블로그 포스트의 내용 유사도 분석을 이용한 스팸 블로그 분류

최승진○ 김성권
중앙대학교 컴퓨터공학과
sjchoi@alg.cse.cau.ac.kr, skkim@cau.ac.kr

Spam Blog Classification using Contents Similarity Analysis of Blog Posts

Seung-Jin Choi○ Sung-Kwon Kim
School of Computer Science and Engineering, Chung-Ang University

1. 서론

블로그는 개인의 생각과 지식을 작성할 수 있는 웹 사이트이다. 최근 UGC(User Generated Contents)의 유행으로 블로그 또한 그만의 문화가 빠르게 정착, 발전되고 있다. 글 작성의 간편함과 코멘트(Comment)나 트랙백(Trackback)을 통해 다른 사용자들과 손쉽게 의견 교환을 할 수 있는 것도 블로그 발전의 이유 중의 하나이다. 블로그의 중요성이 점점 높아지면서 검색 엔진도 블로그 검색 기능을 도입하여 사용자가 작성한 정보들을 다른 사용자들에게 제공해 주고 있다. 하지만 블로그 상에서 점점 스팸 블로그로 인한 피해가 증가하고 있는 추세이다. 이러한 스팸 블로그로 인해 검색 엔진 입장에서는 불필요한 페이지 처리 비용이 증가하게 되고 검색 엔진 랭킹의 신뢰성이 저하될 수 있다. 그리고 사용자 입장에서는 스팸 블로그로 인해 불필요한 정보를 얻게 되고 사용자 간의 의사소통에 불신이 생길 수 있다. 위와 같은 문제점들을 해결하기 위해 스팸 블로그를 탐지하는 연구가 활발히 진행되고 있다. 주로 스팸 블로그가 가진 특성을 도출하여 스팸 블로그를 탐지하거나 많은 블로그들을 통계적 방법이나 기계 학습(SVM)을 통해 정상 블로그와 스팸 블로그를 분류하는 방법들이 쓰이고 있다.

본 논문은 스팸 블로그 포스트가 가지고 있는 내용 유사도를 이용하여 스팸 블로그를 분류하였다. 스팸 블로그 포스트는 기계를 이용하여 다량으로 작성되기 때문에 하나의 스팸 블로그에 있는 포스트들은 그것들의 내용이 매우 유사하다. 이 특성을 바탕으로 실험을 통해 정상 블로그와의 차이점을 도출하였다. 그리고 기존에 많이 사용되는 해외의 데이터 셋을 사용하지 않고 국내의 블로그 포스트들을 자체적으로 수집한 데이터 셋을 이용하여 실험을 진행하였다. 자체적으로 개발한 유사도 분석기를 통해 내용 유사도를 실험한 결과 정상 블로그와 스팸 블로그를 분류할 수 있는 의미 있는 차이점을 도출할 수 있었다.

2. 본론

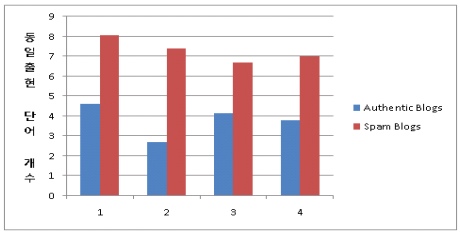
[그림 1] 하나의 스팸 블로그에서 내용이 유사한 포스트들

스팸 블로그의 포스트는 정당하지 않은 방법으로 원하는 페이지를 검색 엔진 결과의 상위에 랭크되게 하거나 사용자들에게 불필요한 광고를 보여주는 것이 목적이다. 이를 위해서 스패머들은 다량의 블로그 포스트를 작성하여 사용자와 검색 엔진에게 많은 노출을 시킨다. 짧은 시간에 많은 블로그 포스트를 작성하기 위해 수작업 대신 기계를 이용하게 된다. 그래서 스팸 블로그 포스트들 간의 내용은 유사한 모습을 보이게 된다. [그림 1]은 실제 스팸

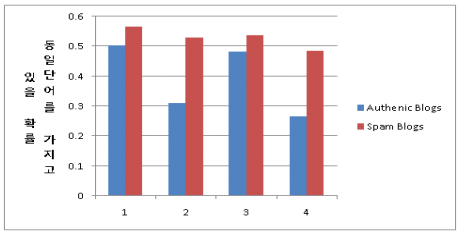
블로그 포스트의 모습을 보여주고 있다. 3개의 스팸 블로그 포스트는 몇 개의 단어가 달라진 것을 제외하고 제목과 본문의 내용이 거의 유사한 것을 볼 수 있다. 대부분의 스팸 블로그는 이와 같이 몇 개의 단어만 바뀐 포스트들을 가지고 있다. 정상 블로그 포스트는 다양한 주제를 가지고 수작업으로 작성되기 때문에 포스트의 내용이 서로 다르다. 따라서 하나의 블로그 내에 있는 포스트 내용을 비교함으로써 정상 블로그와 스팸 블로그의 결과의 차이를 얻을 수 있을 것이다. 그리고 효과적인 결과를 도출하기 위해 포스트 전처리 방법과 포스트의 비교 범위, 비교 방법을 다양화하여 분석하였다. 우선 포스트 전처리는 비교 범위가 설정된 포스트에서 HTML, Script 태그를 제거한 후, 형태소 분석기를 이용하여 가중치가 큰 상위 20개의 명사를 추출하는 방법과 빈도가 높은 상위 20개 명사를 추출하는 방법으로 나누었다. 그리고 비교 범위는 블로그 페이지 전체를 대상으로 전처리를 실행하는 방법과 본문만을 추출하여 전처리를 실행하는 방법으로 나누어 보았다. 마지막으로 각 포스트의 비교 방법은 각각의 포스트 쌍을 비교하는 방법과 전체 포스트를 한 번에 비교하는 방법으로 나누어 보았다. 이제 앞의 실험 방법에서 나올 수 있는 8가지의 경우를 각각 실험한다. 그리고 실험 결과 분석을 통해 어떤 방법이 스팸 블로그 유사도 특성을 효과적으로 도출하는지 살펴본다.

3. 실험 및 분석

실험을 위해 직접 구현한 유사도 분석기는 JDK 1.6.0_03 Eclipse 3.3 환경에서 실행되었다. 그리고 사용된 데이터 셋은 전체 61개의 블로그, 1000개의 페이지를 사용하였다. 여기에서 수작업을 통해 35개 블로그, 583개 페이지는 정상 블로그로 확인되었고 나머지 26개 블로그. 417개 페이지는 스팸 페이지로 확인되었다. 실험 결과 **본문 추출-가중치 단어 선택-포스트 쌍 비교 방법**을 거쳤을 때 정상 블로그와 스팸 블로그의 결과 값이 약 2.7배로 가장 큰 차이를 보여, 이 방법이 스팸 블로그의 내용 유사도를 분석하는데 가장 효율적인 방법으로 분석하였다.



[그림 2] 포스트 쌍 비교 방법 결과



[그림 3] 포스트 전체 비교 방법 결과

[표 1] 포스트 쌍 비교 방법 결과
(정상 블로그 / 스팸 블로그)

	가중치 단어	빈도 단어
페이지 전체 범위	4.601 / 8.067	4.161 / 6.704
본문 범위	2.714 / 7.386	3.778 / 7.016

[표 2] 포스트 전체 비교 방법 결과
(정상 블로그 / 스팸 블로그)

	가중치 단어	빈도 단어
페이지 전체 범위	0.503 / 0.565	0.483 / 0.537
본문 범위	0.309 / 0.529	0.267 / 0.486

4. 결론

본 논문에서는 스팸 블로그 포스트의 유사적 특성을 이용하여 정상 블로그와 스팸 블로그를 분류하는 방법을 제시하였다. 스팸 블로그 포스트는 기계를 통해 대량으로 작성되기 때문에 정상 블로그 포스트에 비해 그 내용이 매우 유사한 형태를 지니고 있다. 실험 결과 본 논문이 제시한 유사적 특성은 스팸 블로그를 효과적으로 분류할 수 있었다. 하지만 실험에 사용된 데이터 셋의 크기가 작아 실험 결과에 큰 신뢰성을 부여하기는 힘들었다. 또한 다양한 내용을 가지고 있는 스팸 블로그에 대해서는 분류가 제대로 되지 않는 문제가 발생하였다.

향후 연구에서는 블로그 포스트의 구조적 유사도를 이용하는 분류 방법을 제시하고자 한다. 스팸 블로그의 포스트 중에는 내용은 다르지만 포스트의 형태가 유사한 형태가 많이 존재한다. 이미지+텍스트, 비디오+텍스트 등의 포스트 구조를 파악하여 포스트의 유사도를 판단하고자 한다. 마지막으로 실험 결과의 신뢰성을 얻기 위해 많은 데이터 셋에 대하여 실험 후 기계 학습(SVM) 알고리즘과 같은 다른 분류 알고리즘과 비교, 평가해보는 것 또한 추후 연구 과제로 남겨둔다.