지지벡터기계를 이용한 스팸 블로그(Splog) 판별 시스템

이성욱*

A Splog Detection System Using Support Vector Systems

Songwook Lee*

요 약

블로그는 인터넷 공간에서 가장 손쉽게 정보 출간, 토론 참여, 커뮤니티 형성하는 수단이다. 그러나 최근에 광고를 유치하거나 페이지 순위를 올리기 위한 목적의 다양한 스팸 블로그가 범람하고 있다. 본 연구의 목적은 웹 환경에서 이러한 스팸 블로그(Splog)를 자동으로 판별하는 시스템을 개발하는 것이다. 먼저 블로그의 HTML을 제거한 후 품사를 부착하였다. 어휘/품사 쌍을 자질로 사용하였으며 카이제곱 통계량을 이용하여 유용한 자질을 선택하였다. 선택된 자질의 가중치를 벡터로 표현한 후, 지지벡터기계(Support Vector Machines)를 학습하여 자동으로 스팸 블로그를 판별하는 시스템을 제안하였으며, SPLOG 데이터 집합으로 실험한 결과 F1척도로 90.5%의 정확률을 얻었다.

ABSTRACT

Blogs are an easy way to publish information, engage in discussions, and form communities on the Internet. Recently, there are several varieties of spam blog whose purpose is to host ads or raise the PageRank of target sites. Our purpose is to develope the system which detects these spam blogs (splogs) automatically among blogs on Web environment. After removing HTML of blogs, they are tagged by part of speech(POS) tagger. Words and their POS tags information is used as a feature type. Among features, we select useful features with $\chi 2$ statistics and train the SVM with the selected features. Our system acquired 90.5% of F1 measure with SPLOG data set.

키워드

스팸 블로그 판별, 스플로그, 지지벡터기계, 카이제곱 통계량

Key word

spam blog detection, Splog, support vector machines, chi square statistics

* 정회원 : 충주대학교 (leesw@cjnu.ac.kr)

접수일자 : 2010. 10. 19
심사완료일자 : 2010. 11. 02

Ⅰ. 서 론

블로그는 하나의 웹페이지에 날짜의 역순으로 여러 개의 엔트리로 구성되어 있으며 인터넷의 발전과 웹의 활성화로 온라인상에는 많은 개인 블로그가 운영되고 있다. 특히 개인이 운영하는 블로그에는 여행, 요리, 제품 사용기, 일기, 신문기사, 정치적 견해 등 다양하고 유용한 정보가 존재한다. 블로그가 활성화되자 기존의 웹 검색엔진과 더불어 블로그 전용 검색 엔진이 개발되었다. 이러한 검색 엔진은 블로그 사이트 인식 및 그 구조 파악, 연관된 메타데이타를 추출하여 그 내용을 색인하는데 특화되어있고, 강인한 스팸 블로그 필터 기능이 필요하다. 그러나 이러한 블로그내에는 기업의 광고성 게시물과 개인이 작성한 게시물이 혼재하고 있어, 많은 게시물 중에 유용한 정보를 담고 있는 게시물을 찾는데 어려움을 겪고 있다. 또한 스팸 블로그는 검색엔진 등의 부하를 가중시키며 사용자 만족도를 저하시키는 원인이 되고 있다.

스팸 블로그는 두 가지 목적에 의해 주로 생성된다. 첫째, 아무 내용이 없는 가짜 블로그를 만들거나 다른 블로그나 뉴스 소스로부터 컨텐츠를 도용하여 가짜 블로그를 만든 후, 여러 상업 광고를 유치하는 것이다. 둘째, 좀더 교묘한 방법으로 가짜 블로그를 만든 후 특정 사이트의 페이지 순위를 높이기 위해 링크 팜(link farm)[10]으로 만드는 것이다. 스팸 블로그의 문제점은 블로그 검색 엔진 및 분석 엔진과 여러 연구 등에 의해 보고 되어 왔으며 스플로그(Splog)라는 신조어를 만들어 냈다[11, 12, 13].

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 살펴보며, 3장에서는 제안 시스템의 구조를 살펴본다. 4장에서는 지지벡터기계의 학습에 사용된 자질과 좋은 자질을 선택하는 방법을 설명하며, 5장에서는 지지벡터기계에 대해 소개하며 어떻게 제안 시스템에 적용하였는지 설명한다. 6장에서는 실험을 통해 제안된 방법의 성능을 보이고 7장에서 결론을 내린다.

Ⅱ. 관련 연구

이러한 스플로그 차단에 관한 연구는 [1]에서 처음 제안되었고, 스팸 메일 필터 연구와 더불어 연구된 [2]의 연구가 있다. [1]과 [2]는 모두 지지벡터기계(Support Vector Machine)를 사용하여 스플로그를 판별하였다.

[1]에서는 블로그내에 존재하는 단어열(Bag-of-words)을 기본 자질로 이용하였고, 앵커(anchor) 태그에 나타나는 텍스트와 제목, 메타 데이터 등을 자질로 사용하였다. 추가로 앵커열(Bag-of-anchors), URL열(Bag-of-urls)과 문자 n-gram열을 자질로 사용하였다. 앵커열 자질은 한 페이지가 포함하고 있는 앵커 텍스트에 나타나는 모든 텍스트를 자질로 사용한 것이고, URL열 자질은 URL에서 "/"등의 기호 문자로 잘라 토큰화한 문자열들을 자질로 사용한 것이다. 문자 ngram열(Bag-of-ngram)은 주로 다중 언어(multi-lingual) 데이터에서 사용하는 자질인데, 문자 길이 "n"을 윈도우 사이즈로 사용하여 인접한 문자열을 추출한 것이다. 예를 들어, "support"라는 단어에 대한 4-gram열은 "supp, uppo, ppor, pport"이 된다. 자질 가중치로는 TF(text frequency)와 이진 가중치를 사용하였고 여러 자질 중 유용한 자질의 선택은 상호정보량(Mutual Information)을 이용하였다.

[2]에서는 단어열(Bag-of-words)과 문자 ngram열을 자질로 사용하였는데 문자 ngram열 자질에 공백을 포함하여 사용한 것이 [1]에서 사용한 자질과 다른 점이다. 자질의 가중치는 이진 가중치를 사용하였으나 자질 선택은 따로 하지 않았고 블로그에 나타난 모든 메타 데이터도 자질에 포함된다.

Ⅲ. 시스템 구조

스플로그를 판별하는 문제의 경우 스팸 메일과 마찬가지로 정상 블로그와 스플로그를 구분하는 이진 분류의 성격을 가지고 있으므로 본 연구에서도 이진 분류기 중에서 가장 성능이 좋다고 알려진 지지벡터기계를 이용하였으며[3,4], 본 연구는 [14]의 연구결과를 확장한 것이다. 기계 학습에서 적절한 자질의 선택은 시스템의 성능에 많은 영향을 끼친다. 자질로는 어휘와 품사 정보를 사용하였고, 그 중에서 카이제곱 통계량을 이용하여 유용한 자질을 선택한다. 다음 그림 1은 제안 시스템의 구조도이다. 제안하는 스플로그 판별 시스템은 크게 두 단계로 나뉜다. 먼저 학습단계에서는 학습용 블로그 데이터로부터 지지벡터기계의 학습에 사용할 수 있는 벡터 자질을 추출하여야 한다.

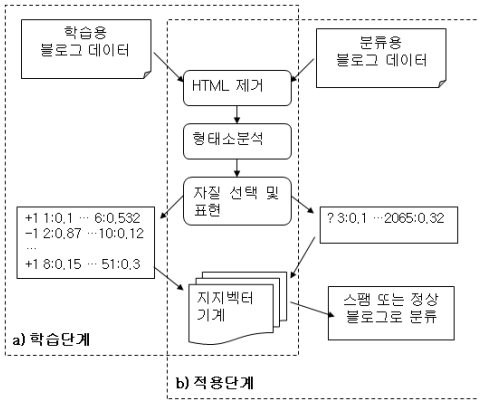

그림 1. 제안 시스템 구조도.
Fig. 1 The system architecture.

학습용 블로그 데이터는 먼저 HTML을 제거한 후, 형태소 분석기에 의해 품사가 부착된다. 우리는 자질로 어휘/품사 쌍을 사용하는데 그 중 유용한 자질들을 카이제곱 통계량을 이용하여 선택한다. 선택된 각각의 자질은 벡터의 차원을 구성하는 축을 이루며 각 자질의 가중치가 해당 차원의 값이 된다. 이렇게 하나의 블로그 데이터는 다차원 공간의 한 점이 되고, 모든 학습 데이터의 벡터가 모두 구성되면 이들로 지지벡터기계를 학습한다. 적용단계에서는 학습 때와 마찬가지로 분류용 블로그 데이터는 형태소 분석 단계와 자질 추출 단계를 거쳐 다차원 공간상의 한 점을 이루는 벡터가 되고 이를 학습된 지지벡터기계가 스팸 또는 정상 블로그로 분류한다.

Ⅳ. 자질과 카이 제곱 통계량

본 연구에서는 SPLOG[1] 데이터 집합을 실험에 이용한다. 블로그 데이터는 HTML을 제거한 후, 형태소분석기로 Montytagger[6]를 사용하여 자동으로 품사를 부착하였다. 다음 그림 2는 블로그 원문과 HTML이 제거된 후, 품사 부착 후의 블로그 데이터의 예를 나타낸다.

기계학습을 위한 자질로는 그림 2의 c)에서와 같은 어휘/품사 쌍을 자질로 사용하였다. 따라서 가능한 자질의 종류는 학습 데이터에서 발견되는 모든 어휘/품사 쌍이 되어 많은 수의 자질을 가지게 된다. 이러한 자질들 중에서는 스플로그를 결정하는 데 유용한 자질이 있는가 하면 그렇지 않은 경우나 오히려 방해가 되는 자질들도 존재한다. 우리는 좋은 자질을 선택하기 위해 카이 제곱 통계량을 이용하며, 카이 제곱 통계량을 계산하는 식은 다음과 같다[8].

$$\chi^2(f,s) = \frac{(A+B+C+D)\times(AD-BC)^2}{(A+B)\times(A+C)\times(B+D)\times(C+D)} \quad (1)$$

A는 스플로그 s 중에 자질 f를 포함하고 있는 문서의 수이고, B는 범주 s 이외의 문서, 즉 정상 블로그 범주에 속해 있는 문서 중에 자질 f를 포함하고 있는 문서의 수이다. 또한, C는 스플로그 s에 속한 문서 중에 자질 f를 포함하지 않는 문서의 수이며, D는 범주 s 이외의 문서 중에 자질 f를 가지고 있지 않는 문서의 수이다.

```
<p class="post-date">Sun 28 Aug 2005</p>
<div class="post-info"><h2 class="post-title"><a
href="http://www.ashwink.net/blog/2005/08/28/unified-feeds/" rel="bookmark" title="Permanent Link: unified feeds">unified feeds</a></h2>
    under <a
href="http://www.ashwink.net/blog/category/general/"
title="View all posts in General" rel="category tag">General</a><br/><a
href="http://www.ashwink.net/blog/2005/08/28/unified-feeds/#comments">No Comments</a> </div>
    <div class="post-content">
            <p>Finally, an integrated feed for my blog, delicious bookmarks and flickr photos. Update your link everyone. <a
href="http://feeds.feedburner.com/bhootakannadi">Feed Link</a></p>
```

(a)

```
Sun 28 Aug 2005
unified feeds
under GeneralNo Comments
Finally, an integrated feed for my blog, delicious bookmarks and flickr photos. Update your link everyone. Feed Link
```

(b)

```
Sun/NNP 28/CD Aug/VB 2005/CD
unified/JJ feeds/NNS
under/IN GeneralNo/NNP Comments/NNS
Finally/RB ,/, an/DT integrated/VBN feed/NN for/IN
my/PRP$ blog/NN ,/, delicious/JJ bookmarks/NNS
and/CC flickr/NN photos/NNS ./. Update/NNP your/PRP$
link/NN everyone/NN ./. Feed/VB Link/NNP
```

(c)

그림 2. 품사 부착 전후의 블로그.
(a) 원문 (b) HTML 제거 후 (c) 품사 부착 후
Fig. 2 Blog data and it's preprocessing.
(a) text (b) HTML removed text (c) Part of Speech tagged text

자질 f와 범주 s가 완전히 독립적이면 0의 값을 갖는다. 어떤 자질에 대한 카이제곱 통계량의 값을 결정하는 방법은 전체 범주에 대한 평균값을 사용하는 방법과 전체 범주에 대해 최대값을 사용하는 방법 등이 있을 수 있다. 우리는 카이제곱 통계량을 이진 분류에 사용하므로 각 자질당 하나의 값만 사용하면 된다.

각각의 자질에 가중치를 부여하는 방법은 여러 가지가 있는데, 본 연구에서는 일반적으로 문서 검색에 가장 많이 사용되는 TF-IDF(Term Frequency-Inverse Document Frequency) 가중치 와 지지벡터기계의 학습에 좋은 성능을 보이고 있는 이중 가중치를 사용하여 실험한다[1,4]. 본 시스템에서 TF-IDF 값을 계산하는 경우, 용어(term)는 자질로, 문서(document)는 블로그로 범주(category)는 스플로그와 정상 블로그로 간주하여 계산한다.

Ⅴ. 지지벡터 기계

지지벡터기계는 두 개의 범주를 구분하는 문제를 해결하기 위해 1995년에 Vapnik[5]에 의해 소개된 학습기법으로 그림 3과 같이 초월공간(hyper-space)에서 두 개의 클래스의 구성 데이터들을 가장 잘 분리해 낼 수 있는 결정면(decision surface)을 찾는 모델이다.

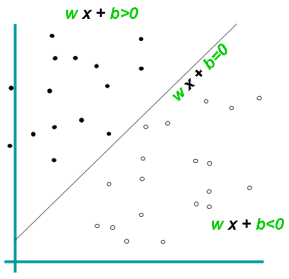

그림 3. 초월 공간에서의 결정면[9].
Fig. 3 Decision plane on hyperspace.

선형 분리가 가능한 공간에서의 결정면은 초월면(hyper-plane) H : y=w·x + b = 0 이며 이 초월면에 평행하고 동일 거리에 있는 두 개의 초월면은 아래 식의 H1,H2와 같으며, H1와 H2사이에 어떠한 데이터 포인트도 존재하지 않는 조건을 만족시키며 H1와 H2사이의 거리는 최대가 된다.

$$H1 : y=w\cdot x + b = +1,$$
$$H2 : y=w\cdot x + b = -1.$$

H1와 H2 사이의 거리를 최대로 만드는 것이 지지벡터 기계의 학습 목적이 된다. 따라서 H1에는 양의 값을 갖는 데이터가 존재하게 되고 H2에는 음의 값을 갖는 데이터가 존재하게 되는데, 이러한 데이터들을 지지벡터라 부르며 이들이 분리 경계면을 결정하는 역할을 한다. 다른 데이터들은 H1와 H2를 교차하지 않도록 분리 경계면 주위로 이동되거나 제거된다. H1와 H2 사이의 거리 M을 최대로 하기 위해서 H1와 H2 사이에 어떠한 데이터 포인트도 존재하지 않도록 하면서 ‖w‖을 최소화시키면 된다.

$$w\cdot x + b \geq +1 \text{ for } yi=+1,$$
$$w\cdot x + b \leq -1 \text{ for } yi=-1.$$

지지벡터기계의 문제는 이러한 w와 b를 찾아내는 문제이며, 이것은 2차 프로그래밍(quadratic programming) 기술에 의해 풀 수 있다[5].

이 지지벡터기계를 우리는 스플로그 판별에 사용하였다. 지지벡터기계는 이진 분류기이므로 우리는 스플로그와 정상 블로그를 분류하기 위해 한 개의 모델만 학습하면 된다. 스플로그인 경우 양(+1)의 자질을 정상 블로그인 경우 음(-1)의 자질을 부여하였다. 지지벡터기계의 학습을 위한 자질은 4장에서 설명한 어휘/품사 쌍의 자질들의 가중치로 벡터를 구성하였다. 본 연구에서는 지지벡터기계의 학습을 위해 LIBSVM[7]을 이용하였고 선형 커널을 이용하여 학습하였다.

Ⅵ. 실험 및 결과

본 연구에서 사용한 SPLOG[1] 데이타는 694개의 SPLOG와 695개의 정상 블로그로 구성되어 있으며. 9:1의 확률로 랜덤추출하여 1,248개의 데이터를 학습 데이터로, 141개의 데이터를 평가 데이터로 사용한다. 제안 시스템의 성능은 정확도(Accuracy), 정확률(Precision)과 재현율(Recall) 등으로 평가하였다.

표1은 제안 시스템의 자질의 개수를 카이제곱 값으로

제한했을 때의 정확도를 나타내며 자질 가중치로 이진 가중치를 사용하였다.

표 1. 자질의 개수에 따른 정확도 비교
Table. 1 Experimental Results of # of features

자질의 개수	정확도(%)	비고
205,691	78.0	모든 자질
13,525	83.7	χ^2>3
9,094	86.5	χ^2>4
7,764	90.1	χ^2>4.5
6,821	87.2	χ^2>5

표 1에서 보는 바와 같이 모든 자질 205,691개를 사용했을 때보다 카이제곱 통계량을 이용하여 7,764개로 제한했을 때 시스템의 정확도가 약 12.1% 더 향상되었다. 이는 카이제곱 통계량이 유용한 자질의 선택에 큰 도움이 되는 것을 의미한다.

다음 표 2는 이진 가중치와 TF-IDF 가중치를 사용했을 때의 성능을 비교한 것이다.

표 2. 가중치 표현에 따른 정확도
Table. 2 Accuracy for the type of feature weight

가중치	정확도(%)	비고
이진	90.1%	χ^2>4.5
TF-IDF	85.8%	

표 2에서와 같이 자질 가중치를 TF-IDF를 사용했을 때보다 이진 가중치를 사용했을 때 제안 시스템이 더 좋은 성능을 보였다.

다음 표 3은 제안 시스템을 동일한 실험 데이터를 사용한 다른 시스템[1,2]과 비교한 것이다.

표 3. 다른 시스템과의 성능 비교
Table. 3 Comparison of other systems.

시스템	P(%)	R(%)	F1(%)	비고
본시스템	90.0	90.9	90.5	어휘/품사
Kolari[1]	89.3	86.9	88.1	어휘+urls
OnSVM[2]	92.8	87.6	90.1	어휘+4gram

표 3에서와 같이 제안 시스템은 다른 SVM을 이용한 시스템보다 나은 성능을 보였는데 비교한 두 시스템의 성능은 교차 검증(n-fold cross validation)의 평균값을 나타낸 것으로 평가데이터와 학습데이터를 랜덤 추출한 제안 시스템의 성능과 직접적인 비교는 어렵다. 그러나 제안 시스템이 얻은 수치적인 성능은 비교 시스템들보다 약간 더 좋은 결과를 얻었다. Kolari[1] 시스템의 경우 어휘 정보와 블로그 페이지의 URL정보를 자질로 이용하였고, OnSVM[2] 시스템의 경우, 4gram 어휘 정보를 자질로 이용하였다. 반면에 제안 시스템은 비교적 간단한 어휘/품사 정보를 자질로 이용하였으나 좋은 성능을 얻은 원인은 카이제곱 통계량을 이용하여 유용한 자질을 선택한 데 있다고 할 수 있다. 시스템의 성능을 더욱 향상시키기 위해서는 어휘/품사의 n-gram과 URL정보를 자질로 추가해서 다양한 자질의 조합으로 실험할 필요가 있다.

표 4는 블로그 데이터의 HTML을 제거한 것과 제거하지 않은 경우의 실험결과를 나타낸다.

표 4. HTML 사용유무
Table. 4 Results of data with/without HTML tags.

시스템	P(%)	R(%)	F1(%)	비고
HTML제거	90.0	90.9	90.5	χ2>4.5
HTML사용	89.0	89.7	89.3	

표 4에서와 같이 본 실험에서는 HTML을 제거한 것이 사용한 것보다 약간 더 좋은 결과를 보였으나 큰 성능 차이는 없었다.

Ⅶ. 결론 및 향후과제

본 논문에서는 웹 환경에서의 스플로그를 판별하기 위해, 먼저 블로그의 HTML을 제거한 후 품사를 부착하였다. 품사가 부착된 후 어휘/품사 쌍을 자질로 사용하였으며 카이제곱 통계량을 이용하여 유용한 자질을 선택하였다. 선택된 자질을 이진 가중치로 표현한 후, 지지벡터기계를 학습하여 자동으로 스플로그를 판별하는 스팸 블로그 판별 시스템을 제안하였다.

실험에는 SPLOG 데이터 집합을 이용하였으며, 실

험 결과 F1 척도로 90.5%의 성능을 얻었다. 더 나은 성능을 위해서는 다른 연구에서 사용하는 블로그 페이지의 URL 자질과 n-gram 자질 등을 추가하는 연구가 필요하다. 실험 데이터가 영어권에서 구축된 데이터를 사용하였으므로 한국어를 사용하는 인터넷 환경에 적용할 수 있도록 한국어로 구성된 다양한 영역에서의 스팸 블로그 데이터 집합의 구축이 필요하다. 향후 스팸 블로그뿐만 아니라 스팸 및 욕설 덧글 등을 모두 판별할 수 있는 시스템을 개발하면 유용한 정보 검색 환경을 제공함과 동시에 건전한 인터넷 문화 확산에 크게 기여할 것이다.

참고문헌

[1] Kolari, P., Finin, T., Joshi, A., "SVMs for the Blogosphere: Blog Identification and Splog Detection", AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006.

[2] D. Sculley, Gabriel M. Wachman. "Relaxed online SVMs for spam filtering," Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp.415-422, 2007.

[3] 이성욱, "카이 제곱 통계량과 지지벡터기계를 이용한 자동 스팸 메일 분류기", 춘계 한국해양정보통신학회 논문집, 2009.

[4] 은종민, 이성욱, 서정연, "지지벡터기계(Support Vector Machines)를 이용한 한국어 화행분석", 정보처리학회논문지, Vol.12-B, No.3, pp.365-368, 2005.

[5] V. Vapnik. The nature of statistical learning theory, Springer, NewYork, 1995.

[6] http://web.media.mit.edu/~hugo/montylingua, 2009.

[7] http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2009.

[8] Yang, Yiming and Jan O. Pedersen. A comparative study on Feature selection in text categorization. In proceedings of the 14th International conference on Machine Learning, 1997.

[9] Martin Law. "A simple introduction to Support Vector Machines," PPT file, 2003.

[10] Wu, B., and Davison, B. D. Identifying link farm spam pages. In WWW ''05: Special interest tracks and posters of the 14th international conference on World Wide Web, 820 - 829. New York: ACM Press. 2005.

[11] Umbria. 2005. Spam in the blogosphere. [Online;http://www.umbrialistens.com/consumer/showWhitePaper].

[12] Cuban,M. 2005. A splog here, a splog there, pretty soon it ads up and we all lose. [Online; accessed 22-December-2005;http://www.blogmaverick.com/entry/1234000870054492/].

[13] Kolari, P.; Java, A.; and Finin, T. 2006. Characterizing the splogosphere. In WWW 2006, 3rd Annual Workshop on the Weblogggging Ecosystem: Aggregation, Analysis and Dynamics.

[14] 이성욱, "지지벡터기계와 카이 제곱 통계량을 이용한 스팸 블로그 판별 시스템", 춘계 한국해양정보통신학회 논문집, 2010.

저자소개

이성욱(Songwook Lee)

1996년 서강대학교 컴퓨터학과 학사
1998년 서강대학교 컴퓨터학과 석사
2003년 서강대학교 컴퓨터학과 공학박사
2004-2005년 LG전자 기술원 선임연구원
2005-2007년 동서대학교 컴퓨터정보공학부 전임강사
2007년-현재 국립충주대학교 컴퓨터과학과 조교수
※관심분야 : 인터넷응용시스템, 한국어정보처리