논문번호 04-02-06

나이브 베이지안 분류기를 이용한 게시물 자동 분류를 위한 eCRM 에이전트 시스템

eCRM Agent System for Articles Automatic Classification System based on Naïve Bayesian Classifier

崔 正 珉*, 李 丙 洙*
Jung-Min Choi* , Byoung-Soo Lee**

요 약

최근 전자 상거래에서 사용하고 있는 게시판은 고객의 능동적인 참여로 운영되며, 게시물은 고객의 직접적인 의사를 들을 수 있는 인 바운드(Inbound)정보로서 다른 eCRM을 위한 고객 접점 채널 과는 성격이 다른 도구이다. 또한 게시판의 효과적인 운영은 게시판 자체의 신뢰도를 향상 시키고 나아가 전자 상거래 전체의 신뢰도를 높여 줄 수 있는 중요한 eCRM 도구이다. 그러나 현재 대부분의 전자상거래에서 운영하는 게시판은 기 분류된 카테고리를 고객이 직접 수동으로 선정하도록 되어 있고, 이렇게 임의로 분류되는 게시물에 대하여 체계적인 처리 과정 없이 답변이 이루어지기 때문에 답변을 하는데 많은 시간이 소요 되고 있으며, 정확한 답변이 이루어지지 않고 있는 실정이다. 따라서, 본 논문에서는 여러 가지 종류의 게시물에 대하여 나이브 베이지안 분류기를 이용하여 게시판의 기존 문제점의 해결과 효과적인 운영 그리고 게시물의 체계적인 분류 관리를 할 수 있는 게시물 자동 분류기를 설계하고 구현하였다. 아울러 문서 분류 학습 기법 중 대표적인 TFIDF. k-NN, 나이브 베이지안 기법들의 게시물 분류 성능을 측정하여 채택한 나이브 베이지안 분류기의 우수성을 확인 하였다.

Abstract

The customer's bulletin board is the important channel to get opinions from customers directly. The effective management of the bulletin board for thecustomer improves the reliance by providing the best replies and by accepting opinions of the customer and furthermore, that can raise the customer's reliance of the whole shopping mall is the important eCRM method. But, the present mostly customer's bulletin board is been replied without any classifying about many kinds of question. Consequently, The shopping mall should do systematic management of the best professional reply about many kinds of question. In order to resolve this problem, we implement a classifier called Naïve Bayesian classifier is classified automatically bulletin board for eCRM of shopping mall.

키워드 : eCRM, 기계학습(Machine Learning), agent, 문서 분류, 게시판,

서론

최근 고성능 개인용 컴퓨터의 보급과 네트워크의 발달로 인하여 인터넷의 보급이 급속히 이루어 지고 있다.

* 仁川大學校 컴퓨터공학과
(Department of Computer Engineering,
University of Incheon)

接受日:2004年 9月 8日, 修正完了日:2004年 12月 14日

전통적인 상거래를 뛰어 넘는 전자 상거래가 대중들에게 나타났고 현재 기업들의 경영환경 변화를 주도 하고 있다. 이러한 전자 상거래 경영환경 하에 기업은 새로운 고객관리방법을 연구하고 있으며, 그 중 eCRM 연구는 기업의 궁극적인 가치가 고객으로부터 나온다는 전제 하에 고객의 성향에 대한 정확한 파악 및 고객과의 최적의 채널 구축 등을 위하여 모든 기업의 관심의 대상으로 떠오르고 있다. 따라서 전자상거래에서는 고객과의 유지 및 관계 구축을 위하여 고객이 원하는 것이 무엇인가를 파악하고 그것을 고객에게 제안하는 여러 가지 고객 채널을 가지고 있는데, 그 중 고객 게시판의 게시물은 고객의 의견을 직접적으로 들을 수 있는 인바운드(Inbound)정보로서 eCRM을 위한 다른 고객 접점 채널 과는 성격이 다른 도구이다. 따라서, 게시판의 효과적인 운영은 고객의 의사를 최대한 수용하여 그에 대한 최적의 답변을 제공함으로써 게시판의 신뢰도를 향상 시키고 나아가 전자상거래 전체의 신뢰도를 높여 줄 수 있는 중요한 eCRM 도구이다. 그러나 현재 대부분의 전자상거래에서 운영하는 게시판은 기 분류된 카테고리 중 고객이 직접 본인 의견에 적합한 카테고리를 수동으로 선정하도록 되어 있고, 이렇게 임의로 분류 되어진 게시물에 대한 답변은 체계적인 처리 과정 없이 답변이 이루어 지고 있기 때문에 게시물에 대하여 정확한 답변을 하는데 어려운 실정이며, 또한 많은 시간이 소요 되고 있다. 따라서, 본 논문에서는 여러 가지 종류의 게시물에 대하여 나이브 베이지안 분류기[1]를 이용하여 기존 전자상거래 게시판에 대한 문제점의 해결과 효과적인 운영 그리고 게시물의 체계적인 분류 관리를 할 수 있는 게시물 자동 분류기를 설계하고 구현하였다. 더불어 문서 분류 학습기법으로 대표적인 TFIDF 기법[3]과 K-NN 기법[2] 그리고 나이브 베이지안 기법의 게시물 분류 성능 실험을 실시 하여 본 논문에서 채택한 나이브 베이지안 기법의 우수성을 증명하였다.

논문의 구성은 다음과 같다. 2장에서는 문서 분류 학습기법의 원리와 기존 분류 시스템들에 대하여 소개하였고 3장에서는 게시물 분류기 전체 시스템의 설계 및 구현에 대해 기술하였다. 4장에서는 문서 분류 학습기법 중 대표적인 3가지 기법에 대한 분류 성능 실험과 결과를 기술하였다. 마지막으로 5장은 결론 및 향후 연구를 기술하였다.

II. 문서 분류 학습기법과 기존 분류 시스템

대표적인 문서 분류 학습기법에는 나이브 베이지안 방법과 최근접 이웃방법인 k-NN(k-Nearest Neighbor)학습방법, TFIDF방법 등이 있다[2][3]. 이 절에서는 이들 분류 학습기법과 대표적인 분류 에이전트에 대해 살펴본다.

2.1 문서 분류 학습기법

2.1.1 나이브베이지안

나이브 베이지안(Naïve Bayesian)학습 기법[4]은 신경망(Neural Network) 또는, 결정 트리(Decision Tree)와 같은 알고리즘 비교연구에서 전자 뉴스기사, 전자편지와 같은 텍스트 문서 분류를 위한 방법으로 나이브 베이지안 학습기법이 가장 효과적인 알고리즘들 중에 하나라고 알려져 있다. 나이브 베이지안 분류 학습

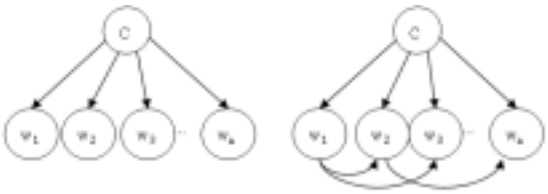

그림 1. 베이지안 네트워크
Fig. 1. Bayesian Network

기법은 베이지안 네트워크을 분류기에 적용한 것으로 베이즈 정리(Bayes Theorem)에 기초한 확률 모델을 이용하는 기법이다.

그림 1은 베이지안 네트웍을 나타낸 그림이며, 노드 C는 클래스를 의미 하고 노드 C에 대한 각각의 특성(feature)을 w_i라 표시했다. (a)는 각각의 특성들 간에는 서로 조건부 독립(conditionally independent)이라는 나이브 베이지안 학습기법을 나타내고, (b)는 특성들 사이에 제한된 종속성을 허용하는 좀더 복잡한 베이지안 학습 기법을 나타내고 있다.

하나의 문서 d가 $w_{1,\ldots,}w_n$의 특성들로 이루어 졌을 때 베이지안 학습기법은 식 (1)과 같이 문서 d에 대한 조건부 확률이 가장 큰 클래스로 분류한다[5].

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

이 식을 승법정리(multiplication theorem)를 이용하여 정리 하면 식 (2)와 같이 된다.

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

식 (2)에서 확률 $P(w_1, w_2, \cdots, w_n)$는 하나의 상수인 정규화 항이므로 우리가 가장 가능성이 높은 하나의 클래스를 결정하는 것에만 관심이 있는 경우 생략 가능하다. 따라서 식 (3)과 같이 된다.

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

여기서 문서 d를 나타내는 특성인 각 w_i는 모든 다른 특성들과 조건부 독립이라는 나이브 베이지안 가정을 적용하면 식 (4)와 같이 된다[1].

$$i=1,n \quad (4)$$

결론적으로 나이브 베이지안 학습기법은 분류 대상 문서 d에 대해 가장 가능성이 높은 분류 클래스를 [식 5]와 같이 계산한다.

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

2.1.2 k-NN

또 다른 학습기법으로는 k-NN(k-Nearest Neighbor)학습기법[1]이 있다. 이 방법은 분류 시에 분류할 문서 $(w_1, w_2, \cdots, w_n)$와 저장된 클래스 별 훈련 예제 $(w_1, w_2, \cdots, w_n)$과의 유클리드 거리(Euclidean distance)를 계산하여 분류대상 문서 와 가장 가까운 훈련 문서 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 클래스로 분류대상 문서 Y가 분류된다. 여기서 i는 클래스의 종류이며 n은 클래스의 개수이다. k값은k-NN기법의 성능을 최적화하기 위하여 일반적으로 교차검증(cross validation) 기법을 사용하여 사전에 결정하며, k=1인 경우를 NN(Nearest Neighbor) 기법이라고 한다.

2.1.3 TFIDF

전통적으로 정보 검색 분야에서 많이 이용되어온 TFIDF 분류 학습기법[2]에서는 각 문서 d를 특성단어(feature word)의 출현 빈도수에 기초한 가중치 벡터(weight vector)로 표현한다. 이 때 각 단어의 가중치는 식 (7)과 같이 문서 d에 나타나는 빈도수인 TF(Term Frequency)와 그 단어가 나타나는 총 문서수에 대한 역수인IDF(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 클래스 별로 그 클래스를 나타내는 프로토타입 벡터(prototype vector)를 구한다. 이때 각 클래스의 프로토타입 벡터 c는 그 클래스에 속한 훈련 문서들의 (TF-IDF)가중치 벡터들의 평균으로 계산한다. 일단 이처럼 각 클래스들이 프로토타입 벡터로 표현되어 있으면, 식 (8)과 같이 분류대상 문서 d의 가중치 벡터와 각 클래스 의 프로토타입 벡터간의 유사성(similarity)을 cosine 규칙을 적용하여 계산한다. 그리고 이와 같은 과정을 거쳐 가장 유사하다고 판단되는 클래스로 문서를 분류한다[16].

(8)

2.2 분류 에이전트 시스템

일반적으로 문서 분류 학습기법을 사용하여 복잡한 분류 작업을 자동으로 수행할 수 있는 자율적인 소프트웨어를 분류 에이전트라 한다. 이와 같은 에이전트를 이용한 분류 시스템으로는 대표적으로 카네기 멜론 대학의 Personal WebWatcher[6]가 있다. 이 시스템은 사용자의 행동을 웹 브라우저를 통해 모니터링하여,

사용자의 관심영역을 학습한 뒤, 브라우징하는 웹 문서내의 링크들에 대해 사용자 관심영역에 속하는 것들과 그렇지 않은 것들을 분류하여 관심있는 링크들만을 제안 해주는 시스템 이다. 또한, 앤더슨 컨설팅 연구실에서 만든 InfoFinder[7]역시 사용자의 관심 프로파일[8]을 바탕으로 온라인 문서들에 대한 분류작업을 통해 사용자가 관심을 가질 문서들을 찾아주는 에이전트 시스템이다. 이외에도 MIT 대학에서 만든 전자우편을 분류하는 Maxims[9], 엔터테인먼트 선별 에이전트인 Ringo, 뉴스 기사 분류 에이전트인 NewT[10] 등이 모두 문서 분류기법을 이용한 대표적인 분류 에이전트 시스템이다.

III. 시스템 설계 및 구현

3.1 기본가정

본 논문의 게시물 분류 시스템은 효과적인 게시물 분류를 위하여 몇 가지 기본 가정을 전제로 하고 있다.

첫째, 일반적으로 웹 게시판 프로그램은 게시물 정보를 데이터베이스에 저장해 두고 이러한 게시물의 정보를 CGI(Common Gateway Interface)에 의해 HTML(Hyper Text Markup Language)형태로 인터넷 상에서 볼 수 있도록 되어있다. 이렇게 이미 웹 게시판 프로그램이 구현되어 있는 상태에서 본 시스템은 이 게시판의 게시물에 접근을 하여 분류작업을 수행한다.

둘째, 게시물의 분류는 단지 게시물의 내용만으로 분석하기 보다는 게시물의 작성자가 만들어 놓은 제목과 내용에 동일한 가중치를 두어 함께 분석함으로써 작업이 이루어진다. 사용자들이 일반적으로 게시판에 글을 기록하기 위해서는 의뢰사항에 대한 함축적인 단어를 제목란에 기입하는 절차를 거치게 되므로 본 시스템에서는 분류작업을 위한 분석 자료로 게시물의 내용뿐만 아니라 제목까지도 고려하여 작업의 효율성을 높이고자 한다.

셋째, 게시물 분류를 위해 적용하는 문서 분류 학습기법은 모두 다수의 훈련 문서(training document)들을 필요로 하는 교사학습(supervised learning)방법들이다. 본 시스템에서 사용하는 문서 분류 기법은 나이브 베이지안 기법, k-NN, TFIDF 등으로 이들은 모두 분류 클래스 별로 충분한 훈련 문서들을 미리 확보하고 있어야 높은 분류성능을 기대할 수 있다.

넷째, 자동 분류 작업 이전에 사용자로부터 직접적으로 충분한 훈련 예들을 얻을 수 없는 경우를 위해 대표적인 인터넷 디렉터리 서비스를 제공하는 eBay 쇼핑 사이트의 도움 카테고리의 클래스와 해당 문서들을 가져와 훈련 예로 이용한다.

3.2 시스템의 구조

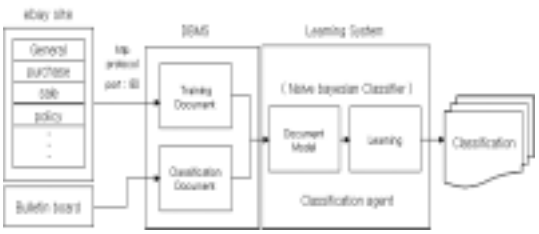

그림 2. 시스템 전체 구조
Fig. 2. System Architecture

그림 2는 전체 시스템의 구조를 보여주고 있으며, 수행 과정은 다음과 같다. 분류 해야 할 게시판의 게시물을 시스템으로 가져오고 분류 알고리즘에 사용할 클래스들과 훈련 문서들을 확보하기 위해 인터넷 쇼핑 서비스를 제공하는 eBay 쇼핑 사이트의 도움 카테고리의 클래스들과 해당 웹 문서들을 수집한다. 이와 같은 과정을 거쳐 분류 작업을 위한 훈련 문서들과 분류 대상 문서들이 수집되면, 적절한 문서 전처리 과정(text/document preprocessing)과 문서 모델화(document modeling)과정을 거친다. 그리고 이러한 분류 클래스들과 훈련 문서들을 바탕으로 사전에 문서 분류 학습기법에 따라 문서 분류기를 학습하고 이것을 바탕으로 분류 대상 문서들을 차례대로 분류한다.

3.3 문서의 수집

게시물의 분류 작업을 수행하기 위하여 본 시스템에서는 훈련 문서로 사용될 웹 문서들을 인터넷으로부터 수집한다. 훈련 문서로 사용되는 웹 문서의 수집을 위해서는 분류가능 클래스들과 클래스에 대응되는 양질의 훈련 예를 확보해야 하는데 기존에 쇼핑 사이트가 운영되고 있는 eBay 쇼핑 사이트의 고객 도움 카테고리의 일반(General), 구매(purchase), 판매(sale),

정책(policy)의 해당 카테고리의 웹 문서를 수집한다. eBay 쇼핑 사이트의 도움 카테고리는 eBay 쇼핑이 몰을 운영하면서 고객들의 질문 중에 빈도수가 높은 질문들을 카테고리 별로 모아서 관리하는 것으로 쇼핑몰에서 발생하는 고객 질의에 대해 비교적 정확한 답변들로 이루어져 있다.

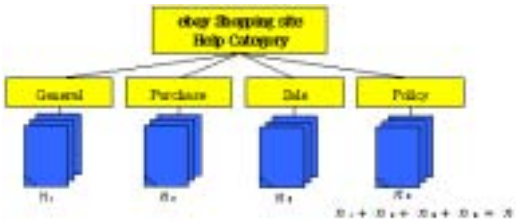

그림3. 문서 수집
Fig 3. Collecting decuments

3.4 웹 문서의 전처리

수집된 웹 문서에 대해 문서 분류 학습기법을 적용하기 위해서는 각 훈련 문서와 분류 대상 문서들을 적절한 특성 단어(feature word)들에 기초한 벡터 모델로 표현하여야 한다. 이를 위해서는 각 웹 문서를 표현하는데 중요한 역할을 하는 의미 있는 특성 단어들을 추출하는 것이 매우 중요한데, 이를 위해 먼저 각 웹 문서에 대한 전처리 과정(preprocessing)이 필요하다. 전처리 과정에서는 각 웹 문서를 구성 단어 별로 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. 그림 4는 이와 같은 웹 문서의 전처리 과정과 이것에 기초한 문서 모델화 과정을 보여주고 있다.

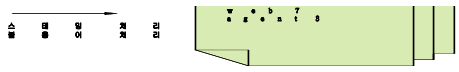

그림 4. 웹 문서의 전처리 과정
Fig. 4. Text preprocessing

3.5 특성 추출 및 모델화

특성 추출은 학습 자원의 중요 속성들을 자원이 구분된 클래스별로 다시 한번 중요도를 정의하는 특성 추출 가중치 설정 기법이다. 이를 위하여 각 학습 자원들의 특성을 고려하여 구분된 클래스들을 대상으로 일련의 구별 작업을 두어 이를 기반으로 한 속성 추출 작업을 수행할 필요가 있다. 이러한 가중치 설정 작업은 해당 키워드가 속해있는 클래스의 정보를 고려하여 이루어지며 이로써 클래스, 즉 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 웹 문서에 대하여 이진 속성 벡터(vector of binary attributes)로 모델화 한다. 수집된 웹 문서에 대해 이와 같은 이진 속성 벡터를 만들기 위해서는 먼저 웹 문서들로부터 각 문서를 표현하는데 사용할 단어들인 특성(features)들을 추출하여야 한다. 특성을 추출하는 방법으로는 정보이론(Information Theory)에 입각한 엔트로피(entropy) 변화량을 기초로 특성 단어를 추출 하는 방법인 정보 획득(Information Gain)방법을 사용한다.[6]

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

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

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 웹 문서에 대해 아래와 같은 모양의 이진 속성 벡터 모델을 만들게 된다.

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

3.6 학습 및 분류

본 연구에서는 문서분류를 위해 확률을 이용한 대표적인 교사학습 알고리즘인 나이브 베이지안 학습기법을 이용한다. 그러나 이외에도 K-NN기법과 TFIDF 기법도 선택적으로 사용될 수 있도록 구현 되었다. C를 식 (12)과 같이 전체 클래스들의 집합이라고 할 때

$$C = \{c_1, c_2, c_3, ... c_k\} \quad (12)$$

나이브 베이지안 분류법은 한 문서 d_i의 각 클래스 c_j에 대한 조건부 확률들을 식 (13)과 같이 구해준다.

$$\Re(d_i) = \{P(d_i \mid c_1), P(d_i \mid c_2), P(d_i \mid c_3), ..., P(d_i \mid c_k)\} \quad (13)$$

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

$$P_{\max}(d_i) = \max_{c_i \in C} P(d_i \mid c_i) \quad (14)$$

$$c_{best}(d_i) = \begin{cases} c_j & \text{if } P_{\max}(d_i) = P(d_i \mid c_j) \geq T \\ c_{unknown} & otherwise \end{cases} \quad (15)$$

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

Ⅳ. 실험 및 평가

이 장에서는 게시물 분류 에이전트의 분류 성능분석을 하였다. 게시물 분류는 웹 문서들의 내용을 파악하여 문서가 속한 정확한 클래스를 정하는 작업으로, 사전에 정해져 있는 클래스들 중에서 어떤 클래스에 웹 문서가 속하는지 판단하는 것이다. 문서를 분류하기 위한 대표적인 방법으로 나이브 베이지안, TFIDF, k-NN 방법이 있다. 우리는 이러한 세가지 문서 분류 학습기법을 이용하여 게시물 분류 성능실험을 하고, 각각의 학습기법에 따른 분류 성능분석을 통해 분류 대상이 게시물일 경우 어떤 학습방법이 가장 우수한 분류성능을 가졌는지에 대하여 비교 분석 한다. 실험 환경은 분류 성능실험을 위한 시스템으로 300MHz 펜티엄Ⅱ 프로세서와 256MB의 주 기억공간을 사용하는 리눅스 환경에서 실험이 이루어졌으며, 실험을 위한 데이터는 기존에 쇼핑 사이트가 운영되고 있는 eBay 쇼핑 사이트의 고객 도움 카테고리의 일반 게시물을 웹에서 받아서 사용하였다. 전체 클래스는 도움 카테고리 중 4개(General, Purchase, Sale, Policy)이며, 각각의 클래스 당 20개씩, 총80개의 게시물을 실험에 사용한다. 실험방법은 20개씩 문서를 가지고 있는 4개의 클래스에서 임의로 클래스 당 5개의 게시물을 분류대상 게시물로 발췌한다. 그리고, 나머지 클래스 당 15개의 문서들 중에서 훈련예제의 개수를 10개 15개로 증가시키면서 세가지 문서 분류 학습기법을 비교하고, 정확도 값을 구한다.

4.1.실험결과

표 1. 실험 결과 ; 분류 정확도(%)
Table 1. Experimental Evaluation : Classification correctness degree(%)

학습기법 / 훈련예		Naive Bayesian	TFIDF	k-NN
10개/클래스	1차	80	75	33
	2차	75	85	40
	3차	80	77	40
	4차	78	80	38
	5차	85	85	53
	평균	**79.6**	**80.4**	**40.8**
15개/클래스	1차	85	75	55
	2차	90	75	45
	3차	85	80	50
	4차	95	85	40
	5차	90	85	45
	평균	**89**	**80**	**47**
	평균	**82.5**	**80.2**	**40.4**

표 1은 본 연구의 실험 결과를 보여주는 표이다. 게시물 분류 에이전트 시스템은 사용하는 분류 학습 기법에 따라 차이가 있기는 하지만 전체적으로 75~90% 정도의 높은 분류 정확도를 보여주었다.

(a)

(b)

그림 5. 분류 성능 실험 결과

Fig. 5. Results of Classification Experiments

그림 5의 (a)와 (b)는 클래스 당 훈련문서의 수를 다르게 했을 때 각각의 분류 성능을 그래프로 나타냈다. 예상대로 훈련 문서의 수가 증가 할수록 분류성능도 조금씩 증가 하는 것을 발견할 수 있었다. 그리고 분류 학습 기법들간의 분류 성능의 차이도 비교적 뚜렷이 나타났다. 본 시스템에서 기본 분류 방식으로 채택하고 있는 나이브 베이지안 학습기법이 90%대의 가장 높은 분류 성능을 보여 주었다. 이에 반해 k-NN 학습기법은 기대와 는 달리 40%~50%의 가장 낮은 분류 성능을 나타냈으며, TFIDF 학습기법은 비교적 나이브 베이지안 학습기법에 필적하는 성능을 보여주었다. 한편, 분류 학습기법들간의 분류 속도 면에서는 나이브 베이지안 기법과 TFIDF 기법 비슷했지만 개체기반 학습기법(instance-based learning)의 하나로서 분류 당시에 많은 계산시간을 필요로 하는 k-NN 기법은 나머지 두 학습기법에 비해 매우 느리게 나타났다.

Ⅴ. 결론

본 논문에서는 학습기법을 적용하여 자동으로 게시판의 게시물을 카테고리 별로 분류하는 eCRM 에이전트 시스템을 설계하고 구현하였다. 이 시스템의 특징은 게시판을 이용하는 사용자의 질의에 대하여 사전에 카테고리 별로 분류 해놓은 웹 문서를 훈련 예로 사용하여 질의에 대한 소속 카테고리를 결정하여 분류 하는 방식의 시스템이다. 여기서, 사전 카테고리 별로 분류해 놓은 웹 문서의 훈련 예는 정제되어 있는 데이터를 사용해야 하는 여러 한계를 극복하기 위하여 쇼핑몰로 유명한 eBay 사이트의 help 카테고리의 해당 웹 문서들을 사용하였다. 본 논문에서 구현된 eCRM 에이전트 시스템의 분류 성능을 평가하는 실험을 통하여 에이전트의 전체적으로 높은 분류 성능을 입증하였고, 특히 분류 학습기법 중 나이브 베이지안 학습 기법의 우수성을 확인하였다. 본 시스템의 성능과 효용성을 높이기 위해서 앞으로 시행되어야 할 향후 연구과제로, 분류 클래스들간의 계층관계 및 중복 관계에 대한 해결, 명확한 훈련 예를 확보할 수 없는 경우를 대비하기 위한 비 교사 학습기법(unsupervised learning)에 대한 도입 등을 검토 하고 있다.

참고 문헌

[1] Tom M. Michell, 'MACHINE LEARNING', the McGraw-Hill Company, 1997.

[2] D. D. Lewis and M. Ringuette, "A Comparison of Two Learning Algorithms for Text Categorization", Proceeding of the 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 81-93, 1994.

[3] Y. Yang and X. Liu, "A Re-Examinationof Text Categorization Methods", Proceedings of

SIGIR-99, 1999.

[4] Sahami, S. Dumais, D. Heckerman and E. Horvitz. "A Bayesian Approach to Filtering Junk e-mail", AAAI Technical Report WS-98-05, 1998.

[5] A. McCallum and K. Nigam, "A Comparison of Event Models for Naïve Bayes Text Classification", In AAAI-98 Workshop on Learning for Text Categorization, 1998.

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

[7] B. Krulwich and C. Burkey. "The InfoFinder agent: Learning user interests through heuristic phrase extraction",IEEE Experts, 12(5): 22-27, 1997.

[8] M. Pazzani and D. Billsus, "Learning and revising user profiles: The identification of interesting Web sites",Machine Learning, 27(3): 313-331, 1997.

[9] P. Maes, "Agent That Reduce Work and Information Overload." Communications of the ACM, Vol.37, No.7, pp.30-40, 1994.

[10] B. Sheth and P. Maes., "Evolveing Agents for Personalized Information Filtering", Proceedings of the 9th IEEE Conference on AI for Applications, 1993.

[11]Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization", Proceedings of ICML-97, 14th International Conference on Machine Learning, PP. 142-420, 1997.

[12]Y. Yang, "An evaluation of statistical approaches to text categorization", Information Retrieval, Vol. 1, No. 1-2, PP. 69-90, 1999.

[13]William W. Cohen. "Learning Rules that Classify E-Mail", AAAI Spring, 1996.

[14]D. Mladenic and M. Grobelink, "Feature Selection for Classification Based on Text Hierarchy", Working notes of Learning from Text and the Web, CONALD-98, 1998.

[15]J.S. Youn, Y. S. Kwan, "Ensemble approach to Naive Bayesian e-mail classifier", Korena Institute of Industrial Engineers, pp. 651-655, 2001.

[16] Myoung, Soon-Hee ; Choi, Jung-Min ; Kim, In-Cheol, "BClassifier: a personal agent for bookmark classification", Proceedings of ICPADS-2001, IEEE Comput. Soc, pp. 713-718, 2001.

저 자 소 개

최 정 민

1999년 경기대학교 화학공학과 (공학사)
2001년 경기대학교 전자계산학과 (이학석사)
2004년 인천대학교 컴퓨터공학과 박사과정 수료

관심분야 : 전자상거래, 인공지능, 데이터마이닝, eCRM, Web Agent, Information Retrival

이 병 수

1976년 단국대학교 전자공학과 (학사)
1980년 동국대학교 대학원 전자정보처리전공 (석사)
1998년 경기대학교 대학원 전자계산전공 (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