동적인 문서 여과에서 나이브 베이즈 분류기와 코사인 유사 계수의 성능 비교

손기준° 임수연 박성배 이상조
경북대학교 컴퓨터공학과
kijunson@msn.com°, nadalsy@hotmail.com, {sjlee, seongbae}@knu.ac.kr

Ki-Jun Son°, Soo-Yeoun Lim, Seong-Bae Park, Sang-Jo Lee

Comparative Between Naive Bayes Classifier and Cosine Similarity Coefficient in Dynamic Document Filtering

Dept. of Computer Engineering, Kyungpook National University

요 약

온라인 정보가 증가함에 따라 많은 양의 정보 중에서 사용자가 원하는 정보를 정확하고 신속하게 찾아주는 문서 여과의 중요성 또한 증가하고 있는 추세이다. 본 논문은 문서 여과 문제를 이진 문서 분류 문제로 보고, 나이브 베이즈 분류기를 동적인 문서 여과 목적으로 사용하였다. 이때 사용자가 자신의 관심 분야에 해당하는 주제를 제대로 여과 받기 위해서 학습 대상으로 삼아야 할 학습문서의 범위와 관련성 있는 문서를 제대로 여과 받기 위해서 체크해야 하는 관련성 표기 비율에 따른 분류기의 성능에 대하여 실험을 하였다. 코사인 유사 계수를 이용한 여과 방법과의 성능도 비교 실험하였다. 실험 결과 나이브 베이즈 이진 분류기는 문서집합의 크기가 일정한 정도일 때 관련성 있는 문서가 모두 표기되지 않더라도 여과에는 큰 영향을 미치지 않음을 볼 수 있었다.

1. 서 론

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

최근 문서 여과에 대한 연구는 인터넷의 유즈넷 뉴스, 전자메일, 웹을 대상으로 진행되어져 왔으며, 다양한 문서 여과 방법이 사용되어지고 있다. 일반적으로 사용되는 분류 알고리즘 중 많이 사용되는 것은 베이즈 분류기, 결정트리, k-NN 등이 있다[2].

본 연구에서는 신문기사 여과를 위해 간단하면서 잘 알려진 전통적인 분류방법으로 문서분류에서 좋은 성능을 보이고 있는 베이즈 분류기를 이용하여 신문기사 여과 시스템을 구현한다. 하지만 웹상의 신문기사 여과 서비스와 같은 실용적인 용도에 기존의 문서분류에서 널리 사용되고 있는 베이즈 분류기를 그대로 적용하는 데는 문제가 있다.

기존에 연구된 방법들은 1년치 혹은 그 이상의 학습문서를 대상으로 학습을 수행하며, 또한 학습대상 문서에 대한 관련성표기가 완전한 문서집합을 사용하여 분류기를 학습시키고 있다[2][3]. 하지만 웹상의 신문기사 여과와 같은 영역에서는 일반 사용자가 기사문을 모두 읽고 관련토픽을 오류 없이 모두 명시 하는 것은 쉽지 않다.

또한 학습대상문서 중 사용자가 관련성 표기를 하지 않은 문서가 비관련 문서집합에 포함되어 있을 수도 있기 때문에, 기본적으로 불완전한 학습문서가 된다. 즉, 신문기사 여과 문제는 불완전한 학습문서들을 대상으로 얼마나 만족할만한 여과 결과를 내는가의 문제로 살필 수 있다.

본 연구에서는 신문기사문에 대하여 여과를 적용하기 위해, 여과 문제를 변형된 문서분류의 문제로 보고 베이즈 이진 분류기를 여과 목적으로 사용할 때 어느 정도의 조건이 갖추어지면 좋은 여과를 행해줄 수 있는지에 대한 연구를 수행 하고자 한다.

본 논문은 다음과 같이 구성되어있다. 2장에서는 내용기반 여과 기법을 이용한 문서 여과에 대하여 설명한다. 그리고 3장에서는 실험 방법 및 결과를 상술하며, 4장에서는 지금까지의 결과를 요약하고 향후 연구 과제를 제시한다.

2. 나이브 베이즈 분류기와 코사인 유사 계수를 이용한 문서 여과

2.1 나이브 베이즈 분류기

나이브 베이즈 분류기는 베이즈 정리에 기초하고 특성들 간의 독립성을 가정한 확률적인 모델이다. 매우 단순하지만 잘 알려진 전통적인 분류방법으로, 텍스트 문서분류에 사용되어 왔다[4][5]. 나이브 베이즈 분류기는 통계적인 알고리즘으로 학습문서의 통계 정보를 학습하고, 이렇게 얻은 통계정보를 이용하여 입력 문서 스트림으로부터 문서를 분류한다.

확률이론을 기계학습에 적용한 것으로, 특정 데이타집합 D를 조사했을 때 가설 h가 사실일 확률은 $P(h|D)$가 된다. 그리고 가설이 사실일 경우 데이타 D의 확률이 $P(D|h)$일때 베이즈 정리는 식 (1)과 같다.

$$P(h|D) = P(D|h)P(h)/P(D) \quad (1)$$

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

최대 확률을 구하기 위해서는 최대사후확률(MAP)을 계산하면 된다. 이 확률은 데이타를 조사했을 때 가장 가능성이 높은 가정으로서 식 (2)을 이용한다.

$$h = \arg\max_{h \in H} P(D|h)P(h) \quad (2)$$

식 (2)에서 나타내는 바와 같이 베이즈 학습방법은 가설집합 H에 포함된 가설 중 가장 큰 확률을 가지는 h를 찾아 최종 가설로 설정하는

것이다.

학습문서집합과 새로운 사례가 $(a_1,a_2,...,a_n)$ 과 같이 속성값들의 벡터로 주어지면, 학습기는 이 사례에 대한 목적함수의 값 혹은 분류를 예측할 수 있다. 새로운 사례에 대한 분류를 예측하는 방법은 주어진 속성벡터 $(a_1,a_2,...,a_n)$에 대응되는 가장 가능성이 높은 목적함수의 값 $vMAP$를 다음 식(3)과 같이 구하는 것이다.

$$vMAP=\arg\max_{v_j\in V} P(v_j\, a_1,a_2,\ldots,a_n,D) \quad (3)$$

식 (3)에 베이즈 정리를 적용하면 식 (4)과 같다.

$$vMAP=\arg\max_{v_j\in V} P(a_1,a_2,\ldots,a_n\, v_j,D)P(v_j D) \quad (4)$$

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

2.2 코사인 유사 계수

벡터 공간 모델은 정보 검색에 기반을 두고 있는 기법으로 모든 색인어는 서로 독립이라는 가정을 하며 문서를 $d_i=(w_{i1},w_{i2},...,w_{in})$로 표현된다. d_i는 문서를 표현하며, w_{ij}는 d_i문서에서의 색인어의 가중치 값이다. 문서의 벡터가 형성된 이후 여과 과정은 벡터의 연산에 의해 이루어진다. 문서 d_1을 $d_1=(w_{11},w_{12},...,w_{1n})$로 표현되고, d_2는 $d_2=(w_{21},w_{22},...,w_{2n})$로 표현되었을 때, 두 문서 사이의 벡터 유사도 측정은 두 벡터 d_1과 d_2사이의 상관도로 구할 수 있다. 이는 두 벡터 간 사이 각의 코사인 값으로 계산될 수 있다[6]. 이는 식 (5)과 같다.

$$sim(d_1,d_2)=\frac{\vec{d_1}\cdot\vec{d_2}}{|\vec{d_1}|\cdot|\vec{d_2}|} \quad (5)$$

문서들 사이의 유사도 측정에 있어서 문서 값은 가중치에 의해 결정되기 때문에 가중치 부여 기법은 $tf\cdot idf$방법을 사용하였다.

2.3 신문기사 여과

문서 여과란 해당 문서집합으로부터 사용자가 필요로 하는 문서를 여과하는 것을 말한다. 본 연구에서는 여과 문제를 변형된 이진 문서 분류 문제로 파악한다. 즉 사용자가 필요로 하는 문서와 그렇지 않은 문서로 분류하는 이진 문서분류 문제로 볼 수 있다. 이에 따라 정보 여과 장치로서 여과 기술을, 신문기사 여과에 대해서 적용 하고자 한다. 신문기사는 동적 정보 문서의 면모를 충분히 지니고 있기 때문에 여과의 대상문서로 적합하다.

동적인 정보 여과 장치로서 나이브 베이즈 이진 분류기를 바로 여과 시스템으로 사용하기에는 몇 가지 문제가 있다. 이는 기본적으로 학습문서가 완전하지 않기 때문에 발생하는 문제들이다. 첫째, 학습대상이 되는 문서의 크기 자체가 충분하지 못하다는 점이다. 하루 혹은 그 이상 분량의 신문기사들을 브라우징하던 사용자가 자신의 관심 대상에 따라 몇 개의 문서들을 선택하는 모델을 고려해보자. 만약 학습의 대상을 사용자가 브라우징하고 있던 기사나 그날 발생한 모든 기사로 두더라도, 충분히 많은 양의학습 문서라고 볼 수는 없다.

둘째, 사용자가 자신의 관심대상 문서를 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 관련성 표기를 해주었다고 볼 수 없다. 따라서 대상 문서들이 많으면 많을수록, 사용자에게 전체 학습대상 문서에 대하여 관련성 표기를 요구하는 것은 실용적인 차원에서 어려운 일이다. 이러한 작업은 일반 사용자에게 큰 부담을 주게 되며, 특히 신문 기사문의 동적 여과와 같은 영역에서 실용적이지 않다.

따라서 학습 대상이 되는 문서 중 사용자가 관련성 표기를 하지 않은 문서가 비관련 문서집합에 포함되어 있을 수 있다. 그러므로 신문기사 여과는 기본적으로 불완전한 학습문서를 대상으로 하게 된다. 결론적으로, 신문기사의 여과문제는 이와 같이 불완전한 학습문서들을 가지고 얼마나 만족할 만한 여과 결과를 내는가의 문제로 살필 수 있다.

3. 실험 및 분석

3.1 실험 및 평가

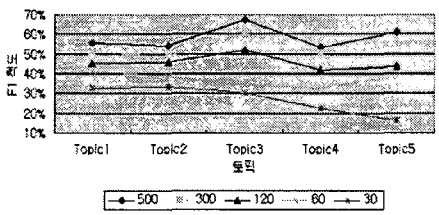
실험대상 문서는 중앙일보 신문기사를 사용한다. 학습문서는 신문의 일면 기사들로 제한하고, 15일간의 신문기사 500건을 사용하였다. 실험대상 문서는 학습 문서를 모두 수집한 다음부터 발생한 7일 간의 표제 기사 353개의 문서를 사용한다. 실험은 5개의 토픽 교육, 경제, 테러, 기업, 환경 관련이 표기된 문서집합을 사용하며 실시간으로 발생한 신문기사를 연속적으로 모은 일정양의 기사를 사용하였다.

문서 분류 시스템의 성능을 평가하는 방법으로는 주로 재현율, 정확률, F1 측정식이 사용되며[6], 본 연구에서는 F1 척도를 사용한다. F1척도는 정확률과 재현율에 동등한 중요도를 부여하는 하나의 평가 방법으로 식 (6)과 같다.

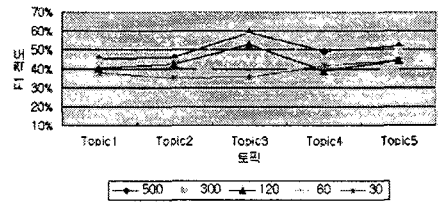
$$F=\frac{2\cdot Precision\cdot recall}{Precision+recall} \quad (6)$$

3.2 학습문서의 수에 따른 성능

본 실험은 나이브 베이즈 이진 분류기(NBC)가 제한된 학습문서집합에서 학습문서의 수에 따라 어느 정도의 성능을 내는지를 살펴보고자 한다. 다섯 개의 토픽이 표기된 학습문서에 대하여 학습문서의 수를 변경하며 NBC와 코사인 유사 계수를 이용한 내용기반 여과 기법(C_V)의 성능을 실험한 결과는 그림 1과 같다.



(a) 나이브 베이즈 이진 분류기(NBC)



(b) 코사인 유사 계수(C_V)

그림 1. 학습문서의 수에 따른 성능 비교

C_V 는 학습문서의 수에 따라 큰 성능의 차이를 보이지 않는다. 하지만 NBC는 학습문서의 수가 증감함에 따라 성능이 향상되는 것을 볼 수 있다. 하지만 초기 학습문서의 수가 60개 이하의 경우 C_V 보다 낮은 성능을 보이고 있다. 이러한 이유는 첫째, 학습 문서의 부족이다. NBC는 학습문서의 수가 60개 이하로 줄어들면 분류식의 값에

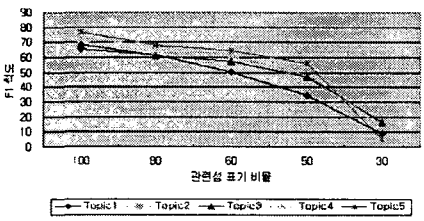
유의미한 영향을 미치는 어휘의 숫자가 평균 10~15개 이하로 떨어지기 때문에 만족할 만한 여과 결과를 성능을 보이지 않았다.

둘째 사용자가 관심을 가지는 정보가 비관련 문서집합으로 잘못 분류 되어진 것을 들 수 있다. 또한, 그림 1을 보면 토픽 3이 다른 토픽들 보다 나은 성능을 보이고 있다. 그 이유는 토픽 3에 등장한 어휘들이 다른 토픽에 나타난 어휘들보다 분별력이 높기 때문이다. 즉, 그 토픽을 분별할 수 있는 분별력이 높은 어휘가 다른 토픽에는 자주 사용되지 않기 때문으로 보인다. 토픽 2,4는 학습문서의 수에 따라 큰 성능의 차이를 보이지 않고 있다. 이러한 이유는 다른 여러 문서들과 분별력을 높일 수 있는 특정 어휘가 사용되지 않기 때문이다.

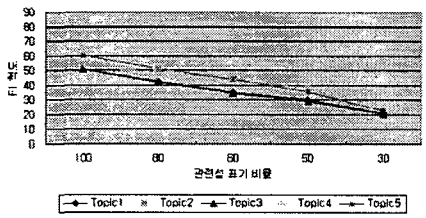
C_V는 문헌의 내용을 기반으로 여과를 한다. 여과 대상 문헌의 내용 분석이 필수적으로 이루어져야 하며, 여과를 위해 이용된 주요 키워드가 낮은 출현 빈도를 갖거나 너무 일반적인 성격의 키워드가 많이 포함된다면 성능이 떨어질 수밖에 없다. 하지만 학습문서의 수에 따라 크게 성능의 변화를 보이지 않는 이러한 성질은 여과 과정의 초기 평가 문제의 해결에 도움이 될 수 있을 것이다.

3.3 사용자의 관련성 표기비율에 따른 성능 비교

본 실험에서는 사용자가 해당 문서집합에서 실제로 여과 대상 주제와 관련 있는 문서를 어느 정도 비율로 선택을 하였을 경우 만족할만한 여과 성능을 내는지에 대한 실험 결과는 그림 2와 같다.



(a) 나이브 베이즈 이진 분류기(NBC)



(b) 코사인 유사 계수(C_V)

그림 2. 관련성 표기 비율에 따른 성능 비교

사용자의 관련성 표기 비율에 따라 다르지만, NBC는 관련성 표기 비율은 높을수록 좋은 성능을 보이고 있으며, 해당문서집합에서 관련성 표기비율이 80%정도까지 줄어들어도 분류기의 성능에는 큰 차이를 보이지 않는다. 하지만 관련성 표기비율이 50%이하로 떨어지면 성능이 현저히 저하되는 것을 볼 수 있다. 즉 관련성표기비율에 영향을 받는다. 토픽 3은 관련성표기 비율에 따라 성능의 차이를 보이지만, 토픽 2와 5는 관련성 표기 비율이 60%이하로 떨어져도 큰 성능의 차이를 보이지 않는다. 즉, 일반적인 주제를 다루는 토픽의 경우는 관련성 표기 비율이 60%이상 일 때 많은 성능의 변화를 보이지 않았지만, 특정 주제와 관련된 토픽(Topic5)은 관련성 표기비율에 따라 성능의 차이를 보임을 알 수 있었다.

C_V은 관련성 표기 비율에 따라 큰 성능의 차이를 보이지 않음을 알 수 있다. 토픽 5는 다른 토픽에 비해 나은 성능을 보인다. 이런 결과는 C_V는 키워드를 통해 유사한 문서를 추천하기 때문에 특정 주제의 문서에는 그 문서를 대표하는 특정적인 단어가 많이 나오기 때문이다.

4. 결 론

본 논문은 나이브 베이즈 이진 분류기를 웹상의 신문기사 여과에 적용하기위한 연구로 학습문서의 수, 관련성 표기 비율에 따른 성능의 변화를 살펴보고, 코사인 유사 계수를 이용한 내용기반 여과 기법과의 성능도 비교 하였다.

실험한 결과, 나이브 베이즈 이진 분류기는 학습문서의 비율이 높으면 높을수록, 관심 주제의 문서 발생 비율이 낮더라도 여과의 성능은 향상을 보였다. 이진 여과 상황에서 나이브 베이즈 이진 분류기는 문서집합의 크기가 일정한 정도일 때 관련성 있는 문서가 모두 표기되지 않더라도 여과에는 큰 영향을 미치지 않음을 볼 수 있었다. 이러한 특징은, 웹상의 신문기사 여과 서비스와 같은 실용적 용도의 여과를 위해 바람직한 것으로 보인다.

코사인 유사 계수를 이용한 내용기반 여과 기법은 학습 문서의 수에 따라 크게 성능의 변화를 보이지 않고 있다. 이러한 특성은 여과 과정의 초기 평가 문제의 해결에 도움이 될 수 있을 것이다.

향후 연구 과제로는 실용적인 신문기사 여과를 위해, 여과 결과에 대한 사용자의 반응으로부터 분류기의 성능을 개선해 나가기 위한 적절한 피드백방법 및 분류기의 결합 방법에 대한 연구가 필요하다.

참고문헌

[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] Resnick, P., Iacovou, N., Suchak, M.,Bergstorm, P. and Ridsl,J., "GroupLens: An open architecture for collaborative filtering of netnews," proc of ACM Conf. on Computer-Supported Cooperative Work, pp.175-186, 1994.

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

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

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

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

[7] 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.