기계학습 기법을 이용한 전자게시판 질문 자동 분류

최형림* · 류광렬** · 강재호*** · 신종일**** · 이창섭*****

An Automatic Question Routing System using Machine Learning

Hyung-Rim Choi* · Kwang Ryel Ryu** · Jaeho Kang*** · Jong-IL Shin**** · Chang-Sup Lee*****

요 약

인터넷의 급격한 발전과 광범위한 보급에 따라 과거 전화, 서신 또는 직접방문을 통하여 해결하던 고객 상담의 상당부분은 인터넷을 이용한 전자우편 및 전자게시판을 이용하는 방향으로 꾸준히 대치되고 있다. 인터넷을 통한 고객과의 접촉방식의 대부분을 차지하는 전자우편과 전자게시판은, 기존의 방식 특히 전화에 비하여 즉각적인 응답을 기대하기가 어렵다는 측면이 고객에게는 가장 큰 불만사항이 되고 있다. 본 논문에서는 문서로 이루어진 전자우편 또는 전자게시판의 고객 상담 내용을 기계학습의 분류기법을 활용하여 담당자를 자동으로 선정함으로써 보다 신속히 고객의 요구에 반응할 수 있는 효과적인 방법을 제안한다. 실제 수집한 다년간의 데이터를 기반으로 다양한 분류기법의 성능을 비교 평가하였으며, 그 결과 k-NN을 이용한 기법이 성능 및 활용도 측면에서 유리함을 보였다. 또한, 인터넷을 통한 질문의 경우 상당 수준의 오탈자 및 띄어쓰기 오류를 내포하고 있는데, 바이그램을 이용한 문서처리방법을 이용함으로써 이러한 상황에 효과적으로 대처할 수 있으며, 바이그램으로 문서 처리 시 발생할 수 있는 시스템의 부담을 큰 성능의 저하 없이 최소화하기 위하여 자주 등장한 단어만을 선정하는 방안이 실용성이 있음을 확인하였다.

Key words : 기계학습, 문서분류, 문서라우팅

1. 서론

오늘날 인터넷의 급격한 발전과 광범위한 보급에 따라 과거 전화, 서신 또는 직접방문을 통하여 해결하던 고객 상담의 상당부분은 인터넷을 이용한 전자우편 및 전자게시판을 이용하는 방향으로 꾸준히 대치되고 있다. 인터넷을 통한 고객과의 접촉방식은 고객 및 기업 모두에게 시간적인 제약을 벗어날 수 있도록 함으로써 편의성을 극대화하고 비용을 절감할 수 있게 하였다. 따라서 효율적인 고객과의 접촉 채널 구축은 인터넷 상에서 정보를 제공하거나 제품을 판매하고자 하는 업체에서는 기본적으로 제공하여야 하는 중요한 서비스의 하나로 손꼽히게 되었다.

인터넷을 통한 고객과의 접촉방식의 대부분을 차지하는 전자우편과 전자게시판은, 기존의 방식 특히 전화에 비하여 즉각적인 응답을 기대하기가 어렵다는 측면이 고객에게는 가장 큰 불만사항이 되고 있다. 인터넷 특히 전자게시판의 경우 고객에 대한 빠른 응답이 어려운 이유는 크게 두 가지로 파악할 수 있다. 첫 번째는 고객을 상담하는 담당자가 맡은 다른 업무가 있어 지속적으로 전자우편과 전자게시판을 확인하고 새로운 상담 내용이 발생시 즉각적으로 대처하는 것이 어려운 경우이다. 두 번째 이유로는 일정 규모 이상의 업체에서는 분야별로 전문 담당자를 두는데, 고객의 상담내용이 파악되어 해당 담당자에게까지 전달되는데 소요되는 시간과 비용이 높기 때문이다. 첫 번째 문제는 업체의 인력과 관련된 문제로 고객상담 인력에 대한 지속적인 교육과 추가 고용을 통해서 해결할 수 있으며, 두 번째는 전화 상담 시 일반적으로 활용하는 ARS 시스템과 같이 고객의 질문을 담당자에게 효과적으로 연결할 수 있는 방안을 시스템적으로 마련함으로써 해결할 수 있다. 본 논문에서는 문서로 이루어진 전자우편 또는 전자게시판의 고객 상담 내용을 기계학습의 분류기법을 활용하여 담당자를 자동으로 선정함으로써 보다 신속히 고객의 요구에 반응할 수 있는 효과적인 방법을 제안하고자 한다.

본 논문에서는 먼저 2장에서는 자동분류기법과 담당자 선정 문제에 적용하는 방안을 설명하고, 3장에서는 이를 활용한 전자게시판 질문 자동 분류 시스템의 구조를 제안한다. 4장에서는 한 대학의 질의응답게시판을 이용한 구체적인 실험환경과 그 결과를 평가하고, 마지막 5장에서 결론 및 향후 연구방향을 도출한다.

2. 문서자동분류

본 장에서는 먼저 기계학습의 자동분류란 어떠한 것인지에 대하여 설명하고, 인터넷 상의 전자게시판에 올려진 문서로 표현된 질문을 자동분류하기 위해서는 어떠한 요소들이 고려되어야 하는지를 서술한다.

* 동아대학교 경영정보과학부, ** 부산대학교 정보컴퓨터공학부, *** 동아대학교 지능형통합항만관리연구센터
**** (주)비투윈, ***** 동아대학교 대학원 경영정보학과

2.1 자동분류

자동분류[1][2][3][4][5][6][7]란 둘 이상의 서로 다른 부류로 나눌 수 있는 학습예제들을 기반으로, 구분에 활용할 수 있는 일관된 특징들을 발견하고 이를 바탕으로 새로운 예제를 자동으로 분류할 수 있는 분류기를 생성하는 기계학습[1][2]의 한 분야이다. [그림 1]에서는 이러한 분류기의 생성과 적용을 간략하게 도시하고 있다. 그림에서 보듯 자동분류는 크게 2단계로 이루어지는데, ① 해답이 이미 있는 학습예제를 이용하여 기계학습기법을 적용함으로써 분류기를 생성해내는 학습단계와 ② 분류하고자 하는 새로운 문제를 분류기에 적용함으로써 해답을 구해내는 적용단계이다.

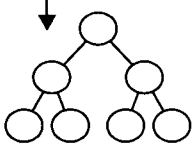
분류기법은 decision tree[3], 규칙[4], 예제기반[6], 인공신경망[1] 등 생성하는 분류를 위한 정보의 표현형태와 그 구체적인 생성방법에 따라 다양한 방법이 있으며, 적용하고자 하는 문제에 따라 성능의 차이가 나는 것이 일반적이다. Decision tree[3]는 나무구조 형태로 분류기를 표현하는 방법으로 각각의 중간 노드에는 예제에서 비교하고자 하는 속성이 기록되어 있다. 주어진 예제의 해당 속성 값에 따라 둘 또는 그 이상의 하위 노드 중 하나로 나누어지게 된다. 즉, 조건이 맞는 경로를 따라 이동하는 것이다. 마지막 단말노드에는 예제의 종류가 표기되게 된다.

[그림 1] 분류기의 생성과 적용 예

학습예제들

예제번호	속성1	속성2	속성3	분류
예제 1	가	2	갑	A
예제 2	나	5	을	B
예제 3	가	4	갑	A
예제 4	나	8	갑	B

① 예제를 이용하여 분류기 생성



② 새로운 문제에 적용

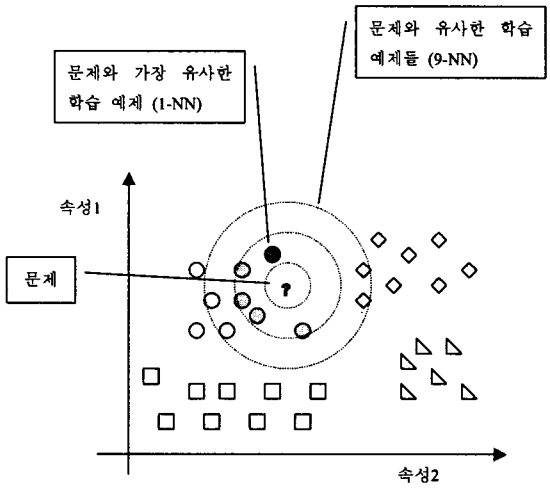
문제예제들

문제번호	속성1	속성2	속성3	분류
문제 1	가	3	을	?
문제 2	나	7	을	?

규칙은 "IF 속성1 = 값a AND 속성2 = 값b THEN 종류 = A" 와 같은 형태로 표현되는데 조건에서 제시한 경우와 동일한 속성값을 가진 예제들을 특정 부류로 분류하는 규칙을 여러 개 만들어 이를 묶어 사용한다. Decision tree와 규칙을 이용한 표현은 명시적으로 분류기를 표현할 수 있어 사용자의 이해를 돕는데 큰 도움이 된다.

예제기반 분류기법은 [그림 2]에서 설명하는 바와 같이 각각의 예제들을 속성들을 축으로 하는 다차원 공간상에 배열하고, 새로운 문제예제와 기존에 해답을 알고 있는 학습예제들간의 거리를 측정하는 방법 등을 통해 유사한 정도를 계산하고 이들 중 가장 유사한 예제와 같은 부류로 새로운 문제를 분류하는 방법이다. 잘못된 예제에 민감하게 반응하는 경우를 최소화하기 위하여 k개의 가장 유사한 학습예제를 찾아내고 그 유사정도에 따른 가중투표형태로 해답을 구하는 경우 k-NN이라고 한다.

[그림 2] 예제기반 학습기법을 이용한 분류의 예



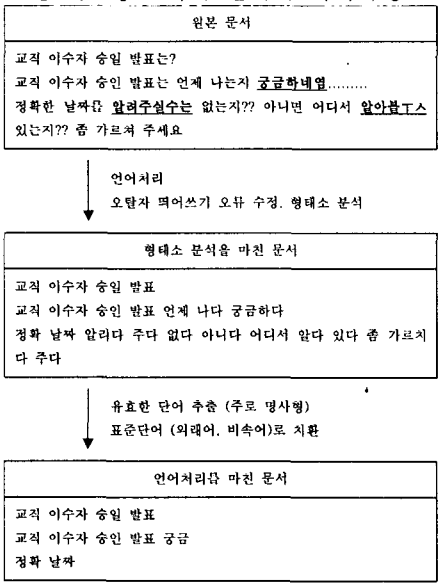
인공신경망은 생물신경세포의 정보전달방식을 기계학습에 응용한 것으로 일반적으로 예제의 속성값 또는 속성값을 전처리한 데이터를 입력으로 하는 입력 계층, 분류된 최종결과를 생성하는 출력 계층 그리고 이들 입출력 계층 사이를 연결하는 숨은 계층으로 이루어져 있으며, 각 계층은 간략화된 복수개의 인공신경으로 모델링된다. 각각의 인공신경은 일반적으로 시그모이드 함수로 표현되며 이들 각 인공신경의 변수값을 결정하는 즉 학습과정으로 역전파 알고리즘을 주로 활용한다. 인공신경망은 잘못된 예제에 민감하지 않으며 많은 응용분야에서 효과적으로 알려져 있으나, 생성된 결과를 직접적으로 이해하기는 어렵다.

2.2 전자게시판 질문 자동 분류를 위한 고려사항

전자우편 또는 전자게시판과 같이 문서로 이루어진 데이터를 대상으로 자동분류를 수행하기 위해서는 우선 문서 데이터를 자동분류에 사용할 수 있는 형태로 변환하여야 할 필요가 있다. 이는 크게 분류에 효과적인 형태로 단어를 표준화하는 언어적인 처리과정과 분류기법을 적용할 수 있는 테이블 형태로 데이터 변환하는 두 가지 단계로 이루어진다. [그림 3]에는 문서의 언어적 처리 과정을 도시하고 있다. 원본 문서는 인터넷 게시판 문서의

특성상 상당수의 오탈자 및 띄어쓰기 오류를 내포하고 있으므로 이를 교정하고, 한글문장을 의미 있는 부분으로 나누는 형태소 분석과정[8]을 수행한다. 또한 문서에 등장한 모든 단어가 분류에 효과적인 것은 아니므로 일반적으로 의미를 가지는 명사만을 추출한다. 이 때 외국어, 비속어 등에 대하여도 표준화 과정을 거치게 되며 필요한 경우 해당 분야의 전문용어사전을 구축하여 활용할 수 있다.

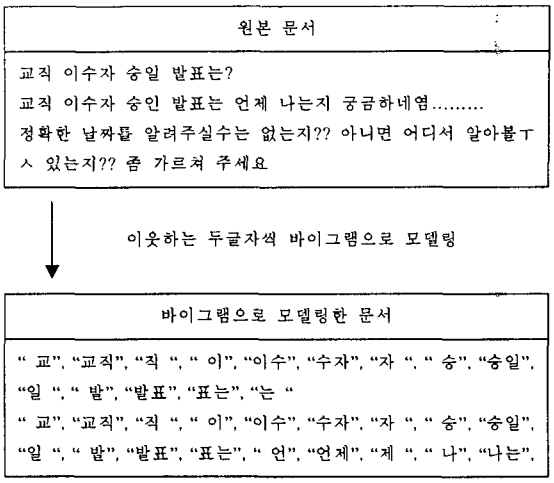
[그림 3] 문서의 언어적 처리과정



형태소 분석을 이용한 문서의 언어적 처리과정은 전통적인 한국어 문서검색에서 널리 활용되어 왔으나, 본 논문에서 대상이 되는 인터넷 게시판의 경우에는 오탈자 및 띄워쓰기 오류가 심각하여 효과적으로 적용하기가 어렵다고 할 수 있다. 이러한 점을 상대적으로 적은 비용으로 해결하는 방안으로 음성인식에서 널리 활용되는 바이그램[9]을 이용하여 문서의 처리할 수 있다. 바이그램은 [그림 4]에서와 같이 모든 이웃하는 두 글자의 조합을 하나의 단어로 취급하는 방법이다.

언어처리단계를 거친 문서는 자동분류기법에 활용할 수 있는 테이블 형태로 변환되어야 하는데, 이러한 표현 방법으로 문서의 경우 정보검색[10]에서 일반적으로 활용되는 벡터공간상의 $tf \times idf$를 이용한 모델링이 가장 널리 활용된다. 벡터공간상의 각 축은 하나의 단어를 의미하게 되며, 특정 축 상의 위치는 해당 단어의 특정문서에서의 중요도를 나타내는 tf값과 전체 문서군에서 상대적인 중요도를 나타내는 idf값을 곱한 값을 이용한다. 문서 i에서 단어 k의 가중치 즉 축에서의 위치는 $w_{ik} = tf_{ik} \times df_k$ = (단어 k가 문서 i에 등장한 횟수) × $\log$(전체문서수 / 단어 k가 등장한 문서수)로 구해진다. 이는 특정 단어가 해당 문서에 많이 등장할수록 그리고 해당 단어가 전체 문서군에서 희귀할수록 중요하다고 가정한 모델이다.

[그림 4] 바이그램으로 모델링한 문서



[그림 5]에는 기계학습을 위하여 문서를 표현한 최종 형태를 보여주고 있다. 예제기반 학습의 경우 이러한 벡터공간상에서 $tf \times idf$를 이용한 문서의 표현 방법을 자연스럽게 적용할 수 있으며, decision tree, 규칙과 인공신경망의 경우에는 해당 단어의 등장여부 또는 등장횟수 정보만을 이용하는 것이 일반적이다. 이는 decision tree, 규칙 및 인공신경망은 학습단계에서 자체적으로 분류에 효과적인 단어를 선정하는 기능을 명시적으로 수행하는데 비해, 예제기반학습은 문서 분류와 같이 예제속성의 개수가 방대한 문제에서는 idf와 같은 척도를 이용하여 가중치를 미리 결정하는 방안이 보다 효율적이기 때문이다. 예제기반학습에서 문서간의 유사한 정도를 비교하는 방법은 각각의 문서를 벡터 d_1, d_2로 표현할 때, 두 벡터간의 각도를 계산하는 코사인 유사도 $(d_1 \times d_2) / (|d_1||d_2|)$를 활용할 수 있다. 각각의 문서들은 전체 단어 중에서 일부의 단어만 등장하므로 구현에서는 희박 행렬 또는 정보검색의 색인어 역파일 구조로 표현된다.

[그림 5] 자동분류를 위한 문서의 최종 표현 형태

ⓐ 예제기반학습을 위한 문서표현
(단어의 문서내 $tf \times idf$ 가중치 부여)

문서번호	단어a	단어b	단어c	담당자
문서1	1.2	-	4.2	갑
문서2	2.4	0.3	-	을
문서3	-	0.6	2.1	병

ⓑ decision tree, 규칙, 신경망을 위한 문서표현
(단어의 등장 횟수를 기입)

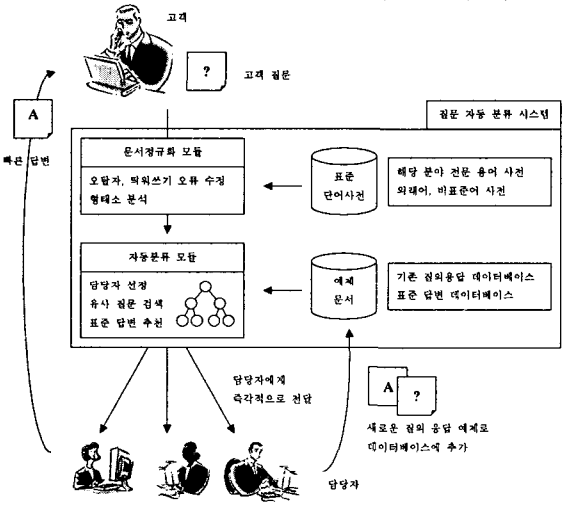
문서번호	단어a	단어b	단어c	담당자
문서1	1	-	2	갑
문서2	2	1	-	을
문서3	-	2	1	병

3. 전자게시판 질문 자동분류 시스템

본 장에서는 고객들의 질문 문서에 분류기법을 활용하여 담당자를 자동으로 지정하는 시스템의 전반적인 구조와 그 기능을 설명한다. [그림 6]에는 본 논문에서 제안하는 시스템의 전반적인 구조를 도시하고 있다. 고객의 질문은 앞서 설명한 문서표준화 과정과 자동분류 단계를 거쳐 담당자를 지정받게 되며, 이때 유사한 과거 질문과 답변을 함께 확인할 수 있도록 함으로써, 해당분야 담당자가 보다 빠르게 고객의 질문에 대응할 수 있도록 하고 있다. 답변된 내용은 해당 질문과 함께 다시 데이터베이스의 예제 문서로 구축하여 향후 유사한 질문에 대해서 보다 정확하게 응답할 수 있어 시스템의 성능을 지속적으로 향상 시킬 수 있다.

담당자 선정은 분류기 생성시 담당자를 분류 속성으로 학습함으로써, 유사 질문을 검색하는 방법으로는 앞서의 k-NN 기법에서 설명한 바와 같이 벡터공간상에서 새로운 질문 문서와 기존 예제 문서와의 유사정도를 계산함으로써 추출할 수 있다. 표준답변을 추천하는 기능 역시 동일한 방법을 적용함으로써 실현할 수 있다. 이러한 단순한 담당자지정을 위한 문서분류기능 이외에 추가의 응용가능성을 감안한다면, k-NN을 이용한 자동분류가 여타 방법에 비해 그 활용용도가 넓다고 할 수 있다.

[그림 6] 질문 자동 분류시스템의 구조와 기능



4. 실험 결과

본 논문에서 전자게시판의 고객 질문에 대한 담당자 자동지정을 위한 분류기법의 성능을 실험하기 위하여 활용한 자료는 D대학교의 종합 질문 및 건의게시판으로 약 2년간에 걸쳐 수집된 1,438건의 문서를 활용하였다. 수집한 대상의 특성상 부서 및 담당자의 수가 많은 편에 속하며 큰 부류로는 14가지(학적과, 수업과, 시설과, 등등), 소분류로는 최대 200여가지(휴학, 복학, 수강신청, 등등)로 나눌 수 있다. 여러 학습기법을 비교실험하기 위하여 동일한 환경에서 다양한 실험이 가능한 WEKA[1] 기계학습 소프트웨어를 이용하였다. 최적의 성능을 얻기 위하여 다양한 분류기법과 해당 분류기법의 설정을 변화시켜가며 실험을 수행하였다.

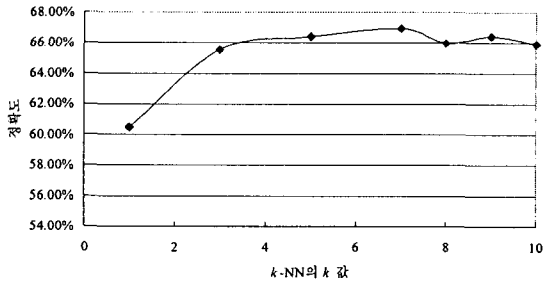
[표 1]에는 소분류 담당자를 대상으로 분류작업을 수행했을 때 기법에 따른 성능을 보여주고 있다. 비교를 위한 주요한 네 가지 분류기법으로 예제기반의 k-NN, decision tree로는 J48 (C4.5와 동일한 방법), 규칙 생성 방안으로 J48.PART (J48 decision tree 결과를 규칙으로 변환하는 방법), 그리고 확률을 이용한 방법으로 Navie Bayes을 선정하였다. 문서의 언어처리 시 해당 업무와 관련된 용어를 수집하여 생성한 전문용어사전을 활용하였으며, 실험은 모두 10-fold cross validation을 이용하였다. 전반적으로 대략 60% 정도의 정확도를 보여주고 있는데, 이는 분류가짓수(대상이 소분류이므로 최대 200여가지)에 비하여 예제 문서의 수가 적으며, 특히 원칙적인 담당자가 아닌 시간적 여유가 있는 다른 담당자(예를 들어 같은 부서에 근무하는)가 본인의 업무와 관련된 질문에 대해서는 답변을 하는 경우가 많아 이러한 데이터가 오류를 증가시키는 요인으로 작용하였기 때문이다.

[표 1] 자동분류 기법에 따른 성능 (소분류)

방법	정확도
k-NN (예제기반,k = 5)	62.86%
J48 (decision tree)	62.00%
J48.PART (규칙)	58.36%
Naive Bayes (확률)	51.69%

분류기법을 비교한 이러한 실험에서 k-NN을 이용한 경우가 성능이 우수하면서 앞서 밝힌 바와 같이 유사 질문 검색 등 추가적인 활용에 매우 유리하므로 이후 실험에서는 이러한 k-NN을 이용한 방법을 보다 최적화하여 적용할 수 있는 방안을 연구하였다. 먼저 k-NN에서 분류를 추정하기 위하여 유사한 문서를 몇 개 정도 사용하는 것이 적절한지 파악하기 위하여 k값을 변화시켜가며 그 성능을 실험하였고 그 결과를 [그림 7]에 보여주고 있다.

[그림 7] k-NN에서 k값의 변화에 다른 성능 변화



이 실험에서 알 수 있듯이 k값이 5이상인 경우에는 전반적으로 성능이 안정됨을 알 수 있는데, 이는 다시 말하면 새로운 분야의 담당자가 등록되었다면 각 부류의 유사 질문이 서너 개(투표 방식으로 분류하므로) 정도 등록된다면 기본적으로 충분한 성능을 나타낼 수 있음을 의미한다. 이 실험

결과에서 성능이 [표 1]에서 보다 높은 이유는 10-fold cross validation이 아닌 개별 예제 각각에 대하여 나머지로 예제로 학습한 후 해당 예제를 평가하고 이를 모두 평균하는 leave-one-out 방식으로 성능을 측정하였기 때문에 상대적으로 많은 양의 문서를 학습예제로 삼을 수 있었기 때문이다.

[그림 8]은 문서의 언어처리 방식과 k값의 변화에 따른 k-NN의 성능 변화를 보여주고 있다. HAM[8]은 국민대학교에서 개발한 형태소 분석기를 이용한 경우이며, BIGRAM은 앞서 설명한 이웃하는 두 글자를 단어로 삼아 이를 추출하는 바이그램 기법을 활용한 경우이다. 바이그램으로 문서를 처리하는 경우에는 두 글자로 조합이 가능한 매우 많은 수의 단어(대략 2500×2500가지)가 생성될 수 있으므로 시스템의 데이터 유지관리 비용을 크게 상승시킬 수 있다. BIGRAM2(dfcut-10)는 이를 최소화 하기 위하여 최소 10개 이상의 문서에 등장하는 단어만을 대상으로 k-NN을 수행하는 것으로 본 데이터에서는 3,106개의 단어들이 선정되었다. WORD는 해당 업무와 관련된 단어만을 선정하여 (예를 들어 휴학, 복학, 수강정정, 전산실, …) k-NN을 수행한 경우이다. 실험 데이터는 앞의 경우와 동일하나 대분류 (부서 기준)에 대하여 적용하였다.

실험결과에서 WORD 방법은 타 기법에 비하여 성능이 낮음을 알 수 있는데 이는 해당분야와 관련된 전문적인 단어를 주의 깊게 추출하여 적용하더라도 일부 유용한 단어가 누락될 수 있으므로, 이보다는 문서 내의 단어들을 자동으로 추출하여 가능한 많은 수의 단어를 활용하는 방안이 보다 효과적임을 보여주고 있다. 형태소 분석 기법을 이용한 경우에 비하여 바이그램을 이용한 경우가 보다 효과적인 이유는 인터넷 문서의 특성상 심각한 오탈자와 띄워쓰기 오류에 적절히 대처하는 것이 어렵기 때문이다. 바이그램의 경우 많은 문서에만 등장하는 일부 단어들만을 대상으로 하더라도 모든 단어를 활용한 경우에 비하여 별다른 성능의 저하없이 적용이 가능하였다. 전반적으로 데이터에 상당한 오류가 있음을 감안한다면 실용적인 수준으로 활용이 가능한 수준의 결과가 도출되었다고 할 수 있다.

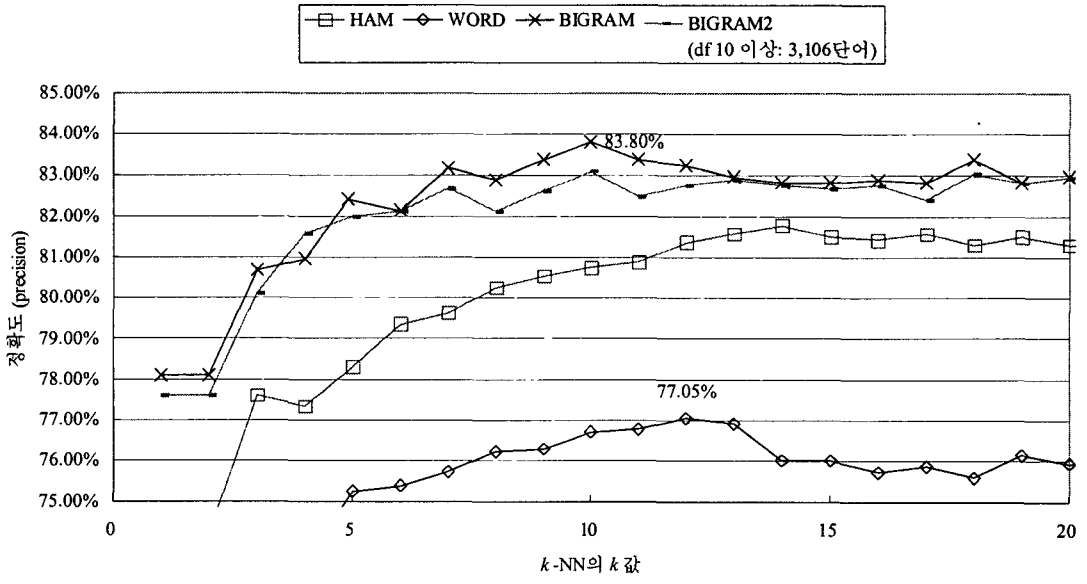
5. 결론 및 향후과제

본 논문에서는 인터넷을 이용한 업무환경이 보편화됨에 따라 늘어나는 고객의 질문에 보다 빠르고 효과적으로 대처할 수 있는 기계학습기법을 이용한 질문자동분류 시스템을 제안하였다. 실제 수집한 다년간의 데이터를 기반으로 다양한 분류기법의 성능을 비교 평가하였으며, 그 결과 k-NN을 이용한 기법이 성능 및 활용도 측면에서 유리함을 밝혔다. 또한, 인터넷을 통한 질문의 경우 상당 수준의 오탈자 및 띄어쓰기 오류를 내포하고 있는데, 바이그램을 이용한 문서처리방법을 이용함으로써 이러한 상황에 효과적으로 대처할 수 있으며, 바이그램으로 문서 처리 시 발생할 수 있는 시스템의 부담을 큰 성능의 저하 없이 최소화하기 위하여 자주 등장한 단어만을 선정하는 방안이 실용성이 있음을 확인하였다. 향후연구로는 본 논문에서 제안한 유사 질문과 응답을 함께 선정하여 담당자에게 제공하는 방안에 대한 연구와, 답변한 질문과 내용을 지속적으로 데이터베이스에 저비용로 추가 유지할 수 있는 기술에 대한 연구를 수행할 필요가 있다.

후 기

본 연구는 동아대학교 산학협력연구센터의 지원에 의한 것입니다.

[그림 8] 문서처리방법 및 k값의 변화에 따른 성능



참 고 문 헌

[1] T. M. Mitchell. (1997), *Machine Learning.* The McGraw-Hill Companies, Inc

[2] Ian H. Witten and Eibe Frank. *Data mining: practical machine learning tools and techniques with Java implementations.* Morgan Kaufmann, 2000. http://www.cs.waikato.ac.nz/ml/weka/

[3] Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

[4] Cendrowska, J. (1987). "PRISIM: An algorithm for inducing modular rules," *International Journal of Man-Machine Studies* 27(4), pp. 349-370.

[5] Langley, P., W. Iba, and K. Thompson (1992), "An analysis of Bayesian classifiers," *Proc. Tenth Conference on Artificial Intelligence,* San Jose, CA. Menlo Park, CA: AAAI Press, pp. 223-228.

[6] Aha, D. (1992) "Tolerating noisy, irrelevant, and novel attributes in instance-based learning algorithms," *International Journal of Man-Machine Studies 36(2),* pp. 267-216.

[7] 이재식, 이종운 (2002) "사례기반 추론을 이용한 한글 문서분류 시스템," 경영정보학연구 제12권 2호, pp. 179-194.

[8] HAM (Hangul Analysis Module), 국민대학교 자연언어 정보검색 연구실, http://nlp.kookmin.ac.kr/

[9] Brown, P.F.. (1990) "A Statistical Approach to Machine Translation," *Computational Linguistics,* 16(2), pp. 79-85

[10] Baeza-Yates, R. and Ribeiro-Neto, B.. *Modern Information Retrieval.* Addison Wesley, Wokingham, UK, 1999