베이지안 분류기를 이용한 문서 필터링

A Study on Document Filtering Using Naive Bayesian Classifier

손기준, 임수연
경북대학교 컴퓨터공학과

Ki-Jun Son(kijunson@msn.com), Soo-Yeon Lim(nadalsy@hotmail.com)

요약

문서 필터링은 어떤 문서가 특정한 주제에 속하는지의 여부를 판별하는 문제이다. 인터넷과 웹이 널리 퍼지고 이메일로 전송되는 문서의 양이 폭발적으로 증가함에 따라 문서 여과의 중요성도 증가하고 있는 추세이다. 본 논문은 문서 필터링 문제를 이진 문서 분류 문제로 보고, 베이지안 분류기를 필터링 목적으로 사용하였다. 그리고 사용자가 관련성 있는 문서를 제대로 필터링 받기 위해서 학습 대상으로 삼아야 할 문서의 범위나 수, 최소한 체크해야 하는 관련성 있는 문서의 수에 대한 값을 구하는 실험을 수행하였다.

■ 중심어 : | 문서 필터링 | 베이지안 분류기 | 문서 분류 | 기계학습 |

Abstract

Document filtering is a task of deciding whether a document has relevance to a specified topic. As Internet and Web becomes wide-spread and the number of documents delivered by e-mail explosively grows the importance of text filtering increases as well. In this paper, we treat document filtering problem as binary document classification problem and we proposed the News Filtering system based on the Bayesian Classifier. For we perform filtering, we make an experiment to find out how many training documents, and how accurate relevance checks are needed.

■ Keyword : | Document Filtering | Relevance Rating | Bayesian Classifier |

I. 서 론

온라인 정보가 급격히 증가함에 따라 많은 양의 정보 중에서 사용자가 원하는 정보를 정확하고, 신속하게 찾아 주는 정보검색과 문서 필터링의 필요성이 커지고 있다[1].

문서 필터링 문제는 어떤 문서가 특정한 사용자의 관심분야에 관련이 있는지 없는지를 결정하는 문제이다. 필터링은 크게 학습과 필터링 과정으로 구성된다. 학습 과정에서는 문서 필터링을 위해 단어들의 나열로 표시된 문서와 그 문서에 대한 특정 사용자의 관심여부로 이루어진 학습데이터를 이용한다. 필터링 과정에서는 문서 필터링을 위한 규칙을 찾아내고 이를 새로운 문서에 적용하여 그 문서가 사용자의 관심영역에 속하는지

접수번호 : #050321-002
접수일자 : 2005년 03월 21일
심사완료일 : 2005년 04월 08일
교신저자 : 손기준, e-mail : kijunson@msn.com

를 판별한다[2]. 필터링 시스템은 사용자가 어떤 정보에 관심이 있는지, 어떤 정보가 유용한지를 예측하는데 초점을 맞춘다. 기존의 필터링 서비스는 주로 문헌의 내용을 분석하여 정보를 걸러내는 내용기반 필터링 기법을 이용한다.

본 논문은 신문기사 필터링을 위한 베이지안 분류기의 적용 방안에 관한 연구이다. 예를 들어 부동산 동향에 특별한 관심을 보이는 어떤 사용자에게 배달되어진 전자우편이나 전자문서가 이 사용자의 관심사에 적합한지 아닌지를 결정해 주는 것과 같은 문제이다. 따라서 본 연구에서는 웹 상의 신문기사에 대하여 필터링을 적용하기 위해, 필터링 문제를 변형된 문서분류의 문제로 보고 베이지안 이진 분류기를 필터링 목적으로 사용할 때 어느 정도의 조건이 갖추어 지면 좋은 필터링을 행해줄 수 있는지에 대한 연구를 수행 하고자 한다.

이진 문서분류의 경우는 관련성표기가 완전하다. 즉 관련 있는 문서와 그렇지 않은 문서의 구분이 명확하다. 하지만, 신문기사문의 동적 필터링과 같은 영역에서는 관련성 표기가 완전하지 않다. 사용자에게 전체 학습대상문서 중에서 받고 싶은 관심 문서를 모두 판별해 달라고 요구하는 일은 실용적인 차원에서 어려운 일이며, 사용자에게 큰 부담을 주게 된다. 또한 학습대상문서 중 사용자가 관련성 표기를 하지 않은 문서가 비관련 문서 집합에 포함되어 있을 수도 있기 때문에, 기본적으로 불완전한 학습문서가 된다. 즉 신문기사 필터링 문제는 불완전한 학습문서들을 대상으로 얼마나 만족할만한 필터링 결과를 내는가의 문제로 살필 수 있다. 따라서, 본 연구에서는 학습 대상문서의 수와 관련성 표기비율에 따른 분류기의 성능에 대하여 실험하고자 한다.

본 논문은 다음과 같이 구성되어있다. Ⅱ장에서는 관련연구에 대하여 기술하고, Ⅲ장에서는 논문에서 사용하는 학습 방법을 살펴보고, 베이지안 분류기를 이용한 문서 필터링에 대하여 설명한다. 그리고, Ⅳ장에서는 실험에 사용한 데이타에 대한 설명과 실험 방법 및 결과를 상술하며, Ⅴ장에서는 지금까지의 결과를 요약하고 향후 연구 과제를 제시한다.

Ⅱ. 관련연구

최근 문서 필터링에 대한 연구는 주로 인터넷의 유즈넷 뉴스(Usenet news), 전자메일(E-mail), 웹을 대상으로 진행되어져 왔으며, 다양한 문서 필터링 방법이 사용되어지고 있다. 일반적으로 사용되는 분류 알고리즘 중 많이 사용되는 것은 베이지안 분류기, 결정 트리, k-NN 등이 있다[4]. 간단하면서도 좋은 성능을 내는 베이지안 분류기는 확률기반의 모델로서 문서를 이루는 각 단어들이 서로 독립이라는 가정을 전제로 한다. 이 가정 때문에 전체 문서집합에 대한 단어별 빈도수와 문서내의 단어빈도수 정보만 있으면 분류를 할 수 있다. 문서에 대하여 가장 좋은 분류 결과인 v_{NB} 를 돌려주는 베이지안 분류기는 식1과 같다.

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i \in positions} P(a_i \mid v_j) \quad (1)$$

위의 식 (1)에서 v_j 는 가능한 클래스 값이고, a_i 는 문서 내에서 발견된 단어를 뜻한다[3].

내용기반 필터링 기법은 정보 검색이나 정보 필터링 연구에서 자연적으로 발전하였다. 내용기반 필터링 기법은 대부분 문서를 추천하기 위해 문서의 내용과 사용자의 정보요구간의 유사도를 측정하고, 그 결과를 순위화 하여 보여준다. 이렇게 문서의 내용을 분석하여 이용자에게 추천하는 기법을 내용기반 필터링 기법이다. 이 기법은 정보 검색에 기반을 두고 있으므로 가중치 기법, 적합성 피드백, 불리안 검색 모형 등을 이용한다.

본 연구에서는 신문기사 필터링을 위해 간단하면서 잘 알려진 전통적인 분류방법으로 문서분류에서 좋은 성능을 보이고 있는 베이지안 분류기를 이용하여 신문기사 필터링 시스템을 구현한다. 하지만 웹상의 신문기사 필터링 서비스와 같은 실용적인 용도에 기존의 문서분류에서 널리 사용되고 있는 베이지안 분류기를 그대로 적용하는 데는 문제가 있다. 기존에 연구된 방법들은 1년치 혹은 그 이상 의 학습문서를 대상으로 학습단계를 수행하며, 또한 학습대상 문서에 대한 관련성표기가 완전한 문서집합을 사용하여 분류기를 학습시키고 있

다. 그러나 웹상의 신문기사 필터링과 같은 영역에서는 일반 사용자가 기사문을 모두 읽고 관련토픽을 오류없이 모두 명시 하는 것은 쉽지 않은 일이므로, 오류가 포함된 학습문서에서 분류기가 어느 정도의 성능을 내는지를 살펴보고자 한다.

III. 베이지안 분류기를 이용한 문서 필터링

3.1 베이지안 분류기

베이지안 분류기는 베이즈 정리(Bayes' theory)에 기초하고 특성들 간의 독립성을 가정한 확률적인 모델이다. 매우 단순하지만 잘 알려진 전통적인 분류방법으로, 텍스트 문서분류에 사용되어 왔다[5][6]. 베이지안 분류기는 통계적인 알고리즘으로 학습문서의 여러 통계정보를 학습하고, 이렇게 얻은 통계정보를 이용하여 입력 문서 스트림으로부터 문서를 분류한다.

확률이론을 기계학습에 적용한 것으로, 특정 데이타 집합 D를 조사했을 때 가설 h가 사실일 확률은 $P(h|D)$가 된다. 그리고 가설이 사실일 경우 데이타 D의 확률이 $P(D|h)$일 때 베이즈 정리는 다음 식 (2)와 같다.

$$P(h|D)=\frac{P(D|h)P(h)}{P(D)} \tag{2}$$

위 식에서 $P(h)$는 데이타에관한 정보가 주어져 있지 않을 때, 가설이 사실일 사전확률(prior probability)이다. 기계학습에서 관심을 가지는 값은 $P(h|D)$인데, 베이지안 학습방법은 가설집합 H에 포함된 가설 중 최대확률을 가지는 가설 h를 구하는 것이다.

최대확률을 구하기 위해서는 최대사후확률(MAP)을 계산하면 된다. 이 확률은 데이타를 조사했을 때 가장 가능성이 높은 가정으로서 식 (3)을 이용한다.

$$h=\arg\max_{h\in H} P(h|D)$$

$$=\arg\max_{h\in H}\frac{P(D|h)P(h)}{P(D)} \tag{3}$$

위의 식 (3)의 $\arg\max P(h|D)$는 확률 $P(D|h)$가 최대가 되는 H 내에서 가설 h를 나타내는데, 이때 분모 $P(D)$는 h와 무관하기 때문에 상수로 간주하여 식(4)과 같이 표현된다.

$$h=\arg\max_{h\in H} P(D|h)P(h) \tag{4}$$

식 (3)의 $P(h|D)$는 학습문서 D가 주어질 때 가설 h가 성립할 확률을 말하는데, 식 (4)에서 나타내는 바와 같이 베이지안 학습방법은 가설집합 H에 포함된 가설 중 가장 큰 확률을 가지는 h를 찾아 최종 가설로 설정하는 것이다.

베이지안 분류자는 베이지안 학습방법 중에서 널리 쓰이는 통계적인 알고리즘이다. 베이지안 학습법은 각 사례 x가 속성값들의 벡터로 표시되고, 기계학습의 결과로서 구하고자하는 목적함수 $f(x)$가 범주들의 유한집합 V의 원소인 경우의 학습에 잘 적용된다. 즉, 기계학습을 위해 사용하는 학습문서인 속성벡터 x 각각에 대하여 가 설$h(x)$가 주어질 때 임의의 속성벡터 x에 대하여 $h(x)=f(x)$인 함수 f를 목적함수라고 하며, 미리 정의된 범주들의 집합 V에서 함수 f를 대신할 수 있는 확률이 가장 큰 원소인 $v\in V$를 구하는 것이 최종 목표이다.

학습문서집합과 새로운 사례가 $(a_1,a_2,\ldots,a_n)$과 같이 속성값들의 벡터로 주어지면, 학습기는 이 사례에 대한 목적함수의 값 혹은 분류를 예측할 수 있다. 새로운 사례에 대한 분류를 예측하는 방법은 주어진 속성벡터 $(a_1,a_2,\ldots,a_n)$에 대응되는 가장 가능성이 높은 목적함수의 값 $vMAP$를 다음 식(5)와 같이 구하는 것이다.

$$vMAP=\arg\max_{v_j\in V} P(v_j|a_1,a_2,\ldots,a_n,D) \tag{5}$$

여기서 V는 미리 정의된 범주들의 집합을 나타낸다. 식 (5)에 베 이즈 정리를 적용하면 식 (6)과 같다.

$$vMAP = \arg\max_{v_j \in V} P(a_1, a_2, \ldots, a_n \mid v_j, D) P(v_j \mid D) \quad (6)$$

이때 학습 문서에 대한 각 범주의 확률 $P(v_j \mid D)$는 식 (7) 과 같이 계산되는데, $|v_j|$가 범주 v_j에 속한 속성들의 개수이고, N_{Total}이 전체 속성데이터의 수이므로 학습 문서 집합전체를 해당 범주에 속한 데이타의 개수로 나누면 된다.

$$P(v_j \mid D) = \frac{|v_j|}{N_{Total}} \quad (7)$$

$P(a_1, a_2, \ldots, a_n \mid v_j, D)$는 추정하기가 쉽지 않기 때문에 베이즈 독립 가정을 이용하여 해결한다. 적용할 베이즈 독립가정은 "주어진 범주에 대해 속성값들은 모두 독립이다." 라는 것이며, 베이즈 독립가정을 적용하면 식 (8)과 같이 표현할 수 있다.

$$P(a_1, a_2, \ldots, a_n \mid v_j, D) = \sum_{i=1} P(a_i \mid v_j, D) \quad (8)$$

베이즈 독립가정인 식 (8)을 식 (6)에 적용하면 식(9)가 구해진다.

$$vMAP = \arg\max_{v_j \in V} \frac{|v_j|}{N_{Total}} \sum_{i=1} P(a_i \mid v_j, D) \quad (9)$$

식 (9)에서 $P(a_i \mid v_j, D)$를 구하기 위해서는 각 범주에 속한 학습 문서집합에서 각각의 속성값이 발생하는 빈도수를 계산해야 한다. 즉,

$$P(a_i \mid v_j, D) = \frac{|a_i|}{\text{범주 } v_j \text{에 속한 사례들의 수}} \quad (10)$$

이다. 위의 식 (10)에서 $|a_i|$는 속성 a_i의 수를 나타낸다.

베이지안 분류기는 문서가 각 범주에 할당될 확률을 계산하여 최대값을 가지는 범주에 문서를 할당한다. 따라서 문서분류 에서 학습문서 수와 그 학습문서를 구성하는 범주의 비율은 실제로 발생할 대상 문서의 성격을 잘 반영할 수 있을 만큼 크고 신뢰성이 있어야 한다.

베이지안 분류기는 문서에 나타난 단어들의 분포는 서로 독립임을 가정하며, 단어가 나타날 확률은 문서 내에서 단어의 위치와도 독립적이라고 가정한다.

3.2 내용기반 필터링 기법

내용기반 필터링 기법은 문헌 집단 내에서 질의 문헌과 다른 문헌간의 유사도를 측정하여 가장 유사도가 높은 문헌을 추천한다. 문헌간의 유사도 측정은 코사인 유사계수 방법을 이용한다.

코사인 유사계수 공식은 식 (11)과 같다.

$$sim(d,q) = \frac{\sum_{i=1}^{n} w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^{n} w^2_{i,j}} \cdot \sqrt{\sum_{i=1}^{n} w^2_{i,q}}} \quad (11)$$

위 식에서 $w_{i,j}$는 문헌 d_j에서 i번째 용어의 가중치를 나타내며, $w_{i,q}$는 질의 문헌 q에서 i번째 용어의 가중치를 나타낸다[7]. 용어의 가중치기법으로는 단어 빈도와 가장 일반적으로 널리 사용되는 역문헌 빈도를 사용하였다.

3.3 문서 필터링

문서 필터링이란 해당 문서집합으로부터 사용자가 필요로 하는 문서를 여과하는 것을 말한다. 본 논문에서는 필터링 문제를 변형된 문서분류 문제로 파악한다. 즉 필터링을 관계있는 문서와 그렇지 않은 문서로 분류하는 이진 문서분류 문제로 보면, 이는 곧 사용자가 필요로 하는 문서의 여과문제가 된다. 따라서 본 논문에서는 정보여과 장치로서 필터링 기술을, 신문기사에 대해서 적용 하고자 한다. 신문기사는 동적 정보 문서의 면모를 충분히 지니고 있기 때문에 필터링의 대상문서로 적합

하다.

본 연구에서 구현한 필터링 시스템의 사용 모델은 온라인상에서 계속적으로 발생하는 문서들을 브라우징하던 사용자가 자신의 관심 대상이 되는 문서 몇 개를 시스템에 제출하는 것으로부터 시작된다. 즉 사용자의 요구사항을 받은 필터링 시스템은 사용자가 구분해준 문서와 나머지 문서들을 대상으로 학습과정을 수행하고, 이어지는 다음 문서의 스트림으로부터 필터링을 행하여 사용자에게 여과된 정보를 제시하여 주게 된다. [그림 1]은 신문기사 필터링 시스템의 모델이다.

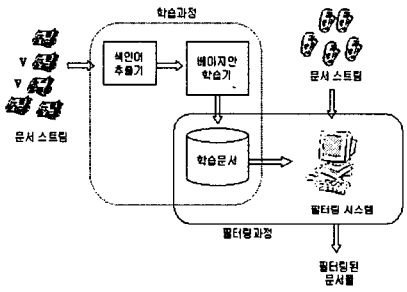

그림 1. 신문기사 필터링 시스템 모델

[그림 1]과 같은 실행 모델에 있어서 베이지안 분류기를 바로 필터링 시스템으로 사용하기에는 몇 가지 문제가 있다. 이는 기본적으로 학습문서가 완전하지 않기 때문에 발생하는 문제들이다. 첫째로 학습대상이 되는 문서의 크기 자체가 충분하지 못하다는 점이다. 하루 혹은 그 이상 분량의 신문기사들을 브라우징하던 사용자가 자신의 관심사에 따라 몇 개의 문서들을 선택하는 모델을 고려해보자. 만약 학습의 대상을 사용자가 브라우징하고 있던 기사 들이나 그날 발생한 모든 기사로 두더라도, 충분히 많은 양의학습 문서라고 볼 수는 없다. 둘째로 사용자가 '이와 같은 문서들을 내게 가져오라'라고 체크하는 예시 문서가 이진분류의 경우와 달리 관련성 표기가 완전하지 않다. 즉 사용자가 일관성을 가지고 몇 개의 문서를 선택했다 하더라도, 사용자가 학습 대상이 되는 문서 전체에서 한 문서도 빠뜨리지 않고 관련문서를 제시해 주었다고 볼 수 없다. 그러므로 대상 문서들이 많으면 많을수록, 사용자에게 전체 학습대상 문서 중에서 받고 싶은 관심문서를 모두 판별해 달라고 요구하는 것은 실용적인 차원에서 어려운 일이다. 이러한 작업은 일반 사용자에게 큰 부담을 주게 되며, 특히 기사문의 동적 필터링과 같은 영역에서 실용적이지 않다. 따라서 학습 대상이 되는 문서 중 사용자가 관련성 표기를 하지 않은 문서가 비관련 문서집합 에 포함되어 있을 수 있다. 그러므로, 신문기사 필터링은 기본적으로 불완전한 학습문서를 대상으로 하게 된다. 결론적으로, 신문기사의 필터링문제는 이와 같이 불완전한 학습문서들을 가지고 얼마나 만족할 만한 필터링 결과를 내는가의 문제로 살필 수 있다.

본 논문은 이와 같은 응용대상에 대해서, 베이지안 이진 분류기를 필터링 목적으로 동적으로 사용할 때 어느 정도의 조건이 갖추어지면 만족할만한 필터링을 행해줄 수 있는지에 대한 연구이다. 또한 기존의 필터링 방법인 내용기반 필터링 기법과의 성능을 비교 실험한다.

Ⅳ. 실험 및 분석

실험은 두 가지 문제에 대해 차례로 수행하였다. 먼저 학습의 대상이 되는 문서의 크기를 변경하며 최소한 어느 정도의 학습문서가 필요한지를 실험하였다. 이어 학습 대상이 되는 문서전체에서 실제로 필터링 대상 주제와 관련 있는 문서를 사용자가 어느 정도의 비율로 선택하였을 경우 만족할만한 필터링 결과를 내는가에 대하여 실험하였다. 실험은 6개의 토픽 교육, 경제, 테러, 기업, 환경, 선거 관련이 표기된 문서집합을 사용하며 실제로 실시간으로 발생한 신문기사를 연속적으로 모은 일정 양의 기사를 사용하였다.

4.1 평가방법

문서분류 시스템의 성능을 평가하는 방법으로는 오류(error), 재현 율(recall), 정확률(precision), F1척도 등이 있다[7]. 본 논문에서는 F1 척도를 사용한다. F1척도는 정확률과 재현율에 동등한 중요도를 부여하는 하나의 평가방법으로 사용하는 것으로 식 (12)와 같다.

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (12)$$

4.2 학습문서의 수에 따른 분류기의 성능

본 실험은 베이지안 이진 분류기가 제한된 학습문서 집합에서 어느 정도의 성능을 내는지를 살펴보고자 한다. 여섯 개의 토픽이 표기된 학습문서에 대하여 학습문서의 수를 변경해가며, 베이지안 이진분류기와 내용기반 필터링 기법의 성능을 실험하였다. 학습문서는 신문의 일면 기사들로 제한하고, 10일간의 신문기사 600건을 사용하였다. 실험대상 문서는 학습 문서를 모두 수집한 다음부터 발생한 7일 간의 표제 기사 353개의 문서를 사용한다.

표 1. 베이지안 분류기의 성능

	학습문서(개)				
토픽	600	300	120	60	30
Topic1	57.5	49.8	43.2	41.0	32.9
Topic2	57.3	48.0	45.7	37.7	33.3
Topic3	64.1	59.3	54.0	33.3	19.6
Topic4	52.4	37.7	39.8	31.6	22.4
Topic5	59.9	53.5	46.5	37.7	14.3
Topic6	66.9	61.4	51.1	43.8	37.3

[표 1]에서 볼 수 있듯이 120여개 이상의 문서를 학습 대상으로 했을 경우 분류기의 성능을 살펴보면 많은 차이를 보이지 않는다. 하지만 문서의 수가 60개 이하로 줄어들면서 분류기의 성능이 낮아짐을 볼 수 있다. 베이지안 분류기는 학습문서의 수가 60개 이하로 줄어들면 분류식의 값에 유의미한 영향을 미치는 어휘의 숫자가 평균 10~15개 이하로 떨어지기 때문에 좋은 성능을 보이지 않았다.

그 결과, 토픽의 발생 빈도에 따라 다르지만, 특정 토픽(테러, 선거 관련)은 문서의 수가 증가할수록 성능의 향상을 보였지만 일반적인 주제를 다루는 토픽(기업, 환경 관련)은 큰 성능의 향상을 보이지 않았다.

또한, [표 1]을 보면 토픽 3이 다른 토픽들 보다 더 나은 성능을 보이고 있다. 그 이유는 토픽 3에 등장한 어휘들이 다른 토픽에 나타난 어휘들보다 분별력이 높기 때문으로 추정된다. 다시 말해 그 주제에 나타난 어휘가 다른 주제에는 자주 사용되지 않기 때문으로 보인다. 예를 들면 테러관련 문서에서 '테러', '부대'와 같은 어휘는 다른 주제에 자주 사용되지 않는 어휘이다. 따라서 이러한 어휘들은 테러관련 문서에 국한되어 나타남으로써 정확률을 높여준다. 토픽별 성능을 살펴보면 토픽 4가 가장 낮은 성능을 보였고, 또한 학습문서의 수에 따라 성능에 큰 변화를 보이지 않았다. 결론적으로 토픽 3과 6이 전반적으로 높은 성능을 보였고, 문서의 특정 부분집합 300~600개 사이에서는 성능의 차이가 미미함을 볼 수 있었다.

학습 문서의 수에 따른 내용기반 필터링 기법의 성능은 [표 2]와 같다.

표 2. 내용기반 필터링 기법의 성능

	학습문서(개)				
토픽	600	300	120	60	30
Topic1	45.5	43.8	40.2	42.0	37.9
Topic2	46.3	44.0	42.7	40.7	35.3
Topic3	60.1	57.3	53.0	38.3	35.6
Topic4	48.4	45.7	39.8	40.6	41.4
Topic5	50.9	49.5	44.5	39.7	44.3
Topic6	55.9	56.4	53.1	42.8	38.3

내용기반 필터링 기법은 문헌의 내용을 기반으로 추천을 한다. 필터링 대상이 문헌일 내용의 분석이 필수적으로 이루어져야 하기 때문에 내용 기반 필터링 추천 기법이 가장 적용하기 쉽고 보편적인 추천 기법 중의 하나라고 할 수 있다. 하지만 내용기반의 필터링 기법의 경우 추천을 하기위해 이용된 주요 키워드가 낮은 출현 빈도를 갖거나 너무 일반적인 성격의 키워드가 많이 포함된다면 성능이 떨어질 수밖에 없다. [표 2]에서 볼 수 있듯이 학습 문서의 수가 많아져도 큰 성능의 향상은 보이지 않는다.

4.3 사용자의 관련성 표기비율에 따른 분류기의 성능

본 실험에서는 사용자가 해당 문서집합에서 실제로 필터링 대상 주제와 관련 있는 문서를 어느 정도 비율로 선택을 하였을 경우 만족할만한 필터링 성능을 내는

지를 살펴본다. 학습문서는 신문기사 120건을 사용한다.

우선, 사용자가 해당 문서집합에서 관련성 있는 모든 문서를 선택하지 않는다고 가정하자. 이것은 사용자가 선택한 범주에 오류가 포함되어 있다는 것을 의미한다. 이러한 상황에서 실제 관련 문서의 어느 정도까지 사용자에 의해 관련성 있다고 체크되어야 신뢰할만한 결과를 보여주는지에 대해 실험을 하였다.

그 결과, 사용자의 관련성 문서 선택 비율에 따라 다르지만, 신문 기사 토픽을 필터링하기 위한 관련성 표기 비율이 높을수록 좋은 성 능을 낸다. [표 3]과 [표 4]는 여섯 개의 토픽에 대한 관련성 표기비율에 따른 성능을 보여준다. [표 3]에서 보면 해당문서집합에서 관련성 표기비율이 80%정도까지 줄어들어도 분류기의 성능에는 큰 차이를 보이지 않는다. 하지만 관련성 표기비율 이 80%와 60%인 경우를 비교해보면 정확률에는 많은 차이를 보이지 않지만 재현율 측면에서는 현저히 차이가 있음을 볼 수 있다. 즉 관련성표기비율에 영향을 받는다. 또한 토픽 3과 6은 관련성표기 비율에 따라 성능의 차이를 보이지만, 토픽 2와 5는 관련성 표기에 따라 큰 성능의 차이를 보이지 않는다. 즉, 일반적인 주제를 다루는 토픽의 경우는 관련성 표기에 많은 영향을 받지 않지만, 특정 주제 와 관련된 토픽은 관련성 표기비율에 따라 성능의 차이를 보임을 알 수 있었다.

내용기반 필터링 기법은 관련성 표기 비율에 따라 큰 성능의 차이를 보이지 않음을 알 수 있다. 토픽 5와 6이 다른 토픽에 비해 나은 성능을 보인다. 이런 결과는 내용기반 필터링 기법이 키워드를 통해 유사한 문서를 추천하기 때문에 특정 주제에 대해서는 성능이 향상됨을 볼 수 있다.

표 3. 베이지안 분류기의 성능

	관련성 표기 비율(%)				
토픽	100	80	60	50	30
Topic1	55.4	47.7	30.2	14.1	9.2
Topic2	60.9	52.3	28.4	16.5	13.5
Topic3	58.6	45.9	33.3	27.1	16.4
Topic4	63.0	53.3	35.9	18.3	15.3
Topic5	66.5	56.0	34.3	16.4	5.5
Topic6	71.0	63.3	31.4	19.4	9.1

표 4. 내용기반 필터링 기법의 성능

	관련성 표기 비율(%)				
토픽	100	80	60	50	30
Topic1	50.4	42.7	35.2	30.1	21.2
Topic2	51.9	49.3	45.4	37.5	12.5
Topic3	50.6	41.9	34.3	29.1	20.4
Topic4	54.0	50.3	43.9	38.3	19.3
Topic5	60.5	51.0	44.3	35.4	22.5
Topic6	62.0	57.3	50.4	33.4	32.1

4.4 필터링 과정 및 인터페이스

신문기사 필터링 시스템은 크게 학습과정과 필터링과정으로 살펴볼 수 있다. 학습과정은 사용자가 자신의 관심대상이 되는 문서를 시스템에 제출하는 것으로부터 시작된다. 즉, 사용자가 문서스트림에서 필터링 받기를 원하는 문서를 선택하면 시스템은 사용자가 구분한 문서와 나머지 문서를 대상으로 학습과정을 수행하게 된다. 필터링과정은 학습과정이후 이어지는 다음 문서의 스트림으로 부터 필터링을 수행하여 사용자에게 여과된 정보를 제시하게 된다.

신문기사 필터링 시스템은 크게 두 부분으로 구성이 된다. 온라인상에서 브라우징을 하던 사용자가 문서 스트림으로부터 관심 문서를 선택하는 부분과 사용자에게 여과된 정보를 제시하는 부분이다. [그림 2]는 직접 사용자와 상호 작용을 통하여 사용자로부터 관심대상 문서를 선택받아 이를 신문기사 추천 시스템에 넘겨주게 된다. 이렇게 사용자가 구분해준 문서를 대상으로 신문기사 추천 시스템은 학습 과정을 수행한다. 이어지는 다음 문서 스트림으로부터 여과된 문서들은 [그림 3]과 같이 사용자에게 제시된다.

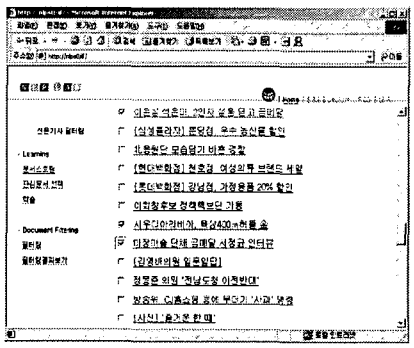

그림 2. 문서 스트림에서 관심문서 선택

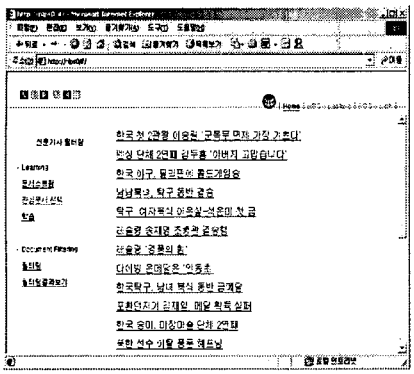

그림 3. 필터링 결과

V. 결 론

본 논문은 베이지안 이진 분류기를 웹상의 신문기사 필터링에 적용하기위한 연구이다. 실험은 크게 두 부분으로 나누어 살펴보았다. 먼저, 학습문서의 수에 따른 성능변화를 살펴보았으며, 다음으로 관련성 있는 문서가 완전히 명기되지 않았을 때 보여지는 베이지안 분류기와 내용기반 필터링 기법의 성능을 실험하였다. 실험한 결과, 베이지안 분류기는 학습문서의 비율이 높으면 높을수록, 관심 주제의 문서 발생 비율이 낮더라도 필터링의 성능은 향상을 보였다. 하지만 60여개 기사(평균 160어절 정도)에서 관심토픽이 10% 이하로 발생할 경우는 필터링에 유의미한 값을 보이는 어휘의 개수가 10~15개 이하로 줄어들어 만족할만한 필터링 결과를 보이지 못했다.

내용기반 필터링 기법의 경우 코사인 유사계수를 이용하였으며, 용어 가중치 기법으로 단어빈도와 역문헌빈도를 이용하였다. 그 결과 학습문서의 수 보다는 관련성 표기 비율에 더 영향을 받는 것을 볼 수 있었다.

따라서 이진 필터링 상황에서 베이지안 분류기는 문서집합의 크기가 일정한 정도 일 때 관련성 있는 문서가 모두 표기되지 않더라도 필터링에는 큰 영향을 미치지 않음을 볼 수 있었다. 일반 독자들이 많은 양의기사문을 모두 읽고 관련토픽을 오류 없이 모두 명시하는 것이 쉽지 않은 일이므로, 어느 정도의 오류 및 누락과 관계없이 동작할 수 있는 이러한 특징은, 웹상의 신문기사 필터링 서비스와 같은 실용적 용도의 필터링을 위해 바람직한 것으로 보인다.

향후 연구 과제로는 실용적인 신문기사 필터링을 위해, 필터링 결과에 대한 사용자의 반응으로부터 분류기의 학습을 개선해 나가기 위한 적절한 피드백방법에 대한 연구가 필요하다.

참 고 문 헌

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

[2] 김유환, 박성배, 장병탁, "BayesBoost를 이용한 한글 문서 라우팅," 제5회 한국 과학기술 정보인프라 워크숍 학술발표 논문집, pp.232-244, 2000.

[3] T.M.Mitchell, Machine Learning, McGraw Hill, 1997.

[4] I.Androutsopoulos, J.Koutsias, K.V.Chandrinos, G.Paliouras and C.D.Spyropoulos, "An Evaluation of Naive Bayesian Anti-Spam Filtering," Workshop on Machine Learning in the New Information Age, pp.9-17, 2000.

[5] D.Lewis, R.Schapire, J.Callan and R.Papka, "Training Algorithm for Linear Text Classifiers," In Proceedings of SIGIR-96, pp.298-306, 1996.

[6] 김진앙, 신상규, "베이지안 학습을 이용한 문서의 자동분류," 정보과학회논문지, Vol.11, No.1, pp.19-30, 2000.

[7] G.Salton and M.J.McGill, Introduction to modern information retrieval, McGraw Hill, 1983.

저 자 소 개

손 기 준(Ki-Jun Son) 정회원

- 2000년 2월 : 상주대학교 컴퓨터공학과(공학사)
- 2003년 2월 : 경북대학교 컴퓨터공학과(공학석사)
- 2003년 3월~현재 : 경북대학교 컴퓨터공학과 박사과정

<관심분야> : 정보검색, 기계학습, 자연어처리

임 수 연(Soo-Yeon Lim) 정회원

- 1988년 2월 : 경북대학교 전자공학과(공학사)
- 1993년 2월 : 경북대학교 컴퓨터공학과(공학석사)
- 2004년 8월 : 경북대학교 컴퓨터공학과(공학박사)
- 2004년 9월~현재 : 경북대학교 컴퓨터공학과 연구원

<관심분야> : 자연어처리, 정보검색, 시멘틱 웹, 기계학습