



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2018년10월30일
(11) 등록번호 10-1913284
(24) 등록일자 2018년10월24일

(51) 국제특허분류(Int. Cl.)
G06Q 50/30 (2012.01) G06Q 50/00 (2018.01)
(52) CPC특허분류
G06Q 50/30 (2013.01)
G06Q 50/01 (2013.01)
(21) 출원번호 10-2017-0162083
(22) 출원일자 2017년11월29일
심사청구일자 2017년11월29일
(56) 선행기술조사문헌
KR101482073 B1*
KR101764696 B1*
KR1020160079863 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
박정희
대전광역시 유성구 은구비남로 56, 908동 1803호
최혁준
대전광역시 유성구 궁동로18번길 17, 202호
(74) 대리인
홍성욱, 심경식

전체 청구항 수 : 총 11 항

심사관 : 송원선

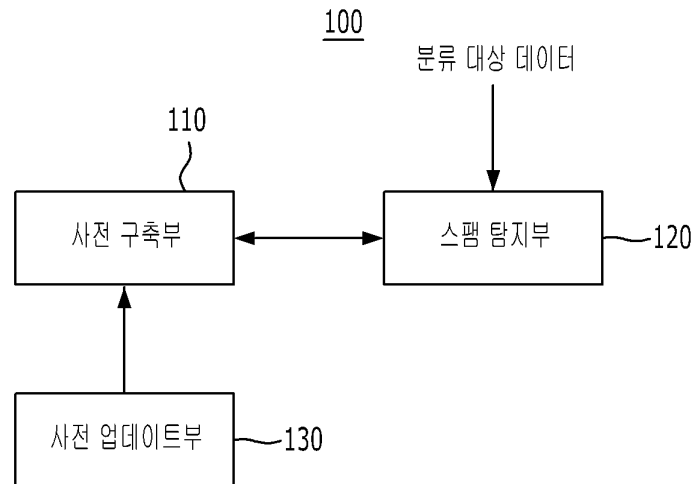
(54) 발명의 명칭 소셜 네트워크 서비스에서 스팸 탐지 방법 및 장치

(57) 요약

소셜 네트워크 서비스에서 스팸 탐지 방법 및 장치가 개시된다.

본 발명의 일 실시예에 따른 소셜 네트워크 서비스에서 스팸 탐지 방법은, 스팸 탐지 장치가 소셜 네트워크 서비스에서 스팸을 탐지하는 방법에 있어서, 트레이닝 데이터의 존재 여부에 따라 n-gram 또는 리트윗 증가율을 이용하여 사전어를 구축하는 단계, 소셜 네트워크 서비스에서 수집된 데이터에 대해 상기 구축된 사전어 기반으로 나이브 베이즈 분류기(Naive Bayes Classification)를 적용하여 스팸을 탐지하는 단계를 포함한다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 1345254481

부처명 교육부

연구관리전문기관 한국연구재단

연구사업명 이공학개인지초연구지원

연구과제명 스트리밍 데이터 마이닝을 위한 시스템 개발

기 여 율 1/1

주관기관 충남대학교

연구기간 2016.11.01 ~ 2017.10.31

공지예외적용 : 있음

명세서

청구범위

청구항 1

스팸 탐지 장치가 소셜 네트워크 서비스에서 스팸(spam)을 탐지하는 방법에 있어서,
 트레이닝 데이터의 존재 여부에 따라 n-gram 또는 리트윗 증가율을 이용하여 사전을 구축하는 단계;
 소셜 네트워크 서비스에서 수집된 데이터에 대해 상기 구축된 사전을 기반으로 나이브 베이스 분류기(Naive Bayes Classification)를 적용하여 스팸을 탐지하는 단계; 및
 상기 사전을 업데이트하는 단계를 포함하되,
 상기 사전을 업데이트하는 단계는,
 일정 시간동안 수집된 데이터에 대하여 리트윗 증가율을 산출하고, 상기 산출된 리트윗 증가율을 기준으로 상기 데이터들을 내림차순으로 정렬하는 단계;
 상기 정렬된 데이터들에 대해 상기 사전을 기반으로 나이브 베이스 분류기를 적용하여 상위 일정 개수 데이터 및 하위 일정 개수 데이터에 대한 스팸 또는 논스팸의 클래스 라벨을 예측하는 단계; 및
 상기 상위 일정 개수의 데이터 중에서 스팸으로 예측된 데이터들을 이용하여 스팸 사전을 업데이트하고, 하위 일정 개수의 데이터 중에서 논스팸으로 예측된 데이터들을 이용하여 논스팸 사전을 업데이트하는 단계를 포함하는 것을 특징으로 하는 소셜 네트워크 서비스에서 스팸 탐지 방법.

청구항 2

제1항에 있어서,
 상기 사전을 구축하는 단계는,
 상기 트레이닝 데이터가 존재하는 경우,
 각 트레이닝 데이터를 전처리하는 단계;
 상기 전처리된 각 트레이닝 데이터에 대해 n-gram의 문자열로 분해하는 단계; 및
 상기 n-gram으로 분해된 각 문자열의 빈도수를 산출하여, 스팸 사전 또는 논스팸(non-spam) 사전을 생성하는 단계를 포함하는 것을 특징으로 하는 소셜 네트워크 서비스에서 스팸 탐지 방법.

청구항 3

제1항에 있어서,
 상기 사전을 구축하는 단계는,
 상기 트레이닝 데이터가 없는 경우,
 일정 시간동안 수집된 데이터에 대하여 리트윗 증가율을 산출하는 단계;
 상기 산출된 리트윗 증가율을 기준으로 상기 데이터들을 내림차순으로 정렬하는 단계;
 상기 정렬된 데이터들에서 상위 일정 개수의 데이터를 스팸 트레이닝 데이터, 하위 일정 개수의 데이터를 논스팸 트레이닝 데이터로 추출하는 단계;
 상기 추출된 스팸 트레이닝 데이터와 논스팸 트레이닝 데이터를 각각 전처리하는 단계;
 상기 전처리된 각 트레이닝 데이터를 n-gram의 문자열로 분해하는 단계; 및
 상기 n-gram으로 분해된 각 문자열의 빈도수를 산출하여, 스팸 사전 또는 논스팸 사전을 생성하는 단계를 포함하는 것을 특징으로 하는 소셜 네트워크 서비스에서 스팸 탐지 방법.

청구항 4

제3항에 있어서,

상기 리트윗 증가율($IR(tid)$)은 아래 수학적식을 이용하여 산출하는 것을 특징으로 하는 소셜 네트워크 서비스에서 스팸 탐지 방법

[수학적식]

$$IR(tid) = \left(\sum_{i=1}^m IR_i(tid) \right) / (followers_{tid} * RC_m(tid))$$

여기서, tid는 원본 트윗에 대한 고유 식별정보, $IR_i(tid)$ 는 i시간 동안 발생한 리트윗 수, $followers_{tid}$ 는 원본 트윗을 작성한 사용자의 팔로워 수, $RC_m(tid)$ 는 해당 시점에 누적된 총 리트윗 수를 의미함.

청구항 5

제1항에 있어서,

상기 스팸을 탐지하는 단계는,

상기 수집된 데이터를 전처리하는 단계;

상기 전처리된 데이터를 n-gram의 문자열로 분해하는 단계;

상기 분해된 각 문자열에 대해 나이브 베이스 분류기를 적용하여 스팸 사전 또는 논스팸 사전에 존재할 확률을 산출하는 단계; 및

각 데이터별로 각 문자열이 스팸 사전에 속할 확률의 곱과 논스팸 사전에 속할 확률의 곱을 비교하여, 해당 데이터를 스팸 또는 논스팸으로 분류하는 단계를 포함하는 것을 특징으로 하는 소셜 네트워크 서비스에서 스팸 탐지 방법.

청구항 6

제5항에 있어서,

상기 산출하는 단계에서,

$$P(W_i|c) = \frac{f_{W_i} + mp_c}{f_c + m}$$

해당 문자열이 사전에 존재하는 경우, $P(W_i|c)$ 를 이용하여 스팸 사전에 속할 확률 또는 논

$$\begin{aligned} P(W_i|spam) &= (1 - \alpha) \\ P(W_i|non-spam) &= (1 + \alpha) \end{aligned}$$

스팸 사전에 속할 확률을 산출하고, 상기 사전에 존재하지 않는 경우, $P(W_i|c)$ 를 이용하여 스팸 사전에 속할 확률 또는 논스팸 사전에 속할 확률을 산출하는 것을 특징으로 하는 소셜 네트워크 서비스에서 스팸 탐지 방법

여기서, f_m 는 사전에 저장된 m 의 빈도수, c 는 클래스 c에 해당하는 사전의 m 의 합, m 은 평가 샘플 크기, i 는 사용자 정의 매개 변수, $P(W_i|spam)$ 는 문자열 (W_i)가 스팸 사전에 속할 확률, $P(W_i|non-spam)$ 는 문자열 (W_i)가 논스팸 사전에 속할 확률, α 는 $0 \leq \alpha < 1$ 인 임의의 값임.

청구항 7

삭제

청구항 8

삭제

청구항 9

제1항에 있어서,

상기 클래스 라벨을 예측하는 단계는,

상기 정렬된 데이터들 중에서 상위 일정 개수의 데이터와 하위 일정 개수의 데이터를 추출하여, 각각 전처리하는 단계;

상기 전처리된 데이터를 n-gram의 문자열로 분해하는 단계;

상기 분해된 각 문자열에 대해 나이브 베이스 분류기를 적용하여 스팸 사전 또는 논스팸 사전에 존재할 확률을 산출하는 단계; 및

각 데이터별로 각 문자열이 스팸 사전에 속할 확률의 곱과 논스팸 사전에 속할 확률의 곱을 비교하여, 해당 데이터를 스팸 또는 논스팸으로 분류하는 단계를 포함하는 것을 특징으로 하는 소셜 네트워크 서비스에서 스팸 탐지 방법.

청구항 10

제1항에 있어서,

상기 업데이트하는 단계는,

상기 상위 일정 개수의 데이터 중에서 스팸으로 예측된 데이터들과 하위 일정 개수의 데이터 중에서 논스팸으로 예측된 데이터들을 각각 전처리하는 단계;

상기 전처리된 데이터를 n-gram의 문자열로 분해하는 단계; 및

상기 n-gram으로 분해된 각 문자열의 빈도수를 산출하여, 상기 사전을 업데이트하는 단계를 포함하는 것을 특징으로 하는 소셜 네트워크 서비스에서 스팸 탐지 방법.

청구항 11

트레이닝 데이터의 존재 여부에 따라 n-gram 또는 리트윗 증가율을 이용하여 사전을 구축하는 사전 구축부;

소셜 네트워크 서비스에서 수집된 데이터에 대해 상기 구축된 사전을 기반으로 나이브 베이스 분류기(Naive Bayes Classification)를 적용하여 스팸을 탐지하는 스팸 탐지부; 및

일정 시간동안 수집된 데이터에 대하여 리트윗 증가율을 산출하고, 상기 산출된 리트윗 증가율을 기준으로 상기 데이터들을 내림차순으로 정렬하며, 상기 정렬된 데이터들에 대해 상기 사전을 기반으로 나이브 베이스 분류기를 적용하여 상위 일정 개수 데이터 및 하위 일정 개수 데이터에 대한 스팸 또는 논스팸의 클래스 라벨을 예측하고, 상기 상위 일정 개수의 데이터 중에서 스팸으로 예측된 데이터들을 이용하여 스팸 사전을 업데이트하며, 하위 일정 개수의 데이터 중에서 논스팸으로 예측된 데이터들을 이용하여 논스팸 사전을 업데이트하는 사전 업데이트부

를 포함하는 것을 특징으로 하는 스팸 탐지 장치.

청구항 12

제11항에 있어서,

상기 사전 구축부는,

트레이닝 데이터가 존재하는 경우, 각 트레이닝 데이터를 전처리하여 n-gram의 문자열로 분해하고, 상기 n-gram으로 분해된 각 문자열의 빈도수를 산출하여, 스팸 사전 또는 논스팸 사전을 생성하고,

트레이닝 데이터가 존재하지 않는 경우, 일정 시간동안 수집된 데이터에 대하여 리트윗 증가율을 산출하고, 상기 산출된 리트윗 증가율을 기준으로 상기 데이터들을 내림차순으로 정렬하며, 상기 정렬된 데이터들에서 상위

일정 개수의 데이터를 스팸 트레이닝 데이터 및 하위 일정 개수의 데이터를 논스팸 트레이닝 데이터로 추출하고, 상기 추출된 스팸 트레이닝 데이터와 논스팸 트레이닝 데이터를 각각 전처리하여 n-gram의 문자열로 분해하며, 상기 n-gram으로 분해된 각 문자열의 빈도수를 산출하여, 스팸 사전 또는 논스팸 사전을 생성하는 것을 특징으로 하는 스팸 탐지 장치.

청구항 13

제11항에 있어서,

상기 스팸 탐지부는,

상기 수집된 데이터를 전처리하여 n-gram의 문자열로 분해하고, 상기 분해된 각 문자열에 대해 나이트 베이스 분류기를 적용하여 스팸 사전 또는 논스팸 사전에 존재할 확률을 산출하며, 각 데이터별로 각 문자열이 스팸 사전에 속할 확률의 곱과 논스팸 사전에 속할 확률의 곱을 비교하여, 해당 데이터를 스팸 또는 논스팸으로 분류하는 것을 특징으로 하는 스팸 탐지 장치.

청구항 14

삭제

발명의 설명

기술 분야

[0001] 본 발명은 소셜 네트워크 서비스에서 스팸 탐지 방법 및 장치에 관한 것으로, 특히 트레이닝 데이터가 제한된 환경에서 n-gram 사전을 구축하고 이를 이용하여 소셜 네트워크 서비스의 스팸을 탐지하는 소셜 네트워크 서비스에서 스팸 탐지 방법 및 장치에 관한 것이다.

배경 기술

[0002] 스마트폰 보급이 활성화되면서 트위터, 페이스북과 같은 소셜 미디어가 급성장하기 시작했다. 그 중에서 트위터는 포스팅 당 140자 제한과 단순한 인터페이스를 통한 용이한 접근성을 바탕으로 대표적인 소셜 미디어로써 자리 잡기 시작했다. 현재는 사용자들이 자신의 일상생활에 관련된 주제뿐만 아니라 정치, 경제, 사회 등 다양한 분야에 대해 자신의 의견이나 감정을 공유하는 하나의 시스템이 되었다. 학계에서는 오래전부터 트위터 데이터를 사용하여 토픽 추출, 이벤트 탐지, 감성 분석 등을 수행하는 방법들을 제안하였다.

[0003] 트위터가 대표적인 소셜 미디어로 성장하게 된 주요 요인은 팔로우와 리트윗 기능이다. 두 기능은 트위터에서 정보를 확산시키는 데에 주요한 역할을 한다. 팔로우 기능은 다른 사용자와 관계를 맺는 기능으로써 팔로우 기능을 통해 자신이 팔로우하고 있는 사용자의 트윗을 실시간으로 받아볼 수 있으며, 리트윗 기능은 이미 존재하는 특정 트윗을 자신의 팔로워들에게 전파하는 역할을 한다. 만약 화제가 되는 사건이 발생할 경우 해당 사건에 대한 트윗이 급증함과 동시에 리트윗 수치 또한 급격히 높아지게 되며, 이에 따라 리트윗된 트윗을 접하는 사용자들이 급속히 증가하게 된다. 이처럼 트위터에서는 팔로우와 리트윗 기능을 통해 정보의 확산이 단순하게 이루어진다.

[0004] 그러나 정보의 확산이 단순하게 이루어진다는 특징은 스팸(spam)들에게 역으로 이용되고 있다. 일반 사용자들은 리트윗 기능을 특정 사건에 대한 정보 전달의 목적을 위해 사용한다. 하지만 스팸들은 리트윗 기능을 자신의 이익을 목적으로 사용하여 스팸 정보가 마치 현재 화제가 되는 사건인 것처럼 위장하기 위해 사용한다. 실제로 트위터의 전체 포스트 중 약 10% 이상을 스팸이 차지하고 있으며, 트위터 사용자의 약 5%는 비정상적인 사용자로 프로그램을 통해 자동적으로 스팸을 작성하고 있다. 또한, 최근 스팸들은 스팸 필터링 시스템을 회피하기 위한 목적으로 일반 사용자의 행동을 따라하려는 경향이 있으며, 스팸인 정보들을 인기 있는 토픽으로 위장하기 위해 트위터의 기능들을 악용하고 있다. 스팸 정보들을 차단하지 않는다면 스팸에 의해 유입된 각종 음란물, 불법 도박, 약물 등의 불건전한 정보를 담고 있는 스팸 트윗들이 확산될 것이며, 일반 사용자들은 이러한 정보에 더욱 쉽게 노출될 것이다. 또한, 스팸에 의해 의도적으로 악용된 행위 때문에 트위터에서 가치 있는 정보를 추출하기 위한 각종 시스템들의 품질 저하를 야기할 수 있고, 이로 인해 트위터에서 실제로 발생하는

사건들에 대한 정보가 아닌 스파머에 의한 정보들로 통계 자료가 구축될 가능성이 있다.

[0005] 그러나 트위터의 스팸 탐지에 대한 연구들은 트윗 내에 포함된 악성 URL을 탐지하기 위한 연구들이 대부분이며, 이러한 탐지 방법들은 음란물, 불법 도박, 약물 등과 같은 불건전한 정보를 담고 있는 스팸 트윗들을 효율적으로 탐지하는데 한계가 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국 등록특허공보 제10-1482073호(2006.04.24.)

발명의 내용

해결하려는 과제

[0007] 본 발명의 목적은, 불건전한 정보를 포함하고 있는 음란물, 불법 도박, 약물에 대한 트윗들을 탐지할 수 있는 소셜 네트워크 서비스에서 스팸 탐지 방법 및 장치를 제공하는 것이다.

[0008] 본 발명의 다른 목적은 초기 트레이닝 데이터가 매우 작거나 존재하지 않은 환경에서 스팸 탐지를 위한 사전(n-gram 사전)을 구축할 수 있는 소셜 네트워크 서비스에서 스팸 탐지 방법 및 장치를 제공하는 것이다.

[0009] 본 발명이 해결하고자 하는 과제는 이상에서 언급한 과제(들)로 제한되지 않으며, 언급되지 않은 또 다른 과제(들)는 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0010] 상기한 목적을 달성하기 위하여, 본 발명의 일 실시예에 따른 소셜 네트워크 서비스에서 스팸 탐지 방법은, 스팸 탐지 장치가 소셜 네트워크 서비스에서 스팸을 탐지하는 방법에 있어서, 트레이닝 데이터의 존재 여부에 따라 n-gram 또는 리트윗 증가율을 이용하여 사전을 구축하는 단계, 소셜 네트워크 서비스에서 수집된 데이터에 대해 상기 구축된 사전을 기반으로 나이브 베이즈 분류기(Naive Bayes Classification)를 적용하여 스팸을 탐지하는 단계를 포함한다.

[0011] 바람직하게는, 상기 사전을 구축하는 단계는, 상기 트레이닝 데이터가 존재하는 경우, 각 트레이닝 데이터를 전처리하는 단계, 상기 전처리된 각 트레이닝 데이터에 대해 n-gram의 문자열로 분해하는 단계, 상기 n-gram으로 분해된 각 문자열의 빈도수를 산출하여, 스팸 사전 또는 논스팸 사전을 생성하는 단계를 포함할 수 있다.

[0012] 바람직하게는, 상기 사전을 구축하는 단계는, 상기 트레이닝 데이터가 없는 경우, 일정 시간동안 수집된 데이터에 대하여 리트윗 증가율을 산출하는 단계, 상기 산출된 리트윗 증가율을 기준으로 상기 데이터들을 내림차순으로 정렬하는 단계, 상기 정렬된 데이터들에서 상위 일정 개수의 데이터를 스팸 트레이닝 데이터, 하위 일정 개수의 데이터를 논스팸 트레이닝 데이터로 추출하는 단계, 상기 추출된 스팸 트레이닝 데이터와 논스팸 트레이닝 데이터를 각각 전처리하는 단계, 상기 전처리된 각 트레이닝 데이터를 n-gram의 문자열로 분해하는 단계, 상기 n-gram으로 분해된 각 문자열의 빈도수를 산출하여, 스팸 사전 또는 논스팸 사전을 생성하는 단계를 포함할 수 있다.

[0013] 바람직하게는, 상기 리트윗 증가율($IR(tid)$)은 아래 수학적식을 이용하여 산출할 수 있다.

[0014] [수학적식]

$$IR(tid) = \left(\sum_{i=1}^m IR_i(tid) \right) / (followers_{tid} * RC_m(tid))$$

[0016] 여기서, tid는 원본 트윗에 대한 고유 식별정보, $IR_i(tid)$ 는 i시간 동안 발생한 리트윗 수, $followers_{tid}$ 는 원본 트윗을 작성한 사용자의 팔로워 수, $RC_m(tid)$ 는 해당 시점에 누적된 총 리트윗 수를 의미함.

[0017] 바람직하게는, 상기 스팸을 탐지하는 단계는, 상기 수집된 데이터를 전처리하는 단계, 상기 전처리된 데이터를 n-gram의 문자열로 분해하는 단계, 상기 분해된 각 문자열에 대해 나이브 베이스 분류기를 적용하여 스팸 사전 또는 논스팸 사전에 존재할 확률을 산출하는 단계, 각 데이터별로 각 문자열이 스팸 사전에 속할 확률의 곱과 논스팸 사전에 속할 확률의 곱을 비교하여, 해당 데이터를 스팸 또는 논스팸으로 분류하는 단계를 포함할 수 있다.

$$P(W_i|c) = \frac{f_{W_i} + mp_c}{f_c + m}$$

[0018] 바람직하게는, 상기 산출하는 단계에서, 해당 문자열이 사전에 존재하는 경우, 를 이용하여 스팸 사전에 속할 확률 또는 논스팸 사전에 속할 확률을 산출하고, 상기 사전에 존재하지 않는 경우,

$$\begin{aligned} P(W_i|spam) &= (1-\alpha) \\ P(W_i|non-spam) &= (1+\alpha) \end{aligned}$$

를 이용하여 스팸 사전에 속할 확률 또는 논스팸 사전에 속할 확률을 산출할 수 있다. 여기서, f_m 는 사전에 저장된 W_i 의 빈도수, i 는 클래스 c 에 해당하는 사전의 W_i 의 합, m 은 평가 샘플 크기, α 는 사용자 정의 매개 변수, $P(W_i|spam)$ 는 문자열 (W_i)가 스팸 사전에 속할 확률, $P(W_i|non-spam)$ 는 문자열 (W_i)가 논스팸 사전에 속할 확률, α 는 $0 \leq \alpha < 1$ 인 임의의 값임.

[0019] 바람직하게는, 상기 소셜 네트워크 서비스에서 스팸 탐지 방법은, 일정 시간 간격으로 수집된 데이터에 대해 리트윗 증가율을 산출하고, 상기 산출된 리트윗 증가율에 기초하여 각 데이터의 클래스 라벨을 예측하여, 상기 사전을 업데이트하는 단계를 더 포함할 수 있다.

[0020] 바람직하게는, 상기 업데이트하는 단계는, 일정 시간동안 수집된 데이터에 대하여 리트윗 증가율을 산출하는 단계, 상기 산출된 리트윗 증가율을 기준으로 상기 데이터들을 내림차순으로 정렬하는 단계, 상기 정렬된 데이터들에 대해 상기 사전을 기반으로 나이브 베이스 분류기를 적용하여 상위 일정 개수 데이터 및 하위 일정 개수 데이터에 대한 스팸 또는 논스팸의 클래스 라벨을 예측하는 단계, 상기 상위 일정 개수의 데이터 중에서 스팸으로 예측된 데이터들을 이용하여 스팸 사전을 업데이트하고, 하위 일정 개수의 데이터 중에서 논스팸으로 예측된 데이터들을 이용하여 논스팸 사전을 업데이트하는 단계를 포함할 수 있다.

[0021] 바람직하게는, 상기 클래스 라벨을 예측하는 단계는, 상기 정렬된 데이터들 중에서 상위 일정 개수의 데이터와 하위 일정 개수의 데이터를 추출하여, 각각 전처리하는 단계, 상기 전처리된 데이터를 n-gram의 문자열로 분해하는 단계, 상기 분해된 각 문자열에 대해 나이브 베이스 분류기를 적용하여 스팸 사전 또는 논스팸 사전에 존재할 확률을 산출하는 단계, 각 데이터별로 각 문자열이 스팸 사전에 속할 확률의 곱과 논스팸 사전에 속할 확률의 곱을 비교하여, 해당 데이터를 스팸 또는 논스팸으로 분류하는 단계를 포함할 수 있다.

[0022] 바람직하게는, 상기 업데이트하는 단계는, 상기 상위 일정 개수의 데이터 중에서 스팸으로 예측된 데이터들과 하위 일정 개수의 데이터 중에서 논스팸으로 예측된 데이터들을 각각 전처리하는 단계, 상기 전처리된 데이터를 n-gram의 문자열로 분해하는 단계, 상기 n-gram으로 분해된 각 문자열의 빈도수를 산출하여, 상기 사전을 업데이트하는 단계를 포함할 수 있다.

[0023] 상기한 목적을 달성하기 위하여, 본 발명의 다른 실시예에 따른 스팸 탐지 장치는, 트레이닝 데이터의 존재 여부에 따라 n-gram 또는 리트윗 증가율을 이용하여 사전을 구축하는 사전 구축부, 소셜 네트워크 서비스에서 수집된 데이터에 대해 상기 구축된 사전을 기반으로 나이브 베이스 분류기(Naive Bayes Classification)를 적용하여 스팸을 탐지하는 스팸 탐지부를 포함한다.

[0024] 바람직하게는, 상기 사전 구축부는, 트레이닝 데이터가 존재하는 경우, 각 트레이닝 데이터를 전처리하여 n-gram의 문자열로 분해하고, 상기 n-gram으로 분해된 각 문자열의 빈도수를 산출하여, 스팸 사전 또는 논스팸 사전을 생성하고, 트레이닝 데이터가 존재하지 않는 경우, 일정 시간동안 수집된 데이터에 대하여 리트윗 증가율을 산출하고, 상기 산출된 리트윗 증가율을 기준으로 상기 데이터들을 내림차순으로 정렬하며, 상기 정렬된 데이터들에서 상위 일정 개수의 데이터를 스팸 트레이닝 데이터 및 하위 일정 개수의 데이터를 논스팸 트레이닝 데이터로 추출하고, 상기 추출된 스팸 트레이닝 데이터와 논스팸 트레이닝 데이터를 각각 전처리하여 n-gram의 문자열로 분해하며, 상기 n-gram으로 분해된 각 문자열의 빈도수를 산출하여, 스팸 사전 또는 논스팸 사전을 생성할 수 있다.

[0025] 바람직하게는, 상기 스팸 탐지부는, 상기 수집된 데이터를 전처리하여 n-gram의 문자열로 분해하고, 상기 분해된 각 문자열에 대해 나이트 베이스 분류기를 적용하여 스팸 사전 또는 논스팸 사전에 존재할 확률을 산출하며, 각 데이터별로 각 문자열이 스팸 사전에 속할 확률의 곱과 논스팸 사전에 속할 확률의 곱을 비교하여, 해당 데이터를 스팸 또는 논스팸으로 분류할 수 있다.

[0026] 바람직하게는, 상기 스팸 탐지 장치는 일정 시간 간격으로 수집된 데이터에 대해 리트윗 증가율을 산출하고, 상기 산출된 리트윗 증가율에 기초하여 각 데이터의 클래스 라벨을 예측하여, 상기 사전을 업데이트하는 사전 업데이트부를 더 포함할 수 있다.

발명의 효과

[0027] 본 발명에 따르면, 초기 트레이닝 데이터가 매우 작거나 존재하지 않은 환경에서 스팸 탐지를 위한 n-gram 사전을 구축할 수 있고, n-gram 사전을 기반으로 하는 분류기를 사용하여 스팸을 탐지함으로써, 불건전한 정보를 담고 있는 스팸 탐지에 성능을 높일 수 있다.

[0028] 또한, 초기에 트레이닝 집합이 존재하지 않을 때, 트위터의 리트윗 특성을 활용하여 의사 라벨을 생성하고, 이를 트레이닝 집합으로 활용할 수 있다. 또한, 국내의 스팸 탐지에 대한 대부분의 연구들은 매우 적은 양의 데이터를 사용하여, 실제 스팸 탐지의 적용에 대한 확신을 갖기 어려운데, 본 발명은 현실적인 스팸 탐지의 효용성을 보인다.

[0029] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

- [0030] 도 1은 본 발명의 일 실시예에 따른 트위터 스팸 탐지 장치를 설명하기 위한 블록도이다.
- 도 2는 본 발명의 일 실시예에 따른 트레이닝 데이터가 존재하는 경우 사전 구축부를 상세히 나타낸 블록도이다.
- 도 3은 본 발명의 일 실시예에 따른 전처리를 설명하기 위한 예시도이다.
- 도 4는 본 발명의 일 실시예에 따른 6-gram 시퀀스를 설명하기 위한 예시도이다.
- 도 5는 본 발명의 일 실시예에 따른 트레이닝 데이터가 없는 경우 사전 구축부를 상세히 나타낸 블록도이다.
- 도 6은 본 발명의 일 실시예에 따른 시간에 따라 누적된 리트윗의 수를 설명하기 위한 예시도이다.
- 도 7은 도 1에 도시된 스팸 탐지부의 구성을 상세히 나타낸 블록도이다.
- 도 8은 도 1에 도시된 사전 업데이트부의 구성을 상세히 나타낸 블록도이다.
- 도 9는 본 발명의 일 실시예에 따른 사전 구축 및 스팸 탐지 방법을 설명하기 위한 흐름도이다.
- 도 10은 본 발명의 일 실시예에 따른 트레이닝 데이터가 존재하는 경우 사전을 구축하는 방법을 설명하기 위한 흐름도이다.
- 도 11은 본 발명의 일 실시예에 따른 트레이닝 데이터가 없는 경우 사전을 구축하는 방법을 설명하기 위한 흐름도이다.
- 도 12는 본 발명의 일 실시예에 따른 스팸 탐지 방법을 설명하기 위한 흐름도이다.
- 도 13은 본 발명의 일 실시예에 따른 사전 업데이트 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0031] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나 이는 본 발명을 특정한 실시 형태에 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.

[0032] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어

들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

- [0033] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.
- [0034] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0035] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0036] 소셜 네트워크 서비스(Social Network Service)는 미니홈피, 블로그, 카페, 트위터, 미투데이, 페이스북 등을 포함하나, 이하에서는 설명의 편의를 위해 트위터를 예로하여 설명하기로 한다.
- [0037] 본 발명에 기재된 사전은 n-gram 사전으로, 스팸 사전과 논스팸 사전을 포함할 수 있다.
- [0038] 이하에서는 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.
- [0039] 도 1은 본 발명의 일 실시예에 따른 소셜 네트워크 서비스에서 스팸 탐지 장치를 설명하기 위한 블록도, 도 2는 본 발명의 일 실시예에 따른 트레이닝 데이터가 존재하는 경우 사전 구축부를 상세히 나타낸 블록도, 도 3은 본 발명의 일 실시예에 따른 전처리를 설명하기 위한 예시도, 도 4는 본 발명의 일 실시예에 따른 6-gram 시퀀스를 설명하기 위한 예시도, 도 5는 본 발명의 일 실시예에 따른 트레이닝 데이터가 없는 경우 사전 구축부를 상세히 나타낸 블록도, 도 6은 본 발명의 일 실시예에 따른 시간에 따라 누적된 리트윗의 수를 설명하기 위한 예시도, 도 7은 도 1에 도시된 스팸 탐지부의 구성을 상세히 나타낸 블록도, 도 8은 도 1에 도시된 사전 업데이트부의 구성을 상세히 나타낸 블록도이다.
- [0040] 도 1을 참조하면, 본 발명의 일 실시예에 따른 소셜 네트워크 서비스에서 스팸 탐지 장치(100)는 사전 구축부(110), 스팸 탐지부(120), 사전 업데이트부(130)를 포함한다.
- [0041] 사전 구축부(110)는 초기 트레이닝 데이터의 존재 여부(또는 크기)에 따라 n-gram 또는 리트윗 증가율을 이용하여 초기 사전을 구축한다. 여기서, 사전은 스팸 사전, 논스팸 사전을 포함할 수 있고, 스팸 사전에는 문자열(n-gram)과 그 빈도수가 저장되고, 논스팸 사전에는 문자열(n-gram)과 그 빈도수가 저장될 수 있다.
- [0042] 스팸 트윗 또는 논스팸 트윗을 분류하기 위해서는 분류 모델, 즉, 스팸과 논스팸에 대한 문자열(n-gram, 단어)에 대한 정보가 저장된 사전이 있어야 한다. 이를 위해, 사전 구축부(110)는 초기 주어진 트레이닝 데이터(또는 트레이닝 집합)를 이용하여 사전을 구축하게 된다. 그러나, 트위터와 같이 대량의 데이터가 실시간으로 유입되는 상황에서 초기 분류 모델을 구축하기 위해 트레이닝 집합을 구성하는 것은 매우 큰 비용을 요구한다. 이로 인해 가용한 트레이닝 데이터(집합)의 크기가 아주 작을 수 있으며, 심지어는 트레이닝 데이터(집합) 자체가 존재하지 않을 수도 있다. 본 발명은 이러한 상황에 대한 문제점을 해결하기 위해 트레이닝 데이터가 존재하는 경우 n-gram을 이용하여 사전을 구축하고, 트레이닝 데이터가 존재하지 않은 경우 리트윗에 대한 순간 증가율을 이용하여 사전을 구축할 수 있다. 즉, 사전 구축부(110)는 트레이닝 데이터의 존재 여부에 따라 다른 방법으로 사전을 구축한다.
- [0043] 먼저, 트레이닝 데이터가 존재하는 경우 사전을 구축하는 방법에 대해 설명하기로 한다. 이때, 트레이닝 데이터는 트윗이 스팸 또는 논스팸인지 미리 구분되어 있다.
- [0044] 사전 구축부(110)는 각 트레이닝 데이터를 전처리하고, 전처리된 각 트레이닝 데이터에 대해 n-gram의 문자열

(또는 문자)로 분해한 후, n-gram으로 분해된 각 문자열의 빈도수를 산출하여, 스팸 사전 또는 논스팸 사전을 구축할 수 있다. 이때, 사전 구축부(110)는 트레이닝 데이터에 포함된 스팸 트윗들을 이용하여 스팸 사전을 구축할 수 있고, 논스팸 트윗들을 이용하여 논스팸 사전을 구축할 수 있다.

[0045] 이러한 사전 구축부(110)는 도 2에 도시된 바와 같이 전처리부(112), n-gram 분해부(114), n-gram 사전 구축부(116)를 포함한다.

[0046] 전처리부(112)는 트레이닝 데이터에서 리트윗 정보, 사용자 멘션정보, URL 정보, 특수문자, 공백 등을 제거하는 전처리를 수행한다. 이때, 전처리부(112)는 정규 표현식을 이용하여 전처리를 수행할 수 있다.

[0047] 트위터에서 사용자들이 작성하는 트윗의 대부분은 맞춤법을 무시하는 경우가 많으며, 또한 과도한 특수문자를 포함하는 경우가 많다. 특히 트위터는 사용자 멘션, 해시태그, 리트윗 등의 기능을 갖고 있어, 사용자가 트윗을 작성할 때 해당 정보들이 트윗 내에 자동으로 포함된다. 이러한 정보들은 트윗 내용 자체에 대한 특성을 반영하지 못하며, 따라서 사전에 제거되어야 한다.

[0048] 따라서, 전처리부(112)는 리트윗 정보, 사용자 멘션정보, URL 정보, 특수문자, 공백 등을 미리 정해진 함수에 넣어서 제거할 수 있다. 여기서, 리트윗 정보는 하나의 트윗에 대해 리트윗이 발생할 경우 “RT @user_id”의 형태로 문장 내에 자동으로 생성된 것이고, 사용자 멘션 정보는 한 사용자가 다른 사용자에게 직접 트윗을 보낼 경우 “@user_id”의 형태로 문장 내에 자동으로 생성된 것이다. URL 정보는 사용자가 트윗 내에 URL을 첨부할 경우 단축된 URL이 트윗 내에 첨부된 것으로, 트위터의 경우 “http://t.co/...”의 형태로 URL이 첨부되며, 단축된 URL 자체는 트윗 내용에 대한 특성을 반영하지 못한다. 특수문자는 스팸어의 경우 문장 사이사이에 특수문자를 섞어서 의미적으로 동일한 단어를 반복적으로 사용하는 경우가 많은데, 특수문자를 제거함으로써 이러한 단어들을 찾아낼 수 있다. 공백 제거는 앞에서의 정보들을 제거한 후 문장 내에 존재하는 공백들을 제거하는 것으로, 하나의 트윗은 사용자가 작성한 문장을 기준으로 공백 없이 한 문장으로 연결된 형태가 된다.

[0049] 전처리에 대해 도 3을 참조하면, (a)와 같은 트윗 문장에서 전처리가 수행되면, (b)와 같이 리트윗 정보, 사용자 멘션정보, URL 정보, 특수문자, 공백 등이 제거된 문장이 생성될 수 있다. 즉, 전처리부(112)는 (a)에서 리트윗 정보인 “RT @user_id”, 특수문자인 “-” 및 “#”, URL 정보인 “https://t.co/...” , “https://t...”, 공백을 제거하여 (b)와 공백없이 이어진 하나의 문장을 생성할 수 있다.

[0050] 한편, 전처리 과정에서 형태소 분석은 사용하지 않는다. 형태소 분석에 따른 시간 비용이 크기 때문에 실시간 탐지에 부적합하기 때문이다. 또한, 스팸어들은 자신들의 이익을 위한 특정 단어만을 사용하는 경향이 있는데, 형태소를 분석할 경우 그러한 단어들의 의미가 사라질 수 있다. 예를 들어 “사설토토추천사이트”에 대해 형태소 분석을 진행하면 “사설토/NNG+하/XSV+도/EC+추천사/NNG+이/JKS”와 같은 형태로 분해될 수 있으며, 이로 인해 단어 자체가 가지고 있는 본래의 의미를 상실할 수 있다.

[0051] n-gram 분해부(114)는 전처리된 각 트레이닝 데이터에 대해 n-gram 단위의 문자열로 분해한다. 즉, n-gram 분해부(114)는 전처리 과정을 마친 트윗에 대해 n-gram의 문자열(또는 문자)로 분해한다. n-gram에 대한 분해는 문자, 음절, 어절 등으로 분해될 수 있는데, 본 발명에서는 문자 단위로 문장을 분해한 경우로 설명하기로 한다.

[0052] 예를 들어, 도 3에서 전처리된 문장에 대해 문자 단위로 6-gram으로 분해하면, 도 4와 같이 6개의 문자들(또는 단어)로 분해될 수 있다.

[0053] n-gram 사전 구축부(116)는 n-gram으로 분해된 문자열들의 빈도수를 산출하여, 스팸 사전과 논스팸 사전을 구축한다. 즉, n-gram 사전 구축부(116)는 n-gram으로 분해된 트레이닝 집합의 트윗들을 이용하여 스팸과 논스팸에 대한 사전을 각각 구성한다. 여기서, 스팸 사전과 논스팸 사전은 n-gram 사전일 수 있다.

[0054] 예를 들어, 도 4과 같이 한 트윗이 n-gram 시퀀스 $(W_0, W_1, \dots, W_k)$ 로 구성되고, $c \in \{spam, non-spam\}$ 를 클래스 정보라고 가정하여 n-gram 사전을 구축하는 방법에 대해 설명하기로 한다. n-gram 사전 구축부(116)는 W_k 가 클래스 c에 대한 사전에서 처음 발생한 경우 W_k 에 대한 빈도수를 '1'로 저장하고, 이미 존재하는 경우 기존 빈도수에 '1'을 더하여 값을 갱신한다. 그러면, 문자열과 그 빈도수에 대한 정보가 저장된 스팸 사전 또는 논스팸 사전이 구축될 수 있다.

[0055] 다음으로, 트레이닝 데이터가 없는 경우에 사전을 구축하는 방법에 대해 설명하기로 한다.

[0056] 사전 구축부(110)는 일정 시간동안 수집된 데이터에 대하여 리트윗 증가율을 산출하고, 산출된 리트윗 증가율에 기초하여 트레이닝 데이터를 생성하며, 생성된 각 트레이닝 데이터를 전처리하고, 전처리된 각 트레이닝 데이터

에 대해 n-gram의 문자열로 분해한 후, n-gram으로 분해된 각 문자열의 빈도수를 산출하여, 스팸 사전 또는 논스팸 사전을 구축할 수 있다. 이 경우, 사전 구축을 위한 트레이닝 데이터가 주어지지 않았기 때문에, 사전 구축부(110)는 트레이닝 데이터를 생성해야 하고, 이를 위해 일정 시간 동안 데이터를 수집해야 한다. 이때, 수집된 데이터는 스팸 또는 논스팸인지에 대한 정보가 없기 때문에, 사전 구축부(110)는 수집된 데이터에서 스팸 사전 구축을 위한 스팸 트레이닝 데이터와 논스팸 사전 구축을 위한 논스팸 트레이닝 데이터를 각각 추출해야 하며, 이를 위해 리트윗 증가율을 이용할 수 있다. 스팸 트레이닝 데이터와 논스팸 트레이닝 데이터가 추출되면, 사전 구축부(110)는 스팸 트레이닝 데이터를 이용하여 스팸 사전을 구축하고, 논스팸 트레이닝 데이터를 이용하여 논스팸 사전을 구축할 수 있다.

[0057] 이러한 사전 구축부(110)는 도 5에 도시된 바와 같이 트레이닝 집합 생성부(111), 전처리부(112), n-gram 분해부(114), n-gram 사전 구축부(116)를 포함한다.

[0058] 트레이닝 집합 생성부(111)는 일정 시간동안 수집된 데이터에 대하여 리트윗 증가율을 산출하고, 산출된 리트윗 증가율을 기준으로 스팸 트레이닝 데이터와 논스팸 트레이닝 데이터를 선택한다. 즉, 트레이닝 집합 생성부(111)는 일정 시간동안 수집된 데이터에 대하여 리트윗 증가율을 산출하고, 산출된 리트윗 증가율을 기준으로 데이터들을 내림차순으로 정렬하며, 정렬된 데이터들에서 상위 일정 개수를 스팸 트레이닝 데이터, 하위 일정 개수의 데이터를 논스팸 트레이닝 데이터로 선택한다. 여기서, 트레이닝 집합은 트레이닝 데이터들의 집합을 의미하는 것으로, 트레이닝 데이터와 동일한 개념으로 사용될 수 있다.

[0059] 이러한 트레이닝 집합 생성부(111)는 리트윗 증가율 산출모듈(107), 의사라벨 생성 모듈(109)을 포함한다.

[0060] 리트윗 증가율 산출모듈(107)은 일정 시간동안 수집된 데이터에 대하여 리트윗 증가율을 산출한다.

[0061] 의사 라벨을 생성하기 위해 트위터의 리트윗에 대한 순간 증가율을 이용할 수 있다. 리트윗 순간 증가율은 특정 사건이 발생했을 때, 해당 사건에 대한 트윗들의 리트윗이 급격히 증가하는 현상을 반영할 수 있다. 예를 들어 재난 소식, 대통령 선거 등과 같은 화제가 되는 사건이 발생하게 되면 이와 관련된 주제를 갖는 트윗이 급증하게 되는데, 이에 따라 리트윗 수도 급격히 증가하게 된다. 리트윗이라는 기능 자체가 트윗을 전파하기 위한 목적을 갖고 있기 때문에 리트윗 수가 급격히 증가할수록 해당 트윗이 화제가 되는 사건을 내포한다고 볼 수 있다. 그러나 스팸머들은 이러한 특성을 악용하여 자신들이 작성한 스팸 트윗이 인기 있는 사건에 대한 내용인 것처럼 위장하기 위해 리트윗 기능을 사용한다.

[0062] 리트윗의 순간 증가율을 활용하여 의사 라벨을 생성하기 위해서는 먼저 트윗의 특성에 대해 알아야 한다. 트위터에서 발생하는 모든 트윗들은 작성한 사용자의 정보와 콘텐츠에 관한 2가지 정보를 기본적으로 포함하고 있다. 사용자 정보는 사용자 아이디, 사용자의 팔로워와 팔로잉 수 등에 관한 정보를 포함하며, 콘텐츠 정보는 트윗의 고유 아이디, 작성된 시간, 현재 시점까지의 리트윗 수 등에 관한 정보를 포함한다. 특정 트윗에 대해 리트윗이 발생하게 되면 리트윗된 트윗은 자기 자신에 대한 정보와 원본 트윗에 대한 정보를 모두 포함하게 된다. 특히, 원본 트윗을 리트윗한 모든 트윗들은 원본 트윗의 트윗 아이디를 갖고 있으며, 그 시점까지의 원본 트윗에 대한 누적 리트윗 수를 갖는다.

[0063] 도 6은 원본 트윗 T(a)에 대해 발생한 리트윗을 시간에 따라 나타낸 예시도이다. Time_i(a)는 트윗이 발생한 시간을 의미하며, RC_i(a)는 Time_i(a)까지 누적된 리트윗 수를 나타낸다. 원본 트윗 T(a)가 발생한 시점인 Time₀(a)에서는 리트윗이 발생하지 않았으므로, 리트윗 수의 값은 '0'이다. Time_i(a)에서의 리트윗 수인 RC_i(a)과 이전 시점인 Time_{i-1}(a)에서의 리트윗 수인 RC_{i-1}(a)을 이용하면 구간에서의 리트윗에 대한 증가율을 산출할 수 있다.

[0064] 특정 구간에서의 리트윗에 대한 순간 증가율(IR_i(tid))은 아래 기재된 수학적 식 1을 이용하여 산출한다.

수학적 식 1

$$IR_i(tid) = \left(\frac{RC_i(tid) - RC_{i-1}(tid)}{Time_i(tid) - Time_{i-1}(tid)} \right) (i=1,2,\dots)$$

[0065]

[0066] 여기서, tid 는 원본 트윗에 대한 트윗 아이디를 나타낸다. $RC_i(tid) - RC_{i-1}(tid)$ 는 i 번째 구간과 $i-1$ 번째 구간에서의 리트윗 수의 차이를 의미하고, $Time_i(tid) - Time_{i-1}(tid)$ 은 i 번째 구간과 $i-1$ 번째 구간의 시간 차이를 나타낸다. 결과적으로 수학적 식 1은 i 번째 구간 동안 발생한 리트윗 수를 나타낸다.

[0067] $IR_i(tid)$ 은 짧은 시간 동안 리트윗이 많이 발생할수록 높은 값을 갖게 된다. 하지만 수학적 식 1은 특정 구간에서의 리트윗 증가율만 반영할 뿐 스팸어의 특징을 제대로 반영하고 있지 않다. 스팸어의 경우 일반 사용자와 팔로우 관계를 맺기 힘들다는 특징이 있으며, 이로 인해 전체 리트윗 수와 팔로워 수는 일반 사용자보다 낮은 값을 갖게 된다. 이러한 스팸어의 특징을 반영하여 리트윗 증가율($IR(tid)$)은 아래 기재된 수학적 식 2와 같이 재정의될 수 있다. 수학적 식 2는 일정 시간 구간 동안 트윗 아이디 tid 를 가지는 원본 트윗에 대한 리트윗에 대한 리트윗 $RT_1(tid), \dots, RT_m(tid)$ 이 수집된 경우, 그 시간 구간 동안 트윗 $T(tid)$ 의 리트윗 증가율을 나타낸다.

수학적 식 2

[0068]
$$IR(tid) = \left(\sum_{i=1}^m IR_i(tid) \right) / (followers_{tid} * RC_m(tid))$$

[0069] 여기서, $followers_{tid}$ 는 원본 트윗을 작성한 사용자의 팔로워 수, $RC_m(tid)$ 는 계산이 이루어진 시점에서 누적된 총 리트윗 수를 의미한다. 결과적으로 수학적 식 1을 $followers_{tid} * RC_m(tid)$ 로 나누게 되면 Spammer에 의한 계산 결과가 일반 사용자보다 높은 값을 갖게 된다.

[0070] 상술한 바와 같이, 리트윗 증가율 산출모듈(107)은 일정 시간동안 수집된 트윗들의 리트윗 증가율을 수학적 식 2를 이용하여 각각 산출한다.

[0071] 의사 라벨 생성모듈(109)은 리트윗 증가율 산출모듈(107)에서 산출된 리트윗 증가율을 기준으로 데이터들을 내림차순으로 정렬하고, 정렬된 데이터들에서 상위 일정 개수를 스팸 트레이닝 데이터, 하위 일정 개수의 데이터를 논스팸 트레이닝 데이터로 추출한다. 여기서, 의사 라벨은 데이터를 스팸 또는 논스팸으로 분류한 것으로, 후술하는 클래스 라벨과 동일한 개념일 수 있다.

[0072] 일정 시간 구간 동안 리트윗이 한 번이라도 발생한 트윗들은 리트윗 증가율($IR(tid)$) 값을 가지게 되고, 이 값들을 내림차순으로 정렬하게 되면 상위에는 대부분 스팸 트윗들이 위치하게 되며, 하위에는 대부분 일반 사용자들의 트윗들이 위치하게 된다. 따라서, 의사 라벨 생성모듈(109)은 리트윗 증가율을 기준으로 내림차순으로 정렬된 트윗들 중에서 상위 일정 개수의 트윗들을 스팸 트윗으로 추출하여 스팸 사전 구축을 위한 스팸 트레이닝 데이터로 사용하고, 하위 일정 개수의 트윗들을 논스팸 트윗으로 추출하여 논스팸 사전 구축을 위한 논스팸 트레이닝 데이터로 사용할 수 있다.

[0073] 전처리부(112)는 트레이닝 집합 생성부(111)에서 생성된 트레이닝 데이터에서 리트윗 정보, 사용자 멘션정보, URL 정보, 특수문자, 공백 등을 제거하는 전처리를 수행한다. 이때, 전처리부는 정규 표현식을 이용하여 전처리를 수행할 수 있다.

[0074] n -gram 분해부(114)는 전처리된 각 트레이닝 데이터에 대해 n -gram 단위의 문자열로 분해한다. 즉, n -gram 분해부(114)는 전처리 과정을 마친 트윗에 대해 n -gram의 문자열(또는 문자)로 분해한다.

[0075] n -gram 사전 구축부(116)는 n -gram으로 분해된 문자열들의 빈도수를 산출하여, 스팸 사전과 논스팸 사전을 구축한다. 즉, n -gram 사전 구축부(116)는 n -gram으로 분해된 트레이닝 집합의 트윗들을 이용하여 스팸과 논스팸에 대한 사전을 각각 구성한다.

[0076] 전처리부(112), n -gram 분해부(114), n -gram 사전 구축부(116)는 도 2에 도시된 전처리부(112), n -gram 분해부(114), n -gram 사전 구축부(116)와 동일하므로, 상세한 설명은 도 2를 참조하기로 한다.

[0077] 다시 도 1을 참조하면, 스팸 탐지부(120)는 소셜 네트워크 서비스에서 수집된 데이터에 대해 사전 구축부(110)에서 구축된 사전에 기초하여 스팸 여부를 탐지한다. 즉, 스팸 탐지부(120)는 분류 대상 데이터를 전처리하고,

전처리된 데이터를 n-gram 단위의 문자열로 분해하며, 분해된 각 문자열이 스팸 사전 또는 논스팸 사전에 존재할 확률을 산출하여, 분석 대상 데이터를 스팸 또는 논스팸으로 분류한다.

[0078] 이러한 스팸 탐지부(120)는 도 7에 도시된 바와 같이 전처리부(122), n-gram 분해부(124), 탐지부(126)를 포함한다.

[0079] 전처리부(122)는 분류 대상 데이터에서 리트윗 정보, 사용자 멘션정보, URL 정보, 특수문자, 공백 등을 제거하는 전처리를 수행한다. 이때, 전처리부(122)는 정규 표현식을 이용하여 전처리를 수행할 수 있다.

[0080] n-gram 분해부(124)는 전처리된 각 분류 대상 데이터에 대해 n-gram 단위의 문자열로 분해한다. 즉, n-gram 분해부(124)는 전처리 과정을 마친 트윗에 대해 n-gram의 문자열(또는 문자)로 분해한다.

[0081] 전처리부(122), n-gram 분해부(124)는 도 2에 도시된 전처리부(112), n-gram 분해부(124)와 동일하므로, 상세한 설명은 도 2를 참조하기로 한다.

[0082] 탐지부(126)는 n-gram 분해부(124)에서 분해된 각 문자열에 대해 스팸 사전 또는 논스팸 사전에 존재할 확률을 산출하고, 산출된 확률에 기초하여 분류 대상 데이터를 스팸 또는 논스팸으로 분류한다. 즉, 탐지부(126)는 분류 대상 트윗별로 각 문자열이 스팸 사전에 속할 확률의 곱과 논스팸 사전에 속할 확률의 곱을 비교하여, 해당 트윗을 스팸 또는 논스팸으로 분류한다. 이때, 탐지부(126)는 스팸 사전에 속할 확률의 곱이 논스팸 사전에 속할 확률의 곱보다 큰 경우 해당 트윗을 스팸으로 분류한다.

[0083] 이하, 탐지부(126)의 동작을 구체적으로 설명하기로 한다.

[0084] 탐지부(126)는 분류 대상 트윗을 스팸 또는 논스팸으로 분류하기 위해, 클래스 조건부 확률을 산출하고, 산출된 클래스 조건부 확률에 기초하여 해당 트윗을 스팸 또는 논스팸으로 분류한다.

[0085] 클래스 조건부 확률($P(c|W_0W_1...W_k)$)은 아래 기재된 수학식 3을 이용하여 산출할 수 있다. 수학식 3은 하나의 트윗이 n-gram 시퀀스 $\langle W_0, W_1, \dots, W_k \rangle$ 로 구성되고, 이에 대한 $c \in \{spam, non-spam\}$ 의 클래스 조건부 확률($P(c|W_0W_1...W_k)$)을 산출하는 식으로, n-gram으로 분해된 트윗에 대해 나이브 베이스 분류기를 적용하여 조건부 확률을 구하는 식일 수 있다.

수학식 3

$$P(c|W_0W_1...W_k) = \frac{P(c) \prod_{i=0}^k P(W_i|c)}{P(W_0W_1...W_k)}$$

[0086]

[0087] 여기서, c는 트윗에 대한 클래스 라벨($c \in \{spam, non-spam\}$)을 의미하고, W_i 는 각각의 n-gram을 의미한다.

$P(W_i|c)$ 는 클래스 c사전(스팸 사전 또는 논스팸 사전)에 문자열이 속할 확률로, 클래스 c사전에 나타난 n-gram들의 전체 빈도수에 대한 W_i 의 빈도수의 비로 산출할 수 있다.

[0088] 탐지부(126)는 수학식 3을 이용하여 스팸과 논스팸에 대한 조건부 확률을 산출하고, 스팸일 조건부 확률이 더 높을 경우 스팸 데이터로 분류하게 된다. 즉, 탐지부(126)는 아래 기재된 수학식 4와 같이 스팸일 조건부 확률이 더 높을 경우 스팸 트윗으로 분류하게 된다.

수학식 4

$$P(spam) \prod_{i=0}^k P(W_i|spam) > P(non-spam) \prod_{i=0}^k P(W_i|non-spam)$$

[0089]

[0090] 한편, 수학식 3에서 $P(W_i|c)$ 를 클래스 c사전에 나타난 n-gram들의 전체 빈도수에 대한 W_i 의 빈도수의 비로 산

출한 경우, 사전에 존재하지 않는 새로운 n-gram인 W_i 가 등장하였을 때 조건부 확률 값이 '0'이 된다는 문제점이 있다. 이를 해결하기 위해 사전에 존재하지 않는 n-gram이 발생했을 때 확률 값을 계산하기 위해 m-estimate를 사용할 수 있다. m-estimate는 속성 값이 존재하지 않을 때 조건부 확률의 계산결과가 '0'이 되는 것을 방지한다. m-estimate를 적용하였을 때의 계산은 아래 기재된 수학적 식 5와 같다.

수학적 식 5

$$P(W_i|c) = \frac{f_{W_i} + mp_c}{f_c + m}$$

[0091]

[0092]

여기서, f_{W_i} 는 사전에 저장된 W_i 의 빈도수, f_c 는 클래스 c에 해당하는 사전의 W_i 의 합을 의미한다. m은 등가 샘플 크기, p_c 는 사용자 정의 매개변수로 이 두 값에 따라서 속성 값이 존재하지 않을 때 기본 확률을 부여할 수 있다. 하지만 m과 p_c 의 값이 지나치게 클 경우 오히려 예측되어야 하는 값과 정반대의 결과를 낼 수 있으며, 따라서 적절한 크기의 값으로 지정해주어야 한다. 본 발명에서는 $m=1$, $p_c=(c \text{에 대한 } n\text{-gram 사전 크기})^{-1}$ 로 주어 확률 계산에 사용하였다. p_c 의 값을 n-gram사전 크기의 역수로 준 것은 $P(W_i|spam)$ 인 경우 스팸 사전에서 발생했을 때의 기본 확률 값을 주는 것이며, $P(W_i|non-spam)$ 인 경우 논스팸 사전에서 발생했을 때의 기본 확률 값을 주는 것이다.

[0093]

한편, 스팸머와 일반 사용자가 사용하는 단어에는 차이가 있다. 예컨대, 대부분의 트윗들이 중복되는 경우가 많고, 스팸머가 작성하는 트윗들은 대부분 유사한 단어들로 구성되어 있다. 즉, 스팸머는 대부분 유사한 단어를 사용하여 스팸 트윗을 작성하며, 이와는 다르게 일반 사용자들은 자신의 감정, 의견 등에 대해 다양한 단어들 사용한다고 유추할 수 있다. 이에, 나이브 베이스 조건부 확률은 표 1과 같이 4가지로 나누어 산출할 수 있다.

표 1

	Existence in a spam dictionary	Existence in a non-spam dictionary
Case 1	O	O
Case 2	O	X
Case 3	X	O
Case 4	X	X

[0094]

[0095]

표 1은 조건부 확률 계산을 위한 모든 경우들을 나타낸다. Case 1~3의 경우 n-gram W_i 가 스팸 또는 논스팸 사전에 존재하는 경우이며, 이러한 경우에는 수학적 식 5를 적용하여 $P(W_i|c)$ 를 계산한다. 하지만 Case 4와 같이 스팸 및 논스팸 사전에 모두 존재하지 않는 경우에는 스팸 사전에서 새로운 n-gram이 발생할 확률보다 논스팸 사전에서 새로운 n-gram이 발생할 확률이 더 높을 것이라 예상할 수 있다. 따라서 Case 4의 경우처럼 n-gram 단어인 W_i 가 두개의 사전에서 모두 발생하지 않은 경우, $0 \leq \alpha < 1$ 인 α 에 대해 아래 기재된 수학적 식 6으로 조건부 확률을 산출한다.

수학식 6

$$\begin{aligned} P(W_i | spam) &= (1 - \alpha) \\ P(W_i | non-spam) &= (1 + \alpha) \end{aligned}$$

[0096]

[0097]

Case 4는 n-gram 단어 w_i 가 spam 또는 non-spam 사전에 모두 존재하지 않는 경우를 의미한다. 스팸 트윗의 경우 대부분 유사한 단어, 문장으로 이루어져 있기 때문에 사전에 모두 존재하지 않는다는 의미는 일반 사용자가 작성한 트윗일 가능성이 높다는 것을 의미한다. 따라서, 이런 경우에는 탐지부(126)은 수학식 6을 이용하여 조건부 확률을 계산할 수 있다. 수학식 6에 의한 계산은 결과적으로 spam일 확률은 낮추고, non-spam일 확률을 높이는 결과를 가져오게 된다.

[0098]

살핀 바와 같이, 탐지부(126)는 문자열이 스팸 사전 또는 논스팸 사전에 존재하는 경우 수학식 5를 이용하여 나이브 베이스 조건부 확률을 산출하고, 문자열이 스팸 사전과 논스팸 사전에 존재하지 않는 경우 수학식 6을 이용하여 나이브 베이스 조건부 확률을 산출할 수 있다. 이렇게 함으로써 스팸 사전과 논스팸 사전 모두에 존재하지 않는 새로운 n-gram이 등장하였을 때 논스팸을 스팸으로 예측하는 경우가 적어지게 되며, 이에 따른 분류기의 오분류율을 감소시킬 수 있다.

[0099]

다시 도 1을 참조하면, 사전 업데이트부(130)는 일정 시간 간격으로 수집된 데이터에 대해 리트윗 증가율을 산출하고, 상기 산출된 리트윗 증가율에 기초하여 각 데이터의 클래스 라벨을 예측하여, 기 구축된 사전을 업데이트한다. 이때, 사전 업데이트부(130)는 주기적으로 사전을 업데이트할 수 있다.

[0100]

즉, 사전 업데이트부(130)는 일정 시간동안 수집된 트윗들에 대하여 리트윗 증가율을 산출하고, 산출된 리트윗 증가율을 기준으로 트윗들을 내림차순으로 정렬한다. 그런 후, 사전 업데이트부(130)는 정렬된 트윗들에 대하여 기 구축된 사전 정보를 기반으로 나이브 베이스 분류기를 적용하여 상위 일정 개수의 트윗, 하위 일정 개수의 트윗에 대한 클래스 라벨 즉, 스팸 또는 논스팸을 각각 예측한다. 그런 후, 사전 업데이트부(130)는 상위 일정 개수의 트윗 중에서 스팸으로 예측된 트윗들과 하위 일정 개수의 트윗 중에서 논스팸으로 예측된 트윗들만을 사용하여 기 구축된 사전을 업데이트한다.

[0101]

이러한, 사전 업데이트부(130)는 도 8에 도시된 바와 같이 리트윗 증가율 산출부(132), 클래스 라벨 예측부(134), 전처리부(136), n-gram 분해부(138), n-gram 사전 업데이트부(140)를 포함한다.

[0102]

리트윗 증가율 산출부(132)는 일정 시간동안 수집된 데이터에 대하여 리트윗 증가율을 산출한다. 이때, 리트윗 증가율 산출부(132)는 수학식 2를 이용하여 리트윗 증가율을 산출할 수 있다.

[0103]

클래스 라벨 예측부(134)는 리트윗 증가율 산출부(132)에서 산출된 리트윗 증가율을 기준으로 트윗들을 내림차순으로 정렬하고, 정렬된 트윗들에 대해 사전 구축부에서 구축된 사전 정보를 기반으로 나이브 베이스 분류기를 적용하여 상위 일정 개수의 트윗, 하위 일정 개수의 트윗에 대한 클래스 라벨 즉, 스팸 또는 논스팸을 각각 예측한다.

[0104]

즉, 클래스 라벨 예측부(134)는 리트윗 증가율을 기준으로 내림차순으로 정렬된 트윗들 중에서 상위 일정 개수의 트윗과 하위 일정 개수의 트윗을 추출한다. 그런 후, 클래스 라벨 예측부(134)는 추출된 상위 일정 개수의 트윗과 하위 일정 개수의 트윗을 각각 전처리하고, 전처리된 트윗들을 n-gram 단위의 문자열로 분해한다. 그런 후, 클래스 라벨 예측부(134)는 분해된 각 문자열에 대해 스팸 사전 또는 논스팸 사전에 존재할 확률을 산출하고, 산출된 확률에 기초하여 각 트윗들을 스팸 또는 논스팸으로 분류한다. 즉, 클래스 라벨 예측부(134)는 각 트윗별로 각 문자열이 스팸 사전에 속할 확률의 곱과 논스팸 사전에 속할 확률의 곱을 비교하여, 해당 트윗을 스팸 또는 논스팸으로 분류한다. 이때, 스팸 사전에 속할 확률의 곱이 논스팸 사전에 속할 확률의 곱보다 큰 경우 해당 트윗을 스팸으로 분류한다. 클래스 라벨 예측부(134)가 트윗을 스팸 또는 논스팸으로 분류하는 동작은 도 7에 도시된 탐지부(126)의 동작과 동일하므로, 상세한 설명은 생략하기로 한다.

[0105]

전처리부(136)는 상위 일정 개수의 트윗 중에서 스팸으로 예측된 트윗들과 하위 일정 개수의 트윗 중에서 논스팸으로 예측된 트윗들 각각에서 리트윗 정보, 사용자 멘션정보, URL 정보, 특수문자, 공백 등을 제거하는 전처리를 수행한다. 이때, 전처리부(136)는 정규 표현식을 이용하여 전처리를 수행할 수 있다.

- [0106] n-gram 분해부(138)는 전처리된 각 트윗에 대해 n-gram 단위의 문자열로 분해한다.
- [0107] n-gram 사전 업데이트부(140)는 n-gram으로 분해된 문자열들의 빈도수를 산출하여, 스팸 사전과 논스팸 사전을 업데이트한다. 즉, n-gram 사전 업데이트부(140)는 클래스 c 에 대한 사전에서 처음 발생한 경우 $\mathbb{W}_k$ 에 대한 빈도수를 '1'로 저장하고, 이미 존재하는 경우 기존 빈도수에 '1'을 더하여 값을 갱신한다. 그러면, 문자열과 그 빈도수에 대한 정보가 저장된 스팸 사전 또는 논스팸 사전이 업데이트될 수 있다.
- [0108] 전처리부(136), n-gram 분해부(138), n-gram 사전 업데이트부(140)는 도 2에 도시된 전처리부(112), n-gram 분해부(114), n-gram 사전 구축부(116)의 동작과 동일하므로, 상세한 설명은 도 2를 참조하기로 한다.
- [0109] 한편, 본 발명에 따른 스팸 탐지 장치(100)는 데이터 수집 처리부(미도시)를 더 포함할 수 있다. 데이터 수집 처리부는 트위터의 Streaming API를 사용하여 트위터 데이터를 수집한다. Streaming API는 트위터에서 실시간으로 발생하는 트윗의 1%를 무작위로 제공하는 API이다. Streaming API를 통해 수집되는 트윗들은 트윗을 작성한 사용자와 트윗에 대한 정보를 JSON 파일의 형태로 반환하게 되며 여기서 필요한 정보만을 파싱하여 사용할 수 있다.
- [0110] 또한, 데이터 수집 처리부는 수집한 트윗들에 대해 기 정의된 개념을 이용하여 스팸 또는 논스팸으로 라벨링을 할 수 있다.
- [0111] 예컨대, 스팸에 대한 주요 카테고리가 성인 만남, 온라인 도박, 불법 약물, 피싱, 스팸 등의 항목으로 구분된 경우, 데이터 수집 처리부는 스팸에 대한 항목을 포함하는 트윗을 스팸으로 라벨링을 할 수 있다.
- [0112] 한편, 본 발명에 따른 스팸 탐지 장치(100)는 트윗의 스팸 탐지 처리를 위한 프로그램 또는 애플리케이션이 저장된 저장부(미도시)를 더 포함할 수 있다. 또한, 저장부는 n-gram 사전 구축 및 스팸 탐지와 관련된 다양한 알고리즘(또는 수학적식)을 저장할 수 있다. 이러한 경우, 제어부는 저장부를 호출하여 필요한 알고리즘을 획득할 수 있다.
- [0113] 또한, 스팸 탐지 장치(100)는 스팸 탐지 장치(100)의 다양한 구성부들의 동작을 제어하는 제어부(미도시)를 더 포함할 수 있다.
- [0114] 본 발명의 일 실시예에 따르면, 사전 구축부(110), 스팸 탐지부(120), 사전 업데이트부(130)는 그 중 적어도 일부가 외부 단말 장치나 외부 서버 등과 통신하는 프로그램 모듈들일 수 있다. 이러한 프로그램 모듈들은 운영 시스템, 응용 프로그램 모듈 및 기타 프로그램 모듈로서 스팸 탐지 장치(100)에 포함될 수 있으며, 물리적으로는 여러 가지 공지의 기억 장치상에 저장될 수 있다. 또한, 이러한 프로그램 모듈들은 스팸 탐지 장치(200)와 통신 가능한 원격 기억 장치에 저장될 수도 있다. 이러한 프로그램 모듈들은 본 발명에 따라 전술한 특정 동작을 수행하거나 특정 추상 데이터 유형을 실행하는 루틴, 서브루틴, 프로그램, 오브젝트, 컴포넌트, 데이터 구조 등을 포괄하지만, 이에 제한되지는 않는다.
- [0115] 한편, 도 1에서는 사전 구축부(110), 스팸 탐지부(120), 사전 업데이트부(130)를 하나의 장치의 구성 요소로 설명하였으나, 사전 구축부(110), 스팸 탐지부(120), 사전 업데이트부(130)를 서로 다른 별개의 장치로 각각 구성할 수도 있다. 또한, 스팸 탐지부(120)와 사전 업데이트부(130)는 일정 시간동안 수집된 데이터에 대해 스팸 탐지와 사전 업데이트 동작을 동시에 수행할 수 있다.
- [0116] 상기와 같이 구성된 스팸 탐지 장치(100)는 불건전한 정보를 포함하고 있는 음란물, 불법 도박, 약물에 대한 트윗들을 탐지할 수 있다. 불건전 정보를 포함하는 트윗의 경우 유사한 단어, 문장을 반복하여 사용하는 경향이 있기 때문에, 문장에 대한 n-gram을 기반으로 하는 분류기를 사용할 수 있다. 즉, 스팸 탐지 장치(100)는 스팸 트윗과 정상 트윗에 대한 n-gram 사전을 각각 구축하고, 나이브 베이스 분류기를 사용하여 스팸 트윗을 탐지한다. 또한, 스팸 탐지 장치(100)는 초기에 트레이닝 집합이 존재하지 않거나 매우 작은 크기의 트레이닝 집합이 있을 때, 트위터의 리트윗 특성을 활용하여 의사 라벨을 생성하고 n-gram 사전 구축은 물론, n-gram 사전을 주기적으로 업데이트할 수 있다.
- [0117] 도 9는 본 발명의 일 실시예에 따른 사전 구축 및 스팸 탐지 방법을 설명하기 위한 흐름도이다.
- [0118] 도 9를 참조하면, 스팸 탐지 장치는 스팸 트윗을 탐지하기 위해 사전을 구축한다(S910). 이때, 스팸 탐지 장치는 트레이닝 데이터의 존재 여부에 따라 다른 방법으로 사전을 구축할 수 있다. 트레이닝 데이터가 존재하는 경우 사전을 구축하는 방법에 대한 상세한 설명은 도 10을 참조하고, 트레이닝 데이터가 없는 경우 사전을 구축하는 방법에 대한 상세한 설명은 도 11을 참조하기로 한다.

- [0119] 단계 S910의 수행 후, 분류 대상 트윗이 수집되면(S920), 스팸 탐지 장치는 구축된 사전을 이용하여 분류 대상 트윗의 스팸 여부를 판단하고, 분석 대상 트윗을 이용하여 사전을 업데이트한다(S930). 이때, 분류 대상 트윗의 스팸 여부 판단동작과 사전 업데이트 동작은 동시에 수행될 수 있다.
- [0120] 스팸 탐지 장치가 분석 대상 트윗의 스팸 여부를 판단하는 방법에 대한 상세한 설명은 도 12를 참조하고, 사전을 업데이트하는 방법에 대한 상세한 설명은 도 13을 참조하기로 한다.
- [0121] 도 10은 본 발명의 일 실시예에 따른 트레이닝 데이터가 존재하는 경우 사전을 구축하는 방법을 나타낸 흐름도이다.
- [0122] 도 10을 참조하면, 스팸 탐지 장치는 트레이닝 데이터를 전처리한다(S1010). 이때, 트레이닝 데이터는 스팸, 논스팸에 대한 정보가 미리 저장되어 있고, 스팸 탐지 장치는 트레이닝 데이터에 포함된 각 트윗에서 리트윗 정보, 사용자 멘션정보, URL 정보, 특수문자, 공백 등을 제거하는 전처리를 수행한다.
- [0123] 단계 S1010이 수행되면, 스팸 탐지 장치는 전처리된 각 트레이닝 데이터에 대해 n-gram 단위의 문자열로 분해한다(S1020). 즉, 스팸 탐지 장치는 전처리 과정을 마친 트윗에 대해 n-gram의 문자열(또는 문자)로 분해한다.
- [0124] 단계 S1020이 수행되면, 스팸 탐지 장치는 n-gram으로 분해된 문자열들의 빈도수를 산출하여, 스팸 사전과 논스팸 사전을 구축한다(S1030). 즉, 스팸 탐지 장치는 사전에 처음 발생한 문자열의 경우 빈도수를 '1'로 저장하고, 사전에 이미 존재하는 문자열의 경우 기존 빈도수에 '1'을 더하여, 문자열과 빈도수로 구성된 스팸 사전 또는 논스팸 사전을 구축할 수 있다. 이때, 트레이닝 데이터에는 스팸 또는 논스팸에 대한 정보가 미리 저장되어 있으므로, 스팸 탐지 장치는 스팸으로 분류된 트윗들을 이용하여 스팸 사전을 구축하고, 논스팸으로 분류된 트윗들을 이용하여 논스팸 사전을 구축할 수 있다.
- [0125] 도 11은 본 발명의 일 실시예에 따른 트레이닝 데이터가 없는 경우 사전을 구축하는 방법을 나타낸 흐름도이다.
- [0126] 도 11을 참조하면, 스팸 탐지 장치는 일정 시간동안 수집된 트윗에 대하여 리트윗 증가율을 산출한다(S1110). 즉, 스팸 탐지 장치는 특정 시간 구간에서 원본 트윗의 리트윗에 대한 순간 증가율을 산출한다. 그런 후, 스팸 탐지 장치는 각 시간 구간에서 산출된 순간 증가율의 합을 원본 트윗을 작성한 사용자의 팔로워 수와 누적된 총 리트윗 수의 곱으로 나누어 리트윗 증가율을 산출할 수 있다.
- [0127] 단계 S1110이 수행되면, 스팸 탐지 장치는 리트윗 증가율 값을 기준으로 트윗들을 내림차순 정렬하고(S1120), 정렬된 트윗들중에서 상위 일정 개수의 트윗과 하위 일정 개수의 트윗을 트레이닝 데이터로 선정한다(S1130). 이때, 스팸 탐지 장치는 정렬된 트윗들에 대해 초기 트레이닝 집합 구성을 위해 상위 일정 개수의 트윗들을 스팸 트레이닝 집합으로 사용하고, 하위 일정 개수의 트윗들을 논스팸 트레이닝 집합으로 사용할 수 있다.
- [0128] 스팸 탐지 장치는 단계 S1110과 단계 S1120을 통해, 사전 구축을 위한 트레이닝 데이터를 생성할 수 있고, 상위 일정 개수의 트윗들은 스팸, 하위 일정 개수의 트윗들은 논스팸으로 의사 라벨을 생성할 수 있다. 이후, 스팸 탐지 장치는 스팸으로 라벨링된 상위 일정 개수의 트윗들을 이용하여 스팸 사전을 구축할 수 있고, 논스팸으로 라벨링된 하위 일정 개수의 트윗들을 이용하여 논스팸 사전을 구축할 수 있다.
- [0129] 단계 S1130이 수행되면, 스팸 탐지 장치는 스팸 트레이닝 데이터와 논스팸 트레이닝 데이터를 각각 전처리하고(S1140), 전처리된 각 트레이닝 데이터를 n-gram의 문자열로 각각 분해한다(S1150).
- [0130] 그런 후, 스팸 탐지 장치는 n-gram으로 분해된 문자열들의 빈도수를 산출하여, 스팸 사전 또는 논스팸 사전을 구축한다(S1160).
- [0131] 도 12는 본 발명의 일 실시예에 따른 분류 대상 트윗의 스팸 탐지 방법을 나타낸 흐름도이다.
- [0132] 도 12를 참조하면, 스팸 탐지 장치는 분류 대상 트윗을 전처리하고(S1210), 전처리된 트윗을 n-gram의 문자열로 분해한다(S1220).
- [0133] 그런 후, 스팸 탐지 장치는 각 문자열에 대해 나이브 베이스 분류기를 적용하여 스팸 사전 또는 논스팸 사전에 존재할 확률을 산출한다(S1230). 즉, 이때, 스팸 탐지 장치는 해당 문자열이 스팸 사전 또는 논스팸 사전에 존재하는 경우, 수학적 5를 이용하여 스팸 사전에 속할 확률 또는 논스팸 사전에 속할 확률을 산출하고, 해당 문자열이 스팸 사전 또는 논스팸 사전에 존재하지 않은 경우, 수학적 6을 이용하여 스팸 사전에 속할 확률 또는 논스팸 사전에 속할 확률을 산출할 수 있다.
- [0134] 단계 S1230이 수행되면, 스팸 탐지 장치는 산출된 확률에 기초하여 분류 대상 트윗을 스팸 또는 논스팸으로 분

류한다(S1240). 즉, 스팸 탐지 장치는 각 문자열이 스팸 사전에 속할 확률의 곱과 논스팸 사전에 속할 확률의 곱을 비교하여, 스팸 사전에 속할 확률의 곱이 논스팸 사전에 속할 확률의 곱보다 큰 경우, 분류 대상 트윗을 스팸으로 분류할 수 있다.

[0135] 도 13은 본 발명의 일 실시예에 따른 사전 업데이트 방법을 나타낸 흐름도이다.

[0136] 도 13을 참조하면, 스팸 탐지 장치는 일정 시간동안 수집된 트윗들에 대하여 리트윗 증가율을 산출하고(S1310), 산출된 리트윗 증가율을 기준으로 트윗들을 내림차순으로 정렬한다(S1320).

[0137] 그런 후, 스팸 탐지 장치는 정렬된 트윗들에 대해 기 구축된 사전 정보를 기반으로 나이브 베이스 분류기를 적용하여 상위 일정 개수의 트윗과 하위 일정 개수 트윗에 대한 클래스 라벨 즉, 스팸 또는 논스팸을 각각 예측한다(S1330). 즉, 스팸 탐지 장치는 내림차순으로 정렬된 트윗들 중에서 상위 일정 개수의 트윗과 하위 일정 개수의 트윗을 추출하고, 추출된 상위 일정 개수의 트윗과 하위 일정 개수의 트윗을 각각 전처리하며, 전처리된 트윗들을 n-gram 단위의 문자열로 분해한다. 그런 후, 스팸 탐지 장치는 분해된 각 문자열에 대해 스팸 사전 또는 논스팸 사전에 존재할 확률을 산출하고, 산출된 확률에 기초하여 각 트윗들을 스팸 또는 논스팸으로 분류한다. 즉, 스팸 탐지 장치는 각 트윗별로 각 문자열에 대해 스팸 사전에 속할 확률의 곱과 논스팸 사전에 속할 확률의 곱을 비교하여, 스팸 사전에 속할 확률의 곱이 논스팸 사전에 속할 확률의 곱보다 큰 경우 해당 트윗을 스팸으로 분류한다.

[0138] 수집된 트윗들은 스팸 또는 논스팸인지에 대한 정보가 없기 때문에, 스팸 탐지 장치는 단계 S1310부터 단계 S1330을 수행하여 각 트윗들을 스팸 또는 논스팸으로 분류할 수 있다. 트윗들이 스팸과 논스팸으로 분류되면, 스팸 탐지 장치는 스팸으로 분류된 트윗들을 이용하여 스팸 사전을 업데이트할 수 있고, 논스팸으로 분류된 트윗들을 이용하여 논스팸 사전을 업데이트할 수 있다.

[0139] 단계 S1330이 수행되면, 스팸 탐지 장치는 상위 일정 개수의 트윗 중에서 스팸으로 예측된 트윗들과 하위 일정 개수의 트윗 중에서 논스팸으로 예측된 트윗들 각각에서 리트윗 정보, 사용자 멘션정보, URL 정보, 특수문자, 공백 등을 제거하는 전처리를 수행한다(S1340).

[0140] 런 후, 스팸 탐지 장치는 전처리된 각 트윗에 대해 n-gram 단위의 문자열로 분해한다(S1350), n-gram으로 분해된 문자열들의 빈도수를 산출하여, 스팸 사전과 논스팸 사전을 업데이트한다(S1360).

[0141] 이제까지 본 발명에 대하여 그 바람직한 실시예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

부호의 설명

[0142] 100 : 스팸 탐지 장치

110 : 사전 구축부

111 : 트레이닝 집합 생성부

112, 122, 136 : 전처리부

114, 124, 138 : n-gram 분해부

116 : n-gram 사전 구축부

120 : 스팸 탐지부

130 : 사전 업데이트부

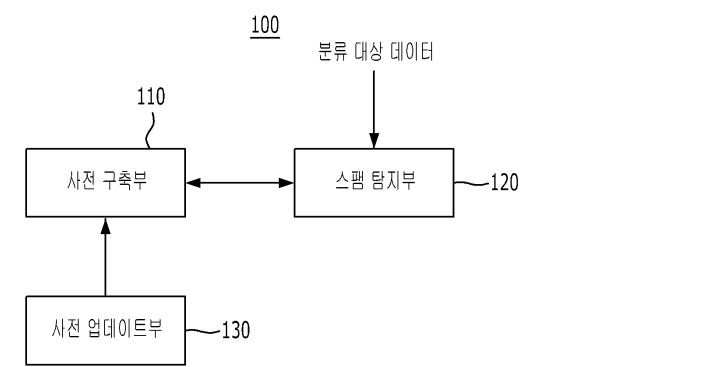
132 : 리트윗 증가율 산출부

134 : 클래스 라벨 예측부

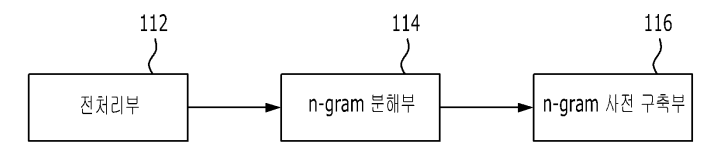
140 : n-gram 사전 업데이트부

도면

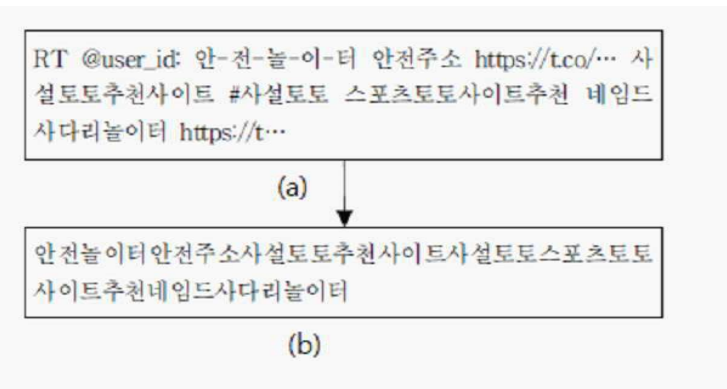
도면1



도면2



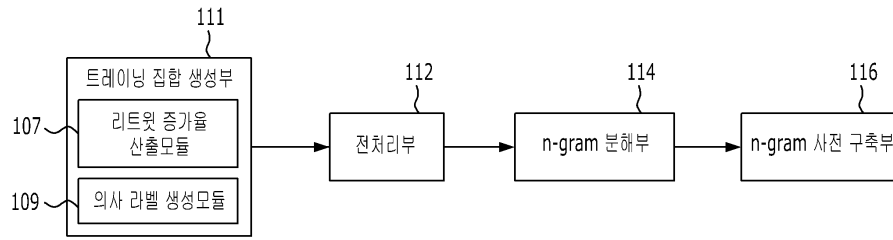
도면3



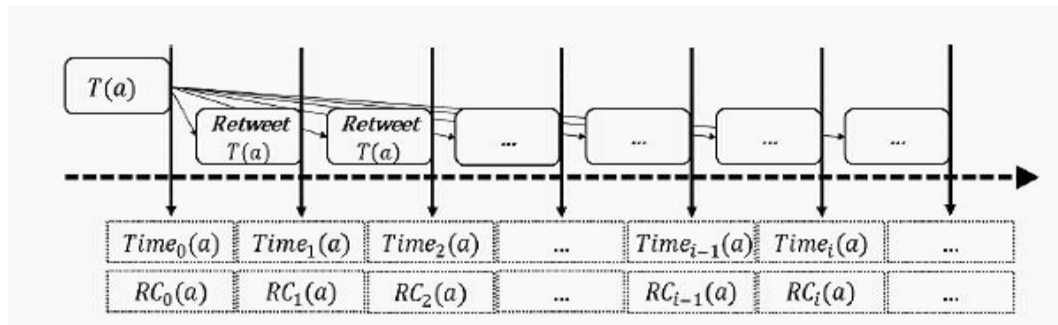
도면4

Sequence	n-gram
W_0	안전놀이터안
W_1	전놀이터안전
W_2	놀이터안전주
W_3	이터안전주소
...	...
W_{k-1}	드사다리놀이
W_k	사다리놀이터

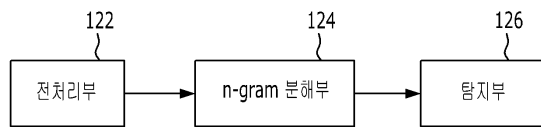
도면5



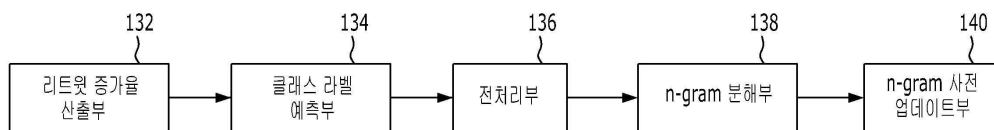
도면6



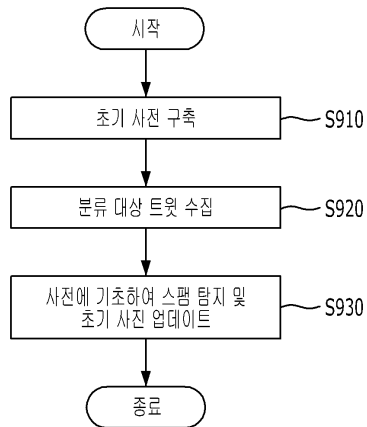
도면7



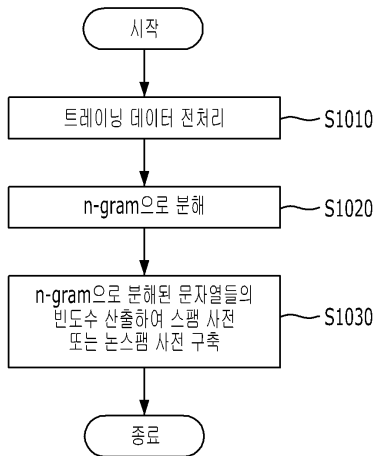
도면8



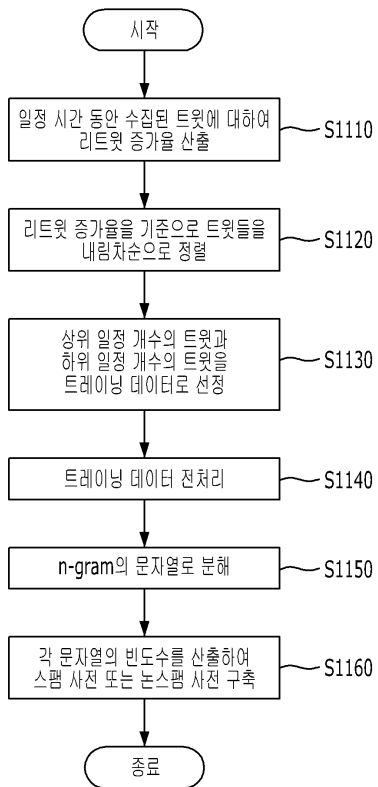
도면9



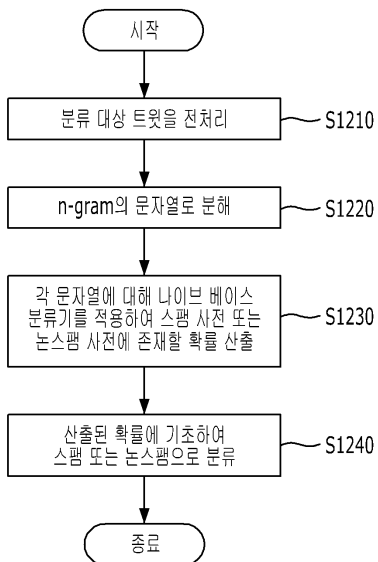
도면10



도면11



도면12



도면13

