Proceedings of KFIS 2001 Spring Conference, 2001

계층적 문서 클러스터링을 이용한 실세계 질의 메일의 자동 분류

Automatic Categorization of Real World FAQs Using Hierarchical Document Clustering

류중원 · 조성배
연세대학교 컴퓨터과학과

Jungwon Ryu and Sung-Bae Cho
Department of Computer Science, Yonsei University
E-mail : rjungwon@candy.yonsei.ac.kr, sbcho@csai.yonsei.ac.kr

ABSTRACT

Due to the recent proliferation of the internet, it is broadly granted that the necessity of the automatic document categorization has been on the rise. Since it is a heavy time-consuming work and takes too much manpower to process and classify manually, we need a system that categorizes them automatically as their contents. In this paper, we propose the automatic E-mail response system that is based on 2 hierarchical document clustering methods. One is to get the final result from the classifier trained seperatly within each class, after clustering the whole documents into 3 groups so that the first classifier categorize the input documents as the corresponding group. The other method is that the system classifies the most distinct classes first as their similarity, successively. Neural networks have been adopted as classifiers, we have used dendrograms to show the hierararchical aspect of similarities between classes. The comparison among the performances of hierarchical and non-hierarchical classifiers tells us clustering methods have provided the classification efficiency.

Keywords : 문서 분류, 계층 클러스터링, 신경망 분류기, 질의 자동응답 시스템.

I. 서 론

컴퓨터의 보급과 함께 인터넷의 대중화로 인해 많은 사람들이 정보통신 기반을 이용하게 되고 많은 정보가 인터넷을 통해 제공되게 되었다. 이러한 방대한 양의 정보를 사람들이 일일이 가공하고 분류하기에는 많은 인력과 시간이 필요하므로 문서 자동 분류의 중요성이 널리 인식되고 있다.

문서 자동 분류란 미리 수집되어 있는 문서 집합을 바탕으로 부류를 나누어 놓고 학습을 통해 새로운 문서를 각 부류에 대응시키는 것이다. 이러한 문서 분류 기법은 인터넷 사이트의 질의 메일을 자동으로 응답하는 시스템을 구축하는데 핵심 기술이 된다. 최근의 문서 자동 분류는 k-최근접 이웃, 의사 결정 트리, 신경망 등의 기계학습 패턴 인식 방법들이 이용되고 있다[1]. 그러나 이러한 분류기들을 효율적으로 사용하려면 데이터에 대한 정성적인 분석이 필요하며, 데이터의 부류를 결정 짓는 내재적 특징을 알아낼 수 있는 도구가 필요하다.

본 논문에서는 신경망 분류기를 사용하여 사용자 질의 자동응답 시스템을 구현하는데 문서의 부류를 계층적으로 분류하는 계층 분류 방법을 제안한다. 한메일넷의 사용자 질의 데이터를 이용하여 실험을 하였다.

II. 한메일넷 질의

질의 메일의 부류를 결정하기 위해 먼저 한 달간 수집된 사용자들의 질의문서 2,204개를 분석하여 68개의 부류로 나누었다. 그러나 분류해야할 문서 부류의 개수가 많으면 그만큼 분류기의 성능이 악화될 수 있으므로 문서의 부류를 적절히 재구성할 필요가 있다. 따라서 사람의 분류 결과를 시스템이 자동으로 응답해야할

문서 부류 20개와, 운영자에게 포워딩하여 개별적으로 답장을 보내주어야 하는 한 개의 부류로 재구성 하였다(표 1).

표 1. 사용자 질의 분포

종 류	부류 개수	질의 개수
응답되어야할 질의	20	1,475(66.9%)
운영자에게 포워딩할 질의	1	729(33.1%)
전 체	21	2,204

그런데 문서들이 각 부류들에 골고루 나누어져 있는 것이 아니라 특정 부류에 지나치게 편중되는 경향이 있고, 21개의 부류들 간에 유사한 부류들이 있어 이들을 단번에 구분해 내는 데에 어려움이 있다. 따라서 효과적인 분류를 수행하기 위하여 문서의 계층분류 시스템을 구현하려면 하나의 분류기를 단층적 분류 시스템보다 좋은 성능을 보임은 다른 연구에서도 입증된바 있다[2, 3].

III. 한메일넷 질의 자동응답

1. 전처리

전체 시스템은 전처리, 수치벡터화 및 특징 추출과 문서 분류 과정으로 나누어진다. 전처리는 텍스트 문서의 특징을 추출하여 기계학습에 적당한 수치벡터로 변환하는 과정이다[4, 5]. 전처리 과정에서는 미리 구축된 색인어 사전을 이용하여 입력 질의 문서로부터 분서 분류에 의미를 갖는 색인어 들을 추출하게 된다. 본 연구에서는 449개의 색인어를 가지고 있는 색인어 사전을 수동으로 구축하여 사용하였다. 색인어 사전은 통신상의 속어나 약어, 동의어에 대한 사전을 유지하여 표현을 정규화 하기나, 맞춤법에 맞지 않더라도 문서 분류에 중요한 키워드일 경우 추출해 내는 기능을 가지고 있다.

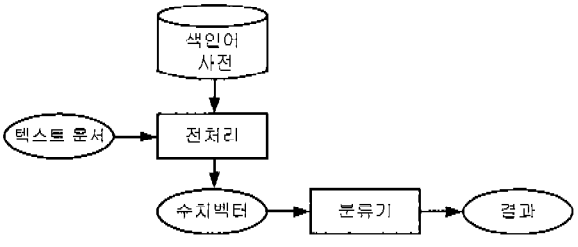

그림 1. 전체 시스템 구조

2. 수치 벡터화 및 특징 추출

위에서 얻어진 색인어 집합을 분류기에 직접 사용할 수는 없고, 신경망 분류기가 이해할 수 있는 수치로 변환하는 과정을 거쳐야 한다. 본 논문에서는 추출된 색인어가 문서에 나타나는 빈도수와 색인어가 나타난 문서의 개수를 기반으로 계산되는 *tfidf* (term frequency and inverse document frequency)값을 사용하였다. 문서 d_i의 j번째 키워드 w_{ij}의 가중치는 다음과 같이 표현된다[6].

$$w_{ij} = tf_{ij} \log\left(\frac{N}{df_j}\right)$$

한편, 문서의 수치화된 특징 벡터의 차원이 너무 크면 패턴인식의 성능이 저하될 수 있으므로, 문서 분류상 중요한 색인어를 추출하여 벡터의 차원을 줄여야할 필요가 있다[4, 5]. 특징 추출 방법으로는 문서 빈도(document frequency), 정보 이득(information gain), 상호 정보량(mutual information), χ^2, 적합성 점수(relevancy score), 교차비(odds ratio), 단순화된 χ^2(simplified χ^2)등의 방법이 있는데, 이들 중에서 일반적으로 가장 좋은 성능을 보이는 χ^2 방법을 사용하여 특징을 추출하였다[5].

$$\chi^2(t,c) = \frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

A는 부류 c에 속한 문서 중 키워드 t를 가진 문서의 개수, B는 부류 c에 속하지 않은 문서들 중 키워드 t가 발생하는 횟수, C는 부류 c의 문서들 중에서 키워드 t가 없는 문서의 개수, 그리고 D는 부류 c에 속하지 않는 문서들에서 키워드 t가 발생하지 않는 횟수이며, N은 전체 문서의 수 이다.

문서에 포함되어있는 각 단어마다 χ^2값을 계산하였고, 가장 높은 값을 나타내는 150개의 색인어를 추출하였다[5].

3. 문서 분류

3.1 단층 분류기

문서에 대한 수치 벡터가 입력으로 들어오면 한 개의 신경망에서 21개 부류 중 하나로 분류하는 방법이다. 분류기의 출력 노드는 전체 부류 개수에 해당하는 21개이다.

3.2 계층 분류기(1)

문서들로부터 부류들 간의 계층적 유사도 구조를 알기 위하여 클러스터 내의 average linkage를 이용한 덴드로그램을 사용하였다. 덴드로그램은 특징 부류간 유사도의 계층 관계를 나무 형태의 그림으로 보여주기 때문에 데이터에 대한 사람의 직관적 분석이 가능하므로 유용하다.

유사도 측정을 위하여, 유클리드 거리(Euclidean distance), 표준 유클리드 거리(standard Euclidean distance), 민코스키 메트릭(Minkowski metric) 및 시티블록 거리(city block distance)를 사용하였는데, 이들 중에서 실험 결과가 가장 좋았던 유클리드 거리 방법

을 사용한 학습 데이터 클러스터링 하였다(그림 2).

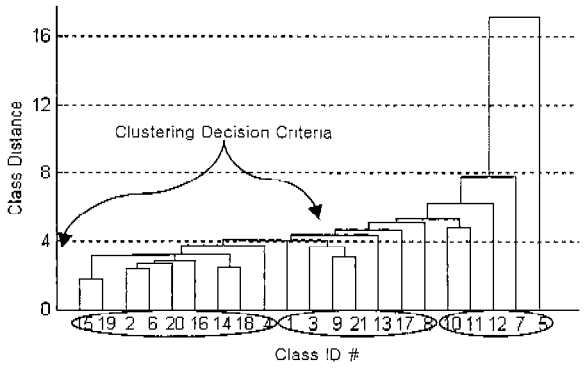

그림 2. 덴드로그램을 이용한 문서집합 클러스터링

그 결과 크게 3개의 전체 21개 부류는 부류간의 유사도가 3개의 그룹으로 군집화 되어있는 양상을 띄고 있음을 알게 되었다. 같은 클러스터 안의 부류들은 서로 낮은 부류간 거리(class distance)를 갖고 있고, 이에 비하여 3개의 각 클러스터간의 거리는 상대적으로 먼 것을 알 수 있다.

위에서 묶여진 그룹을 바탕으로 구축된 문서 분류기는 그림 3과 같다. 대분류기는 들어온 문서가 세 문서 부류 클러스터 중 어느 클러스터에 속하는지를 결정한다. 소분류기는 각 클러스터당 하나씩 있으며 들어온 문서 벡터를 자기 그룹에 있는 부류 중 어느 곳에 속하는지를 판단하게 된다. 두 단계의 신경망을 거쳐야 하는 방법이므로 오차를 줄이기 위해 소분류기는 자신의 그룹에 원래 속하는 부류뿐만 아니라 특징이 비슷하다고 생각되는 이웃 클러스터의 부류도 중첩하였다. 타원안의 숫자는 해당 클러스터의 부류 번호를 의미한다.

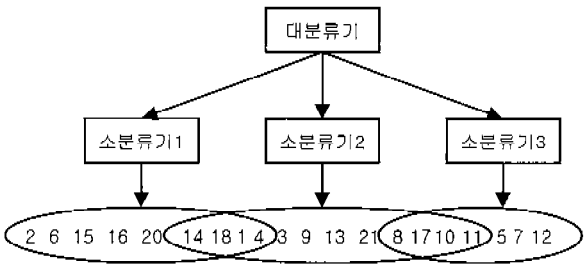

그림 3. 계층 분류기(1) 구조

3.3 계층 분류기(2)

이 방법은 처음에 그려진 덴드로그램을 이용하여 가장 구분되는 몇 개의 부류들부터 순차적으로 우선 분류하여가는 것이다. 최초 덴드로그램을 기준으로하여 가장 유사도 차이가 많이 나는 몇 개의 부류를 하나의 클러스터로 묶는다. 그다음 나머지 부류들만으로 덴드로그램을 그려보면 처음 덴드로그램과 그 모양이 다름을 알 수 있다(그림 4). 이는 먼저 클러스터링된 부류들이 다른 부류들간의 유사도 관계에 영향을 미치기 때문이다. 따라서 현재 남아있는 나머지 부류만으로 계층적 유사도 관계를 다시 본 후 거리가 먼 것들을 반복적으로 클러스터링하여 남아있는 문서 부류들간의 유사도 차이가 이주 미소한 상태로 되면 멈춘다.

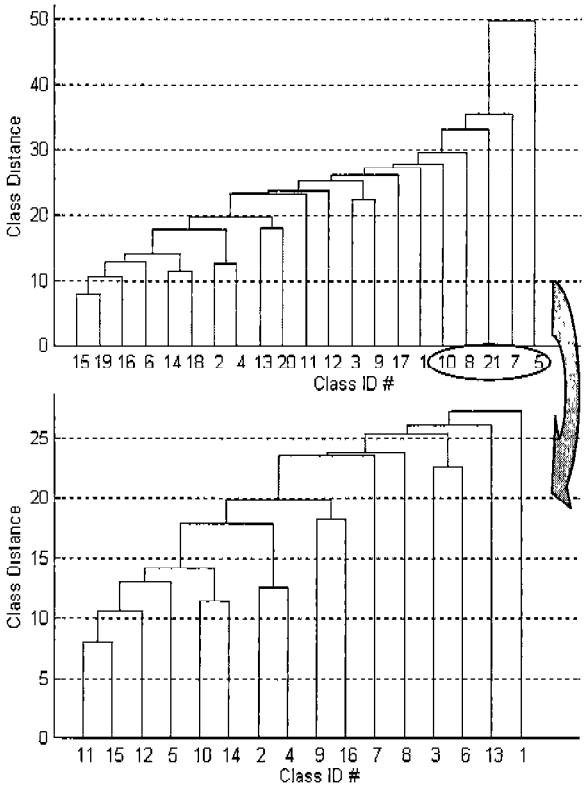

그림 4. 1차 분류 전·후의 클래스 덴드로그램

위의 과정을 반복하여 얻은 분류기 구조는 그림 5와 같다. 전체 부류 개수가 21개임을 감안하여 하나의 클러스터당 최소 4~5개의 부류를 유지하도록 하였고, 분류 단계가 많아질수록 인식률이 낮아질 수 있으므로 3단계로 분류기를 구성하였다.

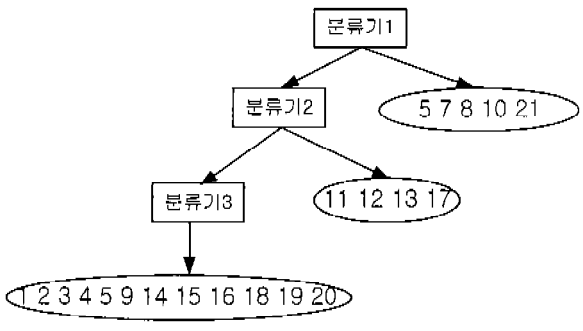

그림 5. 계층 분류기(2) 구조

IV. 실험 결과

1. 실험 환경

약 한 달간 수집된 한메일넷 사용자 질의 2,204개 중에서 1,718개의 문서들을 학습데이터로, 463개의 문서를 성능 평가를 위한 테스트 데이터로 사용하였다.

분류기로는 입력 노드 150개, 은닉층 50개, 분류해야할 부류 개수와 동일한 수를 갖는 출력층을 가지는 역전파 신경망을 사용하였다. 학습 데이터에 대하여 98%의 인식률을 보이거나, 데이터에 대하여 1000번 이상 반복 계산할때까지 학습하였다.

학습률과 모멘텀 등의 여러 파라미터를 조정하여 가장 좋은 인식률을 낸 것을 실험의 최종 결과로 사용하였다.

2. 결과 분석

이 시스템의 성능을 분석하기 위한 지표로 인식률을 사용하였다.

$$\text{인식률} = \frac{\text{분류성공 문서 수}}{\text{전체 문서 수}}(\%)$$

표 2. 실험 결과

분류기		인식률
단층 분류기		21.2%
계층 분류기(1)	대분류기	52.7%
	소분류기(1)	34.7%
	소분류기(2)	32.5%
	소분류기(3)	7.7%
	전 체	19.2%
계층 분류기(2)	1차 분류기	96.1%
	2차 분류기	81.5%
	3차 분류기	42.8%
	전 체	32.6%

계층 분류기(2)가 클러스터링을 적용하지 않은 경우보다 인식률 약 11.4%가 높게 나타났다. 부류들을 여러 단계로 클러스터링하게 되면 한 단계에서 분류해야할 부류의 수가 줄어들고, 비슷한 특징을 갖는 데이터를 군집화 할 수 있으므로 분류 효율 면에서 장점이 될 수 있는 반면, 여러 단계의 분류기를 거쳐야 하기 때문에 각 단계별로 매우 높은 인식률을 유지하지 못하면 에러가 늘어날 수 있다는 단점도 있다.

두 개의 계층 분류 방법 중에서는 중첩 부류를 이용하여 오차율을 줄임으로 단점을 보완한 계층 분류 방법(1)이 더 나음을 알 수 있다.

V. 결 론

본 논문에서는 두 가지의 계층적 문서 클러스터링 방법을 제안하여 실제 문서 분류 시스템에 적용해 보았다. 그 결과 계층 방법을 사용할 때가 단층 분류보다 약 11.4%의 인식률 향상을 가져왔음을 알 수 있었다. 앞으로는 부류들의 연관관계를 파악하여 좀더 효율적으로 부류의 계층 구조를 만들 수 있는 방법에 대한 연구가 필요하며, 다중 신경망등의 좀더 안정적인 분류기 모델에 적용하여 볼 필요가 있다.

VI. 참고문헌

[1] Y. Yang and X. Liu, "A re-examination of text categorization methods," *Proc. of 22nd Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, p. 42~49. 1999.

[2] M. Sasaki and K. Kita, "Rule-Based Text Categorization Using Hierarchical Categories," *IEEE International Conference on Systems, Man and Cybernetics*, p. 2827~2830 ,1998.

[3] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, p. 181~214, 1994.

[6] Q. He, "A review of clustering algorithms as applied in IR," *Technical Report UIUCLIS--1999/6+IRG*, UIUC, 1999.

[4] 이지행, 조성배, "다중 신경망을 이용한 한메일넷 질의 자동 분류 시스템," 제27회 추계학술발표회, 한국정보과학회, p. 232~234, 2000.

[5] 홍진혁, 류중원, 조성배, "실세계의 FAQ 메일 자동분류를 위한 문서 특징추출 방법의 성능 비교," 제28회 춘계학술발표회, 한국정보과학회, p. 232~23 4, 2001.

[6] F. Sebastiani, "Machine learning in automated text categorisation: a survey," *Technical Report IEI-B4-31-1999*, Istituto di Elaborazione dell' Informazione, C. N. R., Pisa, 1999.