고객 질의 문서 자동 분류를 위한 학습 알고리즘 성능 평가

최 정 민[†] · 이 병 수[††]

요 약

최근 인터넷의 보급으로 전자상거래가 대중들에게 나타났고 현재 기업들의 경영환경 변화를 주도하고 있다. 전자상거래에서는 기업이 고객과의 유지 및 관계 구축을 위하여 고객이 원하는 것이 무엇인가를 파악하고 그것을 고객에게 제안하는 여러 가지 고객 채널을 가지고 있는데, 그 중 게시판과 전자메일은 고객의 질의를 직접적으로 들을 수 있는 인바운드(Inbound) 정보로서 매우 중요한 채널로 다루어 지고 있다. 그러나 현재 운영되는 전자상거래의 게시판과 전자메일은 체계적인 관리와 처리과정 없이 질의와 답변이 이루어지고 있는 실정이다. 따라서 본 연구에서는 이러한 문제점의 해결을 위해 인공지능 분야의 문서 분류에서 널리 사용되고 있는 기계학습 알고리즘 중 대표적인 나이브 베이지안(Naive Bayesian), TFIDF, 신경망, k-NN 알고리즘을 도입하여 전자상거래에서 존재하는 여러 가지 고객 질의의 카테고리를 자동으로 분류할 수 있도록 함으로써 관리자가 정확한 답변을 신속하게 처리할 수 있도록 하였다. 그리고 도입한 알고리즘의 고객 질의 문서 자동 분류 성능 실험을 통해 어떤 알고리즘이 우수한 분류 성능을 나타내는지 확인하였으며 실험 결과 나이브 베지이지안 알고리즘이 95%이상의 높은 분류 성능을 나타내는 것을 확인하였다.

키워드 : eCRM, 문서분류, 기계학습

Performance Evaluation on the Learning Algorithm for Automatic Classification of Q&A Documents

Choi Jung Min[†] · Lee Byoung Soo[††]

ABSTRACT

Electric commerce of surpassing the traditional one appeared before the public and has currently led the change in the management of enterprises. To establish and maintain good relations with customers, electric commerce has various channels for customers that understand what they want to and suggest it to them. The bulletin board and e-mail among them are inbound information that enterprises can directly listen to customers'opinions and are different from other channels in characters. Enterprises can effectively manage the bulletin board and e-mail by understanding customers'ideas as many as possible and provide them with optimum answers. It is one of the important factors to improve the reliability of the notice board and e-mail as well as the whole electric commerce. Therefore this thesis researches into methods to classify various kinds of documents automatically in electric commerce; they are possible to solve existing problems of the bulletin board and e-mail, to operate effectively and to manage systematically. Moreover, it researches what the most suitable algorithm is in the automatic classification of Q&A documents by experiment the classifying performance of Naive Bayesian, TFIDF, Neural Network, k-NN

Key Words : eCRM, Document Classification, Machine Learning

1. 서 론

최근 고성능 개인용 컴퓨터의 보급과 네트워크의 발달로 인하여 전통적인 상거래를 뛰어 넘는 전자 상거래가 대중들에게 나타났고 현재 기업들의 경영환경 변화를 주도 하고 있다[1]. 이러한 전자 상거래 경영환경 하에 기업은 새로운 고객관리방법을 연구하고 있으며 전자상거래의 유형 가운데 대표적인 쇼핑 몰에서는 고객과의 유지 및 관계 구축을 위하여 고객이 원하는 것이 무엇인가를 파악하고 그것을 고객에게 제안하는 여러 가지 고객 채널을 가지고 있는데, 그 중 고객 게시판과 전자메일은 고객의 질의를 직접적으로 들을 수 있는 인바운드(Inbound)정보 채널로서 다른 고객 접점 채널 과는 성격이 다른 도구이다[2]. 그러나 현재 대부분의 전자상거래에서 운영되는 게시판과 전자메일은 기 분류된 카테고리를 고객이 직접 질의에 적합한 카테고리를 수동으로 선정하도록 되어 있고, 이렇게 임의로 분류된 고객 질의에 대한 답변은 체계적인 처리 과정 없이 답변이 이루어지고 있는 실정이다. 따라서 이러한 기존 문제점을 해결하기 위해 본 연구에서는 인공지능 분야에서 문서 분류를 위해 널리 사용되고 있는 기계학습 중 대표적인 나이브 베이지안(Naive Bayesian), TFIDF, 신경망, k-NN 알고리즘을

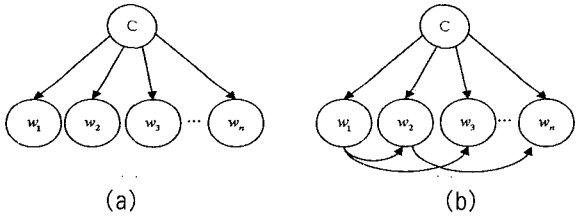
†준 회 원 : 인천대학교 컴퓨터공학과
††종신회원 : 인천대학교 컴퓨터공학과 교수
논문접수 : 2005년 7월 13일, 심사완료 : 2005년 11월 30일

도입하여 고객 질의에 대하여 자동으로 카테고리를 분류할 수 있도록 함으로써 관리자가 정확한 답변을 신속하게 처리할 수 있게 하였다. 이것은 궁극적으로 전자상거래 게시판과 전자메일에 대한 고객 신뢰도를 향상시키고 나아가 전자상거래 전체의 신뢰도를 높일 수 있는 효과가 있다. 본 논문에서는 고객 질의 문서 자동 분류를 위해 도입하는 기계학습 알고리즘의 자동 분류 성능 실험을 현재 인터넷 상에서 운영중인 ebay 사이트의 고객 질의 문서를 수집하여 다양한 조건에서 실험을 실시 하였고 조건에 따른 알고리즘별 성능에 대한 실험 평가를 수행 하였다. 논문의 구성은 다음과 같다. 2장에서는 관련연구로 기계학습 알고리즘에 대하여 소개 하고, 3장에서는 본 연구의 시스템 설계에 대해 기술하였다. 4장에서는 3장에서 설계한 시스템의 분류 수행 절차에 대하여 기술 하였고 5장에서는 네 가지 알고리즘에 대한 분류 성능 측정을 위한 실험과 평가를 기술 하였다. 마지막으로 6장은 결론 및 향후 연구에 대하여 기술 하였다.

2. 관련 연구

2.1 나이브 베이지안 알고리즘(Naive Bayesian Algorithm)

나이브 베이지안 분류 학습기법은 베이지안 네트워크를 분류기에 적용한 것으로 베이즈 정리(Bayes Theorem)에 기초한 확률 모델을 이용하는 기법이다[13, 15]. (그림 2-1)은 베이지안 네트웍을 나타낸 그림이며, 노드 C는 카테고리를 의미하고 노드 C에 대한 각각의 특성(Feature)을 w_n이라 표시했다.



(a) (b)

(그림 2-1) 베이지안 네트워크

(a)는 각각의 특성들 간에는 서로 조건부 독립(Conditionally Independent)이라는 가정하의 나이브 베이지안 알고리즘을 나타내고, (b)는 특성들 사이에 제한된 종속성을 허용하는 좀더 복잡한 베이지안 알고리즘을 나타내고 있다. 하나의 문서 d가 (w_1, w_2, …, w_n)의 특성들로 이루어졌을 때 베이지안 학습기법은 식(1)와 같이 문서 d에 대한 조건부 확률이 가장 큰 카테고리로 분류한다.

$$\arg\max_{c \in C} P(c \mid d) = \arg\max_{c \in C} P(c \mid w_1, w_2, \ldots, w_n) \quad (1)$$

이 식을 승법정리(multiplication theorem)를 이용하여 풀어 내면 식(2)와 같이 된다.

$$\arg\max P(c \mid w_1, w_2, \ldots, w_n) = \frac{P(w_1, w_2, \ldots, w_n \mid c)P(c)}{P(w_1, w_2, \ldots, w_n)} \quad (2)$$

식(2)에서 확률 $P(w_1, w_2, \ldots, w_n)$는 하나의 상수인 정규화 항이므로 우리가 가장 가능성이 높은 하나의 카테고리를 결정하는 것에만 관심이 있는 경우에는 그 값이 1이 된다. 따라서 식은 식(3)과 같이 된다.

$$\arg\max P(c \mid w_1, w_2, \ldots, w_n) = P(w_1, w_2, \ldots, w_n \mid c)P(c) \quad (3)$$

여기서 문서 d를 나타내는 특성인 각 w_i는 모든 다른 특성들과 조건부 독립이라는 나이브 베이지안 가정을 적용하여 식(4)와 같이 된다.

$$P(w_1, w_2, \ldots, w_n \mid c) = \prod_{i=1,n} P(w_i \mid c) \quad (4)$$

결론적으로 나이브 베이지안 알고리즘은 분류 대상 문서 d에 대해 가장 가능성이 높은 분류 클래스를 식(5)와 같이 계산한다.

$$\arg\max_{c \in C} P(c \mid d) = \arg\max_{c \in C} P(c) \prod_{i=1,n} P(w_i \mid c) \quad (5)$$

2.2 k-NN 알고리즘(k-Nearest Neighbor Algorithm)

대표적인 또 다른 알고리즘으로는 k-NN(k-Nearest Neighbor)이 있다. 이 알고리즘은 분류할 문서 d=(w_1, w_2, …, w_n)와 저장된 카테고리별 훈련문서 d'=(w'_1, w'_2, …, w'_n)의 유클리드 거리(Euclidean distance)를 계산하여 분류대상 문서와 가장 가까운 훈련문서 k 개를 선정한다.

$$Dist(d, d') = \sum_{i=1}^{n} \sqrt{(w_i - w'_i)^2} \quad (6)$$

그리고 선정된 k개 중에서 가장 많은 개수의 훈련예제가 소속된 카테고리로 분류대상 문서 Y가 분류된다. 여기서 i는 클래스의 종류이며 n은 클래스의 개수이다. 여기서 k값은 k-NN의 성능을 최적화하기 위하여 일반적으로 교차검증(cross validation) 기법을 사용하여 사전에 결정하며, k=1인 경우를 NN(Nearest Neighbor) 알고리즘이라고 한다[13].

2.3 TFIDF 알고리즘(Term Frequency Inverse Document Frequency Algorithm)

전통적으로 정보 검색 분야에서 많이 사용되어온 TFIDF 알고리즘에서는 각 문서 d를 특성단어(feature word)의 출현 빈도수에 기초한 가중치 벡터(weight vector)로 표현한다. 이 때 각 단어의 가중치 w_i는 식(7)과 같이 문서 d에 나타나는 빈도수인 TF(Term Frequency)와 그 단어가 나타나는 총 문서 수에 대한 역수인 IDF(Inverse Document Frequency)의 곱으로 계산된다. 이것은 한 단어가 특성 문서에 나타나는 빈도수는 높고 다른 문서에 나타나는 빈도수가 낮을수록 다른 문서에 비해 그 문서를 잘 표현해줄 수 있다는 의미를 담고 있다.

$$w_i = TF_i \cdot IDF_i \quad (7)$$

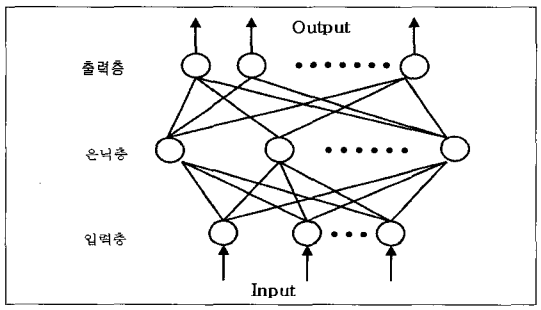
문서 분류작업을 위해서는 각 카테고리별로 그 카테고리를 나타내는 프로토타입 벡터(prototype vector)를 구한다. 이때 각 카테고리의 프로토타입 벡터 c는 그 카테고리에 속한 훈련문서들의 TFIDF 가중치 벡터들의 평균으로 계산한다. 일단 이처럼 각 카테고리들이 프로토타입 벡터로 표현되어 있으면, 식(8)과 같이 분류대상 문서 d의 가중치 벡터와 각 카테고리 c의 프로토타입 벡터간의 유사성(similarity)을 cosine 규칙을 적용하여 계산한다. 그리고 이와 같은 과정을 거쳐 가장 유사하다고 판단되는 카테고리로 문서를 분류한다[13].

$$\arg\max_{c \subset C} \cos(c,d) = \arg\max_{c \subset C} \frac{c}{\|c\|} \cdot \frac{d}{\|d\|} \qquad (8)$$

2.4 신경망 알고리즘(Neural Network Algorithm)

신경망(Neural Network)은 병렬 분산처리(Parallel Distributed Processing) 원리를 근간으로 하여 생물학적인 신경 회로를 수학적으로 모델링하여 구현한 것이다. 신경망의 구조는 입력층(Input Layer), 은닉층(Hidden Layer), 출력층(Output Layer)로 이루어져 있다.

(그림 2-2)는 신경망의 구조를 나타낸 그림이다. 그림에서 입력층은 신경망에서 자료가 제공되는 층이고 출력층은 주어진 입력에 대해 인공 신경망이 반응하여 결과를 내는 층이다. 또한 은닉층은 입력층과 출력층 사이에 존재하는 것으로서 입력 층의 모든 노드들과 완벽하게 연결되어 있다. 은닉층에 존재하는 노드들은 각각의 입력에 상응하는 가중치를 곱함으로써 값을 계산하여 이를 출력층에 넘겨 주는 역할을 한다. 신경망의 학습방법은 각 처리 요소들의 출력 값들은 신경 회로망의 출력 결과를 결정하게 되므로 신경망을 이용하여 원하는 출력 값을 얻기 위해서는 연결 가중치를 조절해야 한다. 모든 신경망은 주어진 자료들을 이용하여 처리 요소들 간의 연결 가중치를 스스로 조정하게 되는데 이 과정을 학습이라 한다[11].



(그림 2-2) 신경망의 구조

3. 시스템의 설계

3.1 기본 가정

본 논문의 고객 질의 문서 자동 분류 시스템은 질의 문서의 특성과 효과적인 분류 작업을 위하여 세 가지 기본 가정을 전제로 운영된다.

첫째, 일반적으로 질의 작성자들이 게시판과 전자메일에 질의를 기록 할 때는 질의에 적합한 함축적인 단어를 제목에 기입하게 된다. 따라서 전자 메일과 게시물의 분류는 단지 질의의 내용만으로 분석하기 보다는 작성자가 만들어 놓은 질의의 제목과 내용에 동일한 가중치를 두어 함께 분석함으로써 작업이 이루어진다. 즉, 분류 작업을 위한 분석 자료로 질의의 내용뿐만 아니라 제목까지도 고려하여 작업의 효율성을 높이고자 한다.

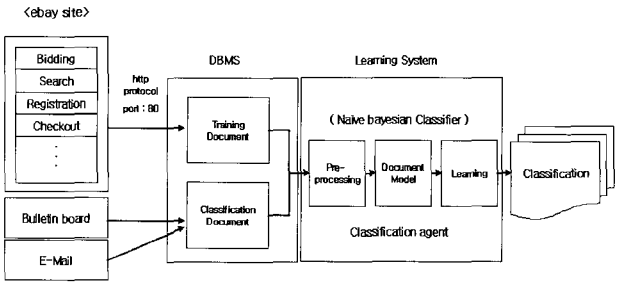
둘째, 고객 질의 문서 분류를 위해 도입하는 분류 학습 알고리즘은 모두 다수의 훈련문서(Training Document)를 필요로 하는 교사학습(Supervised Learning) 방법들이다. 본 시스템에서 사용하는 분류 학습 알고리즘은 나이브 베이지안, k-NN, 신경망, TFIDF이며, 이들은 모두 분류 카테고리별로 충분한 훈련문서들을 미리 확보하고 있어야 높은 분류 성능을 기대할 수 있다.

셋째, 자동 분류 작업 이전에 사용자로부터 직접적으로 충분한 훈련문서들을 얻을 수 없는 경우를 위해 eBay 쇼핑 사이트의 답변 센터(Answer Center)의 카테고리와 해당 문서들을 가져와 훈련문서로 사용한다.

3.2 시스템의 구조

본 연구의 고객 질의 문서 분류 시스템의 구조는 (그림 3-1)과 같다. 시스템의 수행 과정은 먼저 분류 대상 고객 질의 문서를 게시판과 전자메일 서버에서 시스템으로 가져오고 분류 알고리즘에 사용할 카테고리들과 훈련문서들을 확보한다. 이에 본 논문에서는 시스템의 수행과 평가를 위하여 인터넷 쇼핑 서비스를 제공하는 eBay 사이트의 답변 센터의 분류 카테고리와 해당 카테고리에 속한 웹 문서들을 이용하였다.

이와 같은 과정을 거쳐 분류 작업을 위한 훈련문서들과 분류 대상 문서들이 수집되면, 적절한 문서 전처리 과정(Document preprocessing)과 문서 모델화(Document Modeling) 과정을 거친다. 그리고 수집된 분류 카테고리들과 훈련문서들을 바탕으로 문서 분류 학습 알고리즘의 학습이 진행되고 이것을 바탕으로 최종적으로 분류 대상 문서들이 차례대로 분류된다.



(그림 3-1) 시스템 구조

4. 시스템 수행 절차

4.1 문서의 수집

고객 질의 문서 분류 작업을 수행하기 위하여 본 연구에서는 훈련문서로 사용될 문서들을 인터넷으로부터 수집한

다. 훈련문서로 사용되는 문서의 수집을 위해서는 분류 카테고리들과 카테고리에 대응되는 양질의 훈련문서를 사전에 확보해야 하는데 이를 위하여 현재 인터넷상에서 운영되고 있는 eBay 사이트 답변 센터(Answer Center)의 카테고리와 해당 카테고리의 질의 문서를 사용하였다.

훈련문서의 수집을 위하여 eBay 사이트를 선정한 이유는 다른 쇼핑 사이트에서는 관리상 고객 질의를 공개하지 않고 운영하고 있기 때문에 훈련문서를 수집하기가 쉽지 않았다. 그러나 eBay 사이트의 답변 센터에는 총 22개의 카테고리로 쇼핑 사이트에서 발생하는 실제 고객 질의를 분류해 놓았으며 본 시스템에서는 시스템의 자동 분류 능력 수행 평가를 위하여 이 중 10개의 카테고리와 각 카테고리당 50개씩, 총 500개의 고객 질의 문서를 수집하여 훈련문서로 사용하였다. <표 4-1>은 본 연구에서 실험을 위해 선정된 10개의 카테고리를 표로 정리한 것이다.

〈표 4-1〉 시스템 선정 분류 카테고리

Bidding	Escrow_Insurance
Packaging_Shipping	PayPal
PoliciesUser_Agreement	Registration
Trust_Safety	Checkout
Search	Miscellaneous

4.2 질의 문서의 전처리

수집된 질의 문서에 대하여 분류 학습기법을 적용하기 위해서는 각 훈련문서와 분류 대상 문서들을 적절한 특성 단어(Feature Word)들에 기초한 벡터 모델로 표현하여야 한다. 이를 위해서 먼저 질의 문서에 대한 전처리 과정(Preprocessing)이 필요하다. 전처리 과정에서는 각 질의 문서를 구성 단어별로 나누는 작업과 더불어 웹 문서에서 태그(< >)를 제거하는 작업, and, but 등의 문서를 대표할 수 없는 단어들의 집합인 불용어(Stop Word)를 제거하는 작업, 그리고 단어들의 어미 변화에 대한 처리인 스테밍(Stemming) 처리작업 등이 이루어진다.

4.3 특성 추출 및 모델화

특성 추출(Feature Selection)은 학습 자원의 중요 속성들을 자원이 구분된 카테고리별로 다시 한번 중요도를 정의하는 특성 추출 가중치 설정 기법이다. 이를 위하여 각 학습 자원들의 특성을 고려하여 구분된 카테고리들을 대상으로 일련의 구별 작업을 두어 이를 기반으로 한 속성 추출 작업을 수행할 필요가 있다. 이러한 가중치 설정 작업은 해당 키워드가 속해있는 카테고리의 정보를 고려하여 이루어지며 이것으로써 각 카테고리를 대표하는 키워드에게 더욱 높은 가중치가 설정된다. 이러한 특성 추출에 대한 기계학습 방식은 서로 다른 두 카테고리가 존재하는 경우, 각각의 카테고리 별 키워드에 가중치를 주는 것이다. 이러한 과정 후 본 시스템에서는 수집된 모든 질의 문서에 대하여 이진 속성 벡터(vector of binary attributes)로 모델화 한다. 수집된 질의 문서에 대해 이와 같은 이진 속성 벡터를 만들기 위해서는 먼저 질의 문서들로부터 각 문서를 표현하는데 사용할 단어들인 특성(feature)들을 추출하여야 한다. 본 연구에서 문서의 특성을 추출하는 방법으로는 정보이론(Information Theory)에 입각한 엔트로피(entropy) 변화량을 기초로 특성 단어를 추출 하는 방법인 정보 획득(Information Gain)방법을 사용한다.[18]

$$V = \{w_1, w_2, w_3, \ldots, w_n\}$$

$$InfoGain(w) = P(w)\sum_i P(c_i \mid w)\log\frac{P(c_i \mid w)}{P(c_i)} + P(\bar{w})\sum_i P(c_i \mid \bar{w})\log\frac{P(c_i \mid \bar{w})}{P(c_i)} \quad (9)$$

식(9)에 의해 식(10)과 같이 전체 단어집합(V)에서 정보 획득량이 큰 L개의 단어를 추출한다.

$$K = \{w_1, w_2, w_3, \ldots, w_L\} \quad , \quad K \subset V \quad (10)$$

그리고, 추출된 L개의 특징 단어들을 바탕으로 각 질의 문서에 대해 아래와 같은 모양의 이진 속성 벡터 모델을 만들게 된다.

$$d_i = (1,0,1,\ldots,1) \quad (11)$$

4.4 학습 및 분류

본 절에서는 알고리즘의 고객 질의 문서 분류에 대한 학습 및 분류 과정을 설명하기 위하여 네 가지 학습 알고리즘 중 최종 실험에서 가장 우수한 성능을 나타낸 나이브 베이지안 알고리즘의 학습과 분류를 설명하였다. 나이브 베이지안 학습기법은 문서 d_i의 각 카테고리 c_j에 대한 조건부 확률들을 식(13)과 같이 구해준다.

$$\Re(d_i) = \{P(d_i \mid c_1), P(d_i \mid c_2), P(d_i \mid c_3), \ldots, P(d_i \mid c_k)\} \quad (12)$$

결과적으로 분류 대상 문서에 대하여 식(14)와 식(15)에 의해 가장 높은 확률 값을 가지는 카테고리로 분류하게 된다.

$$P_{\max}(d_i) = \max_{c_i \in c}\{P(d_i \mid c_i)\} \quad (13)$$

$$c_{best}(d_i) = \begin{cases} c_j & \text{if } P_{\max}(d_i) = P(d_i \mid c_j) \geq T \\ c_{unknown} & otherwise \end{cases} \quad (14)$$

그러나, 클래스의 확률 값이 일정한 임계 값(threshold) T 이상 되지 않으면 그만큼 분류에 대한 정확도와 신뢰도가 떨어진다, 따라서 이러한 경우에 분류가 자동으로 이루어지지 않고 사용자에게 분류 결정을 양도 하게 된다.

5. 실험 및 성능 평가

5.1 실험 환경

고객 질의 문서 분류는 질의 문서들의 내용을 시스템이 파

악하여 문서가 속한 정확한 카테고리를 정하는 작업으로 사전에 정해져 있는 카테고리들 중에서 어떤 카테고리에 해당 질의 문서가 속하는지 판단하는 것이다. 따라서 이 장에서는 질의 문서 분류를 위해 도입한 알고리즘의 분류 성능을 다양한 실험을 통하여 측정하고, 실험 대상 알고리즘 중 어떤 알고리즘이 가장 우수한 분류 성능을 나타내는지 살펴보고자 한다.

본 논문의 실험 대상 알고리즘은 기존 문서 분류 알고리즘으로 대표적인 나이브 베이지안, TFIDF, 신경망, k-NN의 네 가지 알고리즘을 대상으로 실험을 실시하였다. 실험 환경은 분류 성능 실험을 위한 시스템으로 펜티엄Ⅳ 2.3GHz 프로세서와 512MB의 주 기억공간을 사용하는 리눅스 운영체제 환경에서 실험이 이루어졌으며, 실험을 위한 데이터는 기존에 쇼핑 사이트가 운영되고 있는 eBay 쇼핑 사이트의 답변 센터에 있는 카테고리와 해당 카테고리에 속한 질의 문서를 사용하였다. 전체 실험 카테고리는 답변 센터의 22가지 카테고리 중 비교적 카테고리 소속 질의 문서가 많은 10개의 카테고리를 선정하여 실험에 사용하였으며, 각각의 카테고리당 50개씩, 총 500개의 질의 문서를 실험에 사용하였다. 실험방법은 50개씩 문서를 가지고 있는 10개의 카테고리에서 임의로 카테고리당 10개의 질의 문서를 분류대상 문서로 발췌한다. 그리고, 나머지 40개의 문서를 가지고 훈련문서의 개수를 10개, 20개, 40개로 변화시키면서 네 가지 분류 알고리즘의 성능을 비교하고 정확도 값을 측정하였다.

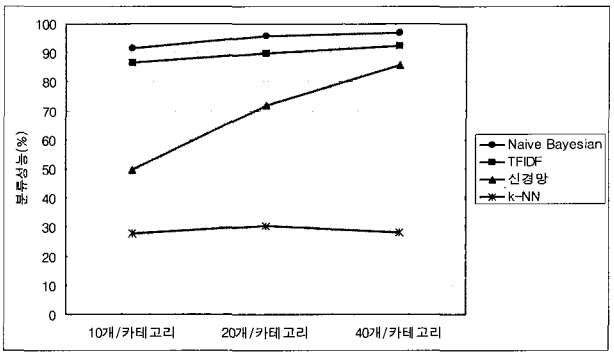
5.2 실험 결과

eCRM을 위한 고객 질의 문서 자동 분류 시스템에서 도입한 네 가지 알고리즘의 실험 결과는 다음 <표 5-1>과 같다. <표 5-1>은 네 가지 알고리즘에 대하여 훈련문서의 개수를 변화하는 실험 조건에 따라 총 5회씩 실험을 실시한 결과값과 평균값을 분류 정확도(%)로 나타내었다.

실험을 실시한 네 가지 학습 알고리즘별 분류 성능의 정

〈표 5-1〉 실험 결과 ; 분류 정확도(%)

훈련문서개수 \ 학습기법		Naive Bayesian	TFIDF	신경망	k-NN
10개/카테고리	1차	94	87	52	25
	2차	92	88	49	20
	3차	93	87	44	33
	4차	89	88	45	25
	5차	90	83	48	36
	평균	91.6	86.6	49.6	27.8
20개/카테고리	1차	97	92	82	32
	2차	95	90	73	18
	3차	97	92	70	37
	4차	94	89	68	37
	5차	96	96	67	28
	평균	95.8	89.8	72.0	30.4
40개/카테고리	1차	98	93	89	29
	2차	98	90	87	26
	3차	97	93	87	27
	4차	96	91	82	30
	5차	96	95	84	28
	평균	97.0	92.4	85.8	28.0



(그림 5-1) 훈련문서의 개수에 따른 알고리즘별 성능변화

확도는 카테고리당 소속되어 있는 50개의 질의 문서를 실험 조건에 따라 임의로 발췌하여 알고리즘별로 재 분류하여 기존에 소속되어 있던 카테고리로 정확히 분류되는 지에 대한 정확도를 측정한 결과이다. 실험 결과 나이브 베이지안이 95%대의 가장 높은 분류 성능을 보여주었으며, TFIDF와 신경망은 훈련문서의 개수가 10개, 20개 일 경우는 낮은 성능을 보였으나 훈련문서의 개수 40개일 경우에는 나이브 베이지안에 필적하는 비교적 높은 성능을 나타냈다. 이에 반해 k-NN은 기대와는 달리 30%~40%의 가장 낮은 분류 성능을 나타냈다. 또한, 나이브 베이지안과 TFIDF, 신경망 알고리즘은 훈련문서의 수가 증가할수록 분류 성능이 조금씩 증가하는 것을 발견할 수 있었다. 그러나 다른 알고리즘에 비해 낮은 분류 성능을 나타낸 k-NN은 훈련문서의 개수에 따른 성능 변화가 정확히 나타나지 않았다.

실험 중 각 알고리즘의 분류 속도는 나이브 베이지안과 TFIDF, 신경망은 비교적 빠른 분류 시간을 보였으나, 개체기반 학습기법(Instance-Based Learning)의 하나로서 분류 당시에 많은 계산시간을 필요로 하는 k-NN 기법은 나머지 세 학습기법에 비해 분류 속도가 매우 느리게 나타났다. (그림 5-1)는 전체 실험 대상 알고리즘의 훈련문서 개수의 변화에 따른 분류 성능을 나타낸 그림이다.

6. 결 론

본 연구는 eCRM을 위한 고객 질의 문서 자동 분류에 대한 연구로써 기업이 전자상거래에서 고객과의 유지 및 관계 구축을 위해 가지고 있는 고객 접점 채널 중 게시판이나 전자메일의 질의 문서를 자동으로 분류하여 정확한 답변을 신속하게 제공할 수 있도록 하기 위함이다. 본 연구에서 중점을 두고 있는 고객 질의 문서는 게시판과 전자메일을 통하여 전자상거래로 유입되는 정보로서 이 두 채널은 고객의 의견을 직접 들을 수 있는 인바운드(Inbound) 정보로 다른 고객 접점 채널과는 성격이 매우 다른 중요한 채널의 정보이다. 그러나 현재 대부분의 전자상거래에서 운영하는 게시판과 전자메일은 체계적인 처리 과정 없이 답변이 이루어지고 있기 때문에 고객은 질의에 대하여 정확한 답변을 기대하기 어려운 실정이며, 또한 고객이 답변을 기다리는데 많은

시간이 소요되고 있다. 따라서 기존의 이러한 문제점을 해결하기 위하여 본 연구에서는 인공지능 분야에서 문서 분류를 위해 널리 사용되고 있는 기계학습 중 대표적인 나이브 베이지안(Naive Bayesian), TFIDF, 신경망, k-NN 알고리즘을 도입하였다. 이것은 여러 가지 고객 질의에 대해 기계학습 알고리즘이 자동으로 카테고리를 분류할 수 있도록 함으로써 관리자가 정확한 답변을 신속하게 처리할 수 있도록 하고 전자상거래 게시판과 전자메일에 대한 고객 신뢰도의 향상과 나아가 전자상거래 전체의 신뢰도를 높일 수 있도록 하였다.

연구 목표를 위해 본 논문에서는 네 가지 알고리즘에 대한 분류 성능 실험을 실시 하여 고객 질의에 대해 어떤 알고리즘이 가장 높은 분류 성능을 갖는지 알아 보았다. 실험은 현재 인터넷 상에서 운영중인 ebay 사이트 답변 센터의 고객 질의 문서를 수집하여 실험 조건에 따라 훈련문서의 개수를 10개, 20개, 40개로 변화하여 동일 조건으로 5회씩 알고리즘별로 수행하였다. 그 결과 각 알고리즘 중 나이브 베이지안이 95%대의 가장 높은 분류 성능을 보여주었으며, TFIDF와 신경망은 나이브 베이지안에 필적하는 비교적 높은 성능을 보여주었다. 이에 반해 k-NN은 30%~40%의 낮은 분류 성능을 나타냈다. 또한, 나이브 베이지안과 TFIDF, 신경망 알고리즘은 훈련문서의 수가 증가할수록 분류 성능이 조금씩 증가하는 것을 발견할 수 있었다. 그러나 다른 알고리즘에 비해 낮은 분류 성능을 나타낸 k-NN은 훈련문서의 개수에 영향을 받지 않았다. 실험 중 각 알고리즘의 분류 속도는 나이브 베이지안과 TFIDF, 신경망은 비교적 빠른 분류 시간을 보였으나, 개체기반 학습기법(Instance-Based Learning)의 하나로서 분류 당시에 많은 계산시간을 필요로 하는 k-NN 기법은 나머지 두 학습기법에 비해 매우 느리게 나타났다.

본 연구의 성능과 효과를 높이기 위해서 앞으로 시행되어야 할 향후 연구과제는 분류 카테고리들 간의 계층 관계 및 중복 관계에 대한 해결과 명확한 양질의 훈련문서를 확보할 수 없는 경우를 대비하기 위한 비교사 학습기법(Unsupervised Learning)에 대한 도입 등을 검토하고 있다.

참 고 문 헌

[1] 김병곤, 최 성, "eCRM 시스템의 개념 및 발전 전망," 정보처리학회지, 제8권, 제6호, pp.7-17, 2001.

[2] 김무엽, "eCRM에서의 고객접점관리와 영업촉진관리," 정보처리학회지, 제8권 제6호, pp.18-24, 2001.

[3] 이경전, "전자상거래 소프트웨어 에이전트," 정보처리학회지, 제6권, 제1호, pp.54-62, 1999.

[4] 이재호, "에이전트 시스템의 연구 및 개발 동향," 정보과학회지 제18권, 제5호, pp.4-9, 2000.

[5] 최중민, "에이전트의 개요와 연구방향," 정보과학회지, 제15권, 제3호, pp.7-16, 1997.

[6] W. Cohen, "Learning Rules that Classify e-mail," Proc. AAAI Spring Symposium Machine Learning and Information Access, pp.18-25, 1996.

[7] R. H. Guttman, A. G. Moukas, and P. Maes, Agents as Mediators in Electronic Commerce, Electronic Markets, Vol. 8, No.1, pp.22-27, 1998.

[8] B. Krulwich and C. Burkey, "The InfoFinder Agent : Learning User Interests through Heuristic Phrase Extraction," IEEE Experts, Vol.12, No.5, pp.22-27, 1997.

[9] D. D. Lewis and M. Ringuette, "A Comparison of Two Learning Algorithms for Text Categorization," Proc. Third Annual Symposium on Document Analysis and Information Retrieval, pp.81-93, 1994.

[10] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," Proc. AAAI-98 Workshop on Learning for Text Categorization, Madison, WI, pp.41-48, 1998.

[11] T. M. Michell, Machine Learning, The McGraw-Hill Company, 1997.

[12] D. Mladenic, "Personal WebWatcher : Design and Implementation," Technical Report IJS-DP-7472, School of Computer Science, Carnegie-Mellon University, Pittsburgh, USA, October, 1996.

[13] S. Myoung, J. Choi, and I. Kim, "BClassifier: A Personal Agent for Bookmark Classification," Proc. International Conference on Parallel and Distributed Systems(ICPADS-2001), IEEE Comp. Soc., pp.713-718, 2001.

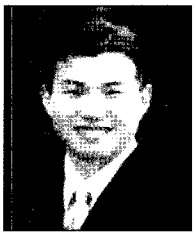
[14] S. Russell and P. Norvig, "Artificial Intelligence : A Modern Approach," Prentice-Hall Series in AI, 1995.

[15] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian Approach to Filtering Junk e-mail," Proc. AAAI Workshop on Learning for Text Categorization, Madison Wisconsin, pp.55-62, 1998.

[16] T. Sandholm, "eMediator : A Next Generation Electronic Commerce Server," AAAI Workshop on AI in Electronic Commerce, Orlando, pp.46-55, 1999.

[17] B. Sheth and P. Maes, "Evolveing Agents for Personalized Information Filtering," Proc. Ninth Conference on Artificial Intelligence for Applications (CAIA-93), IEEE Comp. Soc., pp.345-352, 1993.

[18] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proc. Fourteenth International Conference on Machine Learning(ICML-97), pp.142-420, 1997.



최 정 민

e-mail : cjm@incheon.ac.kr
1999년 경기대학교 화학공학과(학사)
2001년 경기대학교 전자계산학과(석사)
2005년 인천대학교 컴퓨터공학과(박사)
관심분야 : 전자상거래, 인공지능, 데이터마이닝, eCRM, Information Retrival

이 병 수

e-mail : bsl@incheon.ac.kr
1976년 단국대학교 전자공학과(학사)
1980년 동국대학교 대학원전자정보처리(석사)
1998년 경기대학교 전자계산학과(박사)
1981년~현재 인천대학교 컴퓨터공학과 교수
1986년~1989년 인천대학교 전자계산소장
2004년~현재 사단법인 정보처리학회 편집위원회 자문위원
2004년~현재 사단법인 KIPS-IT 인증원 이사
관심분야 : S/W Design, e-Business, 의사결정 시스템, 데이터마이닝, eCRM