1999 年度 大韓電子工學會 夏季綜合學術大會 論文集 第 22 卷 第 1 號 99/6

문서 자동 분류기의 구현을 위한 문서 학습 방법에 관한 연구

선 복 근 이 인 정 한 광 록
호서대학교 컴퓨터 공학부
bksun@shinbiro.com, krhan@office.hoseo.ac.kr

A Study on the Learning Method of Documents for Implementation of Automated Documents Classificator

BokKeun Sun, InJung Lee, KwangRok Han
Dept. of Computer Engineering, Hoseo University

Abstract

We study on machine learning method for automatic document categorization using back propagation algorithm. Four categories are classified for the experiment and the system learns with 20 documents per a category by this method.

As a result of the machine learning, we can find that a new document is automatically classified with a category according to the predefined ones.

1. 서 론

인터넷이 활성화되면서 많은 양의 웹문서가 생성되어지고 있다. 이에 따라 인터넷 사용자가 인터넷을 통해 얻을 수 있는 정보의 양 또한 점점 증가하고 있다. 그러나 이러한 대량의 정보는 사용자에게 유용한 정보를 제공해 주지만 사용자에게 정보의 선별 과정을 거쳐야 한다는 부담을 준다. 다시 말하면 이러한 정보의 양적 증가는 반대로 생각해 보면 정보의 질적 하락으로 이어진다는 것이다. 이러한 상황에서 정보의 여과기능이 점점 중요한 기능으로 인식되어지는 것이 사실이다.[1]

정보의 여과기능은 자동 문서 검색, 분류, 요약 등으로 분류할 수 있다.

이중 특히 자동 문서 분류 기능은 가장 기본적이면서도 중요한 정보 여과 기능으로서 주목 받고 있다.[3][7] 현재 이러한 자동 문서 분류를 위한 연구가 큰 주제로 떠오르고 있느나 국내에서는 이에 관한 연구가 많이 이루어지지 않고 있는 실정이다.

문서분류 서비스를 제공하는 많은 업체의 문서 분류가 현재 수동으로 이루어지고 있다. 그러나 정보량의 증가속도에 대처하지 못하고 있는 실정이며, 과중한 인건비로 인한 경제성에도 문제를 가져온다.

한국어 정보처리 문제등으로 인하여 외국의 기술이 국내에 도입되는 것이 지연되고 있으나 이 또한 멀지않은 시간내에 해결되어진다고 볼 때, 관련 연구가 시급히 요구되어지고 있다.

이에 본 논문에서는 문서의 자동 분류에 관하여 연구하며 문서 분류를 위한 학습은 뉴럴 네트워크의 back propagation 알고리즘을 이용한다.

2. 관련 연구

문서의 분류란 정해진 분류체계 하에서 분류하고자 하는 각 문헌들을 가장 적합한 카테고리에 배정함으로써 문헌을 집단화 하는 작업이다.[1] 이러한 문서의 분류를 자동으로 수행하는 것이 문서의 자동 분류이며 이러한 문서의 자동분류 방법에는 여러 가지 방법과 많은 알고리즘의 응용이 있을 수 있다.

대표적으로 단순한 단어의 매칭을 이용한 방법, 확률을 이용한 방법, 벡터의 유사도를 이용한 방

법, 통계적 기법에 인공지능적 기법을 접목한 방법 등이 있을 수 있다. [2][6]

이중 확률을 이용한 문서분류, 벡터의 유사도를 이용한 문서분류는 통계적 문서분류에 해당하며, 실험 문서 집단에서의 단어의 출현 빈도를 근거로 하여 새로운 문서가 분류될 가능성이 가장 높은 문서집단을 찾아내는 방법으로 일반적으로 많이 이용하고 있다. [1]

2.1 Bayesian 확률을 이용한 방법

가 된다.문서의 색인어를 보고 각 문서분류에 분류될 확률을 계산하는 방법이다.[2][6] 사건 E와 C가 있을 때 이 사건이 동시에 일어날 확률은

$p(E \cap C) = p(E \mid C)*p(C) = p(C|E)*p(E)$

따라서 E가 주어졌을 때, C가 발생할 확률은

$p(C|E) = p(E \mid C)*p(C) / p(E)$가 된다.

분류하려는 문서에 단어 $W_1,W_2,W_3....W_n$이 나타난 경우를 사건 E라 하고, 문서가 카테고리 C_j에 분류되는 것을 사건 C_j라고 하고 각 단어가 나타나는 사건이 독립적이라고 가정하면, 이문서가 카테고리 Cj에 분류될 확률은 다음과 같다

$p(C_j|W_1, W_2,..... W_n) = k*p(C_j)*p(W_1 \mid C_j)*p(W_2 \mid C_j)*p(W_3 \mid C_j)*p(W_n \mid C_j)$

여기서 $p(C_j)$와 $p(W_i|C_j)$는 실험집단으로부터 다음과 같이 계산되어 진다.

$p(C_j)$ = 한 문서가 카테고리 Cj로 분류될 확률 = 카테고리 Cj의 문서수 / 총 문서수

$p(W_i \mid C_j)$ = 카테고리 C_j의 한 문서에 단어 W_i가 나타날 확률 = 카테고리 C_j의 단어 W_i의 빈도수 / 카테고리 C_j의 모든 단어의 빈도수

2.2 벡터의 유사도를 이용한 방법

벡터 유사도를 이용한 방법은 분류하려는 문서와 분류 대상 카테고리들을 색인어들의벡터로 구성하고, 두 벡터 사이의 유사한 정도를 비교하여 유사도가 가장 높은 카테고리로 문서를 분류하는 방법이다. 벡터 사이의 유사도는 두 벡터 사이의 각도를 계산하여 각도가 작은경우가 높은 유사도를 갖도록 한다.[1][2][6]

예를 들어서 문서 D는 색인어 W_1,W_2,W_3,W_6의 색인어를 갖고, 카테고리 C_j는 W_1,W_2,W_6,W_7,W_8의 색인어를 갖는다고 하면, 벡터 D와 카테고리 Cj는 D = (1,1,1,0,0,1,0,0), Cj = (1,1,0,0,0,1,1,1)이 되며 이 두 벡터의 유사도는 다음과 같이 계산된다.

$\text{Similarity}(D, C_j) = \cos\theta = D * C_j / |D| \, |C_j|$

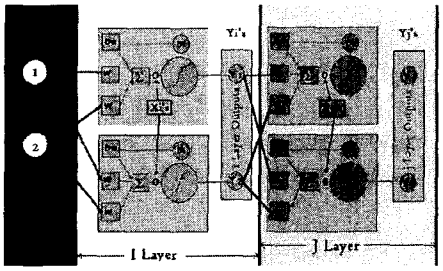
2.3 통계적 기법에 인공지능적 기법을 접목한 방법

위 통계적 기법에 신경망등의 인공지능적 기법을 접목한 방법으로 본 논문의 back propagation 알고리즘이 이에 해당한다.Back propagation에 대한 내용은 3장에 소개된다.

3. Back Propagation 에 의한 학습

본 논문은 뉴럴 네트워크의 back propagation 알고리즘을 이용한다.

3.1 Back Propagation Algorithm



[그림 1] back propagation 네트워크 구성도

Back propagation의 개략적 네트워크 구성도는 위 [그림 1]과 같다.

이 알고리즘은 최소자승 알고리즘의 비선형적 확장으로써 미분의 반복규칙을 여러번 반복적으로 적용하여 네트워크 내부의 weight를 변환함으로써 확률 근사치 프레임워크와 적용함으로써 유도해 낼 수 있다.[4][5][9] Back propagation 네트워크는 입력 학습 데이터로부터 산출된 출력이 원하는 목표치 출력과 같아질 때까지 네트워크 내부의 weight 계산을 반복한다.

네트워크 학습 방법은 아래와 같다.

◆Input layer에서 문서의 vector를 입력 받는다.

◆hidden-layer의 weight가 적용된 합을 계산하고,그 합에 transfer function을 적용해서 hidden-layer의 결과값을 산출한다.

Transfer function은 sigmoid function이며, transfer function의 output은 0 과 1사이의 값을 나타낸다.

$$f(x)=\frac{1}{1+e^{-x}}$$ f : transfer function, x : sum of network layer

◆hidden-layer의 결과값을 output-layer로 전송한다.

◆output-layer의 weight가 적용된 합을 계산하고, 그 합에 transfer function을 적용해서 output-layer의 결과값을 산출한다.

◆실제의 output-layer의 결과값과 target 값을 비교하여 output-layer의 오차에 해당하는 error 값을 계산한다.

$$\delta_o = o_o(1-o_o)(t_o-o_o)$$ o_o : output of output layer, t_o : target of output layer

◆output_layer의 error를 hidden-layer로 back-propagation 시킨다.

◆역전파된 hidden-layer의 error를 weight가 적용된 에러 벡터의 합을 이용하여 계산한다.

$$\delta_h \leftarrow o_h(1-o_h)\sum_{o\in outputs} w_{oh}\delta_o$$ o_h : output of hidden layer, w_{oh} : output layer weights

◆Output-layer와 hidden-layer의 weight를 갱신한다.(즉 weight를 새로운 값으로 바꾼다)

$$w \leftarrow w + \Delta w, \quad \Delta w = \eta\delta_i x$$

i : i layer(hidden layer or output layer)
η : Learning Rate, w : weight, Δw
δ_i : delta value of i layer

◆네트워크의 output이 target값과 정해진 오차만큼의 차이가 날 때까지 반복한다.

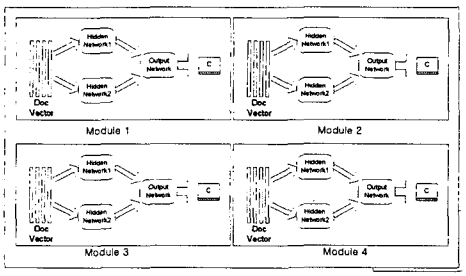
위의 절차는 하나의 입력 벡터에 대한 네트워크 학습 단계이며, 모든 입력 벡터에 대하여 위와 같은 절차를 반복하게 된다.
또한 위 모든 절차는 하나의 모듈에 해당하며 각 모듈은 쓰레드 형태로 병렬 처리되어진다.

4. 문서 자동 분류 시스템 구현

본 논문 에서는 three layer - Input, Hidden, Output - 의 네트워크를 이용하여 back propagation 알고리즘을 적용하였다. Layer의 개수가 많아질수록 보다 정확한 학습을 수행 할 수 있으나, 학습 시간이 상당히 오래 걸리며, 반면 Layer의 개수가 적어질수록 학습 시간은 줄어드나 정확한 학습을 수행 할 수 없다. 어떠한 규칙에 따라서 이 layer의 수가 결정되는 것은 아니며, 경험치에 의해 적당한 layer를 산출하여야 한다.

4.1 학습 Module

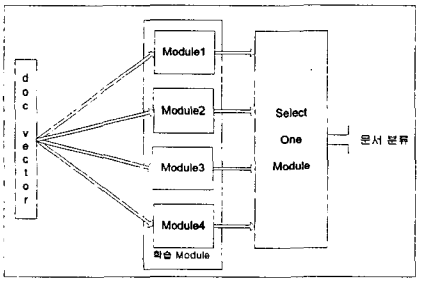
본 논문에서 구현한 문서분류기중 학습 Module의 framwork은 [그림 2]와 같다.



[그림 2] 학습 Module의 Framework

Doc Vector는 문서 벡터의 집합으로써 한 문서 분류에 속하는 예제 문서들을 벡터화 하여 나타낸 것이다. Hiddden Network과 Output Network을 거쳐 나온 출력이 원하는 출력이 될 때까지 Network layer의 weight vector값을 조정하면서 학습을 수행하게 된다.

4.2 문서 분류 Module



[그림 3] 문서 분류 Module의 Framework

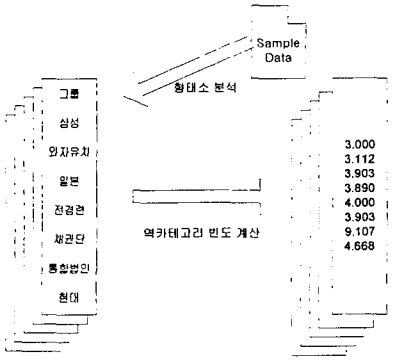
[그림 4]에서와 같이 분류하려는 문서를 벡터화 한 후 학습된 학습 Module에 입력 값으로 주어 [그림 3]과 같이 나오는 출력에 따라 문서를 분류하게 된다.

5. 실험 및 평가

구현된 문서분류기의 실험은 다음과 같이 이루어졌다.

정치, 경제, 정보통신, 스포츠의 4개 문서분류항목을 정하고, 각 문서분류당 5개, 총 20개의 문서를 학습 시켰다.

문서의 벡터화는 형태소 분석과 색인 작업 수행 후 산출된 각 단어의 weight값을 가지고 산출하였으며, 각 문서벡터의 Unit수는 8개로 정하였다. 문서벡터의 추출 과정은 [그림 4]와 같다.



[그림 4] 문서의 벡터화 과정

[그림 4]에서 왼쪽의 단어 데이터는 Sample text 파일에서 형태소 분석 단계를 거쳐 데이터베이스(MS-ACCESS)로 구축 된 것이며, 오른쪽의 벡터 데이터는 구축된 데이터 베이스에서 역 카테고리 빈도계산법에 의해 벡터화 한 Sample.dat 파일이다. 위 작업은 형태소 분석기를 포함한 색인어 생성기를 통해 수행되었다.

위와 같은 단어 벡터 20개를 이용하여 학습을 수행하였다. 학습을 수행하기 위하여 간단한 자동 문서 분류기를 구현하였으며 실험 PC의 환경은 Windows NT 4.0, Peutium 300MHz, 128M RAM이다.

학습 결과의 평가 문서데이터 및 평가 결과는 [표 1]과 같이 나타남으로써 학습이 잘 수행되었음을 보여준다.

[표 1] 학습 데이터 및 학습 결과

Documents Vectors	정치	경제	정통	스포츠
선거구제,총재,파업,한나라 (3.903,4.668,2.602,3.0,0.0,0.0)	0.946	0.3,	0.3 ,	0.3 ,
감독,골프장,농구,박찬호,선수,세이브 (3.112,2.602,3.903,9.107,6,3,0,0)	0.3,	0.3,	0.3,	0.97
ATM,ETRI,PC,데이터,버그,웹,접속,파일 (9.107,2.602,5,2.602,6,4,2.602,2.602)	0.3,	0.3,	0.996	0.3,
모터쇼,법인,원유,자동차,주가,채권, (7.806,9.107,5.204,7.806,2.602,3,0,0)	0.3,	0.977	0.3,	0.3,

6. 결론

본 논문은 최근 주제가 되고 있는 한국어 문서 분류기의 구현하는데 있어서, 학습 방법으로써 뉴럴 네트워크의 back propagation 알고리즘을 이용하는 것에 대해 연구하였다.

본 논문의 실험을 위해 문서 분류기를 구현 하였으며, Sample data 수집과 테스트 시간상의 부족으로 많은 데이터를 이용하여 테스트 하지는 못하였지만, Sample data와 구현된 문서 분류기를 통하여 학습 방법을 테스트 한 결과, 만족할 만한 결과를 얻었다.

실행결과 프로그램 실행시간의 단축, 학습을 위한 문서 데이터의 수집, 알고리즘의 최적화등의 문제가 제기 되었다.

향후 연구과제로써, 웹 문서의 자동 분류를 위해 본 논문에서 보완해야 할 내용으로 웹 문서 태그 정보를 분석한 가중치 적용, 세부 문서 분류체계 설정 등이 남아 있다[8].

참 고 문 헌

[1]조광제, 김준태, "역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류", 정보과학회 봄 학술발표논문집, pp508-510, 1997

[2]정영미, "정보 검색론", 구미무역 출판부,1993

[3]한미성,송영훈,송점동,이정현, "확률벡터간의 교차 엔트로피 계산을 이용한 자동 문서 분류 시스템", 정보처리학회 추계 학술발표 논문집 4권 2호,pp625-630, 1997

[4]김응수 譯,"C로 만든 뇌의 정보 시스템",생능출판사,1996

[5]박문용,최항식 譯,"뉴로 컴퓨터",대영사,1991

[6]W.Frakes and R.Baeza-Yates, "Information Retrieval" Prentice Hall, 1992

[7]D.D.Lewis. "Representation and Learning in Information Retrieval", Ph D. Thesis, 1992

[8]I.Khan, D.Blight, R.D. McLeod, and H.C.Card "Categorizing Web Documents using Competitive Learning", International Conference on Neural Networks(ICNN'97)

[9]D.E Rumelhart. G.E.Hinton, and R.J.Willams, "Learning representations by back-propafating errors", Nature(London),323,1986