인터넷 기반의 스팸 댓글 추적/필터링 시스템

하헌성○ 조경래 마지웅
부산대학교
pooha302@gmail.com hacjkr@gmail.com ne4u7@nate.com

Internet based comment spam tracing/filtering system

Hunsung Ha○ Kyungrae Jo Jiung Ma
Pusan National University

요 약

인터넷에 게시된 글(블로그, 포털 사이트, 게시판 등)에 대한 댓글들 중에서 중복, 광고 등의 스팸성 댓글을 추적하고 이를 필터링 하는 시스템을 개발.

주제어: 스팸 댓글, 추적, 필터링, 인터넷

1. 서 론

전통적으로 인터넷을 이용하는 사람들의 커뮤니케이션 공간으로써 활용되어온 게시판이 생겨난 이후로 다양한 형태의 커뮤니케이션 방법이 나타났다. 하지만 이에 대한 부작용으로 불특정 다수에게 무분별하게 가해지는 스팸성 글들이 등장하게 되었고 이는 인터넷 자원을 낭비하는 동시에 사회적 문제로까지 확대되고 있는 실정이다. 본 연구에서는 인터넷에 게재된 글들 중에서 정상적인 글과 스팸성인 글을 구분하여 스팸으로 판단된 글을 삭제하는 시스템을 개발하는 것을 목표로 한다.

본 시스템에서는 각종 게시판, 블로그, 언론 사이트 등에 게시된 글에 대한 댓글들을 대상으로 한다. 스팸의 여러 가지 형태 중에서 특히 1)동일한 내용의 댓글이 여러 번 출현한 경우와 2)광고성 내용을 담고 있는 댓글을 스팸성 댓글로 정의한다. 이후로 스팸 댓글이라 함은 앞에서 언급한 두 가지 특성을 가지는 댓글을 의미한다.

본 논문에서는 먼저 스팸 댓글에 대한 분석을 통하여 스팸 댓글의 특징을 알아보고 알고리즘을 적용하기 위해서 스팸 댓글을 표준화하는 방법을 설명한다. 그리고 실제로 설계한 알고리즘에 대해서 알아보고 마지막으로 설계한 알고리즘을 이용해서 실제 데이터에 적용한 결과를 나타낼 것이다.

2. 실 험

2.1 스팸 댓글 분석

정상적인 댓글과 스팸 댓글을 구분하기 위해서 스팸 댓글의 특성을 살펴보아야 한다. 첫 번째 특징은 ★,▶, ♡와 같은 특수 문자를 사용하는 경우인데 정상적인 댓글에서는 거의 사용하지 않지만 스팸 댓글의 경우 사람들의 이목을 끌기 위해서 이러한 특수 문자를 과도하게 사용하는 경우가 빈번하다. 두 번째 특징은 댓글 내용 중에 특정 사이트를 홍보하기 위해서 그 사이트의 url을 표기하는 경우이다. 이것은 거의 모든 스팸 댓글에서 발견할 수 있는 특징이다. 세 번째 특징은 동일한 내용의 댓글이 반복해서 나타나는 경우를 의미한다. 마지막으로 스팸 댓글의 특징으로 특정한 단어를 사용한다는 것을 들 수 있다. 스팸 댓글의 목적에 따라 다르겠지만 예를 들어서 음란 사이트를 홍보하는 댓글이라면 내용 중에 <야동>이라는 단어가 들어갈 확률이 높으므로 <야동>이라는 단어가 존재하는 댓글에 대해서는 스팸 댓글이라고 판단할 수 있는 것이다.

표 1. 스팸 댓글 사례1

——SeK1004.CoM——▶▶▶.....
♡ 소녀시대 보도방 ♡
♡SeK1004.CoM♡
아찌 *o* 물 좀 줘잉
.. ┌┐ ┌┐.
┌┴┐ ├┼┐
├─┤ ├─┤ 멋쩌부러~
├─┤ ├─┤
└─┘ └─┘

표 2. 스팸 댓글 사례2

◆◆◆ X X K O . N E T ◆◆◆ 유혹 상상 이제 현실입니다. 분위기 좋은곳에 함께 가고싶어요 음악좋은 자동차안에도 좋구요 전망좋은 창가에서도좋구요 한적한야외에서도 좋아요 마지막 한방울까지 책임집니다◆◆◆ X X K O . N E T ◆◆◆

표 3. 스팸 댓글 사례3

★★ 8282sp.net ★★'주부부터 대학생까지 여성들이 만남원해요..♡

★★ 8282sp.net ★★'주부부터 대학생까지 여성들이 만남원해요..♡

★★ 8282sp.net ★★'주부부터 대학생까지 여성들이 만남원해요..♡

★★ 8282sp.net ★★'주부부터 대학생까지 여성들이 만남원해요..♡

★★ 8282sp.net ★★'주부부터 대학생까지 여성들이 만남원해요..♡

*.

최고에요~ ^0^

2.2 스팸 댓글 수집

실험을 위한 스팸 댓글은 각종 언론사 사이트에서 수집하였으며 아래는 수집한 사이트의 목록이다.

표 4. 댓글 수집 사이트 목록

이름	주소
조선일보	http://www.chosun.com/
동아일보	http://www.donga.com/
서울경제	http://economy.hankooki.com/
SBS TV	http://www.sbs.co.kr/
매일경제	http://www.mk.co.kr
국민일보	http://www.kukinews.com/
MBC TV	http://www.imbs.com
그 외, 다음 미디어(Daum media)와 네이버(Naver) 뉴스 캐스트에 등록된 사이트들.	

2.3 수집한 댓글에 대한 표준화 작업 수행

알고리즘 적용을 위해서 수집한 댓글을 변환하는 표준화 작업이 필요하다. 표준화 작업은 크게 4가지 단계로 진행한다. 각각의 단계는 1)특수문자 제거, 2)불필요한 공백 제거, 3)변형된 단어 표준화, 4)url 추출 이다.

먼저 1)단계는 불필요한 특수문자를 제거함으로써 순수 텍스트만 남겨서 알고리즘 적용을 좀 더 용이하도록 한다. 기본적으로 영문, 숫자, 한글을 제외한 모든 문자들을 제거하며, 점(.), 콜론(:), 슬래시(/)등의 경우는 url 판단을 위해서 제거하지 않는다.

2)단계는 댓글에 의미 없는 공백을 사용하는 경우에 대해서 이러한 공백을 제거하는 단계이다. 예를 들어서 "8 9 2 h a z a . n e t" 와 같이 각각의 문자가 공백으로 띄어져 있는 경우 이것이 url을 의미한다는 것을 기계적으로 판단하기 어렵기 때문에 사이의 공백을 제거하여 "892haza.net" 와 같은 형태로 만들어 줄 필요가 있다.

3)단계에서는 먼저 영문, 숫자를 표현하는데 있어서 특수문자를 사용하여 표현한 경우가 있는데 이러한 경우 각 문자에 해당하는 영문, 숫자로 모두 치환하여 주었다.

표 5. 특수 문자 목록

처리한 특수문자(아래 문자는 모두 특수문자이다.)
* ㅍ + 한자 : A ~ Z, a ~ z
* ㅈ + 한자 : 0 ~ 9
* ㅎ + 한자 : A B E Z H I KM N O P T Y o
* ㅇ + 한자 : ⓐ ~ ⓩ, ① ~ ⑮, ⒜ ~ ⒵, ⑴ ~ ⑮
* 기타 : ⓞ : 12903

두 번째로 한글로 url을 표현하는 경우가 있다. 예를 들면 "www.naver닷컴 "과 같이 영어를 한글로 표현한 경우가 있으며 이것을 모두 영문으로 변환하여 주었다.

표 6. 변형된 url 표기

*컴, 캄, 콤 => com
*넷 => net
*쩜, 닷 => '.'
*씨오 => co
*케이알 => kr

4)단계에서는 앞에서 표준화된 텍스트를 대상으로 정규식을 사용하여 url을 추출한다. 사용한 정규식은 다음과 같다.

표 7. url 판별 정규식

((https*|ftp|mailto)\:\/\/)?([0-9a-zA-Z][.0-9a-zA-Z-]+[ ]*\.)+[ ]*[a-zA-Z]{2,4}

2.4 스팸 단어 사전

스팸 댓글을 판별하는 중요한 기준으로 특정 단어의 출현 여부를 들 수 있다. 이를 위해서 스팸댓글들을 대상으로 자주 출현하는 단어들을 수집하였다. 수집한 단어들은 다음과 같다.

표 8. 스팸 단어 목록

보도방	유혹	한방울	대박	조건	만남
오빠	성인	은밀	즐섹	미소녀	글래머
킹카	미씨	처자	파트너	봉지	풀타임
서양녀	대기	여대생	섹스	오랄	가능
만원	콜	화끈	클릭	사절	비밀
여학생	채팅	나이	사진	프로필	아이디
아뒤	사이트	후불	가입	현금	무료
입금	다운	바다이야기			

표 9. 수집한 스팸 댓글 200개에서 스팸 단어가 출현한 횟수

보도방	유혹	한방울	대박	조건	만남
3	4	3	6	25	47
오빠	성인	은밀	즐섹	미소녀	글래머
24	28	20	1	0	1
킹카	미씨	처자	파트너	봉지	풀타임
2	1	11	11	6	4
서양녀	대기	여대생	섹스	오랄	가능
4	6	5	0	0	63
만원	콜	화끈	클릭	사절	비밀
31	26	25	3	11	11
여학생	채팅	나이	사진	프로필	아이디
14	2	16	16	12	14
아뒤	사이트	후불	가입	현금	무료
6	18	11	29	8	22
입금	다운	바다이야기			
4	4	4			

2.5 알고리즘

본 시스템에서 사용하는 알고리즘은 스팸 댓글의 특성을 파라미터로 하여, 파라미터의 값을 기준으로 일정 값 이상을 가질 경우 스팸 댓글이라고 판단하는 것이다. 알고리즘에서 사용하는 파라미터는 앞에서 스팸 댓글의 특징으로 뽑은 4가지를 이용할 것이다. 즉, 1)특수문자, 2)url, 3)반복출현, 4)스팸 단어 이상 4가지이다. 알고리즘을 실행할 때 입력단위는 하나의 댓글 뭉치이다. 댓글 뭉치라 함은 하나의 글에 달린 댓글들의 집합을 의미한다. 알고리즘의 출력은 입력된 댓글 뭉치에 대해서 각각의 댓글에 대한 판별 결과이다. 댓글 뭉치가 입력 단위가 되는 이유는 여러 댓글들 중에서 동일한 내용의 댓글이 출현하는 지를 알아보기 위해서이다.

파라미터의 경우 각각의 파라미터마다 가중치를 달리하였다. 예를 들어서 특수 문자의 경우는 정상적인 댓글에서도 이모티콘의 형태 등으로 사용될 가능성이 크다. 따라서 특수 문자 각각에 대해서 높은 가중치를 줄 경우 판단 결과가 정확하지 않을 것이다. 반대로 하나의 댓글 뭉치 내에서 같은 내용으로 여러 번 나타난 댓글은 스팸이라고 판단 할 수 있으므로 이 경우에는 해당 파라미터의 가중치를 높게 줄 수 있다. 이처럼 파라미터의 가중치를 실험을 통해서 적절한 값으로 설정해 주어야 한다.

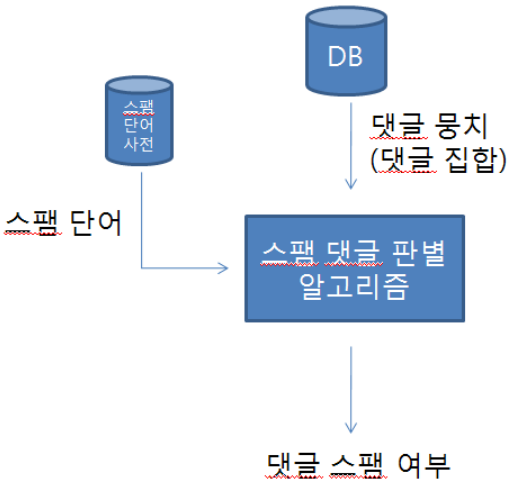

그림 1. 알고리즘의 입력과 출력

알고리즘에서 파라미터의 가중치를 지정하는데 있어서 중요한 기준은 정상적인 댓글을 스팸 댓글로 판정하여 삭제하는 경우이다. 스팸 댓글을 정상 댓글이라고 판별한 경우도 문제이기는 하지만 이것은 알고리즘의 개선을 통해서 수정하거나 직접 삭제할 수 있지만 앞에서 언급한 경우처럼 정상 댓글을 스팸이라고 판단하여 삭제한 경우에는 유효한 정보의 손실을 의미하므로 심각한 문제를 초래할 수 있다. 따라서 파라미터의 가중치를 설정하는 과정에서는 이처럼 정상 댓글을 스팸 댓글로 판단하는 경우를 줄여 줄 수 있는 방향으로 접근하도록 해야 한다.

알고리즘 내부는 다음과 같이 작동한다.

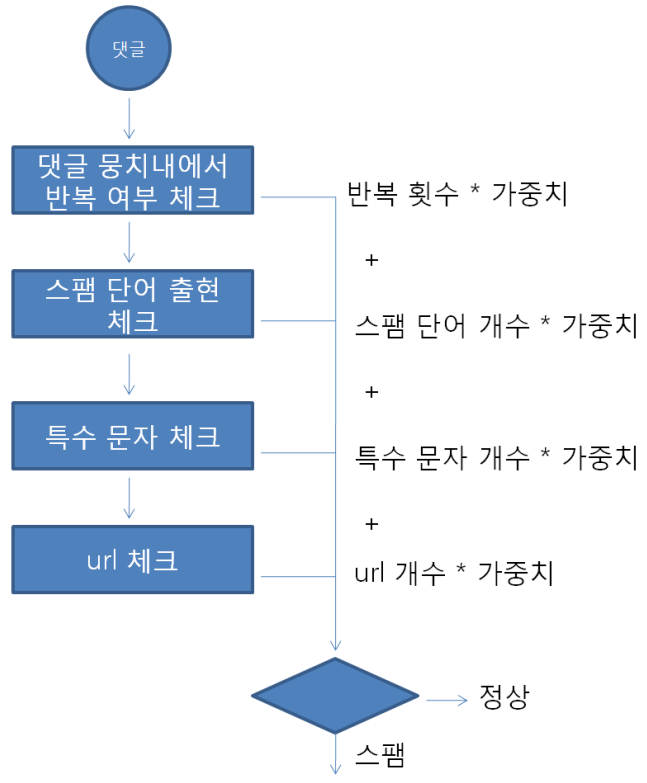

그림 2. 알고리즘 구조

3. 결 론

실험에서 사용한 데이터는 댓글 뭉치가 총 87개이며 포함된 댓글의 수는 모두 2911개이다. 전체 댓글 중에서 정상적인 댓글의 수는 1860개이며 스팸 댓글의 수는 1051개이다. 알고리즘을 적용하기 전에 전체 댓글에 대해서 각각 내용을 검토하여 직접 스팸 댓글과 정상 댓글로 구분하였다. 구분 기준은 처음에 언급했던 스팸 댓글의 정의에 따랐다. 파라미터에 사용된 가중치는 특수 문자에 대해서 1, url에 대해서는 3, 스팸 단어에 대해서는 5,마지막으로 반복 출현된 댓글에는 20을 설정하였다. 그리고 파라미터 값의 총 합이 20이 넘으면 스팸으로 판단하였다. 그 결과 수집한 모든 댓글에 대해서 실험한 결과 98.11%의 정확도를 얻을 수 있었다. 단순한 방법이지만 수집한 댓글에 대해서는 매우 우수한 결과를 얻을 수 있었다.

앞으로 개선할 사항은 스팸 단어 사전의 활용 부분이다. 현재는 이미 입력된 단어를 기반으로 파라미터 값을 구하고 있지만 추후에는 스팸 댓글로 판별된 댓글의 내용을 분석하여 자동적으로 빈번히 출현한 단어에 대해서는 사전에 등록시키고, 기존에 등록된 단어라도 정상 댓글과 스팸 댓글을 판단하는 기준으로 부적합하다고 여겨지면 사전에서 제외시킬 필요가 있다. 또한 이것은 동일한 의미의 여러 형태의 단어에 대해서도 필요한 과정이다. 예를 들어서 “아르바이트”에 대해서 “알바”가 많이 사용되다가 “아르바”라는 형태로 많이 사용된다면 사전에 등록되는 단어도 변경될 필요가 있을 것이다. 또한 단어마다 가중치를 다르게 둘 수도 있다. 이를 위해서 댓글의 내용을 분석하고 특정 단어의 출현 빈도를 측정하는 과정이 필요하다.