한메일넷 질의 자동응답을 위한 이단계 자기구성 지도

김 현돈, 조 성배
연세대학교 컴퓨터과학과

A Two-level Self-Organizing Map for Automatic Response of Hanmail Net Questions

Hyun-Don Kim and Sung-Bae Cho
Computer Science Department, Yonsei University

요 약

컴퓨터가 널리 보급되고 인터넷이 발전함에 따라 많은 정보가 생산되고, 이러한 정보를 가공하여 사용자에게 효율적으로 제공하는 서비스들도 많아지게 되었다. 그러나, 컴퓨터에 익숙하지 않은 사용자들은 쉽게 이러한 서비스를 이용하지 못하기 때문에 사용자들을 돕는 시스템들이 필요하게 되었다. 한메일넷의 경우 전자 우편을 통한 사용자들의 질문에 대해 관리자가 직접 답을 해주는데, 사용자의 증가로 질의응답 업무의 양이 커지고 있다. 따라서, 본 논문에서는 사용자의 질의에 자동으로 응답하는 시스템을 개발하기 위하여 효율적인 이단계 자기구성 지도(SOM)를 제안한다. 이 방법은 다양한 크기의 질의메일을 정형화된 크기로 만들기 위한 데이터 축약 SOM과 이를 실제 해당 답변 클래스로 분류하는 문서 분류 SOM으로 구성된다. 실제 사용되고 있는 2206개의 데이터에 대한 실험 결과, 95%의 분류 성공률을 보여 그 가능성을 볼 수 있었다.

1. 서 론

컴퓨터와 PC통신의 폭넓은 보급과 인터넷에 대한 관심의 확산이 많은 사람들을 정보통신 기반 서비스로 끌어들이고 있다. 그러나, 처음 정보통신을 접하는 사람이 자신이 원하는 바를 얻는 것은 그리 쉽지 않은 일이다. 필요한 프로그램 설치로부터 특정한 기능의 사용방법 등 여러 가지 예기치 못한 상황들이 기다리고 있으며, 이를 혼자만의 힘으로 헤쳐나가는 것은 초보자에게는 물론 어느 정도 익숙한 사람에게도 어려운 일이거나 많은 시간을 요하는 일이다. ISP와 PC통신업체 등 정보통신 서비스 업체는 이러한 문제를 해결하기 위해서 전화상담 창구를 운영하고, FAQ나 게시판의 형태로 유형화된 질문에 대한 답을 제공하기도 하며, 전자우편으로 사용자의 질문에 대한 응답을 준다. 그러나 사용자들의 폭발적인 증가로 인해 답을 제공하는 서비스는 많은 인력을 필요로 하게 되어서, 질의응답의 자동화에 대한 필요성이 대두되었다.

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

본 논문에서는 한메일넷의 사용자 질의 메일을 자동으로 응답하기 위한 이 단계 자기구성 지도(SOM)를 제안한다. 이러한 시스템을 통해 클래스별로 데이터의 개수가 다르고 각 데이터 크기도 다른 전자우편의 자동 응답 가능성을 보이고자 한다.

2. 한메일넷의 사용자 질의

질의 메일들을 분류하기 위해서는 먼저 사용자들의 질의들을 수집하여 분석할 필요가 있다. 표 1은 한 달간 한메일넷 사용자 질의의 분포를 보여준다. 질의의 빈도가 많은 부류가 빈도수의 절반을 차지하므로 집중적으로 학습시켜 분류율을 최대화 시켜야 할 것이다. 개별응답 질의는 분류할 필요없이 관리자에게 포워딩해야할 클래스들이다. 그리고 통계적으로 처리하기 힘든 질의는 빈도수가 너무 적은 클래스들이다.

부류 속성	부류 개수	데이터 개수
질의의 빈도가 많은 부류	6	1002 (44.9%)
개별응답 질의	7	585 (26.2%)
통계적 처리하기 힘든 질의	36	127 (5.7%)
기타	18	518 (23.2%)
계	69	2232 (100.0%)

표 1. 빈도수에 따른 질의 분포

한메일넷 질의의 특징은 다음과 같다. 먼저, 표 1에서 볼 수 있듯이

빈도수가 높은 특정 부류에 사용자의 질의가 편중되는 경향이 있다. 또한, 부류가 정의되지 않았거나 한메일넷 사용과는 관련없는 다양한 질의가 존재하므로 이들의 패턴 추출에 어려움이 있다. 그리고 일반 사용자들이 작성하므로 통신상의 속어나 약어, 맞춤법에 맞지 않는 표현을 많이 포함하고 있다.

3. 한메일넷 질의 자동응답

시스템은 두 부분으로 구성된다. 첫 번째는 전처리라고 할 수 있는 키워드 추출 부분과 문서 분류 SOM의 입력 벡터를 인코딩하는 데이터 축약 부분이다. 두 번째는 문서 분류 SOM을 통해 질의 메일을 분류하여 적절한 답변 메일과 매칭시키는 부분이다. 그림 1은 시스템 구조를 보여준다.

3.1 키워드 추출

질의 메일들은 자연 언어로 이루어져 있다. 질의 메일을 신경망의 입력으로 인코딩하기 위해서는 정규화된 벡터의 형태로 변환해야 한다. 벡터화를 위해서 먼저 질의에서 의미 있는 키워드만을 추출하는 작업이 필요하다. 이 과정을 통해 질의 메일은 조사나 어미 등 문장의 의미에 영향을 미치지 못하는 불용어들과 불필요하게 반복되는 키워드들을 제거한다. 키워드 추출의 예는 그림 2와 같다.

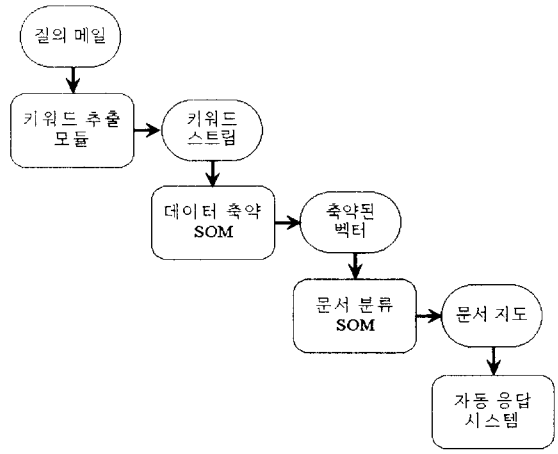

그림 1. 시스템 구조도

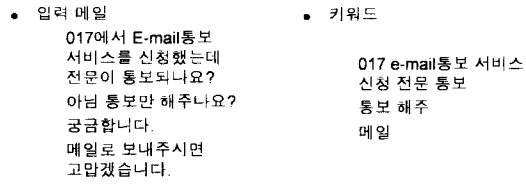

그림 2. 키워드 추출의 예

3.2 데이터 축약 SOM

그림 2와 같이 키워드 추출 과정이 끝난 후에는 키워드의 집합을 수치화된 벡터로 표현하는 작업이 필요하다. 키워드를 수치화된 벡터로 표현하는 방법에는 벡터 스페이스 모델을 포함한 여러 가지 방법들이 있다[1]. 그러나 질의메일들은 키워드의 수가 아주 많기 때문에 기존의 방법들을 통한 인코딩은 적절하지 못하다. 그래서 SOM을 이용한 인코딩 방법을 사용하였다. 여기서 사용되는 자기 구성지도는 쉽게 말해서 동의어 사전으로서의 역할을 한다. 자기구성 지도의 입력값은 각 단어들에 대한 문맥 정보들이 되고 결과는 문맥 정보에 의해 분류된 키워드들이 된다. 이 경우, SOM의 특징에 의해서 유사한 키워드들은 SOM의 같은 노드에 할당되거나 근접한 위치에 할당된다[2,3]. SOM의 학습은 수식 (1)에 의해서 수행된다.

$$m_i(t+1) = m_i(t) + \alpha(t) \times n_{ci}(t) \times \{x(t) - m_i(t)\} \quad (1)$$

여기서 $\alpha(t)$는 학습률을 나타내는 함수, $n_{ci}(t)$는 이웃 함수, $m_i(t)$는 노드의 가중치, $x(t)$는 입력 벡터값이다[4]. $n_{ci}(t)$에서 c는 승리자 노드의 인덱스인데, 승리자는 다음의 수식으로 얻을 수 있다[2].

$$\|x - m_c\| = \min_i \{\|x - m_i\|\} \quad (2)$$

자기 구성지도의 입력으로 다음과 같이 문맥 정보가 이용된다[3].

$$X(i) = \begin{bmatrix} E\{x_{i-1} | x_i\} \\ \varepsilon x_i \\ E\{x_{i+1} | x_i\} \end{bmatrix} \quad (3)$$

수식 (3)에서 키워드 x_i에 대한 입력 벡터 $X(i)$는 x_i의 선행자와 후행자의 평균으로 구성되어있다. 여기서, 선행자의 평균은 질의 자료의 모든 x_i에 대하여 x_i의 바로 앞에 나오는 키워드들의 벡터 값을 합하여 평균을 구한 값이다. 그리고 후행자는 자료의 모든 x_i의 바로 뒤에 나오는 키워드들의 벡터 값을 합하여 평균을 구한 값이다. 선행자와 후행자는 x_i에 대한 특징을 나타내는 값이다. 모든 데이터에 대해서 x_i의 앞과 뒤에 나오는 키워드들을 살펴봄으로써, 문맥 정보를 얻을 수 있다[4]. 이 입력벡터를 SOM에 입력하여 문맥 정보에 의해 분류된 키워드 지도를 얻을 수 있다.

3.3 문서 분류 SOM

일단 데이터 축약 SOM이 만들어 진 다음에는 SOM의 각 노드에 어떤 키워드들이 매핑되어 있는 지를 알 수 있다. 즉, 각 질의 메일의 키워드들이 데이터 축약 SOM의 몇 번째에 해당되는 지 알 수 있게 되고, 각 키워드들에 대한 히스토그램을 구할 수 있게 된다. 예를 들면, 데이터 축약 SOM이 3×3이라고 하고, 질의메일의 키워드들이 각각 (0,0)에 2번, (1,2)에 1번 나타났다고 한다면, 인코딩은

2	0	0	0	0	1	0	0	0

과 같이 된다.

앞에서와 같이 인코딩된 입력 벡터의 차원은 SOM의 크기와 같게 된다. 즉, $m \times n$의 SOM에 의해 생성될 수 있는 입력 벡터는 mn 의 1차원 벡터가 되는 것이다. 그러므로, SOM의 크기를 어떻게 하느냐에 따라 데이터 축약의 차원이 결정되는데, 이렇게 만들어진 히스토그램은 데이터에 민감하다는 단점을 가지고 있다. 질의 응답의 경우,

사람마다 질의하는 방식이 틀리기 때문에 유사하지만 편차를 가지는데, 단순히 빈도수로 히스토그램을 정수화 시킬 경우 특정 경우에 대한 값만을 반영할 수도 있다. 그래서 패턴 인식에서 많이 쓰이는 가우시안 커널[5]로 블러링(bluring)하였다. 사용된 가우시안 커널은 다음과 같다.

0.25	0.5	0.25
0.5	1	0.5
0.25	0.5	0.25

또한 입력 벡터의 분류율을 향상시키기 위해서 키워드에 대한 빈도수에 각 키워드가 얼마나 분류에 중요한 지를 tishs의 엔트로피(Shannon's Entropy)[1]를 통해 보완하였다.

$$V_i = \sum_w (F_w \times \frac{G_i}{E_w}) \quad (4)$$

여기서 V_i는 i번째 벡터값을, G_i는 i 번째 벡터의 커널 값을, F_w 는 키워드 w의 빈도수를, E_w 는 w에 대한 tishs 엔트로피 값을 나타낸다. 블러링과 엔터로피 값에 의한 보완을 통해 만들어진 각각의 히스토그램은 각 질의메일 하나 하나의 고유 입력 벡터가 된다.

입력 벡터를 문서 분류 SOM으로 학습하면, 각 질의메일은 SOM의 특정 노드에 매핑된다. 이 때 SOM의 특성에 의하여, 같은 클래스의 질의들이 같은 노드나 근접한 위치에 있는 노드에 매핑될 것이다. 이렇게 학습된 SOM에 새로이 사용자가 보낸 질의메일을 인코딩하여 입력하면 그 질의 메일이 어떤 노드에 매핑되는 지를 알 수 있게되어, 그 노드가 나타내는 클래스에 대한 답변 메일을 전송하면 된다.

4. 실험 결과

실험은 2232개의 전체 데이터 중에서 분류가 필요없는 2개 클래스, 26개의 데이터를 제외한 67개 클래스, 2206개의 한메일넷 질의 메일을 통해서 이루어 졌다. 데이터 축약 SOM의 크기는 10×10으로 하였다. 즉, 문서 분류 SOM의 입력 벡터는 차원이 100이다. 먼저 2206개의 질의 메일 전체를 학습 데이터로 사용하여, 문서 분류 SOM의 크기를 구하는 작업이 이루어 졌는데, 그 결과는 표 2와 같다.

지도 크기	인식률	
100×100	1553/2206	70.40%
120×120	2014/2206	93.74%
150×150	2098/2206	95.01%
160×160	2075/2206	94.06%

표 2. 학습 데이터에 대한 인식률

표 2에서 나타난 인식률을 150×150의 SOM을 통해 클래스 별로 분석하면 표 3과 같은 결과를 얻을 수 있다. 질의의 빈도수가 많은 부류의 인식률과 기타의 인식률이 낮음을 알 수 있다. 그 이유는 빈도수가 많은 부류에는 특정한 4개의 클래스가 낮은 인식율을 보였기 때문이다. 이것은 한메일넷의 아이디에 관련된 질의인데, 세부적으로 나누면, 아이디 변경, 아이디 삭제, 아이디 중복, 아이디 확인요청 등이다. 내용이 유사하여 키워드로 특징을 추출해내는 것이 힘들기 때문에 분류가 잘 되지 않았다. 그리고 기타 클래스는 16개인데, 나머지 51개 클래스와 연관된 내용을 가지고 있기 때문에 분류가 잘 되지 않았다.

클래스	클래스 수	인식률	
질의의 빈도가 많은 부류	6	954/1002	94.0%
개별응답 질의	7	561/585	95.9%
통계적으로 처리하기 힘든 질의	36	124/127	97.6%
기타	16	459/492	93.3%

표 3. 클래스별 인식률

문서 분류 SOM의 크기를 150×150으로 하여 학습 데이터와 테스트 데이터로 나누어서 두 번째 실험을 하였다. 학습 데이터는 1545개로 하였고 테스트 데이터는 661개로 하였다. 실험은 양자화 오류(Quantization Error)의 임계치 변화에 따라 진행되었다. 기각률은 양자화 오류값에 의해 결정된다. 실험 결과는 표 4와 같이 나타났다.

임계치	기각률		인식률	
0.5	307/661	46.4%	207/354	58.5%
0.4	397/661	60.1%	169/264	64.0%
0.3	482/661	72.9%	148/179	82.7%

표 4. 테스트 데이터의 인식률

4. 결 론

본 논문에서는 실험 결과에서 나타난 바와 같이 학습 데이터에 대해서는 높은 분류율을 보이는 반면 테스터 데이터에 대해서는 분류율은 받아 들일만하지만 기각률이 너무 크다. 향후 연구를 통해서 이점에 대한 보완이 필요하다. 그리고, 학습된 SOM을 이용한 브라우징 시스템을 구축한다면 사용자가 친숙하게 자신의 질의를 검색할 수 있고, 또 검색된 결과 역시 학습 데이터의 분류율 만큼 정확도를 가지므로 효율적일 것이다.

참고문헌

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

[2] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin Heidelberg, 1995.

[3] S. Kaski, T. Honkela, K. Lagus, T. Kohonen, "Creating an order in digital libraries with self-organizing maps," *Proc. World Congress on Neural Networks*, pp. 814-817, 1996.

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

[5] E. Gose, R. Johnsonbaugh, S. Jost, *Pattern Recognition and Image Anaylysis, Prentice Hall PTR*, 1996.