개념으로 추출되므로 개념이 분산되어 해당 전형에 응답을 발견할 수 없었다. 세번째는 어떠한 개념이 실마리가 되어 응답을 찾아내는 경우이다. 예를 들면 통신을 통해 다운받은 파일의 확장자가 '.sit'인 경우 파일은 열리지 않으며, 초보자들은 대게 이를 통신행위에서의 문제점으로 보고 있는데, 실제로 이 문제는 '.sit' 확장자를 가진 파일은 특정 압축 프로그램으로 압축된 파일이라는데 문제의 실마리가 있다. 이러한 문제를 해결하기 위해서는 시소러스 구축시 디스크립터의 성격을 규정하는 방법이 강구되어야 할 것이다.

3. 결 론

본 실험은 이용자의 자연언어 질문을 주제지식으로 분석, 정보요구를 추론하고, 그 모형을 추출하는 것이 가능한 것인지를 타진해 보는 데에 그 목적을 두었으며, 이러한 방법의 질문 분석이 어느 정도 유용성을 가진 것으로 판단된다. 실제로 본 실험결과를 지능형 정보검색시스템에 적용하기 위해서는 시스템 실행 이전에 주제영역에 대한 이용자의 자연언어 질문과 응답을 수집하여 질문/응답구조의 전형을 추출하고 질문/응답구조 전형베이스를 구축하여야 한다. 이를 통해 시스템은 이용자와의 대화 중 탐색전 대화를 통한 이용자 지식과 이용자의 질문중 비주제적인 용어를 인터페이스와 자연언어 처리를 통해 구분하여 정보요구 모형 구축을 위한 작업영역으로 옮기고, 이용자의 질문중 주제어는 질문구조 전형을 이용해 질의어를 확장하며, 이러한 이용자의 전반적인 정보요구 상황을 통하여 이용자의 정보요구 발생시점에서의 정보요구 모형을 생성한다.

질문구조 전형을 이용한 정보요구의 모형화 기법을 실제 시스템에 적용하기 위해서는 몇가지 선결하여야 할 사항이 있는데 그 첫째는 반드시 언어학적 방법과 병행하여 질문을 처리해야 한다는 것이다. 이는 질문내에서 올바른 개념의 파악/추출과 함께, 의문사나 문장의 어미 등 질문내의 언어학적 실마리로써 단순한 사실만을 요구하는 질문을 파악할 수 있기 때문이다. 둘째는 질문분석의 틀인 시소러스의 정교화이다. 특히 개념간의 유사도와 같은 관계 정도와 디스크립터의 성격 등을 규정한 시소러스를 사용하여, 추론될 정보요구 개념의 수를 줄여서 불만족스러운 정보요구 개념이 추론되는 것을 방지하여야 할 것이다. 셋째는 질문구조 내에서 전형을 인지할 수 있는 알고리즘을 개발하여 시스템이 전체 질문구조 내에서 세부적인 전형을 검색할 수 있도록 하여야 할 것이다. 또한 여러 주제영역에 걸친 범용 시스템에서는 주제별로 나뉘어진 시소러스와 질문구조 전형 베이스가 추가되어야 하며, 이를 이용자 질문에 따라 선택적으로 적용하기 위해서 시소러스 및 질문구조 전형 베이스 선정 모듈도 추가되어야 할 것이다.

---

(1) Belkin, N. J., et. al., 1982, "ASK for Information Retrieval : Part 1. Background and Theory", *Journal of Documentation* 38(2) : 61-71.