한메일 FAQ 의 개념적 검색을 위한 계층적 브라우징 시스템

안준현, 김현돈, 조성배
연세대학교 컴퓨터과학과
e-mail : {jhahn, neoace, sbcho}@candy.yonsei.ac.kr

A Hierachical Browsing System for Conceptual Search of Hanmail FAQ

Joon-Hyun Ahn, Hyun-Don Kim and Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

컴퓨터의 보급과 함께 인터넷의 대중화로 많은 정보가 인터넷을 통해 제공되면서 많은 사람들이 정보통신 기반 서비스를 이용하게 되었다. 하지만 이런 서비스에 익숙하지 않은 사용자가 자신이 원하는 정보를 찾는 것은 그리 쉬운 일이 아니다. 그래서 ISP 나 PC 통신 업체들은 사용자들이 겪는 어려움을 해결해 주기 위한 서비스를 제공하고 있다. 그러나 사용자들의 엄청난 증가로 인해 이런 서비스를 유지하는데 많은 인력과 시간이 필요하게 되면서 질의 응답 자동화에 대한 필요성이 대두되었다. 본 논문에서는 ISP 업체 중 하나인 한메일넷의 자동 응답 시스템을 위한 FAQ 브라우징 시스템을 개발하였다. 기존의 많은 검색 서비스가 키워드들을 단순히 나열하고 이 키워드의 링크를 따라가면서 검색을 하게 하였으나 이 방식은 검색 대상에 대한 키워드 정보만을 제공하기 때문에, 문제에 대한 배경 지식이 적거나 검색 서비스 사용에 익숙치 않은 사용자가 이용하기에는 쉽지 않다. 본 시스템에서는 사용자에게 이차원상에 표현된 문서 지도를 제공해서 사용자가 쉽게 전체 검색 자료의 분포를 파악하고 검색하도록 하였다. 또한 단계별 검색이 가능하도록 해서 사용자가 효율적으로 검색할 수 있다.

1. 서론

컴퓨터의 보급과 함께 인터넷의 대중화로 많은 정보가 인터넷을 통해 제공되면서 많은 사람들이 정보통신 기반 서비스를 이용하게 되었다. 하지만 이런 서비스에 익숙하지 않은 사용자가 자신이 원하는 정보를 찾는 것은 그리 쉬운 일이 아니다. 따라서 ISP 나 PC 통신업체 등 정보통신 서비스 업체는 사용자가 접하는 문제들을 해결하기 위해서 전화상담 창구를 운영하고, FAQ 나 게시판의 형태로 유형화된 질문에 대한 답을 제공하기도 하며, 전자우편으로 사용자의 질의에 대한 답을 준다. 그러나 사용자들의 엄청난 증가로 인해 이런 서비스 제공에 많은 인력과 시간이 필요하게 되면서, 질의 응답 자동화에 대한 필요성이 대두되었다.

한메일넷의 경우, 2000 년 현재 500 만명이상의 사용자가 이용하고 있다. 하루 평균 200 통 정도의 사용자 질의를 처리하고 있는데, 이를 실시간으로 자동 응답한다면 사용자에게 만족도 높은 서비스를 제공할 수 있을 뿐만 아니라 관리자도 중복된 일을 피할 수 있으므로 효율적인 일 처리가 가능할 것이다. 따라서, 사용자와 관리자의 편의를 위해 질의 자동 응답 시스템을 개발할 필요가 있다.

본 논문에서는 한메일넷 질의 자동 응답 시스템을 위한 FAQ 브라우징 시스템을 개발하였다. 본 FAQ 브라우징 시스템은 개념적 검색을 위해서 자기구성 지도 인코딩과 색 정보를 사용한다. 자기구성 지도는 각 문서를 서로 연관된 것끼리 모아주는 역할을 하고 색 정보는 그렇게 이루어진 군집을 표시해 주어서, 사용자가 문서 지도의 각 키워드와 군집 정보를 이용해서 효율적 검색을 하게 한다. 또한 사용자가 문서 지도의

특정 영역에 대한 상세 정보를 얻고자 할 경우에는 문서 지도를 확대할 수 있게 하여 효율적 검색이 가능케 했다.

2. 배경

2.1 한메일넷 질의 자동 응답 시스템

한메일넷 질의 자동 응답 시스템은 자동 응답 시스템과 FAQ 브라우징 시스템 두 부분으로 구성이 되어 있다. 두 부분은 공통으로 한메일 사용자의 전자 우편을 통한 질의 데이터를 수집, 분석한 자료를 기반으로 시스템이 개발되었다. 자동 응답 시스템은 이 데이터를 기반으로 비슷한 질문이 들어왔을 때 이를 관리자가 수동으로 처리하는 것이 아니라, 시스템이 자동으로 답장을 보내고 필요한 경우에는 관리자가 직접 답변을 하도록 하는 시스템이다. 하루에 수백통씩 오는 질의 메일에 대한 답변을 자동으로 해 줌으로써 관리자는 반복되는 작업을 줄일 수 있고, 사용자는 빠른 답변을 받을 수 있게 된다. 본 논문에서 구현한 FAQ 브라우징 시스템은 기존의 질의 데이터를 다른 사용자가 검색하도록 하는 시스템으로 기존 검색 서비스와 달리 개념적 검색을 위한 문서 지도를 제공하고 이를 계층적으로 브라우징하도록 한다. 그림 1 은 전체 시스템의 구성을 보여준다.

그림 1. 한메일넷 사용자 질의 자동 응답 시스템 구성

2.2 단계별 검색

단계별 검색은 그림 2 와 같이 문서 지도를 검색하는데 단계별로 사용자가 원하는 영역을 확대해서 보여줌으로써 검색이 보다 효율적으로 이루어지도록 하는 것이다. 사용자는 최상위 단계에서 전체 검색 대상의 분포와 대략적인 정보를 파악하고 자신이 원하는 영역을 선택해서 상세한 정보를 얻고, 다시 자신이 원하는 영역을 선택해서 더 상세한 정보를 얻는 식으로 검색한다.

그림 2 에에 각 원은 하나의 문서를 나타내고 가운데 검은 원은 사용자가 선택한 문서, 주위의 어두운 영역은 다음 단계에서 보여질 문서들을 나타낸다. 그림에서 문서 지도는 총 7X7 의 크기로 단계가 낮아지면서 5X5, 3X3 의 확대된 문서 지도를 보여준다. 사용자는 최상위 단계에서 전체 검색 대상의 분포를 파악하고, 자신이 검색하고자 하는 영역을 선택하면 단계가 낮아지면서 보다 확대된 지도를 보게 되고 더 상세한 정보를 파악한다. 이런 식으로 전체 검색 대상을 효율적으로 파악해가면서 검색을 하게 된다.

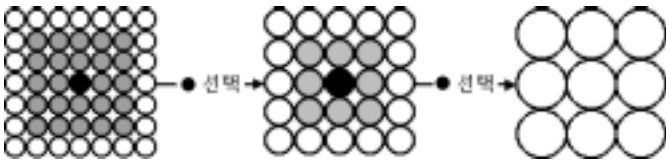

그림 2. 단계별 검색시 문서 지도의 변화

3. 시스템 개요

기존의 FAQ 브라우징은 대부분 서비스 제공자가 임의로 나눈 부류에 따라 키워드들을 나열하고 이 키워드의 링크를 쫓아가면서 정보를 찾도록 되어 있다. 이런 방법은 사용자가 문제에 대한 배경 지식이 있는 경우에는 어렵지 않게 검색을 할 수 있지만, 그렇지 않은 경우에는 효율적인 검색이 어렵다.

본 시스템은 사용자에게 단순히 키워드만을 보여주는 것이 아니라 각 문서의 특징 키워드들을 자기구성 지도[1]를 통해 인코딩해서 이차원 공간 상에 서로 연관된 것들이 모이도록 해서 각 키워드 간의 연관성을 표시한 후, 각 키워드와 영역간의 거리 정보를 색을 통해 표현해서 각 키워드를 중심으로 한 문서 군집을 제시한다. 사용자는 이 이차원 문서 지도를 통해 키워드 간의 연관도와 문서 군집 정보를 파악하고 개념적 검색을 하게 된다. 또한 보다 효율적인 검색을 위해 3 단계의 단계별 검색을 제공하고 있다. 최상위 단계에서는 150X150 크기의 전체 문서 지도를 표시하고, 중간 단계에서는 50X50 크기의 확대된 문서 지도를 보여주며, 마지막 단계에서는 중간 단계에서 선택한 영역을 중심으로 10X10 크기의 더 확대된 문서 지도를 제공하해서 사용자가 이 단계에서 선택한 영역에 대해서 답변을 보여준다.

한메일 FAQ 브라우징 시스템은 실제 한메일 사용자들의 질의 문서들을 자기구성 지도로 인코딩한 질의 문서 지도, 질의 문서 지도의 효율적 검색을 위한 브라우징 인터페이스, 사용자가 찾은 질의 클래스의 답변을 사용자의 웹브라우저에 뿌려주는 답변 서버로 이루어져 있다. 그림 3 은 전체 시스템의 구성을 보여준다.

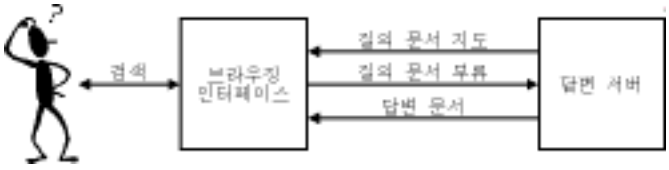

그림 3. 브라우징 시스템 전체 구성

사용자가 처음 FAQ 검색 페이지에 접속하면 웹브라우저는 브라우징 인터페이스와 질의 문서 지도를 읽어오고, 사용자는 브라우징 인터페이스를 사용해서 자신이 생각하는 질의가 있을 만한 위치를 검색하게 된다. 사용자가 최종적으로 질의 문서를 선택하면, 브라우징 인터페이스는 선택된 질의 부류를 답변 서버로 보내게 되고, 답변 서버는 해당 질의 부류에 대한 답변을 사용자의 웹브라우저로 보내준다.

4. 계층적 브라우저

4.1 질의 문서 지도

브라우징 인터페이스를 통해 검색할 문서지도를 만들기 위해서 사용자들의 질의들을 수집하고 분석할 필요가 있다. 표 1 은 한 달간 한메일넷 사용자 질의의 분포를 보여준다. 문서지도는 이 질의 문서들의 분석 자료를 토대로 각 문서의 특징을 추출해서[2] 자기구성 지도로 인코딩한 것이다. 자기구성 지도는 입력 데이터를 서로 연관된 것끼리 모아주기 때문에[1], 질의 문서들을 자기구성 지도로 인코딩을 하면 서로 연관된 질의 문서들끼리 근접한 곳에 모여서 군집을 형성한다. 따라서 사용자는 지도상에 나타난 키워드들의 위치 관계, 군집 모양 등을 고려한 개념적 검색을 하게 된다.

부류 속성	부류 개수	데이터 개수
질의의 빈도가 많은 부류	6	1002(44.9%)
개별응답 질의	7	585(26.2%)
통계적 처리가 힘든 질의	36	127(5.7%)
기타	18	518(23.2%)
계	67	2232(100.0%)

표 1. 빈도수에 따른 질의 분포

4.1.1 키워드 추출

질의 문서들은 자연 언어로 이루어져 있다. 질의 문서를 신경망의 입력으로 인코딩하기 위해서는 정규화된 벡터의 형태로 변화시켜야 한다. 벡터화를 위해서 먼저 질의에서 의미 있는 키워드만을 추출하는 작업이 필요하다. 이 과정을 통해 질의 문서는 조사나 어미 등의 문장의 의미에 영향을 미치지 못하는 불용어들과 불필요하게 반복되는 키워드들을 제거하게 된다.

4.1.2 데이터 축약

키워드 추출 과정이 끝난 후에는 키워드의 집합을 수치화된 벡터로 표현하는 작업이 필요하다. 키워드를 수치화된 벡터로 표현하는 방법에는 기본적으로는 벡터 스페이스 모델을 포함한 여러 가지 방법들이 있다. 그러나 질의 문서들은 키워드의 수가 아주 많기 때문에 기존의 방법들을 통한 인코딩은 적절하지 못하다. 그래서 자기구성 지도를 이용한 인코딩 방법을 사용하였다. 여기서 사용되는 자기구성 지도는 쉽게 말해서 동의어 사전으로서의 역할을 한다. 자기구성 지도의 입력값은 각 단어들에 대한 문맥 정보들이 된고 결과는 문맥 정보에 의해 분류된 키워드들이 된다. 이 경우, 자기구성 지도의 특징에 의해서 유사한 키워드들은 자기구성 지도의 같은 노드에 할당되거나 근접한 위치에 할당되게 된다[3]. 이렇게 완성된 데이터 축약 자기구성 지도는 입력 벡터를 생성하는데 이용된다.

자기구성 지도의 입력으로 사용되는 문맥 정보는 다음의 수식으로 표현될 수 있다[4].

$$X(i) = \begin{bmatrix} E\{x_{i-1} \mid x_i\} \\ \varepsilon x_i \\ E\{x_{i+1} \mid x_i\} \end{bmatrix} \quad (1)$$

(1)에서 키워드 x_i 에 대한 입력 벡터 X(i)는 x_i 의 선행자와 후행자의 평균으로 구성되어있다. 여기서, 선행자의 평균은 질의 자료의 모든 x_i 에 대하여 x_i 의 바로 앞에 나오는 키워드들의 벡터 값을 합하여 평균을 구한 것이다. 그리고 후행자는 자료의 모든 x_i 의 바로 뒤에 나오는 키워드들의 벡터 값을 합하여 평균을 구한 값이다. 선행자와 후행자는 x_i 에 대한 특징을 나타내는 값이다. 모든 데이터에 대해서 x_i 의 앞과 뒤에 나오는 키워드들을 살표봄으로써, 문맥 정보를 얻을 수 있다[5]. 이 방법은 기본적으로 하나의 키워드의 앞과 뒤에 나타날 수 있는 키워드들은 그 키워드의 문법적인 특징과 각 질의 메일의 내용에 따라 항상 거의 일정하게 나타난다는 가정에 의해 도출되었다. 이 입력벡터를 자기구성 지도에 입력하게 되면 결과로 문맥 정보에 의해 분류된 질의 문서 지도가 산출된다.

4.1.3 입력 벡터 생성

입력 벡터는 데이터 축약 자기구성 지도의 결과를 통해서 생성된다. 일단 데이터 축약 자기구성 지도가 만들어 진 다음에는 자기구성 지도의 각 노드에 어떤 키워드들이 매핑되어 있는 지를 알 수 있다. 이 경우, 각각의 질의 메일의 키워드들이 데이터 축약 자기구성 지도의 몇 번째에 할당되는 지를 알 수 있게 되고, 각 키워드들에 대한 히스토그램을 구할 수 있게 된다. 예를 들면, 데이터 축약 자기구성 지도가 3×3 이라고 하고, 질의 문서의 키워드들이 각각 (0,0)에 2 번 (1,2)에 1 번 나타났다고 한다면, 인코딩은

2	0	0	0	0	1	0	0	0

과 같이 된다. 앞에서와 같이 인코딩된 입력 벡터의 차원은 자기구성 지도의 크기와 같게 된다. 즉, m×n 의 자기 구성 지도에 의해 생성될 수 있는 입력 벡터는 mn 의 1 차원 벡터가 되는 것이다. 그러므로, 자기구성 지도의 크기를 어떻게 하느냐에 따라 데이터 축약이 정규화되어 이루어지게 된다. 여기서 만들어진 각각의 히스토그램은 각 질의 문서 하나하나의 고유

한 입력 벡터가 된다.

4.2 브라우징 인터페이스

기존의 단순 키워드 나열을 이용한 검색 방법이 가지는 불편함을 해소하고 일반 사용자도 쉽게 검색을 할 수 있도록 키워드, 키워드간의 위치 정보, 그리고 색을 이용해서 사용자가 보다 쉽게 검색 대상을 파악하고 키워드 간의 지리적 정보와 색 정보를 이용해서 개념적으로 검색할 수 있도록 설계하였다. 또한 기존의 시스템은 대부분 키워드의 링크를 따라가면서 검색을 하도록 되어 있는데, 링크를 통해 검색을 하게 되면 사용자는 링크를 쫓아갈 때마다 매번 웹브라우저와 웹서버간에 네트웍 연결 시간을 낭비하게 되고 웹서버 쪽에도 부담을 많이 주게 된다. 하지만 본 시스템에서는 자바 애플릿을 사용해서 시스템 초기화에 많은 시간을 사용하긴 하지만 일단 초기화되고 나면 최종 질의 문서를 선택해서 답변을 받을 때까지 서버와 분리된 상태로 검색을 할 수 있게 되어 여러 번 검색을 할 경우에는 훨씬 빠른 속도로 검색을 할 수 있다. 또한 검색 과정에서는 웹서버와의 연결없이 진행되므로 서비스를 제공하는 측면에서 웹서버의 부담이 많이 준다.

브라우징 인터페이스는 사용자가 검색 화면에 접속하면 질의 문서 지도와 답변 서버 정보를 읽어들이고, 질의 문서 지도의 내용을 읽어오고, 이 내용을 기반으로 문서 지도 화면을 그리게 된다. 사용자는 이 문서 지도 화면을 이용해서 검색을 하면서 자신 원하는 답변을 찾게 된다. 그림 4 는 브라우징 인터페이스의 순서도를 보여준다.

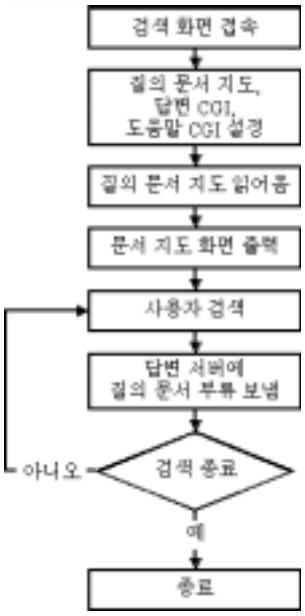

그림 4. 브라우징 인터페이스 순서도

화면구성은 그림 5 와 같이 문서 지도 부분과 답변 출력창 부분으로 나뉘어져 있다. 문서 지도 부분은 자기구성 지도를 통해 인코딩된 질의 문서의 지도를 시각화 한 것으로 사용자가 검색을 하는 영역이고 답변 출력창은 사용자가 마지막 단계에서 선택한 영역에 대한 질의의 답변을 보여준다.

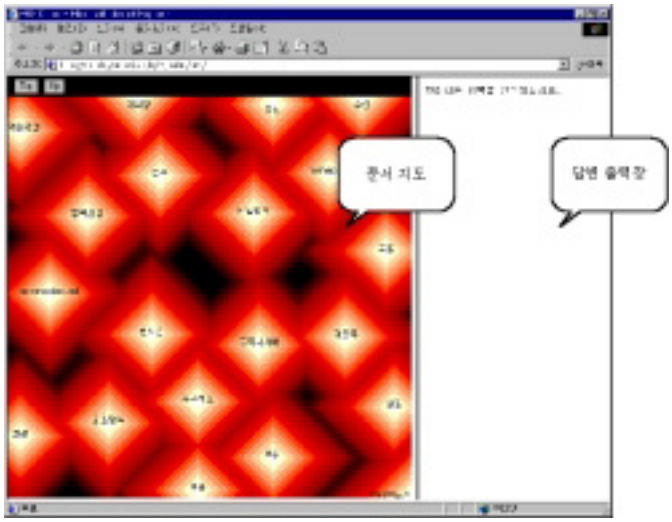

그림 5. 브라우징 인터페이스

그림 6~8 은 브라우징 인터페이스의 단계별 검색 화면을 보여준다. 그림 6 은 문서 지도의 최상위 단계로 150×150 크기의 지도에 전체 질의 문서 데이터 중에서 가장 큰 부류를 이루는 키워드들과 각 키워드를 중심으로 한 군집을 색을 이용해서 표현하였다. 이 단계에서는 전체 질의 문서의 분포를 파악하고 각 키워드와 군집 정보를 보고 개념적으로 자신이 찾고자하는 정보가 있을 것 같은 영역을 선택하게 된다. 그러면 시스템은 문서 지도 중간 단계로 넘어가면서 그림 7 과 같은 화면을 보여준다. 문서 지도 중간 단계는 최상위 단계에서 사용자가 선택한 영역을 중심으로 50×50 크기의 지도를 보여준다. 최상위 단계보다 확대된 문서 지도를 통해 상세한 정보를 보여주게 된다. 여기서 다시 사용자가 특정 영역을 선택하면 그림 8 과 같이 문서 지도 마지막 단계로 넘어간다. 마지막 단계에서는 중간 단계에서 선택한 영역을 중심으로 10×10 크기의 문서 지도를 보여주고, 질의 문서가 존재하는 모든 영역의 대표 키워드값을 보여준다. 사용자가 이 키워드를 선택하면 관련 답변이 출력된다.

키워드는 각 영역의 부류를 가장 잘 대표할 수 있어야 한다. 기본적으로 각 영역에 매핑된 질의 문서들의 키워드의 빈도수를 통해 대표 키워드를 계산한다. 그러나 빈도수가 같다고 하더라도 각 키워드가 가지는 의미는 달라 질 수 있다. 즉 어떤 키워드는 단 한 번 나타나더라도 하나의 부류의 특징을 나타낼 수 있는데, 어떤 키워드는 빈도수는 높지만 그 부류의 특징을 잘 나타내지 못하는 경우가 있다. 그래서 빈도수와 함께 (2)에서 구한 키워드 w 의 중요도 값을 계산해서 각 영역의 레이블 값을 계산하였다.

$$I_w = F \times \frac{DocFin}{DocFout} \qquad (2)$$

여기서, I_w는 키워드 w 의 중요도값, F 는 노드 내의 w 의 빈도수, DocFin 은 노드 내의 w 를 가진 질의 문서의 수, DocFout 은 노드 밖의 질의 문서 중에서 w 를 가지는 문서의 수를 각각 나타낸다.

위 중요도 값은 각 단계별로 군집 영역의 크기를 정해서 각 영역 내에서 계산을 하게 되고 각 영역에서 가장 큰 중요도값을 가진 키워드가 화면에 보여지게 된다. 단계별 군집 영역의 크기는 최상위 단계 30X30, 중간 단계 15X15, 마지막 단계에서는 1X1 의 크기로 정하였다.

색 정보는 키워드값과 해당 영역의 거리 정보를 나타내는데 밝은 영역일수록 키워드와 가까이 있고 연관성이 높은 문서들을 나타내고 어두운 영역일수록 키워드와 멀리 떨어져 있는 연관성이 낮은 문서들이 있음을 나타낸다. 또한 지도 상의 검은 영역은 각 군집간의 경계를 나타낸다.

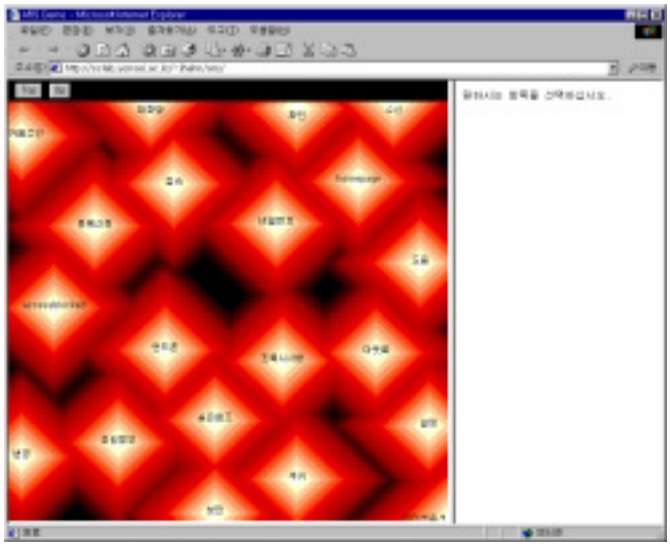

그림 6. 문서 지도 최상위 단계

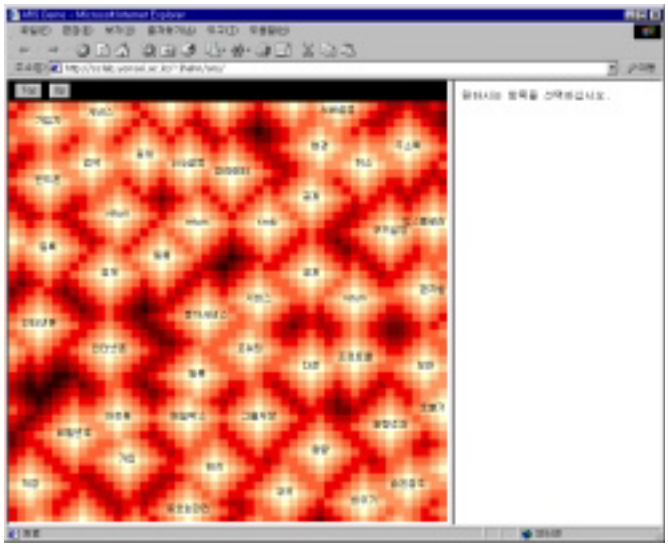

그림 7. 문서 지도 중간 단계

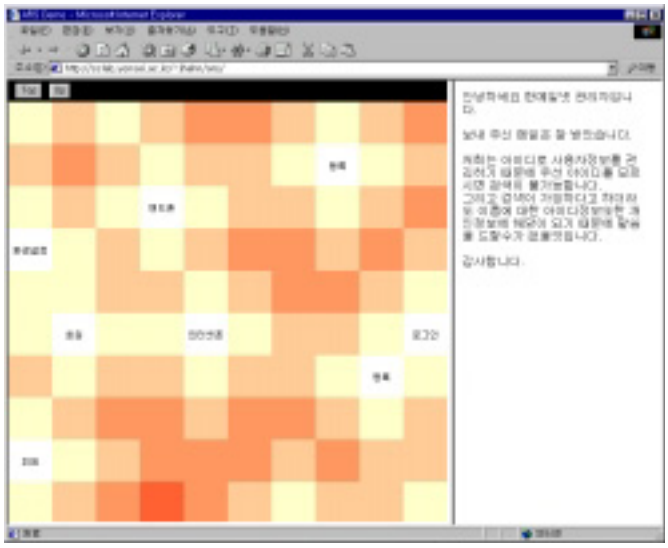

그림 8. 문서 지도 마지막 단계

4.3 답변 서버

답변 서버는 사용자가 검색한 영역의 질의 문서 부류의 답변을 보여준다. 사용자가 문서 지도 마지막 단계에서 선택한 질의 문서 부류 정보를 받으면 사용자가 보낸 데이터 중에서 질의 문서 부류를 파싱하고 이 부류에 대한 답변이 존재하는지를 검사한다. 만약 답변이 존재하지 않으면 관리자에게 직접 문의하라는 도움말을 보여주고 그렇지 않으면 해당 답변을 보여준다.

이 CGI 는 Perl[6]을 이용해서 작성하였는데, Perl 은 문자열 처리가 쉽고 CGI 를 작성하는데도 용이하기 때문에 사용하게 되었다.

그림 9 는 답변 서버의 전체 순서도를 보여준다.

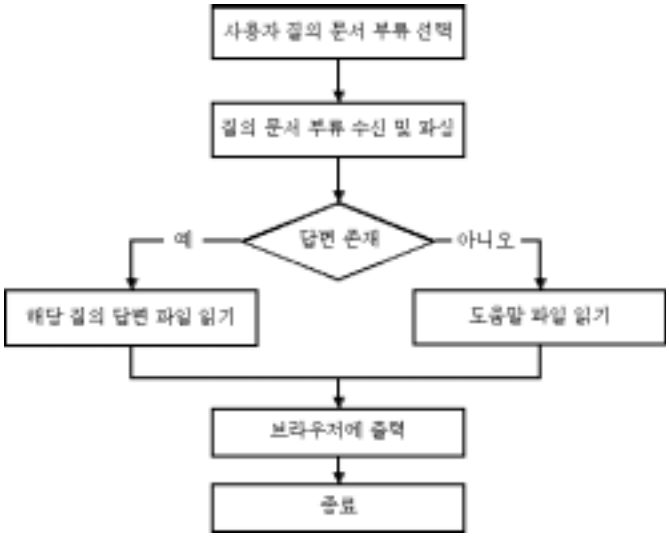

그림 9. 답변 서버 순서도

5. 결론

본 논문에서는 기존의 단순한 키워드 나열을 통한 검색이 아니라 사용자가 문서의 전체적인 분포를 파악하고 능동적, 개념적 검색을 할 수 있는 계층적 브라우징 인터페이스를 설계하고 구현해 보았다. 또한 기존 시스템과 달리 자바 애플릿을 사용해 인터페이스를 구현함으로써 초기 시스템 구동 시간은 좀 많이 걸리지만 검색 시에는 서버와의 접속 없이 검색을 할 수 있게 되어서 여러 번 검색을 할 경우에 좋은 성능을 보일 수 있을 것이다. 또한 FAQ 와 같이 서비스에 익숙하지 않는 사용자나 검색 데이터의 양이 방대하지 않은 경우에 좋은 성능을 기대할 수 있다.

참고문헌

[1] T. Kohonen, “Self-organization of very large document collections: State of the art,” In Niklasson, L., Bodén, M., and Ziemke, T., editors, *Proc. of Int.l Conf. on Artificial Neural Networks,* vol. 1, pp. 65-74. Springer, London, 1998.

[2] K. Lagus and S. Kaski, “Keyword selection method for characterizing text document maps,” *Proc. of Int.l Conf. on Artificial Neural Networks,* vol. 1, pp. 371-376, IEE, London 1999.

[3] G. Salton, *Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer,* Addison-Wesley, 1988.

[4] S. Kaski, T. Honkela, Krista Lagus, T. Kohonen, “Creating an order in digital libraries with self-organizing maps,” World Congress on Neural Networks, pp. 814-817, 1996.

[5] H. Ritter, T. Kohonen, “Self-organizing semantic maps,” *Biol Cyb,* 61:241-254, 1989.

[6] L. Wall, T. Christiansen, R. L. Schwartz, and S. Potter, *Programming Perl,* 2nd Edition, 1996.