(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2011년11월16일
(11) 등록번호 10-1083476
(24) 등록일자 2011년11월08일

(51) Int. Cl.
G06F 17/21 (2006.01) *G06F 17/30* (2006.01)
(21) 출원번호 10-2009-0071254
(22) 출원일자 2009년08월03일
심사청구일자 2009년08월03일
(65) 공개번호 10-2011-0013674
(43) 공개일자 2011년02월10일
(56) 선행기술조사문헌
KR1020070085477 A*
KR1020050027944 A
JP2007183825 A
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
엔에이치엔(주)
경기 성남시 분당구 정자동 178-1 그린팩토리
(72) 발명자
이상호
경기도 안양시 동안구 평촌동 향촌롯데아파트 301동 1201호
박종대
서울시 관악구 남현동 1085-33 우일빌라 B-B02
(뒷면에 계속)
(74) 대리인
특허법인무한

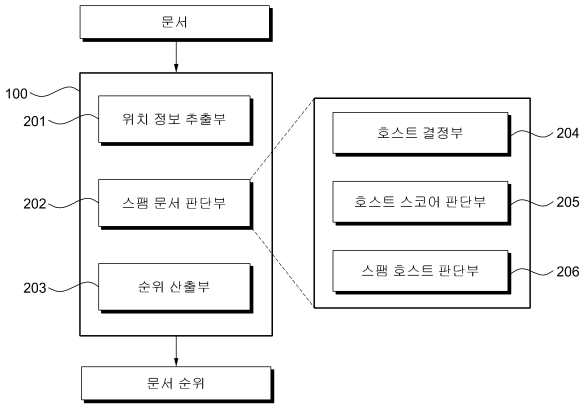
전체 청구항 수 : 총 21 항 심사관 : 천대녕

(54) 문서의 위치 정보를 이용한 문서 순위 산출 시스템 및 방법

(57) 요 약

문서의 위치 정보를 이용한 문서 순위 산출 시스템 및 방법이 개시된다. 문서 순위 산출 시스템은 링크를 통해 제1 문서를 가리키는 적어도 하나의 제2 문서의 위치 정보를 추출하는 위치 정보 추출부, 상기 제2 문서의 위치 정보에 기초하여 상기 제2 문서가 스팸 문서인지 판단하는 스팸 문서 판단부 및 상기 스팸 문서로 판단된 제2 문서를 고려하여 상기 제1 문서의 순위를 산출하는 순위 산출부를 포함할 수 있다.

대 표 도 - 도2



(72) 발명자

배재현

서울시 강남구 역삼동 709 성보아파트 A동 1106호

김훈

경기도 용인시 수지구 죽전동 새터마을 현대홈타운 713동 2301호

김창봉

서울시 송파구 잠실동 44번지 레이크팰리스아파트 130동 2501호

특허청구의 범위

청구항 1

링크를 통해 제1 문서를 가리키는 적어도 하나의 제2 문서의 위치 정보를 추출하는 위치 정보 추출부;

상기 제2 문서의 위치 정보에 기초하여 상기 제2 문서가 스팸 문서인지 판단하는 스팸 문서 판단부; 및

상기 스팸 문서로 판단된 제2 문서를 고려하여 상기 제1 문서의 순위를 산출하는 순위 산출부

를 포함하고,

상기 스팸 문서 판단부는,

상기 제2 문서의 위치 정보에 기초하여 상기 제2 문서를 제공하는 적어도 하나의 호스트를 결정하는 호스트 결정부;

상기 결정된 호스트에 대한 링크 정보와 관련된 정보인 특징 벡터를 추출하여 상기 호스트에 대한 호스트 스코어를 계산하는 호스트 스코어 계산부; 및

상기 호스트 스코어에 기초하여 상기 호스트가 스팸 호스트인지 여부를 판단하는 스팸 호스트 판단부

를 포함하는 문서 순위 산출 시스템.

청구항 2

제1항에 있어서,

상기 위치 정보 추출부는,

상기 제2 문서의 IP 주소, 도메인 정보 또는 호스트 정보 중 어느 하나에 기초하여 상기 제2 문서의 위치 정보를 판단하는 것을 특징으로 하는 문서 순위 산출 시스템.

청구항 3

삭제

청구항 4

삭제

청구항 5

제1항에 있어서,

상기 호스트 스코어 계산부는,

상기 제2 문서를 제공하는 적어도 하나의 호스트로부터 추출한 특징 벡터들 각각에 가중치를 적용하여 호스트 스코어를 계산하는 것을 특징으로 하는 문서 순위 산출 시스템.

청구항 6

제5항에 있어서,

상기 호스트 스코어 계산부는,

스팸 호스트 및 정상 호스트로 라벨링된 학습 데이터로부터 추출된 특징 벡터들 각각의 가중치를 결정하는 것을 특징으로 하는 문서 순위 산출 시스템.

청구항 7

제6항에 있어서,

상기 가중치는,

스팸 호스트의 호스트 스코어가 미리 설정한 최저값이 되도록 결정되고, 정상 호스트의 호스트 스코어가 미리

설정한 최대값이 되도록 결정되는 것을 특징으로 하는 문서 순위 산출 시스템.

청구항 8

제1항에 있어서,

상기 특징 벡터는,

상기 제2 문서를 제공하는 적어도 하나의 호스트 별 수집 문서 정보, 인/아웃 링크 정보, 인/아웃 호스트 정보, 아웃 호스트의 인 호스트 정보, 인 호스트의 아웃 호스트 정보, 아웃 호스트의 인 링크 정보, 인 호스트의 아웃 링크 정보 또는 인/아웃 도메인 정보 중 적어도 하나를 포함하는 것을 특징으로 하는 문서 순위 산출 시스템.

청구항 9

제1항에 있어서,

상기 스팸 호스트 판단부는,

상기 호스트 스코어 중 미리 설정한 기준 스코어 미만인 호스트를 스팸 호스트로 판단하는 것을 특징으로 하는 문서 순위 산출 시스템.

청구항 10

제1항에 있어서,

상기 스팸 호스트 판단부는,

상기 호스트에 대한 아웃 호스트 또는 인 호스트의 호스트 스코어에 기초하여 상기 호스트를 스팸 호스트로 판단하는 것을 특징으로 하는 문서 순위 산출 시스템.

청구항 11

제1항에 있어서,

상기 스팸 문서 판단부는,

상기 호스트 스코어를 적용한 경우 및 적용하지 않은 경우의 순위 변화량을 계산하여, 상기 계산된 변화량을 이용하여 스팸 문서인지 여부를 판단하는 것을 특징으로 하는 문서 순위 산출 시스템.

청구항 12

제1항에 있어서,

상기 순위 산출부는,

동일한 위치 정보를 나타내는 상기 적어도 하나의 제2 문서들을 그룹화하여 상기 제1 문서의 순위를 산출하는 것을 특징으로 하는 문서 순위 산출 시스템.

청구항 13

링크를 통해 제1 문서를 가리키는 적어도 하나의 제2 문서의 위치 정보를 추출하는 단계;

상기 제2 문서의 위치 정보에 기초하여 상기 제2 문서가 스팸 문서인지 판단하는 단계; 및

상기 스팸 문서로 판단된 제2 문서를 고려하여 상기 제1 문서의 순위를 산출하는 단계

를 포함하고,

상기 제2 문서가 스팸 문서인지 판단하는 단계는,

상기 제2 문서의 위치 정보에 기초하여 상기 제2 문서를 제공하는 적어도 하나의 호스트를 결정하는 단계;

상기 결정된 호스트에 대한 링크 정보와 관련된 정보인 특징 벡터를 추출하여 상기 호스트에 대한 호스트 스코어를 계산하는 단계; 및

상기 호스트 스코어에 기초하여 상기 호스트가 스팸 호스트인지 여부를 판단하는 단계

를 포함하는 문서 순위 산출 방법.

청구항 14

제13항에 있어서,

상기 적어도 하나의 제2 문서의 위치 정보를 추출하는 단계는,

상기 제2 문서의 IP 주소, 도메인 정보 또는 호스트 정보 중 어느 하나에 기초하여 상기 제2 문서의 위치 정보를 판단하는 것을 특징으로 하는 문서 순위 산출 방법.

청구항 15

삭제

청구항 16

삭제

청구항 17

제13항에 있어서,

상기 호스트에 대한 호스트 스코어를 계산하는 단계는,

상기 제2 문서를 제공하는 적어도 하나의 호스트로부터 추출한 특징 벡터들 각각에 가중치를 적용하여 호스트 스코어를 계산하는 것을 특징으로 하는 문서 순위 산출 방법.

청구항 18

제17항에 있어서,

상기 호스트에 대한 호스트 스코어를 계산하는 단계는,

스팸 호스트 및 정상 호스트로 라벨링된 학습 데이터로부터 추출된 특징 벡터들 각각의 가중치를 결정하는 것을 특징으로 하는 문서 순위 산출 방법.

청구항 19

제18항에 있어서,

상기 가중치는,

스팸 호스트의 호스트 스코어가 미리 설정한 최저값이 되도록 결정되고, 정상 호스트의 호스트 스코어가 미리 설정한 최대값이 되도록 결정되는 것을 특징으로 하는 문서 순위 산출 방법.

청구항 20

제13항에 있어서,

상기 특징 벡터는,

상기 제2 문서를 제공하는 적어도 하나의 호스트 별 수집 문서 정보, 인/아웃 링크 정보, 인/아웃 호스트 정보, 아웃 호스트의 인 호스트 정보, 인 호스트의 아웃 호스트 정보, 아웃 호스트의 인 링크 정보, 인 호스트의 아웃 링크 정보 또는 인/아웃 도메인 정보 중 적어도 하나를 포함하는 것을 특징으로 하는 문서 순위 산출 방법.

청구항 21

제13항에 있어서,

상기 호스트가 스팸 호스트인지 여부를 판단하는 단계는,

상기 호스트 스코어 중 미리 설정한 기준 스코어 미만인 호스트를 스팸 호스트로 판단하는 것을 특징으로 하는 문서 순위 산출 방법.

청구항 22

제13항에 있어서,

상기 호스트가 스팸 호스트인지 여부를 판단하는 단계는,

상기 호스트에 대한 아웃 호스트 또는 인 호스트의 호스트 스코어에 기초하여 상기 호스트를 스팸 호스트로 판단하는 것을 특징으로 하는 문서 순위 산출 방법.

청구항 23

제13항에 있어서,

상기 제2 문서가 스팸 문서인지 판단하는 단계는,

상기 호스트 스코어를 적용한 경우 및 적용하지 않은 경우의 순위 변화량을 계산하여, 상기 계산된 변화량을 이용하여 스팸 문서인지 여부를 판단하는 것을 특징으로 하는 문서 순위 산출 방법.

청구항 24

제13항에 있어서,

상기 제1 문서의 순위를 산출하는 단계는,

동일한 위치 정보를 나타내는 상기 적어도 하나의 제2 문서들을 그룹화하여 상기 제1 문서의 순위를 산출하는 것을 특징으로 하는 문서 순위 산출 방법.

청구항 25

제13항, 제14항, 제17항 내지 제24항 중 어느 한 항의 방법을 실행하기 위한 프로그램이 기록된 컴퓨터에서 판독 가능한 기록 매체.

명 세 서

발명의 상세한 설명

기 술 분 야

[0001] 본 발명은 문서 순위를 산출하는 시스템 및 방법에 관한 것으로, 보다 자세하게는, 문서의 링크 구조에 따른 위치 정보를 이용하여 문서의 순위를 산출하는 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 사용자가 정보를 검색하기 위해 키워드를 입력하는 경우, 키워드에 대응하는 문서가 도출된다. 이러한 문서는 사용자가 검색하려는 목적에 따라 순위가 산출되어 결과 리스트로 정렬될 수 있다.

[0003] 종래에 문서의 순위는 문서의 중요도에 따라 결정되었다. 구체적으로, 문서의 중요도는 문서와 링크로 연결된 문서의 개수에 기초하여 결정되는 방식이 사용되었다. 그러나, 문서의 순위를 상승시키기 위해, 인위적으로 해당 문서를 가리키는 연결을 가진 무의미한 문서들을 생성하여 악의적으로 문서의 중요도를 증가시키는 문제점이 존재하였다.

[0004] 이러한 스패머들은 경제적인 웹 검색 트래픽을 얻기 위해 무의미한 문서인 스팸 문서를 링크를 통해 무제한적으로 확장시킴으로써 자신의 문서의 순위를 상승시키고, 이로 인해 결과 리스트에 최상위로 노출시키려고 하였다. 이 경우, 사용자는 자신이 검색하고자 하는 문서와는 전혀 무관한 스팸 문서를 획득하기 때문에, 검색 만족도가 떨어질 수 있다.

[0005] 따라서, 특정 문서의 순위를 결정할 때 스팸을 위해 링크로 연결한 문서를 고려하여 보다 정확한 문서 순위를 산출하는 방법이 요구된다.

발명의 내용

해결 하고자하는 과제

[0006] 본 발명은 스팸을 위해 인위적으로 링크를 생성한 대상 문서의 순위를 감소시키기 위해, 대상 문서와 링크로 연결된 연결 문서의 위치 정보를 통해 스팸 문서인지 판단한 후, 스팸 문서로 판단된 연결 문서에 패널티를 부과하는 문서 순위 산출 시스템 및 방법을 제공한다.

[0007] 본 발명은 연결 문서를 제공하는 호스트의 호스트 스코어를 이용하여 스팸 호스트인지 판단하고, 스팸 호스트가 제공하는 연결 문서를 스팸 문서로 판단하여 대상 문서의 순위를 산출하는 문서 순위 산출 시스템 및 방법을 제공한다.

[0008] 본 발명은 스팸 호스트/정상 호스트로 라벨링된 학습 데이터를 통해 기계 학습하여 호스트로부터 추출된 특징 벡터의 가중치를 결정함으로써, 호스트 스코어의 정확도를 향상시키는 문서 순위 산출 시스템 및 방법을 제공한다.

과제 해결수단

[0009] 본 발명의 일실시예에 따른 문서 순위 산출 시스템은 링크를 통해 제1 문서를 가리키는 적어도 하나의 제2 문서의 위치 정보를 추출하는 위치 정보 추출부, 상기 제2 문서의 위치 정보에 기초하여 상기 제2 문서가 스팸 문서인지 판단하는 스팸 문서 판단부 및 상기 스팸 문서로 판단된 제2 문서를 고려하여 상기 제1 문서의 순위를 산출하는 순위 산출부를 포함할 수 있다.

[0010] 본 발명의 일측면에 따르면, 스팸 문서 판단부는 상기 제2 문서를 제공하는 적어도 하나의 호스트를 결정하는 호스트 결정부, 상기 결정된 호스트의 링크 구조에 따라 특징 벡터를 추출하여 상기 호스트에 대한 호스트 스코어를 계산하는 호스트 스코어 계산부 및 상기 호스트 스코어에 기초하여 상기 호스트가 스팸 호스트인지 여부를 판단하는 스팸 호스트 판단부를 포함할 수 있다.

[0011] 본 발명의 일실시예에 따른 문서 순위 산출 방법은 링크를 통해 제1 문서를 가리키는 적어도 하나의 제2 문서의 위치 정보를 추출하는 단계, 상기 제2 문서의 위치 정보에 기초하여 상기 제2 문서가 스팸 문서인지 판단하는 단계 및 상기 스팸 문서로 판단된 제2 문서를 고려하여 상기 제1 문서의 순위를 산출하는 단계를 포함할 수 있다.

[0012] 본 발명의 일측면에 따른 제2 문서가 스팸 문서인지 판단하는 단계는 상기 제2 문서를 제공하는 적어도 하나의 호스트를 결정하는 단계, 상기 결정된 호스트의 링크 구조에 따라 특징 벡터를 추출하여 상기 호스트에 대한 호스트 스코어를 계산하는 단계 및 상기 호스트 스코어에 기초하여 상기 호스트가 스팸 호스트인지 여부를 판단하는 단계를 포함할 수 있다.

효 과

[0013] 본 발명의 일실시예에 따르면, 스팸을 위해 인위적으로 링크를 생성한 대상 문서의 순위를 감소시키기 위해, 대상 문서와 링크로 연결된 연결 문서의 위치 정보를 통해 스팸 문서인지 판단한 후, 스팸 문서로 판단된 연결 문서에 패널티를 부과하는 문서 순위 산출 시스템 및 방법이 제공된다.

[0014] 본 발명의 일실시예에 따르면, 연결 문서를 제공하는 호스트의 호스트 스코어를 이용하여 스팸 호스트인지 판단하고, 스팸 호스트가 제공하는 연결 문서를 스팸 문서로 판단하여 대상 문서의 순위를 산출하는 문서 순위 산출 시스템 및 방법이 제공된다.

[0015] 본 발명은 스팸 호스트/정상 호스트로 라벨링된 학습 데이터를 통해 기계 학습하여 호스트로부터 추출된 특징 벡터의 가중치를 결정함으로써, 호스트 스코어의 정확도를 향상시키는 문서 순위 산출 시스템 및 방법이 제공된다.

발명의 실시를 위한 구체적인 내용

[0016] 이하, 첨부된 도면들에 기재된 내용들을 참조하여 본 발명에 따른 실시예를 상세하게 설명한다. 다만, 본 발명이 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조부호는 동일한 부재를 나타낸다.

[0017] 도 1은 본 발명의 일실시예에 따른 문서 순위 산출 시스템을 통해 문서의 순위를 산출하는 전체 과정을 도시한 블록 다이어그램이다.

[0018] 본 발명의 일실시예에 따른 문서 순위 산출 시스템(100)은 적어도 하나의 문서들(문서 1~문서 N)을 수신하여, 문서 순위를 산출할 수 있다. 일례로, 문서 순위 산출 시스템(100)은 적어도 하나의 문서들 각각의 링크 구조를 고려하여 문서 순위를 산출할 수 있다. 이 때, 문서 순위 산출 시스템(100)은 순위를 산출하고자 하는 문서에 스팸 문서로 판별된 문서가 연결되어 있는 지 판단하여 문서의 순위를 산출할 수 있다.

[0019] 본 발명의 일실시예에 따르면, 링크를 통해 연결된 문서가 스팸 문서인지 여부를 판단하기 위해서 문서 순위 산출 시스템(100)은 문서의 위치 정보를 이용할 수 있다. 일례로, 문서 순위 산출 시스템(100)은 문서의 위치 정보 중 문서를 제공하는 호스트에 기초하여 문서가 스팸 문서인지 여부를 판단할 수 있다. 구체적으로, 문서 순위 산출 시스템(100)은 문서를 제공하는 호스트에 대해 호스트 스코어를 산출하고, 산출된 호스트 스코어를 이용하여 호스트가 스팸 호스트인지 판단할 수 있다. 그러면, 문서 순위 산출 시스템(100)은 스팸 호스트가 제공하는 문서는 스팸 문서로 판단하여, 문서의 순위를 산출할 수 있다.

[0020] 도 2는 본 발명의 일실시예에 따른 문서 순위 산출 시스템의 상세 구성을 도시한 블록 다이어그램이다.

[0021] 도 2를 참고하면, 문서 순위 산출 시스템(100)은 위치 정보 추출부(201), 스팸 문서 판단부(202) 및 순위 산출부(203)를 포함할 수 있다.

[0022] 위치 정보 추출부(201)는 링크를 통해 제1 문서를 가리키는 적어도 하나의 제2 문서의 위치 정보를 추출할 수 있다. 이 때, 제1 문서는 순위를 산출하려는 문서를 의미할 수 있다. 그리고, 적어도 하나의 제2 문서는 제1 문서와 링크로 연결된 문서를 의미할 수 있다. 이 때, 적어도 하나의 제2 문서 각각은 링크를 통해 제1 문서와 다른 문서와 연결된 문서를 의미할 수 있다.

[0023] 일례로, 위치 정보 추출부(201)는 제2 문서의 IP 주소, 도메인 정보 또는 호스트 정보 중 어느 하나에 기초하여 제2 문서의 위치 정보를 판단할 수 있다. 본 발명의 일실시예에 따르면, 문서 정보는 제1 문서와 링크를 통해 연결된 제2 문서의 위치 상 분포를 의미할 수 있다. 예를 들어, 문서 정보는 제2 문서가 어느 IP를 통해 작성되었는 지, 어느 도메인을 통해 노출되는 지 또는 어느 호스트를 통해 제공되는 지 여부를 의미할 수 있다.

[0024] 스팸 문서 판단부(202)는 제2 문서의 위치 정보에 기초하여 제2 문서가 스팸 문서인지 판단할 수 있다. 일례로, 스팸 문서 판단부(201)는 제2 문서의 IP, 도메인 또는 호스트를 분석하여 제2 문서가 스팸 문서인지 판단할 수 있다. 예를 들어, 특정 IP를 통해 미리 설정한 기준 개수 이상의 제2 문서가 제공되면, 상기 IP에 대응하는 제공자를 스패머로 판단할 수 있다. 그러면, 스패머가 제공하는 제2 문서들을 모두 스팸 문서로 판단하여 제1 문서의 순위를 산출할 때 고려할 수 있다.

[0025] 도 2를 참고하면, 스팸 문서 판단부(202)는 호스트 결정부(204), 호스트 스코어 계산부(205) 및 스팸 호스트 판단부(206)를 포함할 수 있다.

[0026] 호스트 결정부(204)는 제2 문서를 제공하는 적어도 하나의 호스트를 결정할 수 있다. 일례로, 호스트 결정부(204)는 제2 문서의 위치 정보를 이용하여 제2 문서를 제공하는 적어도 하나의 호스트를 결정할 수 있다.

[0027] 호스트 스코어 계산부(205)는 결정된 호스트의 링크 구조에 따라 특징 벡터를 추출하여 호스트에 대한 호스트 스코어를 계산할 수 있다. 일례로, 호스트 스코어 계산부(205)는 제2 문서를 제공하는 적어도 하나의 호스트로부터 추출한 특징 벡터를 각각에 가중치를 적용하여 호스트 스코어를 계산할 수 있다.

[0028] 그리고, 임의의 호스트에 대해 호스트 스코어를 계산하기 전에, 호스트 스코어 계산부(205)는 스팸 호스트 및 정상 호스트로 라벨링된 학습 데이터로부터 추출된 특징 벡터들 각각의 가중치를 결정할 수 있다. 즉, 호스트 스코어 계산부(205)는 학습 데이터를 미리 학습하여 특징 벡터를 각각에 대한 가중치를 결정하고, 결정된 가중치를 이용하여 제2 문서의 호스트가 스팸 호스트인 지 여부를 판단할 수 있다.

[0029] 이 때, 가중치는 스팸 호스트의 호스트 스코어가 미리 설정한 최저값이 되도록 결정되고, 정상 호스트의 호스트 스코어가 미리 설정한 최대값이 되도록 결정될 수 있다. 예를 들면, 가중치는 스팸 호스트와 유사한 링크 구조를 가지고 있는 호스트에 대해서는 호스트 스코어가 0.0이 되도록 결정될 수 있고, 정상 호스트와 유사한 링크 구조를 가지고 있는 호스트에 대해서는 호스트 스코어가 1.0이 되도록 결정될 수 있다.

[0030] 일례로, 특징 벡터는 제2 문서를 제공하는 호스트에 대한 링크 정보와 관련된 정보일 수 있다. 구체적으로, 특징 벡터는 제2 문서를 제공하는 적어도 하나의 호스트 별 수집 문서 정보, 인/아웃 링크 정보(in/out link), 인/아웃 호스트 정보(in/out host), 아웃 호스트의 인 호스트 정보, 인 호스트의 아웃 호스트 정보, 아웃 호스트의 인 링크 정보, 인 호스트의 아웃 링크 정보 또는 인/아웃 도메인 정보(in/out domain) 중 적어도 하나를 포함할 수 있다. 이 때, "인"이라는 의미는 링크를 통해 상기 호스트를 가리키는 것을 의미하고, "아웃"이라는

의미는 링크를 통해 상기 호스트로부터 외부로 향하는 것을 의미할 수 있다. 특징 벡터의 예는 호스트의 링크 구조에 따라 달라질 수 있다.

[0031] 스팸 호스트 판단부(206)는 호스트 스코어에 기초하여 제2 문서를 제공하는 호스트가 스팸 호스트인지 여부를 판단할 수 있다. 일례로, 스팸 호스트 판단부(206)는 호스트 스코어 중 미리 설정한 기준 스코어 미만인 호스트를 스팸 호스트로 판단할 수 있다. 예를 들어, 스팸 호스트 판단부(206)는 호스트 스코어가 0.5 미만인 호스트를 스팸 호스트로 판단할 수 있다.

[0032] 다른 일례로, 스팸 호스트 판단부(206)는 호스트에 대한 아웃 호스트 또는 인 호스트의 호스트 스코어에 기초하여 호스트를 스팸 호스트로 판단할 수 있다. 즉, 스팸 호스트 판단부(206)는 제2 문서를 제공하는 호스트에 대한 아웃 호스트 또는 인 호스트들의 호스트 스코어가 어떻게 분포하는 지에 기초하여 상기 호스트가 스팸 호스트인 지 여부를 판단할 수 있다. 예를 들어, 스팸 호스트 판단부(206)는 제2 문서를 제공하는 호스트에 대해 호스트 스코어가 0.5 이하인 아웃 호스트의 비율이 90%이상인 경우, 상기 호스트를 스팸 호스트로 판단할 수 있다.

[0033] 결국, 스팸 문서 판단부(202)는 스팸 호스트가 제공하는 제2 문서를 스팸 문서로 판단할 수 있다. 일례로, 스팸 문서 판단부(201)는 호스트 스코어를 적용한 경우 및 호스트 스코어를 적용하지 않은 경우의 순위 변화량을 계산하고, 계산된 변화량을 이용하여 제2 문서가 스팸 문서인지 여부를 판단할 수 있다.

[0034] 순위 산출부(203)는 스팸 문서로 판단된 제2 문서를 고려하여 제1 문서의 순위를 산출할 수 있다. 일례로, 순위 산출부(203)는 동일한 위치 정보를 나타내는 적어도 하나의 제2 문서들을 그룹화하여 제1 문서의 순위를 산출할 수 있다. 즉, 순위 산출부(203)는 동일한 위치 정보를 나타내는 적어도 하나의 제2 문서들을 하나의 문서로 간주하여 제1 문서의 순위를 산출할 수 있다. 또는, 순위 산출부(203)는 그룹화된 문서들의 링크를 제거하거나 또는 패널티를 부여할 수 있다.

[0035] 그러면, 순위 산출부(203)는 하기 수학식 1에 따라 제1 문서의 순위를 산출할 수 있다.

수학식 1

[0036] $PR(A)=(1-d)+d(PR(T1)/C(T1)+\cdots+PR(Tn)/C(Tn))$

[0037] 이 때, PR(A)은 제1 문서의 순위, PR(Ti)은 제1 문서를 가리키는 제2 문서의 순위, C(Ti)는 제2 문서의 외부로 향하는 연결의 개수, d는 0~1 사이의 변수(완충 팩터)를 의미할 수 있다.

[0038] 도 3은 본 발명의 일실시예에 따른 문서의 링크 구조를 통해 문서의 순위를 산출하는 과정을 설명하기 위한 도면이다.

[0039] 도 3을 참고하면, T0(301)는 순위를 산출하고자 하는 제1 문서를 의미할 수 있다. 그리고, T1(302), T2(303), T3(304), T4(305)는 링크를 통해 제1 문서인 T0(301)를 가리키는 제2 문서를 의미할 수 있다. 이 때, 제2 문서인 T1(302), T2(303), T3(304), T4(305) 각각의 순위가 결정될 수 있다. 예를 들어, 제2 문서인 T1(302)을 링크를 통해 가리키는 문서 T11, T12, T13을 이용하여 T1(302)의 순위가 산출될 수 있다.

[0040] 본 발명의 일실시예에 따른 문서 순위 산출 시스템(100)은 제1 문서인 T0(301)와 연결된 적어도 하나의 T1(302), T2(303), T3(304), T4(305)인 제2 문서의 위치 정보를 추출할 수 있다. 이 때. 제2 문서인 T1(302)과 T4(305)가 동일한 호스트 H1(305)을 통해 제공된다고 가정한다. 이하에서는, 제2 문서인 T1(302)과 T4(305)를 중심으로 설명하도록 한다. 아래 설명은 제2 문서인 T2(303), T3(304)에도 동일하게 적용될 수 있다.

[0041] 그러면, 문서 순위 산출 시스템(100)은 제2 문서의 위치 정보에 기초하여 제2 문서가 스팸 문서인지 여부를 판단할 수 있다. 일례로, 문서 순위 산출 시스템(100)은 제2 문서를 제공하는 적어도 하나의 호스트를 결정할 수 있다. 그리고, 문서 순위 산출 시스템(100)은 결정된 호스트의 링크 구조에 따라 특징 벡터를 추출하여 호스트에 대한 호스트 스코어를 계산할 수 있다. 그리고, 문서 순위 산출 시스템(100)은 호스트 스코어에 기초하여 호스트가 스팸 호스트인지 여부를 판단할 수 있다.

[0042] 도 3을 참고하면, 문서 순위 산출 시스템(100)은 제2 문서인 T1(302)과 T4(305)를 제공하는 호스트 H1(306)을 결정할 수 있다. 그리고, 문서 순위 산출 시스템(100)은 호스트 H1(306)의 링크 구조에 따라 특징 벡터를 추출하여 호스트 H1(306)에 대한 호스트 스코어를 계산할 수 있다.

[0043] 이 때, 문서 순위 산출 시스템(100)은 호스트 H1(306)으로부터 추출한 특징 벡터 각각에 가중치를 적용하여 호스트 스코어를 계산할 수 있다. 여기서, 문서 순위 산출 시스템(100)은 스팸 호스트 및 정상 호스트로 라벨링

된 학습 데이터로부터 특징 벡터를 추출하고, 학습을 통해 특징 벡터 각각에 대해 가중치를 미리 결정할 수 있다. 그러면, 문서 순위 산출 시스템(100)은 호스트 H1(306)으로부터 추출한 특징 벡터 각각에 학습을 통해 결정된 가중치를 적용하여 호스트 H1(306)에 대한 호스트 스코어를 계산할 수 있다. 호스트 스코어를 계산하는 구체적인 일례는 도 5에서 설명된다.

[0044] 이 때, 특징 벡터는 제2 문서를 제공하는 호스트 H1(306)에 대한 수집 문서 정보, 인/아웃 링크 정보(in/out link), 인/아웃 호스트 정보(in/out host), 아웃 호스트의 인 호스트 정보, 인 호스트의 아웃 호스트 정보, 아웃 호스트의 인 링크 정보, 인 호스트의 아웃 링크 정보 또는 인/아웃 도메인 정보(in/out domain) 중 적어도 하나를 포함할 수 있다. 특징 벡터에 대한 가중치를 산출하는 과정은 도 4에서 설명된다.

[0045] 결국, 호스트 H1(306)에 대한 호스트 스코어가 결정되면, 문서 순위 산출 시스템(100)은 호스트 H1(306)이 스팸 호스트인지 여부를 판단할 수 있다. 예를 들어, 호스트 H1(306)의 호스트 스코어가 0.5미만인 경우, 호스트 H1(306)은 스팸 호스트로 결정될 수 있다. 또는 호스트 H1(306)의 아웃 호스트 또는 인 호스트의 호스트 스코어에 기초하여 호스트 H1(306)이 스팸 호스트로 결정될 수 있다.

[0046] 스팸 호스트가 제공하는 문서도 스팸 문서일 가능성이 높다. 따라서, 호스트 H1(306)이 스팸 호스트로 결정된 경우, 문서 순위 산출 시스템(100)은 호스트 H1(306)이 제공하는 제2 문서인 T1(302)과 T4(305)를 제1 문서인 T0(305)에 대한 문서 순위를 산출할 때 패널티를 적용할 수 있다. 예를 들면, 문서 순위 산출 시스템(100)은 제1 문서인 T0(305)와 링크를 통해 연결된 제2 문서인 T1(302)과 T4(305)의 연결을 제거하거나 또는 제2 문서인 T1(302)과 T4(305)을 그룹화하여 하나의 문서로 간주할 수 있다. 또는 수학식 1에서 문서 순위 산출 시스템(100)은 호스트 H1(306)이 제공하는 제2 문서인 T1(302)과 T4(305)에 대한 d를 감소시킬 수 있다.

[0047] 이와 같은 과정을 통해 제1 문서인 제1 문서인 T0(305)에 대한 문서 순위가 산출될 수 있다.

[0048] 도 4는 본 발명의 일실시예에 따른 호스트 스코어 산출을 위해 특징 벡터 별 가중치를 산출하는 과정을 설명하기 위한 도면이다.

[0049] 문서 순위 산출 시스템(100)은 기존에 알려진 정상 호스트 및 스팸 호스트로 라벨링된 학습 데이터(401)으로부터 특징 벡터(402)를 추출할 수 있다. 앞에서 이미 언급했듯이, 특징 벡터(402)는 호스트 별 수집 문서 정보, 인/아웃 링크 정보(in/out link), 인/아웃 호스트 정보(in/out host), 아웃 호스트의 인 호스트 정보, 인 호스트의 아웃 호스트 정보, 아웃 호스트의 인 링크 정보, 인 호스트의 아웃 링크 정보 또는 인/아웃 도메인 정보(in/out domain) 중 적어도 하나를 포함할 수 있다. 이 때, 특징 벡터(402)는 호스트의 링크 구조를 통해 추출될 수 있다.

[0050] 도 4에서 볼 수 있듯이, 호스트에 대해 특징 벡터(402) 1,2,3,4,5,6이 추출되었다고 가정한다. 그러면, 문서 순위 산출 시스템(100)은 학습 데이터(401)를 통해 기계 학습을 수행하여 특징 벡터(402) 각각에 대해 특징 벡터 별 가중치(403)를 산출할 수 있다.

[0051] 그러면, 문서 순위 산출 시스템(100)은 임의의 호스트에 대해 특징 벡터를 추출하고, 기계 학습을 통해 산출된 특징 벡터 별 가중치를 적용하여 상기 호스트에 대한 호스트 스코어를 계산할 수 있다.

[0052] 도 5는 본 발명의 일실시예에 따른 호스트에 대한 링크 구조를 설명하기 위한 도면이다.

[0053] 도 5를 참고하면, 호스트 A(500)가 제공하는 문서를 문서 1(507), 문서 2(508), 문서 3(509)이라고 가정한다. 이 때, 문서 1(507), 문서 2(508), 문서 3(509)은 문서의 순위를 결정하고자 하는 특정 문서와 링크를 통해 연결된 제2 문서를 의미할 수 있다.

[0054] 문서 순위 산출 시스템(100)은 호스트 A(500)를 링크를 통해 가리키는 호스트 X(501), 호스트 Y(502) 및 호스트 Z(503)를 호스트 A(500)에 대한 인 호스트로 설정하여, 특징 벡터를 추출할 수 있다. 그리고, 문서 순위 산출 시스템(100)은 호스트 A(500)가 링크를 통해 외부로 향하는 호스트 1(504), 호스트 2(505) 및 호스트 3(506)을 호스트 A(500)에 대한 아웃 호스트로 설정하여 특징 벡터를 추출할 수 있다. 그리고, 문서 순위 산출 시스템(100)은 호스트 A가 제공하는 문서 1(507), 문서 2(508) 및 문서 3(509)을 수집 문서 정보로 설정하여 특징 벡터를 추출할 수 있다.

[0055] 또한, 문서 순위 산출 시스템(100)은 호스트 A(500)에 대한 아웃 호스트인 호스트 1(504), 호스트 2(505) 및 호스트 3(506) 각각의 인 호스트 정보도, 호스트 A(500)의 특징 벡터로 설정할 수 있다. 그리고, 문서 순위 산출 시스템(100)은 호스트 A(500)에 대한 인 호스트인 호스트 X(501), 호스트 Y(502) 및 호스트 Z(503) 각각의 아웃

호스트 정보도 호스트 A(500)의 특징 벡터로 설정할 수 있다.

[0056] 그러면, 문서 순위 산출 시스템(100)은 호스트 A(500)에 대한 특징 벡터를 추출하고, 기계 학습을 통해 결정된 특징 벡터별 가중치를 적용하여 호스트 A(500)의 호스트 스코어를 계산할 수 있다.

[0057] 도 6은 본 발명의 일실시예에 따른 문서 순위 산출 방법의 전체 과정을 도시한 플로우차트이다.

[0058] 문서 순위 산출 시스템(100)은 링크를 통해 제1 문서를 가리키는 적어도 하나의 제2 문서의 위치 정보를 추출할 수 있다(S601). 일례로, 문서 순위 산출 시스템(100)은 제2 문서의 IP 주소, 도메인 정보 또는 호스트 정보 중 어느 하나에 기초하여 제2 문서의 위치 정보를 판단할 수 있다.

[0059] 문서 순위 산출 시스템(100)은 제2 문서의 위치 정보에 기초하여 제2 문서가 스팸 문서인지 판단할 수 있다(S602).

[0060] 일례로, 문서 순위 산출 시스템(100)은 제2 문서를 제공하는 적어도 하나의 호스트를 결정할 수 있다. 이 때, 문서 순위 산출 시스템(100)은 제2 문서의 위치 정보에 기초하여 상기 제2 문서를 제공하는 호스트를 결정할 수 있다.

[0061] 그리고, 문서 순위 산출 시스템(100)은 결정된 호스트의 링크 구조에 따라 특징 벡터를 추출하여 호스트에 대한 호스트 스코어를 계산할 수 있다. 일례로, 특징 벡터는 제2 문서를 제공하는 적어도 하나의 호스트 별 수집 문서 정보, 인/아웃 링크 정보, 인/아웃 호스트 정보, 아웃 호스트의 인 호스트 정보, 인 호스트의 아웃 호스트 정보, 아웃 호스트의 인 링크 정보, 인 호스트의 아웃 링크 정보 또는 인/아웃 도메인 정보 중 적어도 하나를 포함할 수 있다. 이 때, 문서 순위 산출 시스템(100)은 제2 문서를 제공하는 적어도 하나의 호스트로부터 추출한 특징 벡터들 각각에 가중치를 적용하여 호스트 스코어를 계산할 수 있다.

[0062] 문서 순위 산출 시스템(100)은 호스트 스코어를 산출하기 이전 또는 동시에, 스팸 호스트 및 정상 호스트로 라벨링된 학습 데이터로부터 추출된 특징 벡터들 각각의 가중치를 결정할 수 있다. 이 때, 가중치는 스팸 호스트의 호스트 스코어가 미리 설정한 최저값이 되도록 결정되고, 정상 호스트의 호스트 스코어가 미리 설정한 최대값이 되도록 결정될 수 있다.

[0063] 그런 후, 문서 순위 산출 시스템(100)은 호스트 스코어에 기초하여 상기 호스트가 스팸 호스트인지 여부를 판단할 수 있다. 일례로, 문서 순위 산출 시스템(100)은 호스트 스코어 중 미리 설정한 기준 스코어 미만인 호스트를 스팸 호스트로 판단할 수 있다. 다른 일례로, 문서 순위 산출 시스템(100)은 호스트에 대한 아웃 호스트 또는 인 호스트의 호스트 스코어에 기초하여 호스트를 스팸 호스트로 판단할 수 있다.

[0064] 최종적으로, 문서 순위 산출 시스템(100)은 스팸 호스트가 제공하는 제2 문서를 스팸 문서로 판단할 수 있다. 일례로, 문서 순위 산출 시스템(100)은 호스트 스코어를 적용한 경우 및 적용하지 않은 경우의 순위 변화량을 계산하여, 상기 계산된 변화량을 이용하여 스팸 문서인지 여부를 판단할 수 있다.

[0065] 문서 순위 산출 시스템(100)은 스팸 문서로 판단된 제2 문서를 고려하여 제1 문서의 순위를 산출할 수 있다(S603). 일례로, 동일한 위치 정보를 나타내는 적어도 하나의 제2 문서들을 그룹화하여 상기 제1 문서의 순위를 산출할 수 있다.

[0066] 도 6에서 설명되지 않은 사항은 도 1 내지 도 5의 설명을 참고할 수 있다.

[0067] 또한 본 발명의 일실시예에 따른 문서 순위 산출 방법은 다양한 컴퓨터로 구현되는 동작을 수행하기 위한 프로그램 명령을 포함하는 컴퓨터 판독 가능 매체를 포함한다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0068] 이상과 같이 본 발명은 비록 한정된 실시예와 도면에 의해 설명되었으나, 본 발명은 상기의 실시예에 한정되는 것은 아니며, 이는 본 발명이 속하는 분야에서 통상의 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및

변형이 가능하다. 따라서, 본 발명 사상은 아래에 기재된 특허청구범위에 의해서만 파악되어야 하고, 이의 균등 또는 등가적 변형 모두는 본 발명 사상의 범주에 속한다고 할 것이다.

도면의 간단한 설명

[0069] 도 1은 본 발명의 일실시예에 따른 문서 순위 산출 시스템을 통해 문서의 순위를 산출하는 전체 과정을 도시한 블록 다이어그램이다.

[0070] 도 2는 본 발명의 일실시예에 따른 문서 순위 산출 시스템의 상세 구성을 도시한 블록 다이어그램이다.

[0071] 도 3은 본 발명의 일실시예에 따른 문서의 링크 구조를 통해 문서의 순위를 산출하는 과정을 설명하기 위한 도면이다.

[0072] 도 4는 본 발명의 일실시예에 따른 호스트 스코어 산출을 위해 특징 벡터 별 가중치를 산출하는 과정을 설명하기 위한 도면이다.

[0073] 도 5는 본 발명의 일실시예에 따른 호스트에 대한 링크 구조를 설명하기 위한 도면이다.

[0074] 도 6은 본 발명의 일실시예에 따른 문서 순위 산출 방법의 전체 과정을 도시한 플로우차트이다.

[0075] <도면의 주요 부분에 대한 부호의 설명>

[0076] 100: 문서 순위 산출 시스템

[0077] 201: 위치 정보 추출부

[0078] 202: 스팸 문서 판단부

[0079] 203: 순위 산출부

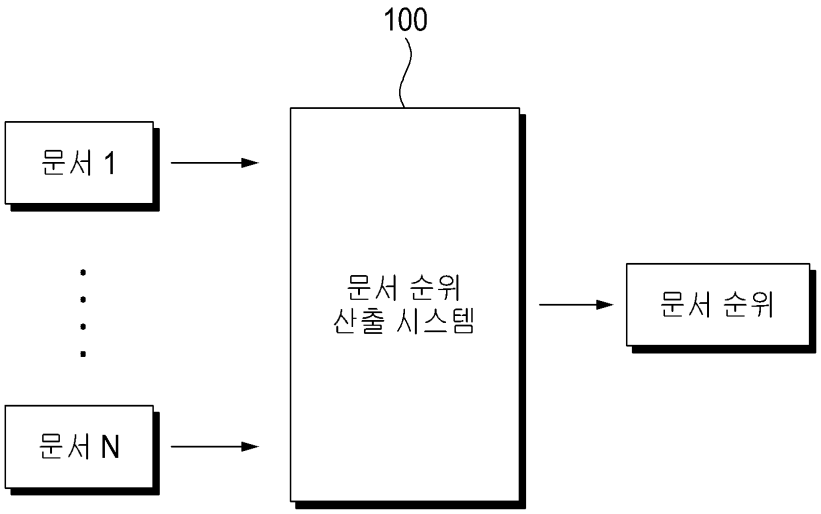
[0080] 204: 호스트 결정부

[0081] 205: 호스트 스코어 판단부

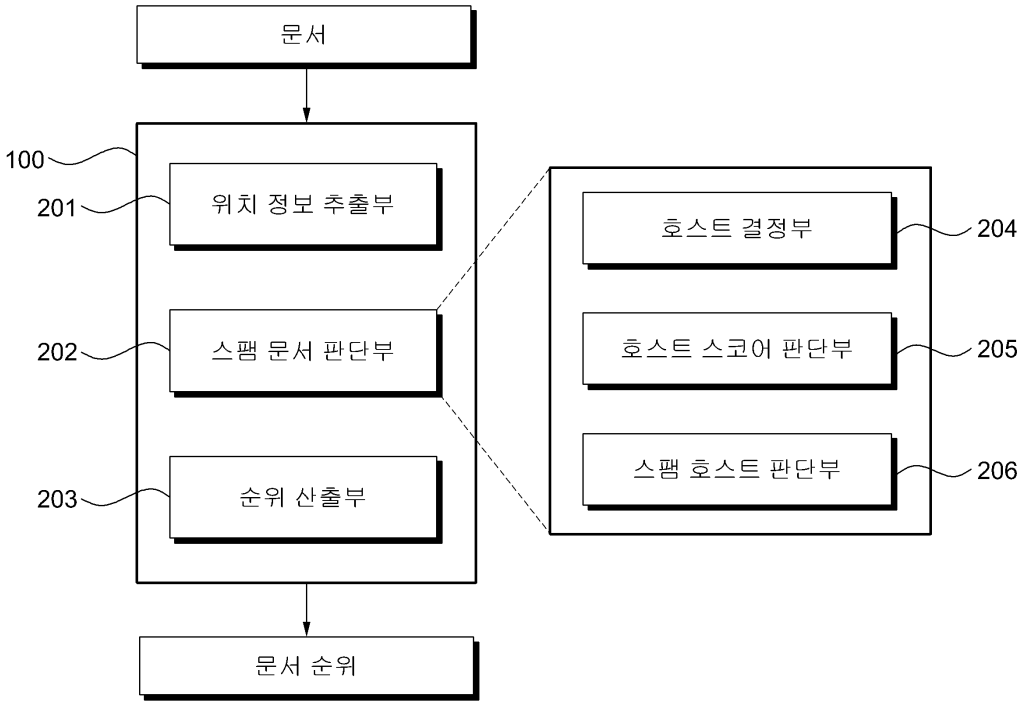
[0082] 206: 스팸 호스트 판단부

도면

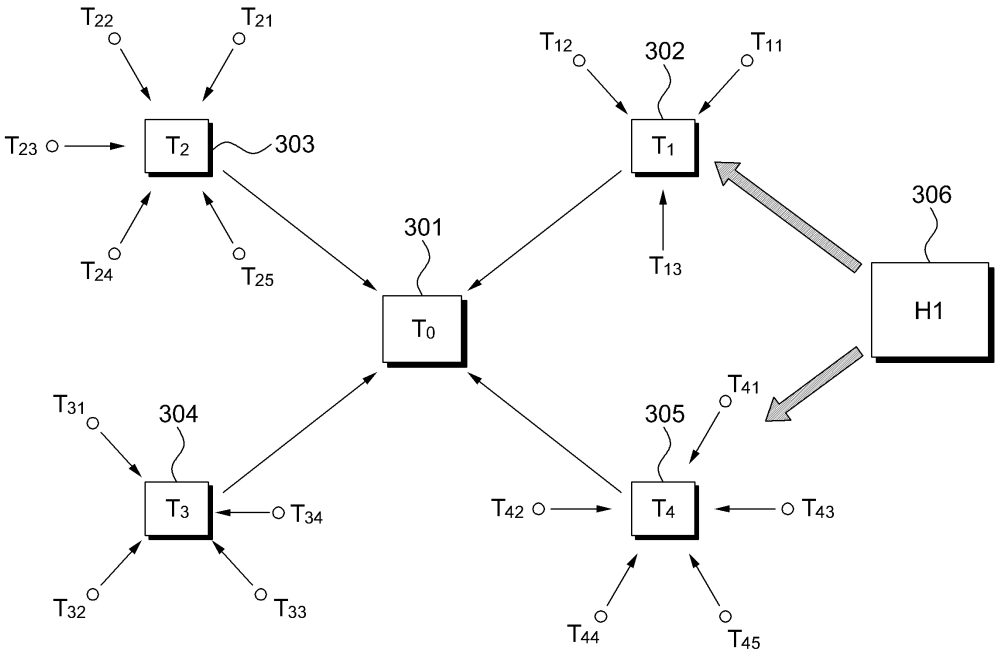
도면1



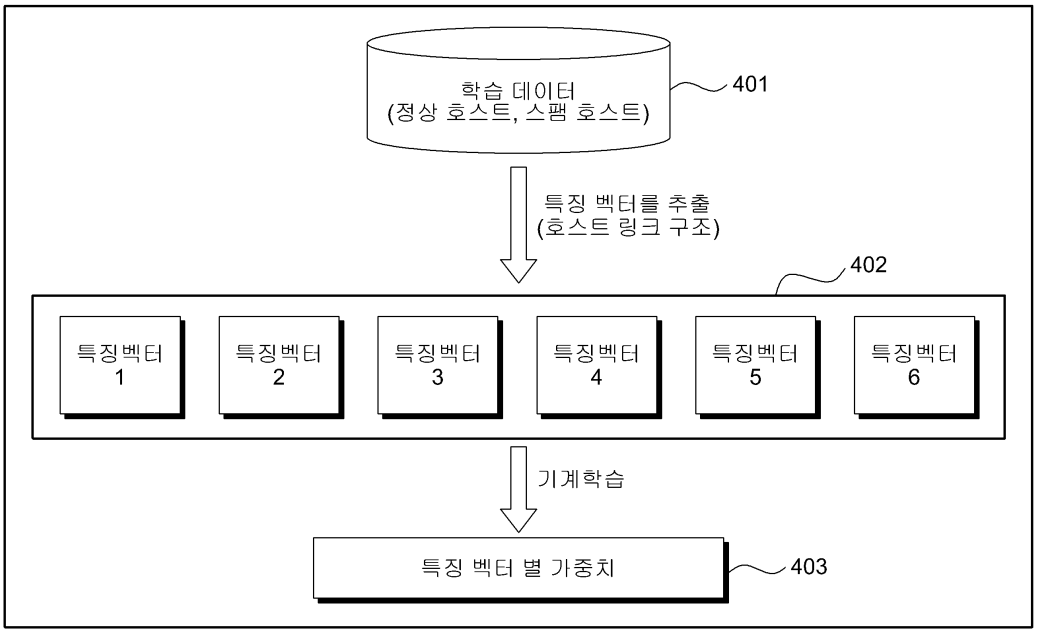
도면2


문서
100
위치 정보 추출부
201
스팸 문서 판단부
202
순위 산출부
203
호스트 결정부
204
호스트 스코어 판단부
205
스팸 호스트 판단부
206
문서 순위

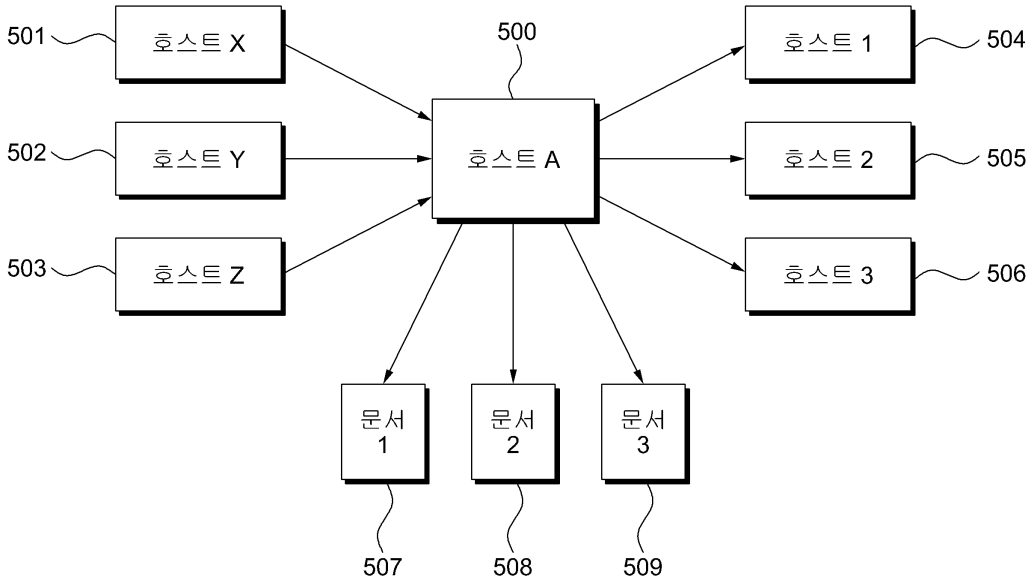
도면3


T22
T21
T23
T2
303
T24
T25
301
T0
T12
302
T11
T1
T13
306
H1
T41
T31
304
T3
T34
T32
T33
305
T42
T4
T43
T44
T45

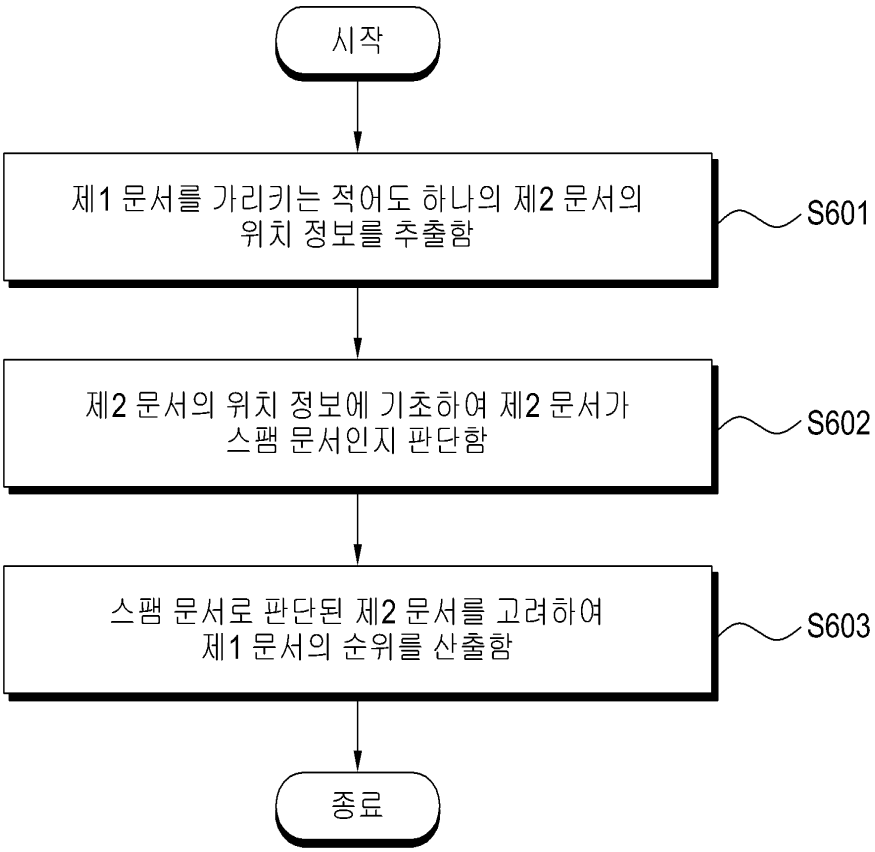
도면4


학습 데이터
(정상 호스트, 스팸 호스트)
401
특징 벡터를 추출
(호스트 링크 구조)
402
특징벡터 1
특징벡터 2
특징벡터 3
특징벡터 4
특징벡터 5
특징벡터 6
기계학습
특징 벡터 별 가중치
403

도면5


501
호스트 X
502
호스트 Y
503
호스트 Z
500
호스트 A
호스트 1
504
호스트 2
505
호스트 3
506
문서 1
문서 2
문서 3
507
508
509

도면6


시작
제1 문서를 가리키는 적어도 하나의 제2 문서의 위치 정보를 추출함
S601
제2 문서의 위치 정보에 기초하여 제2 문서가 스팸 문서인지 판단함
S602
스팸 문서로 판단된 제2 문서를 고려하여 제1 문서의 순위를 산출함
S603
종료