지지 벡터 기계와 토픽 시그너처를 이용한 댓글 분류 시스템

언어에 독립적인 댓글 분류 시스템

Comments Classification System using Support Vector Machines and Topic Signature

배민영, Minyoung Bae*, 은지현, Jihyun En**, 장두성, Dusung Jang***, 차정원, Jeong-Won Cha****

요약 댓글은 일반적인 글에 비해 작성가능한 문장의 길이가 짧고, 띄어쓰기나 마침표를 잘 쓰지 않는 등 비정형화된 형식 구조를 가진다. 이러한 댓글의 악성 여부를 판별하기 위하여 본 논문에서는 문장을 n-gram으로 나누고 문서요약이나 문서분류에서 자질 선택에 많이 사용되는 토픽 시그너처(Topic Signature)를 이용하여 자질을 추출한다. 또한 지지 벡터 기계(Support Vector Machines)을 사용하여 댓글의 악성 여부를 판별한다. 본 논문에서는 한글과 영어 댓글에 대한 악성 여부를 판별하는 실험을 통하여 복잡한 전처리과정을 요구하는 기존에 제안된 방법들 보다 우수한 성능을 보이는 것을 확인할 수 있었다.

Abstract Comments are short and not use spacing words or comma more than general document. We convert the 7-gram into 3-gram and select key features using topic signature. Topic signature is widely used for selecting features in document classification and summarization. We use the SVM(Support Vector Machines) as a classifier. From the result of experiments, we can see that the proposed method is outstanding over the previous methods. The proposed system can also apply to other languages.

해심어: comment classification, machine learning, topic signature, support vector machines, n-gram

*주저자 : 창원대학교 컴퓨터공학과 e-mail: nikismy@changwon.ac.kr
**공동저자 : KT 음성언어연구부 HCI연구담당 e-mail: jh06@kt.com
***공동저자 : KT 음성언어연구부 HCI연구담당 e-mail: dschang@kt.com
****교신저자 : 창원대학교 컴퓨터공학과 교수 e-mail: jcha@changwon.ac.kr

1. 서론

인터넷의 활성화에 따라 사용자간의 의견을 자유롭게 주고 받을 수 있는 댓글 문화가 활성화 되었다. 그러나 댓글의 익명성을 악용한 무분별한 악플은 개인에서 사회적 문제로 대두되고 있다. 방송통신위원회의 '인터넷 권리 침해 현황' 에 따르면 인터넷을 통한 권리 침해가 2006년 1,595건에서 2007년 1만 2,959건으로 급증했고 2008년 9월까지 5,842건을 기록한 것으로 나타났다[1]. 또한, 2007년 시행된 '제한적 본인 확인제' 도입 후 악플 감소는 2% 정도에 불과한 것으로 나타났다. 2007년 MIT Spam Conference[2]에서 상당수의 주제가 스팸 이었으며, '선플 달기 운동본부' [3]가 출범하는 등 악플에 관한 관심이 꾸준히 증가하고 있으며, 악플 차단을 위한 다양한 방법[4,5]이 제시되었다.

악플은 작성가능한 문장의 길이가 제한적이고 악플 등록을 위한 시스템 자제의 등록 방지를 피하기 위해 정형화된 문장 구조를 가지지 않는 등 일반적인 글과는 다른 몇 가지 특징을 가진다[6]. 따라서 본 논문에서는 악플과 비악플 모두를 학습하여 중요한 단어가 낮은 출현빈도에도 자질로 추출 가능한 토픽 시그너처를 사용하여 자질을 선택한다. 문서 분류는 지지 벡터 기계를 사용한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 연구에 대한 조사를 하고 3장에서는 제안하는 시스템에 대해 설명한다. 4장에서는 기존의 실험과의 비교 실험을 통하여 시스템의 성능을 분석한다. 마지막으로 5장에서는 본 시스템에 대한 결론 및 향후 연구 과제를 제시한다.

↲

2. 관련연구

인터넷 상의 주관적이고 감정적인 표현이 많은 문서에 대해 지지 벡터 기계와 같은 기계학습을 이용한 연구가 활발하다[7]. 감정 분류(Sentiment Classification)의 경우 국외에서 더 많은 연구가 이루어지고 있으며, 정확성 향상을 위하여 단어에서, 문서로 자질에 대한 범주를 확대한 연구가 이루어지고 있다[8-11]. 국내의 경우 등록된 댓글의 악성여부를 판별하기 위하여 지지 벡터 기계를 이용하는 방법[12]과 역 카이 제곱 통계량을 이용하는 방법[13]이 있다. 그러나 댓글은 일반적인 글과는 다른 몇 가지 특징을 가지므로 품사태거 혹은 명사추출기를 이용하는 기존의 방법에서는 자질 추출에 있어 발생하는 오류가 문서 분류에까지 전파될 수 있다.

3. 댓글 분류 시스템

본 시스템의 이전 논문에서는 한글과 영어 댓글의 악성여부를 판별하기 위하여 한글 음절(character) 단위와 영어 단어(word) 단위 n-gram에 대해 베이지안(Naive Bayes) 분류기를 이용한 실험이 이루어졌다[14]. 본 논문에서는 지지 벡터 기계를 이용한 실험을 수행하였다.

3.1 토픽 시그너처(Topic Signature)

짧은 문장 길이의 댓글에서 반복되는 자질을 추출하기에는 많은 어려움이 존재한다. 따라서 학습하고 하는 각 문서 집합 내 출현하는 모든 단어에 대해 학습 가능한 토픽 시그너처를 이용하여 자질을 추출한다. Chin-yew Lin[15]에 의해 제안된 Log-likelihood Ratio 기반의 토픽 시그너처는 단어 추출(Term Extraction) 방법을 사용한다. [표 1]의 테이블에서 토픽 시그너처는 식 (1)과 같다.

표 1. 토픽 시그너처의 Contingency 테이블

	악성댓글	비악성댓글
t	V_{11}	V_{12}
~t	V_{21}	V_{23}

$$TS_s(t) = 2 \times (v_{11} + v_{12} + v_{21} + v_{22}) \times \left(\frac{v_{11}}{(v_{11} + v_{21}) \times (v_{11} + v_{12})} \right) \quad (1)$$

여기서 $TS_S(t)$는 단어 t가 악성댓글에 속할 경우 토픽 시그너처 값이며, 현재 문서집합에서의 출현빈도가 높고 반대 문서집합에서의 출현빈도가 낮을수록 상위에 위치한다. 최종적으로 순위화된 단어의 리스트를 이용해서 자주 나타나지 않은 단어(하위 순위)에 대해 평탄화(smoothing) 작업을 거친 후 자질로 선택되어 진다.

3.2 libSVM

지지 벡터 기계는 두 집단간의 분류 경계를 구하는 이진 분류기로 분류에 관련된 다양한 연구 분야에서 널리 응용되어 높은 성능을 보이고 있다. libSVM은 R.E. Fan, P.H. Chen, C.J. Lin에 의해 개발된 다중 집합 분류기(multi-class classifier)이다[16]. libSVM은 train과 predict 프로세스로 구성되어 있으며, 본 논문에서는 제공되는 옵션 중 svm_type은 C-SVC를, kerner_type은 linver를 이용하였다.

↲

4. 실험 및 토의

4.1 실험 데이터 및 평가 방법

본 논문에서는 한글과 영어 댓글에 대한 악성여부를 판별하기 위하여 한글은 정치 뉴스 분야 기사의 댓글을 다른 기간 동안 무작위로 수집하였으며, 영어는 비교 실험을 위하여 [13]과 [14]에서 사용된 문서집합을 이용하였다. 비교 대상은 [14]의 결과를 바탕으로 한다. 시스템 평가를 위한 문서집합은 [표 2]와 같으며, 성능 평가의 방법으로는 정확도(Precision), 재현율(Recall), F_1-measure를 이용하였다.

표 2. 학습과 테스트에 사용된 문서집합

	한글 데이터				영어 데이터	
	일반 댓글	악성댓글			일반댓글	악성댓글
		문서량	구간수	단어수		
학습데이터	1,300	1,300	2,216	14,801	10,000	19,586
평가데이터	170	130	-	-	329	612
총 댓글 수	1,470문서 / 1,430문서				10,329 /	20,198

$$P(\mathrm{Precision}) : \frac{\text{악플로 분류된 실제 악플 수}}{\text{댓글 중 악플로 분류된 총 수}} \quad (2)$$

$$R(Recall) : \frac{\text{악플로 분류된 실제 악플 수}}{\text{전체 악플 수}} \quad (3)$$

$$F_1(F_1-measure) : \frac{2 \times Recall \times \mathrm{Precision}}{Recall + \mathrm{Precision}} \quad (4)$$

4.2 실험 결과

[표 3]은 한글은 음절 n-gram으로 자질로 추출하고 토픽 시그너처와 카이 제곱 통계량(Chi-Square)과 베이지언 모델을 이용하여 출현 빈도에 따른 댓글의 악성여부 판별 실험 결과이다. [표 4]는 영어는 단어 n-gram에 대한 [14]의 실험 결과이다.

표 3. 한글 댓글에 대한 베이지안 모델 분류기와 토픽 시그너처/ 카이 제곱 통계량을 이용한 성능 비교 실험. 한글 음절 구간 7-gram, 음절 자질 3-gram 이용. f는 출현빈도를 의미

평가(%) / Model		P	R	F1
Topic Signature	f=10	81.4570	94.6154	**87.5445**
Chi-square	f=9	80.6667	93.0769	86.4286

표 4. 영어 댓글에 대한 베이지안 모델 분류기와 토픽 시그너처/ 카이 제곱 통계량을 이용한 성능 비교 실험. n은 단어의 수를 의미

Model / 평가(%)		P	R	F1
Topic Signature	n=2	76.8448	93.3539	**84.2987**
Chi-square	n=4	79.7160	60.7419	68.9474

[표 3]의 한글 음절 7gram - 3gram 에 대한 실험 결과에서 확인할 수 있듯이 출현빈도가 높은 자질만을 이용한 경우 나타나는 모든 3-gram을 자질로 이용한 경우보다 높은 성능을 나타내었다. 영어의 경우 공백 단위로 문장을 분류하고 단어 단위 n-gram의 자질을 이용하므로 unigram이나 5-gram등 n의 값을 변경해 실험해 보았으나 토픽 시그너처의 경우 bigram에 대한 실험이 가장 높은 성능을 나타내었고, 카이 제곱 통계량의 경우 4-gram에서 가장 높은 성능을 나타내었다.

[표 5]와 [표 6]은 한글과 영어 댓글 문서에 대해 [표 3]과 [표 4]의 자질을 이용하여 지지 벡터 기계의 입력 데이터 생성 시 상위 N개의 자질만을 사용한 경우에 대한 실험 결과이다. 한글의 경우 음절 n-gram의 출현 빈도에 따라 상위 N개의 자질을 다시 추출하여 실험을 수행하였다.

표 5. 한글 댓글에 대한 토픽 시그너처와 카이 제곱 통계량을 이용한 자질 추출 후 지지 벡터 기계로 문서를 분류한 성능 비교 실험. N은 자질 중 상위 N개, f는 음절 n-gram의 출현빈도 f개 이상을 의미

Model / 평가(%)		P	R	F1
Topic Signature	f=10, N=7500	92.86	100.00	96.30
	f=10, N=8000	92.20	100.00	95.94
	f=10, N=8500	95.59	100.00	**97.74**
	f=10, N=9000	92.20	100.00	95.94
	f=10, N=9500	91.55	100.00	95.59
Chi-square	f=1, N=1000	73.99	98.46	**84.49**
	f=1, N=1500	71.91	98.46	83.12
	f=1, N=2000	66.15	97.69	78.88
	f=1, N=2500	69.23	34.62	46.15

표 6. 영어 댓글에 대한 토픽 시그너처와 카이 제곱 통계량을 이용한 자질 추출 후 지지 벡터 기계로 문서를 분류한 성능 비교 실험. n은 단어의 수, N은 자질 중 상위 N개를 의미

Model / 평가(%)		P	R	F1
Topic Signature	n=4, N=9000	100.00	60.18	75.61
	n=4, N=9500	100.00	60.79	75.14
	n=4, N=10000	100.00	61.40	**76.08**
	n=4, N=10500	100.00	61.40	76.08
Chi-square	n=1, N=7000	33.85	99.70	50.54
	n=1, N=7500	33.71	100.00	50.42
	n=1, N=8000	33.81	100.00	50.54
	n=1, N=8500	33.81	100.00	50.54
	n=1, N=9000	33.81	100.00	50.54
	n=2, N=1000	100.00	19.45	32.57
	n=2, N=1500	33.81	100.00	**50.54**
	n=2, N=2000	33.74	100.00	50.46
	n=2, N=2500	33.74	100.00	50.46

n-gram의 출현빈도 별 상위 N개의 n-gram을 자질로 이용하여 각 실험을 수행하였다. N은 500개씩 증가시키며 실험하였다. [표 5]와 [표 6]에서 보이는 것과 같이 1차적으로 선택된 자질을 모두 사용하는 것 보다 상위 N개를 다시 선택하여 사용하는 것이 더 좋은 성능을 나타내었다.

한글과 영어 모두 토픽 시그너처를 이용하여 자질을 추출한 경우 가장 높은 성능을 나타냈다.

5. 결론 및 향후 연구과제

사회적, 법적 제제에도 날로 증가하는 악플은 사회적 문제로 그 심각성이 증가하고 있다. 이러한 악플을 판별하기 위한 다양한 연구와 시스템이 개발되었으나 일반적인 글과는 다른 특징을 가지는 댓글의 악성여부를 판별하기란 쉽지 않은 실정이다.

본 논문에서는 악플의 특징을 이용하여, n-gram으로 자질을 생성하고 댓글의 악성여부를 판별할 수 있다는 것을 보였다. n-gram의 방식은 특정 언어에 상관없이 적용 가능하며, 한글의 경우 본 논문에서 분류기로 사용한 지지 벡터 기계가 기존의 베이지안 모델보다 약 10% 정도의 성능 향상을 보이는 것을 [표 5]를 통해 확인할 수 있었다. 그러나 영어의 경우 오히려 베이지안 모델을 이용하는 경우 더 높은 성능을 보였다. 이는 성능 평가를 위해 사용된 영어 댓글 문서의 특징상 반복되는 사이트 주소나 특정 단어가 많아 학습과 평가에 사용된 문서에서 나타날 수 있는 동일 자질이 적어 지지 벡터 기계용 코퍼스 생성 시 자질이 존재하지 않는 경우가 발생하기 때문으로 보여 진다.

또한, 자질 추출에 있어서도 단어의 출현 빈도에 따라 자질을 선택해 보았는데 출현 빈도가 높을수록 댓글의 악성여부를 판별하는데 더 좋은 성능을 보임을 확인할 수 있었다. 또한 지지 벡터 기계를 이용한 분류 과정에서도 상위 N개의 자질만을 재선택 함으로써 불필요한 자질이 제거됨에 따라 성능이 향상되는 것을 확인할 수 있었다.

본 논문에서 제안하는 방법은 자질 선택을 위해 한글의 경우 2단계 과정을 거쳐야함에 따라 속도상의 문제가 제기될 수 있다. 만약 특정 구간에 대한 악성 여부 판별 전처리 과정이 존재한다면 더 빠른 속도로 댓글의 악성여부를 판별할 수 있을 것으로 생각된다.

역 카이 제곱이나 TF-IDF등을 이용하고 최대 엔트로피(Maximum Entropy) 등을 이용한 실험은 차후 과제로 남겨둔다.

참고문헌

[1] 방송통신위원회 '인터넷 권리 침해 현황'
http://www.kcc.go.kr/

[2] MIT Spam Conference 2007
http://www.spamconference.org/

[3] 선플 달기 운동 본부
http://www.sunfull.or.kr/index.php

[4] Movable Type Black Filter, with content filtering
http://www.jayallen.org/projects/mt-blacklist/

[5] Mishne G., D. Carmel, "Blocking Blog Spam with Language Model Disagreement". 1st International Workshop on Adversarial Information Retrieval on the Web. pp. 1-6, 2005.

[6] 배민영, 차정원, "Topic Signature를 이용한 댓글 분류 시스템", 한국정보과학회 2008 종합학술대회 논문집, 제35권, 제1호(A), pp. 81-82, 2008.

[7] P.D. Tueney, M.L. Littman, "Unsupervised Learning of Semantic Orientation from a Hundred-billion-word Corpus", National Research Council, Institute for Information Technology, ERP-1094, NCR-44929, 2002.

[8] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", EMNLP, pp. 79-86, 2002.

[9] Soo-Min Kim, Eduard Hovy, "Automatic Detection of Opinion Bearing Words and Sentences", IJCNLP, pp. 61-66, 2005.

[10] Soo-Min Kim, Eduard Hovy, "Determining the Sentiment of Opinions", COLING, pp. 1367-1373, 2004.

[11] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, "A Sentimental Education: Sentiment Analysis using Subjectivity Summarization Based on Minimum Cuts", ACL, pp. 271-278, 2004.

[12] 김묘실, 강승식, "SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현", 한글 및 한국어 정보처리 학술대회, 18th, pp. 285-289, 2006.

[13] 전희원, 임해창, "본문과 덧글의 동시출현 자질을 이용한 역 카이 제곱 기반 블로그 덧글 스팸 필터 시스템", 한글 및 한국어 정보처리 학술대회, 19th, pp. 122-127, 2007.

[14] 배민영, 차정원, "Topic signature와 n-gram을 이용한 댓글 분류 시스템", 한글 및 한국어 정보처리 학술대회, 20th, pp. 188-193, 2008.

[15] Chin-Yew Lin, Eduard Hovy, "The Automated Acquisition of Topic Signatures for Text Summarization", COLING, 18th, pp. 495-500, 2000.

[16] R.E. Fan, P.H. Chen, C.J. Lin, "Working set selection using second order information for training SVM", Journal of Machine Learning Research 6, pp. 1889-1918, 2005.