인터넷 문서 자동 분류 시스템 개발에 관한 연구

한 광 록[†] · 선 복 근[††] · 한 상 태[†††] · 임 기 욱[††††]

요 약

본 논문은 인터넷 문서 자동 분류 시스템의 구현에 대하여 논한다. 문서 자동분류 알고리즘을 설정하고, 역전파 학습 모델을 이용하여 문서의 범주화를 수행하는 시스템을 구축한다. 문서학습을 위해서 범주별 인터넷 문서들을 수집하고 수집한 문서에 대하여 카이제곱(χ^2) 검정을 수행함으로써 범주화 자질을 추출한다. 이 범주화 자질을 바탕으로 하여 학습 및 분류 벡터 집합을 생성한다. 실험 결과의 평가로부터 본 논문에서 구현한 시스템이 유사도 계산을 이용한 문서의 분류 시스템보다 성능이 향상된 것을 알 수 있었다.

A Study on Development of Automatic Categorization System for Internet Documents

Kwang Rok Han[†] · B. K. Sun[††] · Sang Tae Han[†††] · Kee Wook Rim[††††]

ABSTRACT

In this paper, we discuss the implementation of automatic internet text categorization system. A categorization algorithm is designed and the system is implemented by back propagation learning model. Internet documents are collected according to the established categories and tested by Chi-squre(χ^2) for the document leaning, and the category features are extracted. The sets of learning and separating vector are producted by these features. As a result of experimental evaluation, we show that this system is more improved in the performance of automatic categorization than the nearest neigbor method.

1. 서 론

인터넷 상에서 문서의 검색 및 분류 서비스는 현재 가장 많이 이루어지는 인터넷 서비스의 한 영역으로 자리잡고 있다. 날로 증가하는 인터넷의 이용자와 정보량으로 인해 서비스의 활용도는 더 높아지고 있으며, 인트라넷 문서의 정보 검색 서비스, 북마크 분류 서비스등 문서의 정보를 이용하여 제공되어지는 서비스가 증가하고 있다. 자동 문서 분류란 특정 문서의 내용에 기반하여 컴퓨터가 자동으로 이 문서를 미리 정의되어 있는 분류 목록에 할당하는 작업을 말하며, 자동 문서 분류를 위한 학습 프로그램은 특정 문서가 적합한 분류 목록에 들어갈 수 있도록 규칙을 배우게 된다[1]. 문서의 자동분류는 문서 영역의 특성상 그 작업이 매우 어렵다. 문서는 벡터의 형태로 표현되어 있지 않으며, 문서에 수많은 특징이 존재하며, 각 문서별로 많은 변화를 일으킬 가능성을 내포하고 있다. 우리가 관심을 가지는 것은 문서내에 포함된 단어이다. 하지만 문서는 자연언어로 씌어져 있으므로 많은 중의성을 포함하고 있다. 이는 문서 영역에 에러공간이 상당히 넓게 분포하고 있음을 의미하며 숫자, 생략기호, 약어등이 이러한 에러공간에 해당한다. 이렇게 볼 때 고차원적

† 종신회원 : 호서대학교 벤처전문대학원 교수
†† 준 회 원 : 호서대학교 대학원 벤처전문대학
††† 정 회 원 : 호서대학교 수학과 교수
†††† 종신회원 : 선문대학교 산업공학과 교수

논문접수 : 2000년 6월 26일, 심사완료 : 2000년 9월 1일

이고 에러공간이 넓은 영역의 특성상 문서의 분류를 기계학습을 통해 수행하는 것은 그만큼 더 어려움이 존재한다[13].

문서의 분류를 기계학습에 적용한 Apte'와 Damerau에 의해 한가지 주목할 만한 결과가 나왔는데 이들은 규칙기반 증명을 통하여 Reuters-22173 집합에서 문서 분류를 실시한 결과 80.5%의 정확률(precision)과 재현율(recall)을 얻어냈다[6, 13]. 이 결과를 얻기 위하여 10,000개의 정리된 예제가 사용되어졌다. 일반적으로 감독학습을 통해서 좋은 결과를 얻기 위해서는 수 천개 정도의 많은 정리된 예제가 필요하다. Castelli와 Cover는 정리된 데이터와 정리되지 않은 데이터의 상대적 가치를 베이시안 정리를 통해서 계산해 보았을 때, 정리된 데이터가 정리되지 않은 데이터보다 지수적으로 더욱 가치가 있다고 결론을 내렸다[15]. 이는 정리된 예제를 가지고 학습을 수행하는 것이 학습을 하는데 계산적 노력을 훨씬 줄여준다는 사실을 나타내고 있다. 그러나 문서는 그 자체로 정리되어 있지 않으며 이 문서의 정리 작업은 사람의 몫으로 남게 된다. "지수적으로 더욱 가치가 있다"라는 측면에서 보면 이는 효율성을 위해 전처리 과정을 거쳐야 되는 것임을 알 수 있으며, 이는 돈과 시간의 투자, 사람의 노력으로 이루어져야 한다. 이러한 사실은 전처리 과정을 통해서 학습 작업을 적게 할 것인가, 전처리 과정 없이 많은 학습 작업을 할 것인가를 시스템의 설계자로부터 선택하게 만든다.

현재 대부분의 자동 문서 분류방법은 규칙기반[12], 확률기반[4, 17], 통계/학습 기반[9]으로 이루어지고 있다[1]. 본 논문에서는 이러한 문서자동 분류 시스템에서 사용되어지는 분류 방법을 참고로 하여 이중 통계적 기법과 신경망 학습을 접목하여 시스템을 구현하고 이를 평가하고자 한다. 본 논문에서 구현하는 문서자동 분류 시스템은 문서의 정리작업을 수행하는 학습 및 분류 벡터의 생성과정, 문서의 분류를 위한 문서 학습 및 분류과정으로 이루어져있다. 인터넷 문서를 학습기의 입력으로 만들어주는 벡터 생성에는 태그 분석, 형태소 분석, 카이제곱 통계량 측정, 벡터 생성의 과정을 거치며, 학습 및 분류 과정에서는 역전파 네트워크로 구성된 학습 네트워크를 이용한다. 2장에서는 관련 연구로써 현재 사용되어지는 문서의 분류 방법에 대해 알아보며, 3장에서는 전체 시스템의 구조를 설계, 구현하고 4장에서는 실험을 통해 시스템을 평가, 5장에서 결론을 내린다.

2. 관련연구

2장에서는 문서의 분류에 많이 사용되어지는 Naive-Bayes 분류기[10], 유사도 계산법에 의한 분류기[2], 결정트리[16]에 대하여 논한다.

2.1 Naïve-Bayes 분류

Carnegie Mellon 대학에서 개발한 Rainbow 문서 자동 분류 시스템은 Naïve-Bayes 이론을 이용하여 개발되었다[10]. 문서 범주 $C = c_1,\ldots,c_m$ 가 있다고 하자. 새로 주어진 분류되지 않은 문서 D가 들어왔고 그 문서 D의 단어 리스트는 $W(w_1\text{-}w_d)$까지 있을 때, 문서 D를 범주 C_{NB}에 할당하는 계산식은 식 (1)과 같다.

$$C^*_{NB} = P(c_j)\ \Pi\ ^{d}_{i=1}\ P(W_i \mid c_j) \qquad (1)$$

$P(c_j)$는 범주c_j로 분류될 사전 확률이며, $P(w_i|c_j)$는 주어진 범주 c_j에 w_i가 들어있을 사후 확률이다. 이 Naïve-Bayes 분류는 문서에 단어가 나타나는 사건이 각각 독립사건이라는것을 가정하고 있다.

학습 시킬 문서의 집합이 작을 경우, $P(w_i|c_j)$는 정확하지 못할 수 있다. 만일 범주화 할 문서 데이터에서 학습 범주 내에 속한 단어가 하나도 들어 있지 않다면 확률값은 0이 될 것이다. 이러한 $P(w_i|c_j)$ 값에 0이 포함될 경우를 생각하여 식 (2)처럼 사후 확률을 계산할 수도 있다.

$$P(w_i \mid c_j) = \frac{n_{ij}+1}{n_j+k_j} \qquad (2)$$

N_j는 분류 c_j내의 단어의 총 개수이며, N_{ij}는 분류 c_j에서 단어 w_i의 출현 빈도수, 그리고 k_j는 분류 c_j의 어휘수이다. 이 값으로 계산 하였을때 앞에서의 Bayesian의 확률계산에서와 거의 비슷한 확률값을 얻을 수 있으면서 확률값으로 0가 산출되는 것을 막을 수 있다[17].

2.2 유사도 계산에 의한 분류

이 분류방법은 문서 D를 범주 C_j로 새로 분류하려 할 때, 범주 C_j에 가장 근접한 패턴을 가진 문서가 범주에 할당된다. 많은 경우 TF-IDF(Term Frequency/Inverse Document Frequency)[8]를 통한 가중치 계산법과 코사인 유사도계산을 이용하여 문서의 범주 할당

이 이루어진다[2].

문서 D의 가중치 벡터를 $T_d=(w_1,....,w_n)$라 하고 범주 C의 가중치 벡터를 $T_c=(w_1,....,w_n)$라 할 때, 문서와 범주간의 유사도 S는 식 (3)과 같다.

$$S(T_d, T_c) = \frac{T_d T_c}{|T_d||T_c|} \quad (3)$$

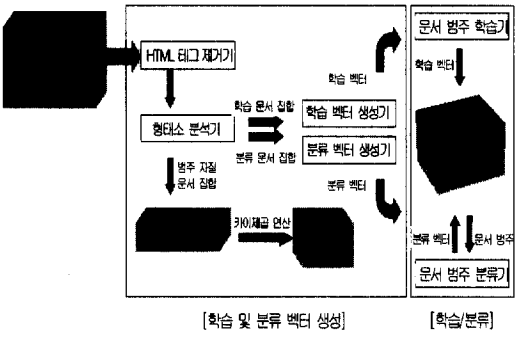
2.3 결정트리(Decision Tree) 분류

결정트리는 정보이론에 기반한 귀납적 유도 학습 방법으로 가장 많이 사용되어지는 것중 하나로써 1949년 Shannon과 Weaver에 의해 처음으로 소개되었다. 잡음이 있는 데이터에 유리하며, 표현을 구별하는 학습 능력이 뛰어난 점이 결정트리의 특성이며 이 특성을 이용하여 문서의 범주화에 많이 사용되어진다.[17] 영국의 Timberlake사에서 만든 CART (Classification And Regression Trees) 시스템은 결정트리 알고리즘을 이용해서 만든 범주 데이터 및 연속 데이터의 분류 시스템의 한 예이다[11].

가장 잘 알려진 결정트리 알고리즘으로 ID3가 있으며 ,이의 뒤에 나온 C4.5, C5 등이 있다[16]. C5 결정트리는 정확도, 속도, 메모리 사용측면에서 ID3나 C4.5보다 성능이 많이 향상되었을뿐 아니라 부스팅 알고리즘[17]도 포함되어 있다. 부스팅의 기본 알고리즘은 하나의 분류자 대신에 사용자가 지정한 n(n>1)개의 분류자를 산출해내는 것이다. i번째 분류자는 (i-1)번째 분류자에의해 만들어진 에러의 계산을 통해 만들어진다. 새로운 문서가 분류되어질 때, 이 n개의 분류자에 기초를 둔 시스템에서 문서의 최후 범주를 선택하여 분류하게 된다.

3. 시스템 설계 및 구현

2장의 범주화 알고리즘들은 단독으로 쓰이거나 하나 이상의 알고리즘이 병합되어 사용되어지기도 한다. 본 논문에서는 문서 범주화의 효율을 높이는 방안으로 신경망 중 역전파 알고리즘을 사용하였으며, 통계적 방법으로 TF/IDF 계산 알고리즘의 변형인 TF/ICF[2] 계산 알고리즘과 카이제곱 연산[3]을 접목하여 보다 정확한 범주별 자질 집합을 구성하였다. 자동 문서 범주화의 평가를 위해 위에서 언급한 알고리즘 중 유사도 계산에 의한 분류 알고리즘과의 범주화 성능을 비교하여 본 논문에서 구현한 자동 범주화 시스템의 성능을 측정하고자 한다. 이를 위하여 인터넷 문서를 자동으로 분류하기위해 본 논문에서 구현한 시스템의 구성은 (그림 1)과 같다.



(그림 1) 전체 시스템 구성

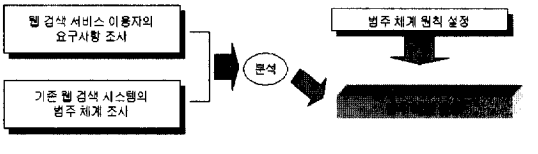
범주별 문서집합은 문서의 분류를 위한 범주 체계를 설정하고, 수집된 문서를 범주체계에 맞게 분류함으로써 만들어진다. 학습 및 분류 벡터를 생성하는 과정에서는 학습/분류에 사용될 문서를 학습/분류 네트워크의 입력형태인 벡터로 변환해 주는 작업을 수행하며, 이 벡터를 이용하여 학습/분류과정에서는 학습 네트워크를 이용하여 문서의 학습을 수행하고, 학습된 결과에 따라 문서를 분류하게 된다.

다음의 각 절에 범주별 문서 집합의 구성 원칙 및 범주 체계와 시스템의 세부 구성 및 알고리즘에 대해 기술한다.

3.1 범주별 문서 집합의 구성

3.1.1 문서 범주 체계

문서의 범주체계 설정을 위한 기본 절차는 (그림 2)와 같다.



(그림 2) 범주체계 설정 과정

시스템의 구축을 위한 문서의 범주 계층은 이용자의 요구사항과 기존 시스템의 범주 체계를 조사하여 분석한 결과와 아래 4가지 범주 체계의 원칙에 기초하여 작성한다.

① 다른 시스템에서의 활용 가치
② 학습 및 실험 문서 수집의 용이함
③ 범주간 계층 구조가 쉽게 구성
④ 계층적 범주 구조 체계

범주 계층은 분석결과와 범주 체계 원칙에 따라 8개의 주 분류 목록과 그 아래 90개의 중,하위 분류 목록으로 이루어진다.

3.1.2. 범주별 학습 문서 수집

본 논문에서 제시한 시스템의 문서 학습을 위해서 문서를 수집하는 것은 웹 에이전트를 이용하여 자동으로 수행하였으며, 그 이외에 분류될 문서가 학습 데이터로 유효한가의 판단과 그 판단에 따른 학습 문서의 분류는 정확도를 기하기 위해 모두 수작업을 통해 이루어졌다. 이는 모든 문서에 대하여 전처리 과정을 수행함으로써 학습에 소요되는 비용을 줄이고 학습 데이터의 가치를 보다 높이고자 이루어진 작업이다.

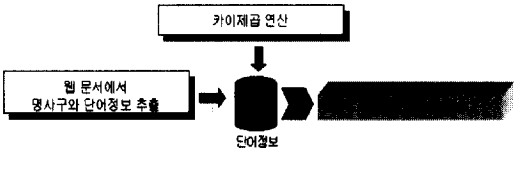
위 작업을 통하여 각 분류 목록별로 60개의 문서를 수집하여 범주별 자질의 추출, 문서 범주 학습, 문서 분류 성능 평가를 위한 문서집합으로 각 20개씩 할당하여 총 90개 목록에 5,400개의 문서를 할당한다.

3.2 학습 및 분류 벡터 생성

학습 및 분류 벡터 생성 과정에서는 범주의 자질 집합 추출하고 학습/분류기의 입력을 위한 문서의 벡터화 작업을 수행한다.

3.2.1. 범주별 자질 추출

범주별 자질 추출은 문서 범주를 대표할 수 있는 단어를 추출하는 과정이다. 각 범주별로 자질 추출이 이루어지며, 이 범주별 자질은 입력벡터가 만들어질 때, 입력 문서의 벡터자질을 추출하기위한 기본 데이터로 사용되어진다. 범주별 자질을 추출하는 과정은 (그림 3)과 같다.



(그림 3) 범주별 자질 추출

학습 및 분류 벡터를 만들기 위한 기초 자료가 되는 범주의 자질 추출은 문서에 대한 형태소 분석 과정과 카이제곱 검정을 통해 이루어진다[3]. 태그 제거기는 입력받은 html 문서를 텍스트 문서로 변환하여 형태소 분석기로 보내게 된다. 형태소 분석기는 태그 제거기로부터 받은 텍스트 문서에서 명사어구만을 추출하여 문서내 빈도수와 분류 목록내 빈도수를 계산하여 분류 목록별 데이터베이스를 형성한다.

이렇게 형성된 데이터베이스는 각 90개 하위 분류 목록별로 20개씩 총 1800개 문서에 대한 명사어구, 빈도수를 가진 단어정보를 포함하고 있다. <표 1>은 [가족,가정]이라는 하나의 범주에 대한 단어정보의 예 보여준다.

〈표 1〉 단어 정보의 예[가족,가정]

단어	단어 빈도수	출현 카테고리 수
제 사	76	7
신 랑	58	6
회 갑	46	5
혼 례	41	4
신 부	65	13
혼 인	34	5
고 인	29	4

이러한 단어정보 안에서 분류 목록별로 상위 200개의 학습 자질 집합을 추출해내기 위해서 카이제곱 검정을 수행한다.

카이제곱 검정 방법은 각 단어의 출현빈도에 따른 통계량을 계산하여 90개의 분류목록별로 각 단어가 통계적으로 그 분류목록을 대표할만한 특징을 가진 단어인가를 알아보기 위해 수행된다. 이때 카이제곱 검정 통계량을 수식으로 표현하면 자유도 89(I-1)인 식 (4)가 된다.

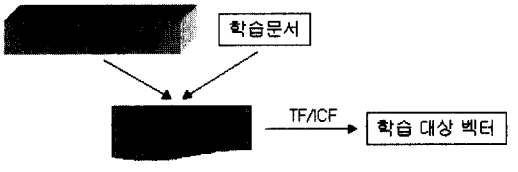
$$\chi^2 = \sum_{i=1}^{90(=I)} \frac{(O_i - E_i)^2}{E_i} \qquad (4)$$

여기서 I 는 분류 목록수, Ei는 각 분류 목록에서 출현하는 각 단어의 기대도수, Ft는 각 분류목록에서 출현된 관찰도수를 의미한다.

이 카이제곱 검정을 통해 90개의 분류목록별로 통계적으로 대표성을 가지고 있는 단어를 추출한 다음, 각 분류목록별로 카이제곱 값 크기순으로 200개의 단어를 추출하였다. 이 때 추출된 단어들을 각 분류 목록에 해당하는 자질 집합으로 결정하였다.

3.2.2. 학습 벡터 생성

학습 벡터 생성을 위해 카이제곱 검정을 통해서 생성된 범주별 자질집합과 비교하여 범주 목록 자질로 인정된 단어 200개를 추출한다. 추출된 단어를 벡터로 만들기 위해서 각 단어의 TF/ICF를 계산하여 그 결과를 벡터의 원소로 하며 벡터파일로 저장하게 된다. 범주당 학습문서 20개에 대응하는 20개의 학습 벡터가 만들어지며, 만들어진 벡터파일은 문서 범주 학습기의 입력으로 사용된다. (그림 4)는 학습 벡터의 생성 과정을 보여준다. 학습벡터(W)의 i번째 원소 wi를 만들기 위해 TF/ICF를 사용한 계산식은 식 (5)[2]와 같다.



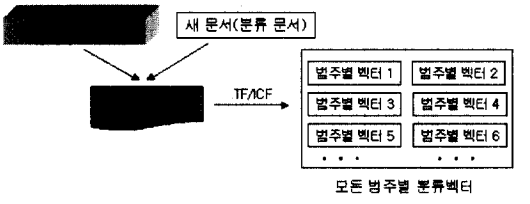
(그림 4) 학습벡터 생성과정

$$W_i = tf * \log(N/cf) \quad (5)$$

여기서 tf는 문서에 나오는 단어의 빈도수를 의미하며, 총 범주의 개수는 N, 단어가 출현하는 범주의 개수가 cf이다. cf가 작아질수록 wi의 값은 커지게 된다. 즉, 문서에 나오는 단어가 다른 범주에는 거의 나오지 않는다면 cf의 값은 작아지고 wi의 값은 커지게 된다. 이렇게 계산함으로써 벡터의 원소를 wi의 값이 큰순서로 정렬하여 선택함으로써 범주의 대표성을 띄는 벡터를 생성하게 된다.

3.2.3. 분류 벡터 생성

분류 벡터 생성을 위해 분류 벡터 생성기는 분류하려는 문서 하나를 입력으로 하여 학습 벡터 생성기에서와 같이 범주별 자질 집합과 비교, TF/ICF를 계산하여 그 결과를 벡터파일로 저장한다. 그러나, 이때의 문서는 어느 범주에 속한 문서인지 모르는 상태이다. 따라서 이 문서를 벡터화하기 위한 범주별 자질 집합 비교 과정은 모든 자질 집합과의 매칭을 통해 이루어져야 한다. 따라서 분류문서 하나당 총 범주의 개수와 같은 90개의 벡터파일이 만들어지게 되며 이 벡터 파일이 문서 범주 분류기의 입력으로 사용된다. (그림 5)는 분류 벡터의 생성 과정을 보여준다.



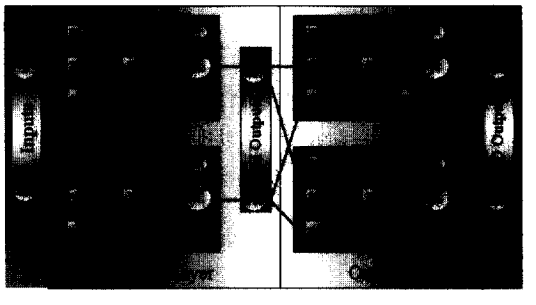
(그림 5) 분류 벡터 생성과정

3.3 시스템 구현

학습 및 분류기는 전처리기를 통해 생성된 학습/분류 벡터를 역전파 네트워크의 입력으로 사용하여 문서의 학습과 범주화를 수행하게 된다[5][14].

3.3.1. 학습 네트워크 모델

본 시스템의 학습 네트워크는 뉴럴 네트워크의 한 분야인 역전파 네트워크를 이용하여 설계하였다. 역전파 네트워크 모델은 은닉층(hidden layer)에 충분한 유닛(units)이 주어진다면, 어떠한 연속함수 모델에 대해서도 학습이 가능하여 예측과 분류 시스템에 많이 사용되고 있다[12]. 본 시스템의 학습 네트워크의 기초가 되는 역전파 네트워크의 개념도는 (그림 6)과 같다.



(그림 6) 역전파 네트워크 모델

(그림 6)에서 입출력 계층(input/output layer)의 유닛의 크기는 적용하게 되는 응용분야마다 다를 수 있으며, 은닉층의 유닛의 크기 역시 구체적으로 명시된 것이 없다. 일반적으로 입력 유닛의 크기와 출력 유닛의 크기사이의 크기로 은닉층의 유닛크기를 정하게 된다. 역전파 네트워크의 내부에서 사용되는 네트워크 제어 상수 역시 실험을 통하여 최적화를 이루어야 한다. 역전파 네트워크는 새로운 응용분야로의 적용이 용이하지만 상당히 오랜 기간의 학습 시간을 요구한다[12].

본 논문에서 제시한 시스템의 역전파 네트워크는 입력벡터를 입력값으로 하여 학습을 수행하게 되며, 다층 퍼셉트론(multi-layer perceptron)을 역전파 알고리즘으로 훈련하여 사용하였다. 역전파 알고리즘은 다층 퍼셉트론의 실제 출력과 원하는 출력 사이의 평균 자승 오차를 최소화 시키는 최소 평균 자승(least mean square)알고리즘이며 입력벡터 p에 대한 오차 E_p를 계산하는 식은 식 (5)와 같으며, 모든 입력값에 대한 오차 E는 E_p의 합으로 구해진다[6, 12].

$$E_p = \frac{1}{2}\sum_j (t_{pj} - O_{pj})^2 \quad (5)$$

출력층의 오차계산식은 식 (6)과 같으며, 이는 위 식의 미분을 통해 얻어진다.

$$\delta_0 = o_0(1-o_0)(t_0-o_0) \quad (6)$$

O_o, t_o는 각각 출력층의 결과와 출력층의 목표값이다.

이 식 (6)의 결과 출력층의 오차를 은닉층으로 역전파 시킨다. 중간층에서는 역전파된 오차를 받아 가중치가 적용된 오류를 계산하며 이 과정은 식 (7)과 같다.

$$\delta_h \leftarrow o_h(1-o_h)\sum_{o\in outputs} w_{oh}\delta_o \quad (7)$$

여기서 O_h는 중간층의 출력값이며, W_{oh}는 출력층의 네트워크 가중치 값이다. 이 값에 의하여 식 (8)과 같이 네트워크의 가중치가 변하게 된다.

$$w_{ji} \leftarrow w_{ji} + \Delta w_{ji}, \Delta w_{ji} = \eta\delta_j x_{ji} \quad (8)$$

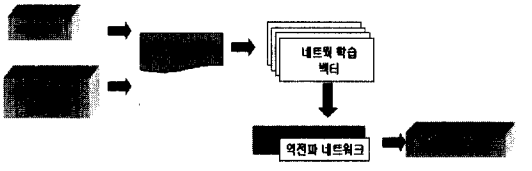
여기서 η는 학습 파라메터, w는 가중치 이며, x는 벡터의 유닛이다.

3.3.2. 문서 범주 학습

문서 범주의 학습은 입력 벡터 생성기로부터 생성된 벡터를 역전파 네트워크의 입력으로 사용하여 (그림 7)과 같은 과정을 통해 이루어진다.

각 범주별 네트워크 학습 벡터는 학습 문서와 범주당 200개 자질 집합의 비교를 통해 입력벡터 생성기에서 만들어진다. 이 학습벡터가 역전파 네트워크의 입력으로 사용되어진다.

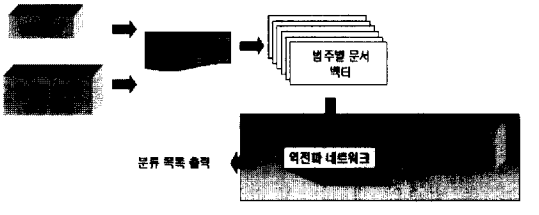
문서 범주 학습기는 역전파 네트워크의 내부 파라메터, 네트워크 상태등을 초기화하며, 학습이 끝난후의 내부 파라메터와 네트워크 상태를 저장하여 문서 분류기에서 학습 네트워크의 사용시 이용할 수 있도록 한다. 학습기에서의 학습은 모든 범주의 학습이 끝날때까지 계속해서 반복 수행된다.



(그림 7) 범주 학습 수행 절차

3.3.3. 문서 범주 분류

문서 분류시 학습 네트워크는 문서 범주 학습기를 통해 생성된 학습결과 네트워크를 로딩함으로써 만들어지며 분류 과정은 (그림 8)과 같다.



(그림 8) 범주 분류 수행절차

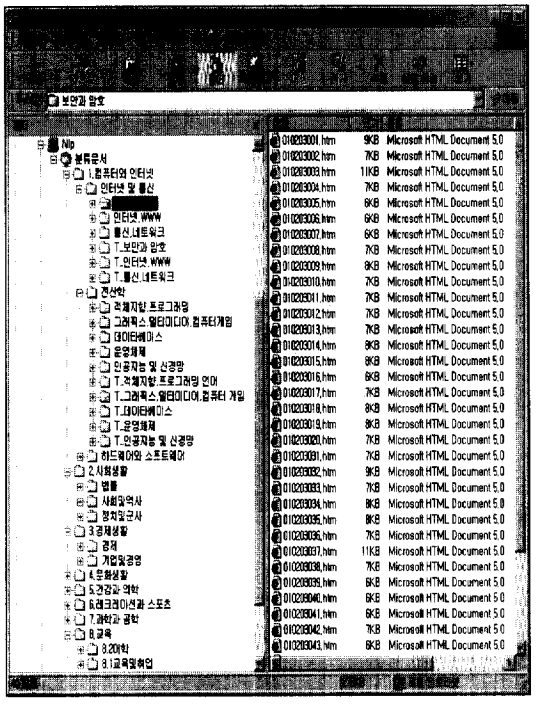
하나의 문서의 범주를 분류하기위해서는 범주별 문서 벡터를 만들어야 한다. 위 그림에서 범주당 200개의 자질집합의 개수는 총 범주 개수인 90개이다. 입력벡터 생성기에서 분류하려는 문서와 이 자질집합의 비교를 통해 각 범주별 문서벡터를 생성하게 된다. 분류하려는 문서는 현재 어떠한 범주에 속하는지 알수가 없으므로 벡터의 생성또한 특정 범주의 자질집합과 비교할 수 없으며 모든 범주의 자질과 비교하여 범주별 벡터를 생성하여야 한다.

범주별 문서 벡터는 범주의 개수인 총 90개의 벡터가 만들어지며, 이 벡터들이 네트워크 학습결과가 로딩된 역전파 네트워크의 입력값으로 들어가게 된다. 학습된 역전파 네트워크는 하나의 벡터가 입력으로 들어오게 되면, 네트워크를 통해 그 값을 통과시키고 90개의 범주에 대한 출력값을 0과 1사이의 실수로 나타낸다. 이 출력 값이 가장 높은 출력유닛이 가지고 있는 범주목록이 입력벡터에 대한 범주로 할당된다. 네

트워크는 하나의 분류문서에 대해 총 90개의 출력값인 분류 목록을 문서 범주 분류기에 넘긴다. 분류기는 90개의 출력값중 가장 높은 값을 가진 출력값을 가진 범주에 문서를 할당하게 된다.

4. 실험 및 고찰

본 논문에서 구현한 문서 분류 시스템은 문서 분류를 위한 문서의 범주를 총 90개로 설정하였으며, 범주당 문서의 범주 학습과 평가를 위하여 20개의 문서를 할당하였다. 논문을 투고하는 시점에서의 평가를 위해 20개의 범주에 대하여 학습을 수행하고 이에 해당하는 평가 문서를 분류해 보았다. (그림 9)에서 개발된 문서 분류 계층 구조와 하나의 분류 목록 디렉토리에 수집된 문서 파일들을 보여주고 있다.



(그림 9) 수집 문서의 예

(그림 9)에서 나타난 바와 같이 90개 분류 목록이 각 계층을 이루고 있으며, "보안과 암호" 등의 디렉토리는 각각 40개의 HTML문서들이 들어있어 자질추출에 20개, 문서 학습에 20개의 문서가 사용되어지고, "T_보안과 암호" 등의 디렉토리는 실험을 위한 테스트 문서들이 들어있다. 그림의 오른쪽의 HTML문서들은 "보안과 암호" 분류 목록에 수집되어 저장된 실제 문서들을 보여주고 있다.

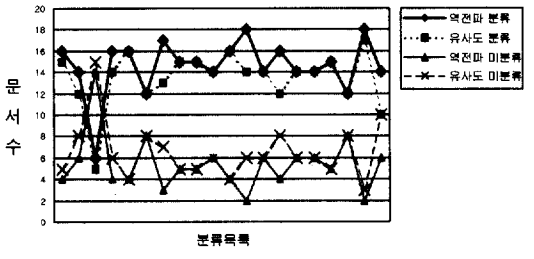
위 문서들을 가지고 실험한 결과 학습 네트워크의 내부 파라메터의 변환, 입력벡터의 유닛수의 변환등 여러 요소에 의해 학습 시간과 그 효율에 많은 차이가 남을 발견할 수 있었으며, 여러 번의 실험 결과 가장 적절한 내부 파라메터와 입력 벡터의 유닛수를 찾아내 학습을 수행하였다. 학습을 위한 문서는 각 분류 목록당 20개, 총 400개의 문서를 역전파 네트워크에 학습을 수행하고 같은 수의 평가 문서를 분류하였다.

분류에 사용된 네트워크의 은닉층의 개수는 2개로 설정하였으며, 제 1은닉층의 노드수는 15개이며, 제 2은닉층의 노드수는 5개이다. 입력층과 은닉층,출력층은 모두 완전연결로 이루어진다. 실험에 사용된 네트워크는 momentum rate 0.9, learning rate 0.7, total error 0.01, individual error 0.001의 네트워크 상수값을 가진다.

2장에서 언급한 유사도 계산에 의한 분류를 똑 같은 조건에서 수행함으로써 본 논문에서 구현한 시스템의 재현율[8]을 비교하였으며 그 결과를 <표 2>와 (그림 10)에 나타내었다.

〈표 2〉 역전파 네트워크와 유사도 계산의 범주화 재현율 비교표

분류목록	문서수	분류		미분류		재현율(%)	
		역전파(1)	유사도(2)	역전파(3)	유사도(4)	역전파	유사도
가족가정	20	16	15	4	5	80	75
객체지향	20	14	12	6	8	70	60
건강관리상담	20	6	5	14	15	30	25
건축환경토목	20	16	14	4	6	80	70
게임도박	20	16	16	4	4	80	80
경영,경제학	20	12	12	8	8	60	60
교육진학	20	17	13	3	7	85	65
교통수송	20	15	15	5	5	75	75
군사북한안보	20	15	15	5	5	75	75
그래픽스,멀티미디어	20	14	14	6	6	70	70
기계항공자동차	20	16	16	4	4	80	80
기업산업무역	20	18	14	2	6	90	70
날 시	20	14	14	6	6	70	70
날 씨	20	16	12	4	8	80	60
날씨기상	20	14	14	6	6	70	70
노 동	20	14	14	6	6	70	70
놀이공원유원지	20	15	15	5	5	75	75
농업농학	20	12	12	8	8	60	60
뉴스서비스	20	18	17	2	3	90	85
다이어트체중조절	20	14	10	6	10	70	50
총 계	400	292	269	108	131	73	67.25



(그림 10) 재현율 비교 그래프

정확률(Precision)-재현율(recall)은 임의 테이블(contingency table) 모델[7]을 기초로하여 측정하였으며, 위의 400개 문서를 기초로 측정된 평균치로써 <표 3>에 나타난바와 같다.

〈표 3〉 정확률-재현율 결과

	평균 정확률 (Precision)	평균 재현율 (Recall)
역전파 네트워크	82 %	73 %
유사도 계산	76 %	67.25 %

실험에서 살펴본 바와 같이 벡터의 유사도 계산을 통한 문서 분류방법보다 신경망의 역전파 네트워크 학습 시스템을 사용하여 좀더 좋은 분류 결과를 가져올 수 있었다.

하지만 본 시스템에서 학습한 데이터는 400개의 문서로써 서론에서 언급한 Apte'와 Damerau의 문서 데이터[6]에 비해 많은 부분 부족한 상태이며, 이 또한 20개의 범주학습으로 제한하였다. 향후 90개의 범주로 학습과 테스트를 확대해야 할 과제가 남아 있으며, 시스템 구현중 문제점 및 해결해야 할 사항이 몇가지 발견되었다.

첫째, 분류 자질의 지속적인 업데이트가 이루어져야 한다. 분류 자질이 고정되어 있다면, 계속 변화하는 인터넷 문서와 사용되는 개념, 단어에 대응하지 못하는 시스템이 될 것이다.

둘째, 범주의 기준이 되는 자질의 추출을 카이제곱 검정을 통해 추출하지만, 사람이 문서를 선택하는 단계에서 전문가 집단의 확실한 검정을 받아야 한다. 본 논문의 실험중 [건강관리상담] 범주는 30%이하의 재현율을 나타내었다. 이는 분석결과 범주 자질의 선택이 잘못된 것이 여러 원인중 하나인 것을 확인 할 수 있었다.

5. 결 론

본 논문에서는 뉴럴 네트워크의 역전파 학습 알고리즘을 문서 자동 분류 시스템의 학습 및 분류 네트워크에 응용하여 이를 구현하고, 시스템의 분류 성능을 평가하였다. 카이제곱연산을 이용하여 자질집합을 구성함으로써 자질집합의 신뢰도가 향상되었으며, 이를 바탕으로 하여 시스템을 학습, 학습 결과를 바탕으로하여 분류성능을 시험해본 결과, 벡터 유사도 계산을 통한 분류와 비교하였을 때 근소한 성능우위를 나타내었다.

여러가지의 문서 범주화 알고리즘이 모두 그렇듯이 정리된 범주화 문서 데이터의 중요성을 본 논문을 통해 새삼 느끼게 된다. 범주화 알고리즘의 성능 개선에 의한 범주화 성능 개선도 중요하지만 전처리를 위한 문서 데이터 역시 그 중요도가 높다는 것을 인식하였으면 한다.

역전파 알고리즘은 local maxima에 빠질 수 있는 가능성과 가중치가 수렴되지 않고 발산할 염려가 존재하며, 학습의 시간이 오래 걸린다는 단점이 존재한다.[1,6] 이러한 점들을 향후 연구과제로 하여 좀더 유연하고 효율성이 있는 시스템 구축에 노력할 것이다.

참 고 문 헌

[1] 김상범, 임해창, 윤덕호, 한광록, 이미영, " 범주간 관계의 고려를 통한 자동 문서 범주화의 개선", HCI 2000 학술발표 논문집, 2000.

[2] 조광제, 김준태, "역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류", 정보과학회 봄 학술발표논문집, 4권 2호, pp.508-510,1997.

[3] 최종후, 한상태, "정보 통계학 입문", pp.244-247, 자유아카데미, 1999.

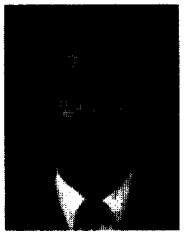
[4] 한미성, 송영훈, 송점동, 이정현, "확률 벡터간의 교차 엔트로피 계산을 이용한 자동 문서 분류 시스템", 정보처리학회 추계학술발표논문집, 제4권 제2호, pp. 625-630,1997.

[5] Adam Blum, "Neural Networks in C++", John Wiley &Sons, INC, pp.55-65, 1992.

[6] Chidanand Apte, Fred Damerau, "Automated Learning of Decision Rules for Text Categorization", ACM TOIS, Vol.12, No.3, pp.233-251, 1994.

[7] David. D. Lewis, "Evaluating Text Categorization", Proceedings of the Speech and Natural Language Workshop, pp.312-318, Asilomar, 1991.

[8] Gerard Salton, "Automatic Text Processing," Addison Wesley.INC, pp.275-280, 1989.
[9] I .Khan, D. Blight, "Categorizing Web Documents Using Competitive Learning", ICNN' 97, Vol.1, pp.96-99, 1997.
[10] Introduction to Rainbow URL : http ://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html.
[11] J. P. Bigus, J. Bigus, "Constructing Intelligent Agents with Java," Wiley&Sons INC, pp.127-130, 1997.
[12] M. Sasaki, K. Kita, "Rule-Based Text Categorization Using Hierarchical Categories," IEEE SMC'98, Vol. 3, pp.2827-2830, 1998.
[13] Ray Liere, Prasad Tadepalli, "The Use of Active Learning in Text Categorization," Working notes of the AAAI Spring Symposium on Machine Learning, Stanford, 1996.
[14] S. Y. Kung, "Digital Neural Networks," Prentice Hall, pp.184-187, 1993.
[15] Vittorio Castelli, Thomas M.Cover, On the Exponential Value of Labeled Samples , Pattern Recognition Letters, Vol.16, No.1, pp.105-111, 1995.
[16] W. Pedrycz, Z.A.Sosnowski, "Designing Decision Trees with the Use of Fuzzy Granulation," IEEE TSMC Part A, Vol.30, No.2, pp.151-159, 2000.
[17] Y. H. Li, A. K. Jain, "Classification of Text Documents," The Computer Journal Vol.41, No.8, pp.537-546, 1998.



한 광 록

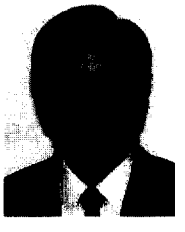
1984년 인하대학교 전자공학과 졸업(공학사)
1986년 인하대학교 대학원 정보공학전공(공학석사)
1989년 인하대학교 대학원 정보공학전공(공학박사)
1989년~1991년 한국체육과학원 선임 연구원
1991년~2000년 현재 호서대학교 컴퓨터공학부 교수
1999년~현재 호서대학교 벤처전문대학원 컴퓨터응용기술팀장
관심분야 : 정보검색, 자연언어처리, 기계번역, HCI, 지능형 에이전트 등

선 복 근

1999년 호서대학교 컴퓨터공학과 졸업 (공학사)
2000년 호서대학교 벤처전문대학원 컴퓨터응용기술팀 석사과정
관심분야 : 정보검색, 에이전트, Mobile Network 등

한 상 태

1987년 고려대학교 통계학과 졸업 (경제학학사)
1991년 고려대학교 대학원 졸업 (통계학 석사)
1995년 고려대학교 대학원 졸업 (통계학박사)
1997년~현재 호서대학교 자연과학부 수학전공 교수
관심분야 : 데이터마인닝, 통계자료 분석, 정보분석 시스템 개발 등



임 기 욱

1977년 인하대학교 공과대학 전자공학과 졸업
1987년 한양대학교 전자계산학 석사
1994년 인하대학교 전자계산학 박사
1977년~1983 한국전자기술연구소 선임연구원
1983년~1988년 한국전자통신연구소 시스템소프트웨어 연구실장
1988년~1989년 미 캘리포니아 주립대학(Irvine) 방문연구원
1989년~1997년 한국전자통신연구원 시스템연구부장 주전산기(타이컴) III, IV 개발 사업책임자
1997년~2000년 정보통신연구진흥원 정보기술전문위원
2000년~현재 선문대학교 교수
관심분야 : 실시간 데이터베이스시스템, 운영체제, 시스템구조 등