다중 신경망을 이용한 한메일넷 질의 자동분류 시스템

이 지행, 조 성배
연세대학교 컴퓨터과학과

An Automatic Classification System for Hanmail Net Questions Using Multiple Neural Networks

Jee-Haeng Lee and Sung-Bae Cho
Computer Science Department, Yonsei University

요 약

최근들어 정보의 양이 날로 방대해 짐에 따라 이를 자동으로 분류해 줄 수 있는 문서 자동분류의 중요성이 널리 인식되고 있다. 문서 자동분류는 새로운 문서를 미리 정의된 부류로 대응시키는 일련의 작업을 말하며, 각종 패턴인식 기법들을 이용하여 시도되고 있다. 본 논문에서는 수많은 사용자들의 질의들을 분류하여 자동으로 응답하는 시스템에 적용할 수 있는 자동 질의 분류시스템을 제안한다. 실험은 500만명 이상이 사용하고 있는 한메일넷의 실제 사용자 질의를 수집하여 수행하였으며, 자동분류 방법으로는 다중 신경망을 이용하였다. 또한 효율적인 특징추출 기법과 결과 결합방법을 적용하여 분류의 정확율을 높이고자 하였다. 2204개의 실제 질의메일에 대한 실험결과, 91.1%까지의 정확율을 얻어 제안한 시스템이 실제 한메일넷의 자동응답 시스템에 효과적으로 적용될 수 있음을 알 수 있었다.

1. 서론

컴퓨터가 널리 보급되고 인터넷이 발전함에 따라 수없이 많은 정보들이 생산되고 있다. 이러한 정보를 사람이 일일이 가공하고 분류하기에는 한계가 있으므로, 정보검색이론을 바탕으로 한 문서 자동분류의 중요성이 널리 인식되고 있다[1,2,3]. 새로운 뉴스를 분류하거나[1], 회사로 들어오는 방대한 양의 이메일을 관련부서로 라우팅시키는 시스템[2] 등이 그 좋은 예라 할 수 있다. 최근의 문서 자동분류는 k-Nearest Neighborhood, Decision Tree, Naive Bayesian 등의 확률적 모델, 해석적 방법론, 신경망 등의 패턴인식 기법이 이용되어 그 성능을 입증받고 있다[3].

본 논문에서는 다중 신경망을 이용하여 (주)다음커뮤니케이션에서 제공하는 한메일넷의 사용자 질의 자동분류를 수행하였다. 사용자들의 질의 데이터를 이용하여 실험한 결과는 실제 한메일넷의 자동응답 시스템에 효과적으로 적용될 수 있음을 보여준다.

2. 배경

2.1 질의 자동분류

문서 자동분류는 새로운 문서를 미리 정의된 부류로 대응시키는 일련의 작업을 말한다[4]. 따라서, 문서 자동분류시스템은 미리 수집한 문서집합을 이용한 패턴학습과 새로운 문서에 대한 분류를 수행하는 과정이 필요하다.

이와같은 관점에서 사용자 질의 자동분류 문제는 문서 자동분류의 하나로 간주될 수 있다. 사용자들의 질의들을 수집하여 분류한 후 학습과정을 거치면, 새로운 사용자의 질의를 미리 정의된 하나 또는 하나 이상의 부류로 대응시킬 수 있다. 이와같은 시스템은 관련 답변들과 담당자가 처리해야할 질의 등으로 분류하여, 사용자는 즉각적인 응답을 받을 수 있으며 시스템 운영의 효율을 높일 수 있다는 장점이 있다.

2.2 한메일넷 사용자 질의 분석

한메일넷은 (주)다음커뮤니케이션에서 제공하는 포탈 시스템의 이름이다. 2000년 2월 현재 500만명 이상의 사용자가 이용하고 있어서, 운영자가 일일이 사용자들의 이용관련 질의에 답변하기에는 많은 어려움이 있다. 표 1은 한달간의 빈도수에 따른 한메일넷 사용자질의의 분포를 보여준다.

부류 속성	부류 개수	데이터 개수
빈도가 많은 질의	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

한메일넷 사용자 질의 자동분류 문제는 기존의 문서 자동분류 문제와는 몇가지 다른 특징을 가진다. 첫째, 사용자에게 정확한 답변을 해야하므로 재현율보다는 정확율에 초점을 맞추어야 하는 문제이다. 둘째, 표 1에서 볼 수 있듯이 빈도수가 높은 특정 부류에 사용자의 질

의가 편중되는 경향이 있다. 셋째, 부류가 정의되지 않았거나 한메일넷 사용과는 관련없는 다양한 질의가 존재하므로 이들의 패턴 추출에 여려움이 있다. 마지막으로 일반 사용자들이 작성하므로 통신상의 속어나 약어, 맞춤법에 맞지 않는 표현이 많이 포함된다는 점이 있다.

3. 다중 신경망을 이용한 질의 자동분류

한메일넷 질의의 자동분류를 위하여 본 논문에서는 다중 신경망을 이용한 분류를 수행하였다. 사용자의 질의는 전처리 과정을 거쳐 의미있는 키워드의 집합으로 분류 시스템에 전달된다. 특징추출 모듈을 통해 정규화된 다차원 벡터는 각 신경망 분류기의 입력으로 쓰이며, 각 분류기의 결과를 결합하여 최종적인 결과를 도출한다. 그림 1은 전체적인 시스템의 구성을 보여준다.

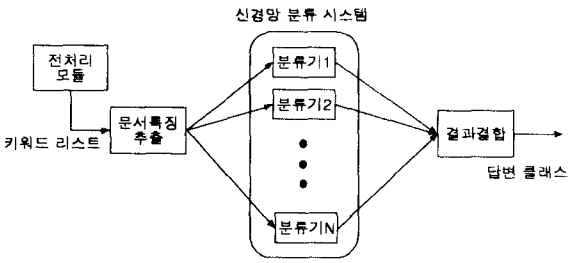

그림 1. 시스템 구성

3.1 특징추출

3.1.1 전처리

전처리 과정에서는 사용자의 질의를 의미있는 키워드의 집합으로 추출한다. 형태소 분석을 거치지 않고, 키워드의 사전을 미리 만들어 대응되는 단어를 추출해 내는 방법을 사용하였다. 이때 한메일넷 질의의 분류에 중요한 단서가 될 수 있는 부사나 형용사 등도 추가된다. 또한, 통신상의 속어나 약어, 동의어에 대한 사전을 유지하여 정규화시키며, 맞춤법에 맞지 않더라도 문서분류에 중요한 키워드일 경우 추출해 내는 기능을 가지고 있다.

3.1.2 Vector Space 모델

키워드의 집합을 패턴인식을 위한 수치화된 벡터로 표현하는 모듈이다. 정보검색의 대표적인 방법론이며, 대부분의 정보검색 시스템이나 문서분류 시스템에서 기본적인 특징추출 방법으로 사용하고 있다[4,5]. 이 방법에서 문서 i의 j번째 키워드의 가중치는 다음과 같이 표현된다.

$$w_{ij} = tf_{ij}\log(N/df_j) \quad (1)$$

여기서 tf_{ij}는 j번째 키워드의 문서 i에서의 가중치이며, df_j는 키워드 j가 전체 문서집합에서 나타나는 문서의 개수이다. (1)에서와 같이 Vector Space 모델은 문서에 자주 나타나면서 다른 문서에 나타나는 빈도가 적은 키워드에 가중치를 두어 문서의 특징으로 표현한다.

3.1.3 특징축약

사전에 등록된 모든 키워드를 분류에 이용할 경우 벡터의 차원이 너무 커지게 되어 패턴인식의 성능 및 속도의 저하를 가져온다. 본 논문에서는 χ^2-statics 방법을 이용하여 각 부류별로 중요한 키워드를 추출하고 분류에 잡음으로 작용하는 키워드를 제거하였다.

χ^2-statics 방법은 일반적으로 계산량이 적고, 성능이 우수한 특징축약 방법론으로 분서분류 시스템에 널리 쓰이고 있다[6,7]. 특정 부류 내의 키워드 하나의 부류 중요도는 다음과 같이 표현된다.

$$\chi^2 = \frac{N(N_{r+}N_{n-} - N_{r-}N_{n+})^2}{(N_{r+}+N_{r-})(N_{n+}+N_{n-})(N_{r+}+N_{n+})(N_{r-}+N_{n-})} \quad (2)$$

여기서 N은 문서의 개수이며 r과 n은 문서가 부류내에 속하는지 아닌지를, +와 -는 그 키워드가 문서내에 존재하는지 아닌지를 나타낸다. 본 논문에서는 각 부류별로 상위의 키워드를 특징으로 사용하며, 선택된 키워드의 가중치는 신경망의 입력으로 사용된다.

3.2 분류 시스템

3.2.1 다중 신경망 모듈

본 논문은 문서분류를 위한 인식기로 다중 신경망 시스템을 사용하였다. 신경망 시스템은 패턴인식 문제에서 널리 쓰이고 있는 분류기로 일반화 능력이 높으며 안정적인 학습 알고리즘을 가지고 있다[8]. 하지만, 한메일넷 질의분류 문제의 경우, 불균등한 부류별 빈도수 분포와 많은 잡음 문서가 존재하므로 하나의 신경망으로 전체 분류를 수행하기에는 많은 문제가 따른다. 따라서 각 부류를 인식할 수 있는 신경망을 구성하여 학습한 후, 그 결과를 결합하는 방식의 모듈형 시스템을 구성하였다.

따라서 각 신경망은 해당 부류에 속하는지 아닌지의 출력을 내게 된다. 하지만 이렇게 신경망을 구성할 경우 식(3)에서와 같이 특정 부류 에 속하는 데이터 개수 N_{pos}와 속하지 않는 데이터 개수 N_{neg}의 심한 불균형을 초래하게 된다. 이와같은 데이터의 불균형은 신경망 학습의 성능 저하를 가져온다고 알려져 있다[9].

$$\frac{N_{neg}}{N_{pos}} = \frac{N - N_{pos}}{N_{pos}} \gg 1 \quad (\because N \gg N_{pos}) \quad (3)$$

본 논문에서는 식 (4)와 같이 N_{pos}와 N_{neg}의 차만큼 특정 부류에 속하는 데이터를 추가로 임의 추출하여 데이터의 균형을 맞추는 샘플링 기법을 사용하였다.

$$\frac{N_{neg}}{N_{pos}'} = \frac{N_{neg}}{N_{pos} + (N_{neg} - N_{pos})} = 1 \quad (4)$$

이와같은 샘플링 기법은 학습 데이터의 개수가 커지게 되어 학습속도의 저하를 가져올 수 있지만, 모든 데이터를 사용하여 신경망의 정확율을 높인다는 관점에서 의의가 있다.

3.2.2 결합 모듈

각 신경망에서 출력된 결과를 종합하여 최종적으로 분류하는 모듈이다. 결합 모듈은 기본적으로 가장 큰 출력을 낸 신경망으로 분류결과를 결정하는 방법을 사용하였다. 또한 한개 이상의 분류결과를 출력하는 다중분류, 기각 임계치를 이용한 분류기각 방법 등을 이용하여 분류의 정확율을 향상시킨다.

4. 실험결과

4.1 실험환경

실험은 약 한달간 수집된 질의 2204개를 대상으로 이루어 졌으며, 질의 집합중 임의로 선택하여 1718개의 문서를 학습문서로, 463개의 문서를 성능 평가를 위한 테스트 문서로 사용하였다. 또한 질의의 특성상 직접 분류가 필요 없이 운영자에게 포워딩해야할 부류가 존재한다. 개별답장이 필요한 질의부류, 답장이 필요없는 질의부류, 재 질의 부류, 질의의 빈도가 현저히 떨어지는 부류 등은 운영자에게 포워딩해야할 부류로 간주하였다. 그 분포는 표 2와 같다.

	부류 개수	질의 개수
응답되어야할 질의	22	1475 (66.9%)
운영자에게 포워딩해야할 질의	46	729 (33.1%)

표 2. 응답종류에 따른 분류

신경망은 기본적으로 각 부류별로 구성하였으며, 운영자에게 포워딩해야할 부류집합도 각각의 부류별로 신경망을 구성하였다. 또한, 질의의 빈도수가 작아 부류의 특징을 표현하기 힘든 부류는 관련있는 부류를 합쳐 3개의 대부류로 묶는 작업을 수행하였다. 결과적으로 답변되어야할 부류 22개, 운영자에게 포워딩해야할 분류 9개인 총 31개에 대한 개별신경망을 구성하여 학습을 수행하였다. 각 신경망의 인식율의 향상이 없을 때까지 학습을 수행하였는데, 모든 경우에서 200세대 안에 학습이 완료되었다.

4.2 결과분석

전체 시스템의 성능 분석은 문서 자동분류기의 성능 기준이 되는 정확율과 재현율을 이용하였다[3].

정확율 (Precision) =

$$\frac{\text{분류가 이루어진 질의이면서 정확히 분류된 질의의 수}}{\text{분류가 이루어진 질의의 수}} \quad (5)$$

재현율 (Recall) =

$$\frac{\text{분류가 이루어진 질의이면서 정확히 분류된 질의의 수}}{\text{분류되어야 할 질의의 수}} \quad (6)$$

첫 번째 실험은 답변 분류결과의 수에 따른 결과를 살펴보았다. 각 신경망의 출력을 종합하여 큰 출력을 낸 순으로 정렬하여 선택하는 방법이다.

답변부류수	정확율	재현율
1	69.8%	71.8%
2	77.2%	79.3%
3	78.2%	80.3%

표 3. 다중부류에 따른 정확율과 재현율

표 3은 답변 후보의 수가 많을 수록 정확율과 재현율이 향상됨을 보여준다. 하지만 응답시스템에 적용할 경우 사용자의 만족도가 낮아진다는 단점이 있다.

두 번째 실험은 기각 임계치에 따른 결과를 살펴보았다. 질의분류는 정확율을 최대한 높여 사용자에게 응답하는 것을 목표로 한다. 기각 임계치는 신경망의 출력이 임계치를 넘지 못할 경우 기각시킴으로써 분류의 기각율을 높여 정확율을 높일 수 있다. 표 4는 기각 임계치에 따른 결과를 보여준다. 하지만 기각율이 높아질수록 재현율이 떨어져 운영자가 처리해야할 질의의 양이 늘어날 수 있다.

임계치	정확율	재현율
0.9	86.4%	59.3%
0.95	87.8%	55.0%
0.99	91.1%	41.0%

표 4. 기각 임계치에 따른 정확율과 재현율

5. 결론

본 논문에서는 다중 신경망을 이용한 문서분류 기법을 이용하여 한메일넷의 사용자 질의를 분류하는 시스템에 대하여 설명하였다. 분류의 정확율을 향상시키기 위해 특징 추출 및 다중분류, 기각임계치를 이용한 결과결합을 수행하여 그 결과를 살펴보았다. 실험 결과는 자동 질의분류 시스템이 실제 응답시스템에 효과적으로 적용될 수 있음을 보여준다. 앞으로 다중 신경망의 체계적인 구성방법과 불균등한 데이터 집합에서의 효과적인 분류에 대한 연구가 필요하며, TREC과 같은 문서집합으로 시스템의 성능을 객관적으로 검증할 필요가 있다.

참고문헌

[1] Y. Yang, et al., "Learning Approaches for Detecting and Tracking News Events," *IEEE Intelligent System*, pp. 32~43, July/August 1999.

[2] S.M. Weiss, et al., "Maximizing Text-Mining Performance," *IEEE Intelligent System*, pp. 63~69, July/August 1999.

[3] Y. Yang and X. Liu, "A Re-examination of Text Categorization Methods," *Proc. SIGIR'99*, pp. 42~49. 1999.

[4] 정영미, *정보검색론*, 구미무역출판부, 1993.

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

[6] H.T. Ng, W.B. Goh, and K.L. Low, "Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categrization," *Proc. SIGIR'97*, pp. 67~73, 1997.

[7] H. Schütze, D.A. Hull and J.O. Pedersen, "A Comparison of Classifiers and Document Representations for the Routing Problem," *Proc. SIGIR'95*, pp. 229~237, 1995.

[8] R.P. Lippmann, "An Introduction to Computing with Neural Networks," *IEEE Acoustics, Speech, and Signal Processing Magazine*, vol. 4, no. 2, pp. 4~22, 1987.

[9] R. Anand, K.G. Mehrotra, C.K. Mohan and S. Ranka, "An Improved Algorithm for Neural Network Classification of Imbalanced Training Sets," *IEEE Transaction on Neural Networks*, vol. 4, no. 6, pp. 962~969, 1993.