베이지안 분류기를 이용한 신문기사 필터링

손기준[0] 노태길 이상조
경북대학교 컴퓨터공학과
{kjson,nayas}@sejong.knu.ac.kr sjlee@knu.ac.kr

A Study On Filtering of Newspaper Article by Using Bayesian Classifier

Ki-Jun Son[0] Tae-Gil Noh Sang-Jo Lee
Dept. of Computer Engineering, Kyungpook National University

요 약

본 논문에서는 필터링 문제를 이진 문서 분류 문제로 보고 신문기사 필터링에 베이지안 분류자를 사용한다. 신문 기사 필터링 문제에서 베이지안 분류자를 사용할 경우 학습 문서가 고정되어 있지 않기 때문에 여러 가지 파라미터를 사용하여 실험을 하였다. 실험 결과 베이지안 이진 분류기는 제한된 학습 문서에서 더 나은 성능을 보였고, 해당 문서 집합에서 10%이상 비율의 문서를 사용자가 선택해야 함을 알 수 있었다.

1. 서 론

문서 필터링이란 해당 문서 집합으로부터 사용자가 필요로 하는 문서를 여과하는 것을 말한다. 인터넷의 발전과 더불어 온라인상의 정보량이 증가 하고, 이러한 정보 중 사용자가 원하는 정보만을 얻기가 어렵다. 정보의 종류가 다양해지고 정보의 양이 증가할수록 사용자가 필요로 하는 정보를 찾기 위한 시간과 노력이 증가하게 된다. 이러한 정보과잉문제에서 사용자가 필요로 하는 정보를 여과하기 위한 필터링에 대한 연구로는 베이지안 네트워크를 이용한 정크 메일 필터링[1], 카이제곱 통계량 을 이용하여 단어의 가중치를 구하고, 이를 로그(Log) 단어 가중치 공식에 적용하여 스팸메일을 필터링하는 연구가 있다[2]. 그리고 TREC(Text REtrieval Conference)도 필터링 트랙을 도입하여 많은 연구를 수행 하고 있다[3].

본 연구에서는 신문기사에 대해서 필터링을 적용하기 위해, 필터링 문제를 변형된 문서 분류의 문제로 보고, 베이지안 이진 분류기를 필터링 목적으로 사용할 때, 학습 대상으로 삼아야 할 문서의 범위나 수 및, 사용자가 관련성 있는 문서를 제대로 필터링 받기 위해서 최소한 체크해야 하는 관련성 있는 문서의 수에 대한 값을 얻고자 한다. 학습문서의 수에 따른 분류기의 성능과, 사용자가 선택한 문서의 수에 따른 분류기의 성능에 대하여 실험을 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 베이지안 분류기에 대하여 설명하고, 3장에서는 신문기사 필터링에 대하여 살펴보고, 4장에서는 실험에 대하여 기술할 것이며 5장에서는 결론 및 향후 연구과제를 제시한다.

2. 베이지안 분류기

베이지안 분류기는 잘 알려진 전통적인 분류방법으로 텍스트 문서 분류에 사용되어 왔으며, 통계적인 알고리즘으로 학습 문서의 여러 통계 정보를 학습한다. 그리고, 이렇게 얻은 통계정보를 이용하여 새로운 문서를 분류한다.

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

베이지안 분류기는 범주가 출현할 사전 확률을 기반으로 하여 특정 범주가 문서에 할당될 확률과, 특정 단어가 문서에서 발생할 조건 확률을 계산 분류하려는 문서 D에 단어 Voc_i가 출현한 경우 문서가 범주 C_j에 분류될 확률을 아래와 같이 계산된다.

$$P(C_j \mid D) = \arg\max P(C_j) \cdot \sum_{i=1} P(Voc_i \mid C_j)$$

이 경우 $P(C_j \mid D)$은 문서 D가 범주C_j에 할당될 확률이고, $P(C_j)$와 $P(Voc_i \mid C_j)$학습 문서에서 아래 와 같이 계산할 수 있다.

$$P(C_j) = \frac{C_j\text{에 할당된 학습문서의 수}}{\text{모든 학습문서의 수}}$$

$$P(Voc_i \mid C_j) = \frac{C_j\text{에 할당된문서에서}Voc_i\text{가 발생한횟수}}{C_j\text{에 할당된문서에나타난모든단어의발생횟수}}$$

따라서, 베이지안 분류기는 문서가 각 범주에 할당될 확률을 계산 최대값을 가지는 범주에 문서를 할당
한다. 이와 같은 문서 분류에서 학습 문서 수와 그 학습 문서를 구성하는 범주의 비율은, 실제로 발생할 대상 문서의 성격을 잘 반영할 수 있을 만큼 크고 신뢰성 있어야 한다.

3. 신문 기사 필터링

문서 필터링이란 해당 문서 집합으로부터 사용자가 필요로 하는 문서를 여과하는 것을 말한다. 필터링 문제를 변형된

문서 분류 문제로 파악할 수 있다. 즉 관계 있는 문서와 그렇지 않은 문서로 분류하는 이진 문서 분류 문제로 보면, 이는 곧 사용자가 필요로 하는 문서의 여과 문제가 된다.

본 논문에서는 정보 여과 장치로서 필터링 기술을, 신문기사에 대해서 적용하고 있다. 신문 기사는 필터링 대상으로 동적 정보 문서의 면모를 충분히 지니고 있다. 본 연구에서 구현한 필터링 시스템의 사용 모델은 온라인 상에서 계속 적으로 발생하는 문서들을 브라우징 하던 사용자가, 자신의 관심 대상이 되는 내용을 담은 문서 몇 개를 시스템에 제출하는 것으로 시작된다. 사용자의 요구사항을 받은 필터링 시스템은 사용자가 구분해준 문서와 나머지 문서들을 대상으로 학습을 거치고, 이어지는 다음 문서의 스트림으로부터 필터링을 행하여 사용자에게 여과된 정보를 제시하여 주게 되는 모델이다. 아래 그림1은 신문기사 필터링 시스템의 모델이다.

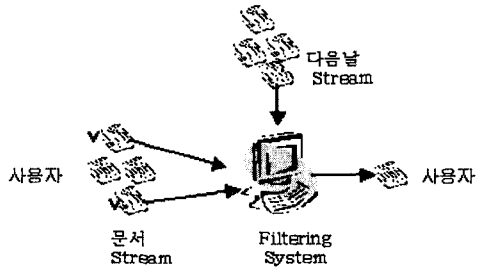

그림1 신문기사 필터링 시스템의 모델

위와 같은 실행 모델에 있어서 베이지안 분류기를 바로 필터링 시스템으로 사용하기에는 몇 가지 문제가 있다. 이 문제는 기본적으로 학습 문서가 완전하지 않다는 점이다. 첫째로 학습 대상이 되는 문서의 크기 자체가 충분하지 못하다. 전체 문서들 하루 혹은 그 이상 분량의 신문 기사들을 브라우징 하던 사용자가 자신의 관심사에 따라 몇 개의 문서들을 선택하는 모델을 고려해 보자. 만약 학습의 대상을 사용자가 브라우징 하고 있던 기사들이나, 그날 발생한 모든 기사로 두더라도, 충분히 많은 양의 학습 문서라고 볼 수는 없다. 둘째로 사용자가 ' 이와 같은 문서들을 내게 가져오라' 라고 체크 하는 예시 문서가, 이진 분류의 경우와 달리 완전하지 않다. 사용자가 일관성을 가지고 몇 개의 문서를 선택했다 하더라도, 그가 학습 대상이 되는 문서 전체에서 한 문서도 빠뜨리지 않고 관련 문건을 제시해 주었다고 볼 수 없다. 대상 문서들이 많으면 많을수록, 사용자에게 전체 학습 대상 중에서 받고 싶은 관심 문서를 모두 판별해 달라고 요구하는 것은 실용적인 차원에서 어려운 일이다. 이러한 작업은 일반 사용자에게 큰 부담을 주게 되며, 기사문의 동적 필터링과 같은 영역에 실용성이 없다. 학습 대상이 되는 문서중에서 사용자가 관련성을 표기하지 않은 문서 중에도 관련성 있는 문서가 있을 수 있으므로, 기본적으로 불완전한 학습 문서가 된다.

따라서 신문 기사의 필터링 문제는 이와 같이 불완전한 학습 문서들에서 어떻게 만족할 만한 필터링 결과를 내는가의 문제로 살필 수 있다. 본 논문은 이와 같은 응용대상에 대해서, 베이지안 이진 분류기를 필터링 목적으로 동적으로 사용할 때에, 어느 정도의 조건이 갖추어 지면 좋은 필터링을 행해줄 수 있는지에 대한 연구이다. 학습 문서의 수가 어느 정도까지 줄어들어도 신뢰할 결과를 보여주는지, 또 학습 문서 중에 포함된 실제 관련 문서의 어느 정도까지 사용자에 의해 관련성 있다고 체크되어야 신뢰할 만한 결과를 보여주는지를 실험 하였다.

4. 실험

실험은 두가지 문제에 대해 차례로 수행하였다. 먼저 베이지안 필터링에서 학습의 대상이 되는 문서의 크기를 변경하며 최소한 어느 정도의 학습문서와 문서내의 관련문서가 필요한지를 실험하였고, 이어 학습 대상이 되는 문서 전체에서 실제로 필터링 대상 주제와 관련 있는 문서를 사용자가 어느 정보의 비율로 선택 하였을 경우 만족할 만한 필터링 결과를 내는가에 대하여 실험하였다. 실험은 실제로 실시간으로 발생한 신문기사를 연속적으로 모은 일정 양의 신문기사를 대상으로 이루어졌다. 실험 대상으로는 중앙일보 신문기사 2002년 3월의 기사들을 사용하였다.

4.1 베이지안 이진 분류기의 성능에 대한 실험

실시간으로 발생한 신문기사에 대해 동일한 주제에 대한 필터링을, 학습문서의 양을 사흘간 발생한 분량에서부터, 사용자가 한번에 브라우징 할 만한 적은 크기의 문서 수까지 크기를 변경하며 실험해 보았다. 이 때 관련 문서는 모두 표기되어 있다.

첫번째 실험은 베이지안 이진 분류기가 제한된 학습 문서 집합에서 어느 정도의 성능을 내는지에 대한 실험이다. 선거 관련, 환경문제, 교육관련과 같이 3개의 토픽이 표기된 학습 문서에 대하여 학습 문서의 수를 변경하여 가며 베이지안 이진 분류기의 성능을 실험하였다. 학습문서는 신문의 1면 기사들로 제한하고, 5일간의 문서 120건을 학습문서로 사용하였다. 실험 대상은 학습 문서를 모두 수집한 뒤부터 발생한 6일 동안의 표제 기사 353개중에서 관련 문서를 골라내는 실험을 수행하였다. 각각 필터링 작업을 수행하고 어떤 성능을 보이는지를 실험하였다.

토픽의 발생 빈도에 따라 다르지만, 신문기사 토픽을 필터링 하기 위한 학습 문서로는, 60여개 이상의 문서가 필요한 것으로 보인다. 이 정도의 이하의 양으로 줄어들면, 분류식의 값에 유의미한 영향을 미치는 어휘의 숫자가 평균 15개 이하로 줄어들고, 신뢰할 만한 결과를 얻을 수가 없었다. 60개 정도의 문서에서 관련 문서가 모두 명시 되어 있으면 베이지안 분류기는 더 많은 문서를 학습 대상으로 삼은 것과 큰 차이가 없는 성능을 보였다. 아래 그림2와 3은 베이지안 이진 분류기의 정확률과 재현율을 나타낸다.

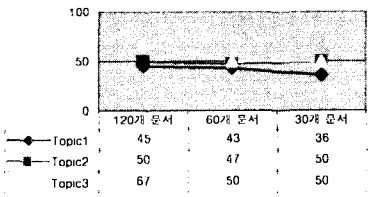

그림2 베이지안 이진 분류기의 정확률

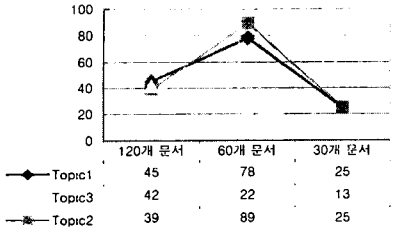

그림3 베이지안 이진 분류기의 재현율

4.2 사용자가 선택한 문서 수에 따른 분류기의 성능

사용자가 해당 문서 집합에서 모든 문서를 선택하지 않는다고 가정하자. 그러한 상황에서 학습 문서 중에 포함된 실제 관련 문서의 어느 정도까지 사용자에 의해 관련성 있다고 체크되어야 신뢰할 만한 결과를 보여주는지에 대해 실험하였다. 학습 문서에서 토픽과 유관하다고 표기된 문서를 차례로 빠뜨려가며 실험한 결과는 아래의 그림4와 같다. 그림4에서 보는 것과 같이 해당 문서 집합에서, 60% 정도까지 선택 비율이 줄어도, 크게 차이가 없는 성능을 보이고 있다. 그림4, 5는 관련성 표기가 완전하지 않은 경우 분류기의 정확률과 재현율을 나타낸다.

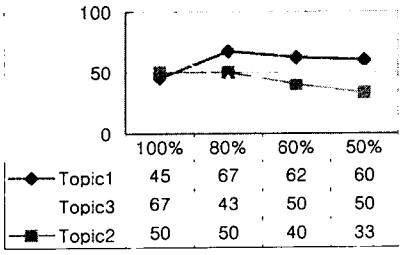

그림4 관련성 표기가 완전하지 않은 경우별 실험에서 분류기의 정확율

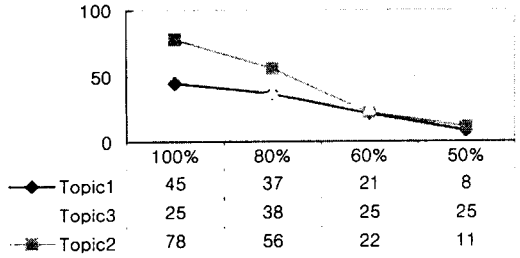

그림5 관련성 표기가 완전하지 않은 경우 실험에서 분류기의 재현율

5. 결론 및 향후 연구과제

본 연구는 베이지안 이진 분류기에서 학습 문서의 수에 따른 분류기의 성능과, 관련성 있는 문서가 완전히 명기 되지 않을 때에 분류기의 성능에 대하여 실험하고, 그 결과를 기술하였다. 신문기사에 필터링을 위한 입력을 받을 때, 최소한 60여개 이상의 문서를 보여주고 그 집합에서 원하는 토픽의 문서를 선택하게 할 필요가 있는 것으로 보인다. 학습 문서의 비율이 높으면 높을 수록, 관심 주제의 문서 발생 비율이 낮아도, 신뢰성 있는 필터링을 보였다. 60여개 기사(평균 160여절 정도) 정도에서는 관심 토픽이 10% 이하로 발생할 경우에는 좋은 필터링을 보이지 못했다. 문서 집합에 대해 10% 이하의 발생 빈도를 지니는 토픽에서는, 60개의 문서에서 학습된 어휘들로는 필터링에 유의미한 값을 보이는 어휘가 새로운 문서 중에 10-15개 이하로 매우 적은 수였다.

이진 필터링 상황에서, 문서 집합의 크기가 일정한 정도일 때 관련성 있는 문서가 모두 표기되지 않아도 필터링은 큰 차이가 없이 수행 되었다. 일반 독자들에게는 60여개의 기사문을 모두 읽고 관련 토픽을 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.

향후에, 불완전한 학습문서로 인한 베이지안 분류기의 범주 발생확률(카테고리확률)의 오류를 추론하는 방법을 살펴 볼 필요가 있다. 또 실용적인 신문기사 필터링을 위해, 필터링 결과에 대한 사용자의 반응으로부터 부가적으로 분류기의 학습을 개선해 나가는 적절한 피드백 방법을 연구할 필요가 있다.

[참고문헌]

[1] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E, " A Bayesian Approach to Filtering Junk E-Mail" , In Learning for Text Categorization: Papers from the 1998 Workshop. AAAI Technical Report WS-98-05, 1998

[2] 김호성, 정경호, 황도삼," 단어 가중치를 이용한 스팸메일 필터링" , 제13회 한글 및 한국어 정보처리 학술대회, 2001

[3] E. Voorhees, D. Harman," Overview of the Eighth Text Retrieval Conference (TREC-8)" , 1999

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