(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2013년02월07일
(11) 등록번호 10-1230687
(24) 등록일자 2013년01월31일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01)
(21) 출원번호 10-2007-7011999
(22) 출원일자(국제) 2005년10월26일
심사청구일자 2010년10월25일
(85) 번역문제출일자 2007년05월28일
(65) 공개번호 10-2007-0085477
(43) 공개일자 2007년08월27일
(86) 국제출원번호 PCT/US2005/038619
(87) 국제공개번호 WO 2006/049996
국제공개일자 2006년05월11일
(30) 우선권주장
11/198,471 2005년08월04일 미국(US)
60/623,295 2004년10월28일 미국(US)
(56) 선행기술조사문헌
GYONGYI Z., GARCIA-MOLINA H., PEDERSEN J.: "Combating Web Spam with TrustRank" TECHNICAL REPORT(STANFORD UNIVERSITY), 2004.3.11.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
야후! 인크.
미국, 94089 캘리포니아, 써니베일, 퍼스트 애브뉴 701
(72) 발명자
베킨, 파벨
미국 94087 캘리포니아 써니베일 턴스톤 웨이 1378
기용이, 졸탄 아이.
미국 94305 캘리포니아 스탠포드 올름스테드 로드 27 아파트먼트101
페데르센, 얀
미국 94022 캘리포니아 로스 알토스 힐스 조세파 레인 25750
(74) 대리인
백만기, 양영준

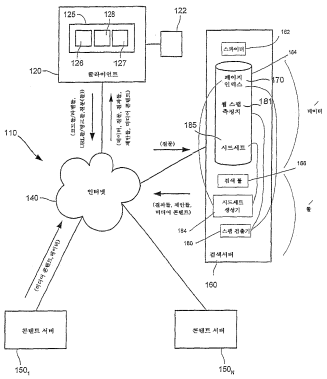
전체 청구항 수 : 총 10 항 심사관 : 이명진

(54) 발명의 명칭 **링크 바탕 스팸 검출**

(57) 요 약

컴퓨터 실행 방법은 검색 결과 세트에서 검색 히트들의 순위를 결정하기 위하여 제공된다. 상기 방법은 사용자로부터 질문을 수신하고 질문에 관련된 히트들의 리스트를 생성하고, 여기서 히트들 각각은 질문에 대한 관련성을 가지며, 히트들은 히트들을 가리키는 하나 또는 그 이상의 상승하는 링크 도큐먼트들을 가지며, 상승하는 링크 도큐먼트들은 질문에 대한 히트들의 관련성에 영향을 미친다. 상기 방법은 각각을 가리키는 상승하는 링크 도큐먼트들의 수를 나타내는 측정치를 히트들과 연관시킨다. 상기 방법은 임계값과 측정치를 비교하고, 부분적으로 비교들을 바탕으로 변형된 리스트를 형성하기 위하여 히트들의 리스트를 처리하고, 사용자에게 변형된 리스트를 전송한다.

대 표 도 - 도2



특허청구의 범위

청구항 1

검색 결과 세트에서 검색 히트들의 순위를 매기는 방법으로서, 상기 방법은 컴퓨팅 시스템에 의해서 수행되며,

사용자로부터 질문을 수신하는 단계;

상기 질문에 관련된 히트들의 리스트를 생성하는 단계 - 상기 히트들의 리스트의 각각의 히트들은 상기 질문에 대한 관련성(relevance)을 가지며, 적어도 하나의 히트가 부스팅(boosting) 도큐먼트내의 링크에 의해 가리켜지고, 상기 부스팅 도큐먼트내의 링크는 상기 질문에 대한 상기 적어도 하나의 히트의 관련성을 인위적으로 높임 -;

상기 적어도 하나의 히트에 대한 제1 측정치(first measure)를 결정하는 단계 - 상기 제1 측정치는 상기 적어도 하나의 히트에 대한 링크-기반 인기도 측정치임 -;

상기 적어도 하나의 히트에 대한 제2 측정치를 결정하는 단계 - 상기 제2 측정치는, 상기 적어도 하나의 히트가 우수한(reputable) 도큐먼트일 가능성을 나타내는, 상기 적어도 하나의 히트에 대한 신뢰도 측정치임 -;

상기 제1 측정치와 제2 측정치간의 불일치에 적어도 부분적으로 기초하여, 상기 적어도 하나의 히트에 대한 측정치(metric)를 생성하는 단계 - 상기 측정치(metric)는, 상기 질문에 대한 상기 적어도 하나의 히트의 관련성을 인위적으로 높이는, 상기 적어도 하나의 히트로의 링크들을 포함하는 다수의 부스팅 도큐먼트들을 나타냄 -;

임계값을, 적어도 부분적으로 상기 측정치(metric)에 기초하는 값과 비교하는 단계;

변형된 리스트를 형성하기 위해 상기 히트들의 리스트를 처리하는 단계 - 상기 처리하는 단계는, 적어도 부분적으로 상기 측정치(metric)에 기초한 상기 값이 상기 임계값보다 크다는 것을 결정하는 단계에 응답하여,

상기 변형된 리스트로부터 상기 적어도 하나의 히트를 제외시키는 단계; 또는,

상기 변형된 리스트 내의 상기 적어도 하나의 히트를 강등시키는 단계

중 어느 하나를 수행하는 단계를 더 포함함 -; 및

상기 변형된 리스트를 상기 질문에 대한 응답으로서 상기 사용자에게 송신하는 단계

를 포함하는, 검색 히트들의 순위를 매기는 방법.

청구항 2

제1항에 있어서, 상기 측정치(metric)를 생성하는 단계는, 상기 질문을 수신하는 단계 이전에 수행되는, 검색 히트들의 순위를 매기는 방법.

청구항 3

제1항에 있어서, 상기 제2 측정치를 결정하는 단계는,

우수한 도큐먼트들의 시드(seed) 세트를 형성하는 단계 - 상기 우수한 도큐먼트들의 시드 세트는 다른 도큐먼트들에 대한 링크들을 포함함 -;

상기 시드 세트내의 각각의 도큐먼트들에 신뢰 값을 할당하는 단계;

상기 시드 세트내의 도큐먼트들 중 적어도 하나에 의해 가리켜지는 복수의 도큐먼트들의 각각에 상기 신뢰 값을 전달(propagating)하는 단계; 및

상기 시드 세트내의 도큐먼트들 중 적어도 하나에 의해 가리켜지는 복수의 도큐먼트들의 각각에 배분된(prorated) 신뢰 값을 할당하는 단계를 포함하는, 검색 히트들의 순위를 매기는 방법.

청구항 4

제3항에 있어서,

상기 시드 세트를 형성하는 단계는, 제2 복수의 도큐먼트들 각각에 대해, 상기 제2 복수의 도큐먼트들 각각에

포함된 다수의 아웃링크들을 각각 나타내는 아웃링크 측정치(outlink metric)를 결정하는 단계;

상기 아웃링크 측정치(outlink metric)를 이용하여 상기 제2 복수의 도큐먼트들의 순위를 매기는 단계;

상기 제2 복수의 도큐먼트들 내에서 가장 높은 순위의 도큐먼트들의 세트를 식별하는 단계;

상기 시드 세트에 포함되도록 상기 가장 높은 순위의 도큐먼트들의 세트로부터 하나 이상의 도큐먼트들을 식별하는 입력을 수신하는 단계;

상기 입력에 기초하여 가장 높은 순위의 도큐먼트들의 변형된 세트를 형성하는 단계; 및

상기 가장 높은 순위의 도큐먼트들의 변형된 세트를 이용하여 상기 시드 세트를 형성하는 단계를 포함하는, 검색 히트들의 순위를 매기는 방법.

청구항 5

제1항에 있어서,

상기 제1 측정치를 결정하는 단계, 상기 제2 측정치를 결정하는 단계, 및 상기 측정치(metric)를 생성하는 단계는 상기 히트들의 리스트내의 각각의 히트에 대해 수행되고,

상기 방법은,

상기 히트들의 리스트내의 각각의 히트에 대한 상기 측정치(metric)를 생성한 후에, 상기 히트들의 리스트내의 히트들에 대해 생성된 상기 측정치들(metrics)에 기초하여 상기 히트들의 리스트를 정렬(sort)하여, 정렬된 리스트를 생성하는 단계;

상기 정렬된 리스트의 최상부(top portion)를 식별하는 단계 - 상기 최상부내의 히트들은 상기 최상부내에 있지 않은 히트들보다 더 높은 측정치들(metrics)과 연관됨 -;

상기 정렬된 리스트의 최상부내의 각각의 히트에 대해, 상기 제1 측정치에 대한 상기 측정치(metric)의 비(ratio)에 기초하여 상기 히트를 스팸으로 분류할지 여부를 결정하는 단계를 더 포함하는, 검색 히트들의 순위를 매기는 방법.

청구항 6

검색 결과 세트에서 검색 히트들의 순위를 매기는 명령어들을 저장한 컴퓨터 판독가능 저장 매체로서,

상기 명령어들은 컴퓨팅 시스템에 의해서 수행되며,

상기 명령어들은,

사용자로부터 질문을 수신하는 단계;

상기 질문에 관련된 히트들의 리스트를 생성하는 단계 - 상기 히트들의 리스트의 각각의 히트들은 상기 질문에 대한 관련성을 가지며, 적어도 하나의 히트가 부스팅 도큐먼트내의 링크에 의해 가리켜지고, 상기 부스팅 도큐먼트내의 링크는 상기 질문에 대한 상기 적어도 하나의 히트의 관련성을 인위적으로 높임 -;

상기 적어도 하나의 히트에 대한 제1 측정치를 결정하는 단계 - 상기 제1 측정치는 상기 적어도 하나의 히트에 대한 링크-기반 인기도 측정치임 -;

상기 적어도 하나의 히트에 대한 제2 측정치를 결정하는 단계 - 상기 제2 측정치는, 상기 적어도 하나의 히트가 우수한 도큐먼트일 가능성을 나타내는, 상기 적어도 하나의 히트에 대한 신뢰도 측정치임 -;

상기 제1 측정치와 제2 측정치간의 불일치에 적어도 부분적으로 기초하여, 상기 적어도 하나의 히트에 대한 측정치(metric)를 생성하는 단계 - 상기 측정치(metric)는, 상기 질문에 대한 상기 적어도 하나의 히트의 관련성을 인위적으로 높이는, 상기 적어도 하나의 히트로의 링크들을 포함하는 다수의 부스팅 도큐먼트들을 나타냄 -;

임계값을, 적어도 부분적으로 상기 측정치(metric)에 기초하는 값과 비교하는 단계;

변형된 리스트를 형성하기 위해 상기 히트들의 리스트를 처리하는 단계 - 상기 처리하는 단계는, 상기 측정치(metric)에 적어도 부분적으로 기초한 상기 값이 상기 임계값보다 크다는 것을 결정하는 단계에 응답하여,

상기 변형된 리스트로부터 상기 적어도 하나의 히트를 제외시키는 단계; 또는,

상기 변형된 리스트 내의 상기 적어도 하나의 히트를 강등시키는 단계

중 어느 하나를 수행하는 단계를 더 포함함 -; 및

상기 변형된 리스트를 상기 질문에 대한 응답으로서 상기 사용자에게 송신하는 단계

를 수행하는 명령어들을 포함하는, 컴퓨터 판독가능 저장 매체.

청구항 7

제6항에 있어서, 상기 측정치(metric)를 생성하는 단계는, 상기 질문을 수신하는 단계 이전에 수행되는, 컴퓨터 판독가능 저장 매체.

청구항 8

제6항에 있어서, 상기 제2 측정치를 결정하는 단계는,

우수한 도큐먼트들의 시드 세트를 형성하는 단계 - 상기 우수한 도큐먼트들의 시드 세트는 다른 도큐먼트들에 대한 링크들을 포함함 -;

상기 시드 세트내의 각각의 도큐먼트들에 신뢰 값을 할당하는 단계;

상기 시드 세트내의 도큐먼트들 중 적어도 하나에 의해 가리켜지는 복수의 도큐먼트들의 각각에 상기 신뢰 값을 전달하는 단계; 및

상기 시드 세트내의 도큐먼트들 중 적어도 하나에 의해 가리켜지는 복수의 도큐먼트들의 각각에 배분된 신뢰 값을 할당하는 단계를 포함하는, 컴퓨터 판독가능 저장 매체.

청구항 9

제8항에 있어서,

상기 시드 세트를 형성하는 단계는, 제2 복수의 도큐먼트들 각각에 대해, 상기 제2 복수의 도큐먼트들 각각에 포함된 다수의 아웃링크들을 각각 나타내는 아웃링크 측정치(outlink metric)를 결정하는 단계;

상기 아웃링크 측정치(outlink metric)를 이용하여 상기 제2 복수의 도큐먼트들의 순위를 매기는 단계;

상기 제2 복수의 도큐먼트들내에서 가장 높은 순위의 도큐먼트들의 세트를 식별하는 단계;

상기 시드 세트에 포함되도록 상기 가장 높은 순위의 도큐먼트들의 세트로부터 하나 이상의 도큐먼트들을 식별하는 입력을 수신하는 단계;

상기 입력에 기초하여 가장 높은 순위의 도큐먼트들의 변형된 세트를 형성하는 단계; 및

상기 가장 높은 순위의 도큐먼트들의 변형된 세트를 이용하여 상기 시드 세트를 형성하는 단계를 포함하는, 컴퓨터 판독가능 저장 매체.

청구항 10

제6항에 있어서,

상기 제1 측정치를 결정하는 단계, 상기 제2 측정치를 결정하는 단계, 및 상기 측정치(metric)를 생성하는 단계는 상기 히트들의 리스트내의 각각의 히트에 대해 수행되고,

상기 명령어들은,

상기 히트들의 리스트내의 각각의 히트에 대한 상기 측정치(metric)를 생성한 후에, 상기 히트들의 리스트내의 히트들에 대해 생성된 상기 측정치들(metrics)에 기초하여 상기 히트들의 리스트를 정렬하여, 정렬된 리스트를 생성하는 단계;

상기 정렬된 리스트의 최상부를 식별하는 단계 - 상기 최상부내의 히트들은 상기 최상부내에 있지 않은 히트들보다 더 높은 측정치들(metrics)과 연관됨 -;

상기 정렬된 리스트의 최상부내의 각각의 히트에 대해, 상기 제1 측정치에 대한 상기 측정치(metric)의 비에 기초하여 상기 히트를 스팸으로 분류할지 여부를 결정하는 단계를 수행하는 명령어들을 더 포함하는, 컴퓨터 판독

가능 저장 매체.

명 세 서

기 술 분 야

[0001] 본 발명은 일반적으로 검색 시스템 및 특히 최종 세트에서 검색 히트들(hit)의 순위를 매기는 검색 시스템에 관한 것이다.

배 경 기 술

[0002] 검색은 전체적인 코퍼스(corpus)가 흡수될 수 없고 목표된 아이템들에 대한 정확한 포인터가 존재하지 않거나 가능하지 않은 경우 유용하다. 일반적으로, 검색은 검색 질문을 공식화하거나 수용하고, 도큐먼트들의 코퍼스로부터 매칭 도큐먼트들의 세트를 결정하고 만약 상기 세트가 너무 크면 상기 세트의 세트 또는 몇몇 하부세트를 리턴하는 처리이다. 본 명세서를 제한하지 않는 특정 실시예에서, "웹(Web)"이라 불리는 하이퍼링크된 도큐먼트들의 세트를 검색하는 것을 고려하자. 코퍼스는 이후 페이저들, 또는 보다 일반적으로 도큐먼트들이라 불리는 많은 검색 가능한 아이템들을 포함한다. 검색 엔진은 통상적으로 검색 질문의 수신에 앞서 미리 생성된 인덱스를 사용하여 검색 질문과 매칭하는 코퍼스로부터 도큐먼트들을 식별한다. "매칭"은 많은 것들을 의미할 수 있고 검색 질문은 다양한 형태들일 수 있다. 일반적으로, 검색 질문은 하나 또는 그 이상의 단어들 또는 용어들을 포함하는 문자열이고 매칭은 도큐먼트가 검색 질문 문자열로부터 하나 또는 그 이상의 단어들 또는 용어들(또는 이들 모두)을 포함할 때 발생한다. 각각의 매칭 도큐먼트는 히트라 하고 히트들의 세트는 결과 세트 또는 검색 결과들이라 한다. 코퍼스는 데이터베이스 또는 다른 데이터 구조 또는 구성되지 않은 데이터일 수 있다. 도큐먼트들은 종종 웹 페이지들이다.

[0003] 웹 페이지들의 통상적인 인덱스는 수십억의 엔트리들을 포함하고, 따라서 일반적인 검색은 수백만의 페이지들을 포함하는 결과 세트를 가진다. 명확하게, 상기 상황들에서, 검색 엔진은 질문자(통상적으로 인간 컴퓨터 사용자이지만, 상기 경우일 필요는 없음)에게 리턴되는 것이 합리적인 크기 이도록 추가 결과 세트를 포함하여야 한다. 상기 세트를 억제하는 한 가지 방법은 사용자가 순차적 검색 결과들에서 보다 높게 나타나는 상위의 작은 수의 히트들만을 판독하거나 사용한다는 가정을 가지는 순서로 검색 결과들을 제공하는 것이다.

[0004] 이런 가정으로 인해, 많은 웹 페이지 저자들은 그들의 페이지들이 순차적 검색 결과들에서 상위에 나타나는 것을 원한다. 검색 엔진은 관련 페이지들의 다양한 특성들에 의존하여 가장 높은 품질만을 선택하고 리턴한다. 질문 결과 리스트에서 상위 위치들(높은 순위)이 비지니스에 장점들을 제공하기 때문에, 특정 웹 페이지들의 저자들은 페이지들의 순위를 부당하게 부스트(boost) 시키기 위하여 시도한다. 인위적으로 부스트된 순위를 가진 페이지들은 소위 "웹 스팸" 페이지들이라 하고 집합적으로 "웹 스팸"으로서 공지되었다.

[0005] 웹 스팸과 관련된 다양한 기술들이 있다. 그중 하나는 많은 질문들에 의해 선택되도록 하기 위하여 적당하게 웹 페이지를 만드는 것이다. 이것은 핵심 콘텐트에 관련되지 않고 작거나 보이지 않는 폰트(font)들로 렌더되는 다량의 용어들을 가진 페이지를 증가시킴으로서 달성된다. 상기 증가는 페이지가 보다 많이 노출되게 하지만(즉, 잠재적으로 보다 많은 질문들에 관련됨), 임의의 특정 질문에 대한 관련성을 진정으로 개선시키지 못한다. 이것과 관련하여, 스팸 저자들은 다른 기술을 사용한다: 스팸 저자들은 다른 것들에 의해 보다 자주 인용되는 페이지들이 검색 엔진들에 의해 일반적으로 바람직한(보다 높은 관련성) 것으로 생각된다는 관찰 결과를 바탕으로 페이지에 많은 인입(하이퍼) 링크들, 소위 인링크(inlink)들을 부가한다. 실제로 보다 상위의 값으로 인해 많은 다른 것들에 의해 인용되는 고품질의 페이지들과, 많은 인링크들을 가진 웹 스팸을 구별하는 것은 어렵다.

[0006] 웹 스팸 페이지들과 검색 결과 리스트에서 추후 강등물의 식별은 검색 엔진에 의해 형성된 대답 품질을 유지 또는 개선하는데 중요하다. 따라서, 웹 스팸 검출은 검색 엔진에 필요한 임무이다. 인간 에디터들은 검색 엔진 인덱스에 제공된 다수의 페이지들을 조사함으로써 웹 스팸을 식별하기 위하여 주로 사용되지만, 이것은 종종 실행하기 어렵다.

발명의 상세한 설명

[0007] 그러므로, 웹 스팸을 극복하고 도큐먼트 저자들의 조작들과 조화하기 보다 사용자들이 원하는 것과 보다 잘 조화하는 검색 결과들을 제공하는 개선된 검색 처리가 필요하다.

[0008] 본 발명의 실시예들은 검색 결과 세트를 형성하는 순위 히트들을 포함하는 검색 결과들을 처리하기 위한 시스템 및 방법들을 제공한다. 히트들은 특정 페이지에 대한 스팸 팜 크기 측정치인 유효 양(mass), 및 다른 파라미터들을 사용하여 순위가 매겨진다.

[0009] 일실시예에서, 본 발명은 검색 결과 세트에서 검색 히트들의 순위를 매기는 컴퓨터 실행 방법을 제공한다. 컴퓨터 실행 방법은 사용자로부터 질문을 수신하고 질문과 관련된 히트들의 리스트를 생성하는 것을 포함하고, 여기서 각각의 히트들은 질문과 관련성을 가지며, 히트들은 히트들에 대한 하나 또는 그 이상의 부스팅 링크된 도큐먼트들을 가지며, 부스팅 링크된 도큐먼트들은 질문에 대한 히트들의 관련성에 영향을 미친다. 그 다음 상기 방법은 히트들의 적어도 하나의 하위 세트 각각에 대한 측정법과 연관되고, 상기 방법은 히트들의 적어도 하나의 하위 세트 각각을 가리키고 히트들의 관련성을 인위적으로 부풀리는 부스팅하는 링크된 도큐먼트들의 수를 나타낸다. 그 다음 상기 방법은 히트들을 가리키는 스팸 팜(farm)의 크기를 나타내는 측정치와 임계값을 비교하고, 부분적으로 비교를 바탕으로 변형된 리스트를 형성하기 위하여 히트들의 리스트를 처리하고, 변형된 리스트를 사용자에게 전송한다.

[0010] 일측면에서, 상기 측정치는 제 1 측정치 및 제 2 측정치의 결합이다. 히트에 대한 제 1 측정치는 히트들의 링크 인기를 나타내고, 제 2 측정치는 히트가 신뢰성 있는 도큐먼트일 가능성의 측정치이다.

[0011] 다른 측면에서, 제 2 측정치는 링크중인 도큐먼트들인 신뢰성 있는 도큐먼트들의 시드(seed) 세트를 형성하고, 시드 세트의 도큐먼트들의 각각에 신뢰 값을 할당하고, 상기 신뢰 값을 링크중인 도큐먼트들에 의해 지시되는 링크된 도큐먼트들의 각각에 전달하고, 할당된 신뢰 값을 링크된 도큐먼트들의 각각에 할당함으로써 생성된다.

[0012] 다른 측면에서, 신뢰성 있는 도큐먼트들의 시드 세트는 다수의 도큐먼트들의 각각에 대해 도큐먼트들의 각각의 아웃링크들의 수를 나타내는 아웃링크 측정치를 결정하고, 아웃링크 측정치를 사용하여 다수의 도큐먼트들의 순위를 매기고, 한 세트의 가장 높은 순위의 도큐먼트들을 식별하고, 가장 높은 순위 도큐먼트들의 품질을 평가하고, 가장 높은 순위 도큐먼트들로부터 부적당한 것으로 생각되는 도큐먼트들을 제거함으로써 변형된 도큐먼트들의 세트를 형성하고, 변형된 유지 세트를 사용하는 시드 세트를 형성함으로써 형성된다.

[0013] 첨부 도면들과 함께 다음 상세한 설명은 본 발명의 성질 및 장점들을 보다 잘 이해할 수 있게 제공될 것이다.

실 시 예

[0017] 정의되지 않으면, 여기에 사용된 모든 기술 및 학술 용어들은 본 발명이 속하는 기술의 당업자에 의해 일반적으로 이해되는 의미를 가진다. 여기에 사용된 바와 같이, 다음 용어들은 다음과 같이 정의된다.

[0018] 페이지순위(PageRank)는 검색 엔진에 의해 인덱스된 하이퍼링크된 도큐먼트들(또는 웹 페이지들 또는 웹 사이트들)에 수치 웨이티들을 할당하기 위한 잘 공지된 알고리듬들의 일종이다. 페이지순위는 웹상 도큐먼트들에 글로벌 중요성 스코어들을 할당하기 위한 링크 정보를 사용한다. 페이지순위 처리는 특허되었고 미국특허 6,285,999에 기술된다. 도큐먼트의 페이지순위는 웹상 도큐먼트의 링크 바탕 인기도의 측정치이다.

[0019] 신뢰순위(TrustRank)는 페이지순위에 관련된 링크 분석 기술이다. 신뢰순위는 웹 스팸으로부터 웹상 신뢰성 있고, 우수한 페이지들을 분리하기 위한 방법이다. 신뢰순위는 웹상 우수 도큐먼트들이 거의 스팸에 링크하지 않는 가능성을 바탕으로 한다. 신뢰순위는 두 단계들, 즉 시드 선택 및 스코어 전달을 포함한다. 도큐먼트의 신뢰순위는 도큐먼트가 신뢰성 있는(즉, 스팸없음) 도큐먼트일 가능성의 측정치이다.

[0020] 링크 또는 하이퍼링크는 일반적으로 다른 페이지, 다른 사이트 또는 동일한 페이지의 다른 부분을 유도하는 웹 페이지상 클릭 가능한 도큐먼트를 말한다. 그러므로 클릭 가능한 콘텐트는 동일한 페이지의 다른 페이지/사이트/부분에 대한 링크라 한다. 스파이더들(spider)은 웹 사이트들을 인덱스할 때 하나의 페이지에서 다음 페이지로 크롤하기 위한 링크들을 사용한다.

[0021] 인바운드 링크 또는 인링크/아웃바운드 링크 또는 아웃링크. 사이트 A가 사이트 B에 링크할 때, 사이트 A는 아웃바운드 링크이고 사이트 B는 인바운드 링크이다. 인바운드 링크들은 링크 인기도를 결정하기 위하여 카운트된다.

[0022] 웹, 또는 월드 와이드 웹("WWW", 또는 간단히 "웹")은 리소스들이라 불리는 관심있는 아이템들이 소위 유니폼 리소스 아이덴티파이어들(URI)이라 불리는 글로벌 식별자들에 의해 식별되는 정보 공간이다. 용어 웹은 종종 인터넷에 대한 동의어로서 사용된다; 그러나, 웹은 실제로 인터넷상에서 동작하는 서비스이다.

[0023] 웹 페이지 또는 웹페이지는 일반적으로 HTML/XHTML 포맷(파일 확장부들이 통상적으로 htm 또는 html임)이고 하

나의 페이지로부터 또는 섹션으로부터 다른 페이지 또는 섹션으로 네비게이션할 수 있게 하는 하이퍼텍스트 링크를 가진 월드 와이드 웹의 페이지 또는 파일이라 한다. 웹페이지들은 종종 도면을 제공하기 위하여 연관된 그래픽 파일들을 사용하고, 이들은 또한 클릭 가능한 링크들일 수 있다. 웹페이지들은 웹 브라우저를 사용하여 디스플레이되고, 종종 모션, 그래픽, 대화, 및 사운드를 제공하는 애플릿들(페이지내에서 작동하는 것보다 오히려 서브프로그램들)을 사용할 수 있게 설계된다.

[0024] 웹 사이트는 단일 폴더 또는 웹 서버의 관련 서브폴더들내에 저장된 웹페이지들의 수집물이라 한다. 웹 사이트는 일반적으로 index.htm 또는 index.html이라 불리는 프론트 페이지를 포함한다.

[0025] 웹 호스트는 서버 공간, 웹 서비스들 및 자신의 웹 서버들을 가지지 않는 개인들 또는 회사들에 의해 제어되는 웹 사이트들에 대한 파일 유지를 제공하는 비지니스이다. 많은 인터넷 서비스 제공자들(ISP)은 가입자들에게 개인 웹 페이지를 호스트하기 위한 작은 공간의 서버 공간을 허용한다.

[0026] 스팸은 대량으로 분배되는 일반적으로 상업적 성질의 원하지 않는 도큐먼트들 또는 이메일들을 말한다.

[0027] 웹 스팸은 웹상의 스팸 페이지들이라 한다. 웹 스팸을 생성하는 행위는 웹 스패밍이라 한다. 웹 스패밍은 받을 가치가 있는 보다 높은 순위의 몇몇 도큐먼들을 제공하기 위하여 검색 엔진들을 잘못 인도하게 하는 행위들을 말한다. 웹상 스팸 페이지들은 몇몇 형태의 스패밍의 결과들이다. 스패밍의 한 가지 형태는 링크 스패밍이다.

[0028] 스팸 페이지는 순위 스코어시 불법의 부스팅을 수신하고 그러므로 상위 검색 결과들에서 나타날 수 있고 검색 엔진을 잘못 인도하고자 하는 웹 도큐먼트이다.

[0029] 링크 스패밍은 종종 상호접속된 스팸 도큐먼트들의 생성 및 소위 스팸 팜들이라 불리는 그룹들 형성을 말하고, 이것은 다수의 부스팅 도큐먼트들이 하나 또는 약간의 타겟 페이지들의 링크 바탕 중요도 순위를 증가시키도록 형성된다.

[0030] 스팸 팜은 특정 타겟 페이지들의 링크 바탕 중요도 스코어(예를들어, 페이지순위 스코어들)를 부스트하기 위하여 생성된 상호링크된 스팸 페이지들의 그룹을 말한다.

[0031] 개요

[0032] 본 발명의 실시예들은 링크 바탕 스팸의 검출을 위한 방법들 및 시스템들에 관한 것이다. 검색 질문에 응답하여 형성된 검색 결과들은 유효 히트들 양을 결정하기 위하여 처리된다. 유효 히트 양은 히트를 가리키고 히트의 관련 중요성을 인위적으로 부스트하기 위하여 생성된 스팸 팜의 크기 측정치이다. 본 발명의 실시예들에 따른 방법 및 시스템은 유효 히트들의 수를 사용하고 히트들을 나타내고, 상기 히트들의 유효 양은 링크 바탕 스팸에 의해 인위적으로 부스트 되게 한다. 주어진 웹 도큐먼트에 대한 유효 양의 결정은 주어진 웹 도큐먼트의 링크 바탕 인기도(예를들어, 페이지순위) 및 신뢰가치(예를들어, 신뢰순위) 사이의 불일치를 부분적으로 평가하는 기술의 결합에 따른다. 주어진 웹 도큐먼트의 유효 양의 결정을 위한 기술들은 이후 상세히 기술된다.

[0033] 네트워크 실행

[0034] 도 1은 본 발명의 실시예들을 실행하기 위하여 사용될 수 있는 하나 또는 그 이상의 클라이언트 시스템들(20_{1-N})을 포함하는 정보 검색 및 통신 네트워크(10)의 일반적인 개요를 도시한다. 컴퓨터 네트워크(10)에서, 클라이언트 시스템(들)(20_{1-N})은 인터넷(40), 또는 다른 통신 네트워크(예를들어, 임의의 로컬 영역 네트워크(LAN) 또는 광역 네트워크(WAN) 접속)를 통하여 임의의 수의 서버 시스템들(50_1 내지 50_N)에 결합된다. 여기에 기술될 바와 같이, 클라이언트 시스템(들)(20_{1-N})은 예를들어, 미디어 콘텐트 및 웹 페이지들 같은 다른 정보에 액세스, 수신, 검색 및 디스플레이하기 위하여 임의의 서버 시스템들(50_1 내지 50_N)과 통신하도록 본 발명에 따라 구성된다.

[0035] 도 1에 도시된 시스템의 몇몇 엘리먼트들은 여기에 상세히 설명될 필요가 없는 통상적인 잘 공지된 엘리머트들을 포함한다. 예를 들어, 클라이언트 시스템(20)은 데스크톱 퍼스널 컴퓨터, 워크스테이션, 랩톱, 퍼스널 디지털 어시스탄트(PDA), 셀 폰, 또는 임의의 WAP 실행 가능 장치 또는 인터넷에 직접적으로 또는 간접적으로 인터페이싱할 수 있는 임의의 다른 컴퓨팅 장치를 포함할 수 있다. 클라이언트 시스템(20)은 통상적으로 마이크로소프트의 인터넷 익스플로러™ 브라우저, 네스케이프 네비게이터™ 브라우저, 모질라™ 브라우저, 오페라™ 브라우

저, 애플의 사파리™ 또는 셀 폰, PDA 또는 다른 무선 장치의 경우 WAP 실행 가능 브라우저, 또는 등등 같은 브라우징 프로그램을 운용하여, 클라이언트 시스템(20_{1-N})의 사용자가 인터넷(40)을 통하여 서버 시스템들(50_1 내지 50_N)로부터 이용 가능한 정보 및 페이지들에 액세스, 처리 및 뷰잉하게 한다. 클라이언트 시스템(20)은 통상적으로 서버 시스템들(50_1 내지 50_N) 또는 다른 서버들에 의해 제공된 페이지들, 형태들 및 다른 정보와 관련하여 디스플레이(예를들어, 모니터 스크린, LCD 디스플레이, 등등)상 브라우저에 의해 제공된 그래픽 사용자 인터페이스(GUI)와 인터페이싱하기 위한 키보드, 마우스, 터치 스크린, 펜 또는 등등 같은 하나 또는 그 이상의 사용자 인터페이스 장치들(22)을 포함한다. 본 발명은 네트워크들의 특정 글로벌 관련 세트에 관한 인터넷에 사용하기에 적당하다. 그러나, 다른 네트워크들이 인트라넷, 익스트라넷, 가상 사적 네트워크(VPN), 비 TCP/IP 바탕 네트워크, 임의의 LAN 또는 WAN 또는 등등 같은 인터넷 대신 또는 상기 인터넷에 부가하여 사용될 수 있다는 것이 이해된다.

[0036] 일실시예에 따라, 클라이언트 시스템(20) 및 모든 구성요소들은 인텔 펜티엄™ 처리기, AMD 애슬론™ 처리기, 애플의 파워 PC, 또는 등등 또는 다중 처리기들 같은 중앙 처리 유니트를 사용하는 컴퓨터 소프트웨어 운용을 포함하는 애플리케이션을 사용하여 구성할 수 있는 오퍼레이터이다. 여기에 기술된 바와 같은 데이터 및 미디어 콘텐트를 통신, 처리 및 디스플레이하기 위해 클라이언트 시스템(20)을 동작 및 구성하기 위한 컴퓨터 소프트웨어는 바람직하게 하드 디스크에 다운로드되고 저장되지만, 전체 프로그램 코드, 또는 그의 일부들은 ROM 또는 RAM 같은 잘 공지된 임의의 다른 휘발성 또는 비휘발성 메모리 매체 또는 장치에 저장될 수 있거나, 컴팩트 디스크(CD) 매체, 디지털 다기능 디스크(DVD) 매체, 플로피 디스크 및 등등 같은 프로그램 코드를 저장할 수 있는 임의의 매체상에 제공될 수 있다. 부가적으로, 전체 프로그램 코드, 또는 그의 일부들은 임의의 통신 매체 및 프로토콜들(예를들어, TCP/IP, HTTP, HTTPS, 이더넷 또는 다른 종래 매체 및 프로토콜들)을 사용하여, 인터넷을 통해 소프트웨어 소스, 예를들어 서버 시스템들(50_1 내지 50_N)중 하나로부터 클라이언트 시스템(20)으로 전송되고 다운로드되거나, 임의의 다른 네트워크 접속(예를들어, 익스트라넷, VPN, LAN 또는 다른 종래의 네트워크들)을 통하여 전송될 수 있다.

[0037] 본 발명의 측면들을 실행하기 위한 컴퓨터 코드가 C, C++, HTML, XML, 자바, 자바스크립, 등등, 코드, 또는 임의의 다른 적당한 스크립트 언어(예를들어, VBScript), 또는 클라이언트 시스템(20)에서 실행되거나 클라이언트 시스템(20), 또는 시스템들(20_{1-N})에 컴파일될 수 있는 임의의 적당한 프로그램 가능 언어일 수 있다는 것이 인식되어야 한다. 몇몇 실시예들에서, 코드는 클라이언트 시스템(20)에 다운로드되고, 필요한 코드는 서버에 의해 실행되거나, 클라이언트 시스템(20)에 이미 제공된 코드는 실행된다.

[0038] 검색 시스템

[0039] 도 2는 본 발명의 일실시예에 따라 미디어 콘텐트를 통신하기 위한 다른 정보 검색 및 통신 네트워크(110)를 도시한다. 도시된 바와 같이, 네트워크(110)는 클라이언트 시스템(120), 하나 또는 그 이상의 콘텐트 서버 시스템들(150), 및 검색 서버 시스템(160)을 포함한다. 네트워크(110)에서, 클라이언트 시스템(120)은 인터넷(140) 또는 다른 통신 네트워크를 통하여 서버 시스템들(150 및 160)에 통신적으로 결합된다. 상기에 논의된 바와 같이, 클라이언트 시스템(120) 및 그의 구성요소들은 인터넷(140) 또는 다른 통신 네트워크들을 통하여 서버 시스템들(150 및 160) 및 다른 서버 시스템들과 통신하도록 구성된다.

[0040] 1. 클라이언트 시스템

[0041] 일실시예에 따라, 클라이언트 시스템(120)에서 실행하는 클라이언트 애플리케이션(모듈 125로서 표현됨)은 서버 시스템들(150 및 160)과 통신하고 상기 서버시스템들로부터 수신된 데이터 콘텐트를 처리 및 디스플레이하기 위한 클라이언트 시스템(120) 및 그의 부품들을 제어하기 위한 명령들을 포함한다. 클라이언트 애플리케이션(125)은, 비록 클라이언트 애플리케이션 모듈(125)이 상기 논의된 바와 같이 플로피 디스크, CD, DVD 등 같은 임의의 소프트웨어 저장 매체상에 저장될 수 있지만, 원격 서버 시스템(예를들어, 서버 시스템들 150, 서버 시스템 160 또는 다른 원격 서버 시스템) 같은 소프트웨어 소스로부터 클라이언트 시스템(120)으로 전송 및 다운로드된다. 예를들어, 일측면에서, 클라이언트 애플리케이션 모듈(125)은 인터넷(140)을 통하여 다양한 오브젝트들, 프레임들 및 윈도우들의 데이터를 조작 및 렌더링하기 위하여, 예를들어, 내장된 자바스크립트 또는 액티브엑스 컨트롤들 같은 다양한 컨트롤들을 포함하는 HTML 레퍼(wrapper)의 클라이언트 시스템(120)에 제공될 수 있다.

[0042] 부가적으로, 클라이언트 애플리케이션 모듈(125)은 검색 요구들 및 검색 결과 데이터를 처리하기 위한 검색 모

듈(126), 예를들어 브라우저 윈도우즈 및 다이얼로그 박스들 같은 텍스트 및 데이터 프레임들의 데이터 및 미디어 콘텐트 및 액티브 윈도우즈들을 렌더링하기 위한 사용자 인터페이스 모듈(127), 및 클라이언트(120)상에 실행하는 다양한 애플리케이션들과 인터페이싱 및 통신하기 위한 애플리케이션 인터페이스 모듈(128) 같은 데이터 및 미디어 콘텐트를 처리하기 위한 다양한 소프트웨어 모듈들을 포함한다. 애플리케이션 인터페이스 모듈(128)이 인터페이스하기 위하여 바람직하게 구성되는 클라이언트 시스템(120)상에서 실행하는 다양한 애플리케이션들의 예들은 다양한 이메일 애플리케이션들, 인스탄스 메시징(IM) 애플리케이션들, 브라우저 애플리케이션들, 도큐먼트 관리 애플리케이션들 및 등등을 포함한다. 게다가, 인터페이스 모듈(127)은 클라이언트 시스템(120)상에서 구성된 디폴트 브라우저 또는 다른 브라우저 같은 브라우저를 포함할 수 있다.

[0043] 2. 검색 서버 시스템

[0044] 일실시예에 따라, 검색 서버 시스템(160)은 검색 결과 데이터 및 미디어 콘텐트를 클라이언트 시스템(120)에 제공하기 위하여 구성된다. 클라이언트 서버 시스템(150)은 예를들어 검색 서버 시스템(160)에 의해 제공된 검색 결과 페이지들에서 선택된 링크들에 응답하여, 웹 페이지들 같은 데이터 및 미디어 콘텐트를 클라이언트 시스템(120)에 제공하도록 구성된다. 몇몇 변형들에서, 검색 서버 시스템(160)은 콘텐트에 대한 링크들 및/또는 다른 레퍼런스들뿐 아니라, 또는 대신 콘텐트를 리턴한다.

[0045] 일실시예에서 검색 서버 시스템(160)은 인덱스된 페이지들, 등등을 나타내는 페이지들, 페이지들에 대한 링크들, 데이터가 거주되는 다양한 페이지 인덱스들(170)을 인용한다. 페이지 인덱스들은 자동 웹 크로울러(crawler)들, 스파이더들, 등등뿐 아니라 계층 구조내 웹 페이지들을 분류 및 순위 결정하기 위한 자동 또는 반자동 분류 알고리듬 및 인터페이스들을 포함하는 다양한 수집 기술들에 의해 생성될 수 있다. 이들 기술들은 페이지 인덱스(170)를 생성하고 검색 서버 시스템(160)에 이를 이용하게 하는 검색 서버 시스템(160) 또는 독립된 시스템(도시되지 않음)에서 실행될 수 있다.

[0046] 검색 서버 시스템(160)은 검색 모듈(126) 같은 클라이언트 시스템으로부터 수신된 다양한 검색 요구들에 응답하여 데이터를 제공하도록 구성된다. 예를들어, 검색 서버 시스템(160)은 주어진 질문(예를들어, 질문시 검색 용어들의 발생 패턴들에 의해 측정되는 논리적 관련성; 문맥 식별자들; 페이지 스폰서쉽; 등등을 바탕으로)에 관련하여 웹 페이지들을 처리 및 순위를 매기기 위한 검색 관련 알고리듬들로 구성될 수 있다.

[0047] 링크 바탕 스팸 검출

[0048] 도 2에 도시된 바와 같이, 검색 서버 시스템(160)은 변형된 검색 리스트를 리턴하는 링크 바탕 스팸 검출기(180)와 결합하여 작동하고 그 출력(결과들, 제안들, 미디어 콘텐트, 등등)을 제공하고, 여기서 웹 스팸 페이지들은 리스트로부터 강등되거나 제거되었다. 검색 서버 시스템(160)은 본 발명의 실시예들에 따라 검색 엔진을 동작시키기 위하여 구성된다. 검색 엔진은 3 부분들: 하나 또는 그 이상의 스파이더들 162, 데이터베이스 163 및 툴들/애플리케이션들 167로 구성된다. 스파이더들(162)은 인터넷 수집 정보에서 크롤한다; 데이터베이스(163)는 스파이더들이 수집하는 정보뿐 아니라 다른 정보를 포함한다; 및 툴들/애플리케이션들(167)은 데이터베이스를 통하여 검색하기 위하여 사용자들에 의해 사용되는 검색 툴(166) 같은 애플리케이션들을 포함한다. 데이터베이스(167)는 검색 툴에 의해 사용되는 페이지 인덱스(170)를 포함한다. 게다가, 본 발명의 실시예에 따른 검색 엔진은 스팸 검출기(180)를 포함한다. 스팸 검출기(180)는 하기에 기술된 다양한 알고리듬들을 실행하고, 페이지 인덱스(170)의 페이지들을 위해 웹 스팸 측정치(181)를 저장한다. 상기된 바와 같이, 본 발명의 실시예들에 따른 스팸 검출기(180)는 유효 히트들의 양에 해당하고 검색 툴(166) 및 페이지 인덱스(170)와 결합하여 작동하는 측정치를 평가하고 유효 히트들의 양이 링크 바탕 스팸에 의해 인위적으로 부스트되는 히트들을 강등시킨다. 주어진 웹 도큐먼트에 대한 유효 양의 결정은 주어진 웹 도큐먼트의 링크 바탕 인기도(예를들어 페이지순위) 및 신뢰가치(예를들어, 신뢰순위) 사이의 불일치를 부분적으로 평가하는 기술들의 결합에 의존한다. 일실시예에서, 웹 스팸 검출기(180)는 인덱스의 페이지들에 대한 웹 스팸 측정치(181)를 계산하기 위하여 페이지 인덱스(170)의 모든 페이지들을 처리하고 데이터베이스(163)에 웹 스팸 측정치(181)를 저장한다. 측정치(181)는 도큐먼트가 검색 결과들에 포함되게 하는 검색 질문에 무관하다.

[0049] 주어진 웹 도큐먼트에 대한 스팸 검출기(180)에 의한 스팸 팜의 유효 양의 결정은 부분적으로 주어진 웹 도큐먼트의 링크 바탕 인기도(예를들어, 페이지순위) 및 신뢰가치(예를들어, 신뢰순위) 사이의 차 평가에 의존한다. 주어진 웹 도큐먼트의 신뢰가치의 결정은 부분적으로 주어진 페이지가 신뢰가치있는(즉, 스팸없음 도큐먼트들) 것으로 알려진 웹 도큐먼트의 초기 시드 세트로부터 얼마나 떨어져 있는가에 의존한다. 따라서, 본 발명의 실시예들에 따른 검색 엔진은 신뢰성 있는 웹 도큐먼트들의 초기 시드 세트(185)를 형성하기 위하여 페이지 인덱스(170)와 결합하여 작동하는 시드 세트 생성기(184)를 포함한다. 웹 스팸 측정치(181)를 형성하는 스팸 검출

기(180)의 동작 및 시드 세트(185)를 형성하는 시드 세트 생성기(184)의 동작은 추후에 상세히 기술된다.

[0050] 스팸 팜, 페이지순위 및 신뢰순위

[0051] 이 섹션에서, 스팸 팜, 인링크 페이지 순위(일반적으로 "페이지순위"라 함), 및 신뢰 순위의 개념들은 기술된다. 스팸 팜은 중요성을 부스트하기 위하여 스팸 타겟 페이지를 가리키는 인위적으로 생성된 페이지들의 세트이다. 신뢰 순위("신뢰순위")는 고품질 페이지들의 서브세트에 대한 특정 텔레포테이션(즉, 점프들)을 가진 페이지순위의 형태이다. 여기에 기술된 기술들을 사용하여, 검색 엔진은 나쁜 페이지들(웹 스팸 페이지들)을 자동으로 발견할 수 있고 보다 구체적으로 인위적 스팸 팜들(인용 페이지들의 수집물들)의 생성을 통하여 중요성을 부스트하기 위하여 생성된 웹 스팸 페이지들을 발견한다. 특정 실시예들에서, 균일한 텔레포테이션 및 신뢰 순위 처리를 가진 페이지순위 처리는 수행되고 그 결과들은 페이지의 "스팸성" 또는 페이지들의 수집물의 "스팸성"의 검사의 일부로서 비교된다. 게다가, 신뢰순위 처리에 대한 입력들을 구성하는 새로운 방법은 하기에 기술된다.

[0052] 본 발명의 일측면은 둘러싸는 하이퍼링크 구조의 분석을 바탕으로 스팸 페이지들(적어도 일부)의 식별에 관한 것이다. 특히, 스팸 팜 크기들을 평가하는 새로운 방법은 사용된다. 스팸없음 페이지들이 스팸을 거의 가리키지 않기 때문에, 신뢰순위의 특정 권한 분배는 스팸없음 페이지들 및 스팸 페이지들 사이의 일정 분리도를 유발하고; 고품질 스팸없음 웹 페이지들은 신뢰순위에 의해 할당된 가장 높은 스코어들을 가지는 것으로 기대된다.

[0053] 신뢰순위는 지시된 다른 페이지들의 스코어들에 따라 각각의 웹 페이지에 수치 스코어들을 할당하는 잘 공지된 웹 분석 알고리듬, 페이지순위에 관련된다. 페이지순위는 텔레포테이션 기술을 사용한다: 총 스코어의 특정 양은 일반적으로 균일한 분배인 소위 텔레포테이션 분배에 따라 몇몇 또는 모든 페이지들에 전달된다. 균일한 텔레포테이션 분배를 사용하는 대신, 신뢰순위는 신뢰성 있는(스팸없음) 웹 페이지들(즉, 소위 "시드 세트")에만 텔레포테이션을 제공한다. 이것은 실제 시드 세트로부터 다른 페이지들에게 스코어들을 분배하는 것을 유발한다.

[0054] 하기 설명들은 웹 페이지들을 인용한다. 그러나, 논리, 실행 및 알고리듬들은 (1) 사이트들의 웹(웹 콘텐트/페이지들의 논리 그룹들 및 하나의 권한과 관련된 다른 형태의 웹 도큐먼트들), (2) 호스트들 사이의 그래프 에지들의 몇몇 정의를 가지는(예를들어, 만약 두 개의 호스트들이 하이퍼링크에 의해 접속된 적어도 하나의 페이지를 가진 두 개의 호스트들이 하나의 링크를 가지는 호스트 그래프, 또는 다른 검사들) 호스트들의 웹(호스트랭크)에 의해 표현되는 사이트들의 웹 근접도, (3) 임의의 다른 웹 페이지 그래프 집합, 및/또는 (4) 소개의 강도를 반영하는 웨이트들과 연관된 링크들의 수집물에 똑같이 응용할 수 있다.

[0055] 스팸 팜

[0056] 스팸 팜은 중요성을 부스트하기 위하여 스팸 타겟 페이지를 가리키는 인위적으로 생성된 페이지들의 세트(또는 선택적으로 호스트들)이다. 도 3A-3B는 두 개의 간단한 스팸 팜들을 도시하는 예시적인 도면들이다.

[0057] 도 3A는 스팸 팜이 타겟 스팸 페이지(s)를 가리키는 모두 m 페이지들을 가진다는 것을 도시한다. 스팸 팜 크기의 우수한 평가를 얻기 위한 처리는 하기에 기술된다. 매 페이지(i)에 대해, 수(M_i)는 계산되고, 여기서 수(M_i)는 페이지의 "유효 양"이라 한다. 웹 스팸 페이지들에 대해, M은 페이지를 부스트하는 스팸 팜의 크기의 우수한 평가치로서 사용한다.

[0058] 간단한 스팸 팜의 경우, 유효 양은 m에 근접한다. 보다 복잡한 팜에 대해, 예로서 도 3b에 도시된 스팸 팜에서, 유효 양(M)은 표시자로서 사용하고, 높은 M 값은 스팸 팜을 가리킨다. 상기 설명이 웹 페이지들을 인용하지만, 개념들은 페이지들, 호스트들 및 등등의 그룹들에 적용될 수 있다는 것이 인식되어야 한다.

[0059] 페이지순위 및 신뢰순위

[0060] 페이지순위의 개념은 웹 페이지들의 분석시 사용한다. 페이지순위에 대한 많은 가능한 정의들 중에서, 다음 페이지 순위의 선형 시스템 정의는 사용된다:

[0061] $x = cT^{T}x + (1-c)v$ (방정식 1)

[0062] 방정식 1에서:

[0063] T는 페이지 i로부터 페이지 j로 향하는 링크(i→j)가 있다면 엘리먼트들이 $T_{ij} = 1/\mathrm{outdeg}(i)$인 전이 매트릭스이고, 그렇지 않으면 영이다. 여기서, outdeg(i)는 매트릭스 T 확률을 형성하기 위하여 표준화 인자로서 사용하

는 페이지(i)상 아웃링크들의 수이고,

[0064] c는 범위(0.7-0.9)에서 선택되는 텔레포테이션 상수이고,

[0065] $x = (x_i)$는 권한 벡터이고, 여기서 인덱스 i는 모두 n 페이지들상에서 운용되고, i = 1:n(n은 웹 페이지들의 수),

[0066] $v = v(v_i)$는 가능성 분포도인 것으로 가정된 텔레포테이션 벡터이고, $0 \leq v_i \leq 1$, $v_1 + \ldots + v_n = 1$.

[0067] 방정식 1을 해결하기 위한 반복 방법들은 공지되었다. 방정식 1은 텔레포테이션 벡터에 관련하여 선형인 권한 벡터를 정의하는 장점을 가진다.

[0068] 페이지순위에 대해, p는 단일 텔레포테이션에 해당하는 방정식 1의 솔루션을 제공하는 권한 벡터이다(즉, $v_i = 1/n$일 때). 신뢰순위에 대해, t는 특정 텔레포테이션에 해당하는 방정식 1의 솔루션을 제공하는 권한 벡터이다(즉, v의 k 엘리먼트들이 영이 아니고 나머지가 영이도록 하는 v, 영이 아닌 엘리먼트들은 신뢰성 있는 세트에서 대응 인덱스들(i)을 가짐).

[0069] 유효 양의 평가

[0070] 웹 페이지의 유효 양은 웹 페이지가 스팸 페이지를 결정하는 것을 돕기 위한 표시자로서 사용된다.

[0071] 평가치 구성

[0072] 잠재적인 스팸 페이지(s)에 대해, 임의의 웹 페이지(i)중에서,

[0073] $p_s - t_s = b_s^{\text{상승}} + b \cdot p_s^{\text{누설}} + (1-c)/n$ (방정식 2)인 것이 수학적으로 도시되고, 여기서 방정식의 우측 제 1 항은 지원 스팸 팜(상기 팜은 스팸없음 페이지들의 경우 비어있거나 존재하지 않음)으로부터 페이지에 도달하는 부스트로 인한 것이고, 제 2 항은 스팸 페이지들에 때때로 잘못 지적한 스팸없음 페이지들로부터 권한 누설로 인한 것이다. 이런 누설은 웹의 나머지로부터 주어진 페이지로 여러가지 우연한 하이퍼링크들을 나타내는 점선 화살표로서 도 3A-B에 도시된다. 스팸 페이지(s)에 대해, 제 1 항은 스팸 팜을 생성하는 스팸 생성자에 대한 동기가 높은 s의 페이지순위를 형성하기 때문에 매우 우세하다. 간단한 팜에 대해,

[0074] $p_s^{\text{상승}} = m \cdot c(1-c)/n$ (방정식 3)

[0075] 유사한 식은 다른 구조의 팜에서 유효하다. 예를들어, 백 링크들을 가진 팜에 대해,

[0076] $p_s^{\text{상승}} = m \cdot c(1-c)/(1-c^2)n$ (방정식 4)

[0077] $p_s^{\text{누설}} << p_s^{\text{상승}}$ (방정식 5)인 조건에서, 간단한 스팸 팜의 크기(m)에 대한 우수한 평가치는 방정식(2) 및 (3)으로부터 다음과 같이 구성된다.

[0078] $M_s = n(p_s - t_s)/c(1-c)$ (방정식 6)

[0079] 방정식 6은 임의 웹 페이지(i)에 대해 계산될 수 있는 유효 양(M_i)을 정의한다. 상기된 바와 같이, 만약 i가 간단한 스팸 팜에 의해 부스트된 스팸 페이지이면, M_i는 실제 팜 크기(m)에 가까워지고, 다른 구조의 팜들에 대해, 방정식 4에 의해 도시된 바와 같이 실제 팜 크기로부터 하나의 상수만큼 다르다. 상기 차이는 실제 스팸 팜들이 오히려 크다는 사실의 측면에서 중요하지 않다(예를들어, 수백만의 부스트 페이지들은 부정적으로 생성된다).

[0080] 스팸없음 페이지에 대해, M_i는 절대항들 또는 p_i에 관련하여 크지 않은 약간의 수일 것이다. 본 발명의 실시예들에 따른 링크 바탕 스팸 검출은 이것을 발견하고 표시자로서 M_i를 바탕으로 잠재적인 웹 스팸 페이지로서 상기 페이지를 지명하지 않는다.

[0081] 스팸 검출 처리

[0082] 다음 예시적인 처리는 링크 바탕 스팸을 검출하기 위하여 사용된다. 상기 처리는 매우 간단하고 효과적이고, 가장 높은 유효 양을 가진 페이지들을 발견하는 것을 목적으로 한다. 그러나, 유효 양은 만약 방정식(5)이 만

족되면 스팸 크기에 대한 우수한 근접치를 제공하여, 신뢰성 있는 웹 페이지들로부터의 인기도가 스팸 페이지들에 의한 인위적인 부스트로 인한 페이지의 링크 바탕 인기도 보다 매우 작은 할당으로 인한 페이지의 링크 바탕 인기도를 보장한다. 방정식 5의 조건하에서, 스팸 검출 처리는 합법적으로 인기있는 페이지들과 링크한 스팸 팜에 의하여 인기가 만들어진 페이지들 사이를 구별할 수 있다. 본 발명의 실시예들에 따른 기술은 방정식 5의 조건이 충족되는 것을 보장한다. 이것은 $\eta>1$이 임계치로서 사용하는 알고리듬 파라미터인 하기 단계(C)에서 수행된다. C의 큰 비율들이 방정식 5를 만족하는 페이지들에 해당하는 것을 알 수 있다. 전체적으로, 예시적인 처리는 다음과 같다:

[0083] A. 모든 페이지들(호스트들, 등등)에 대해, 리스트(예를들어, 질문과 관련된 히트들의 리스트, 또는 페이지 인덱스)의 i는 방정식 6에 따라 유효 양(M_i)을 발견한다.

[0084] B. M_i의 감소 순위에서 페이지들(i) 분류 및 분류된 리스트(sorted list)의 상위 부분을 유지 또는 식별. 선택적으로, 전체 리스트는 너무 많은 리소스들을 요구하지 않더라도 유지될 수 있고, 그러므로 낮은 M_i을 유지하는 것이 보다 효과적이지 않다. 이런 식별 및/또는 유지가 임의의 단계에서 수행될 수 있다. 선택 처리 부분은 높은 M_i 및 높은 M_i/P_i 모두를 가진 페이지들을 선택하는 것에 관한 것이다.

[0085] C. 리스트에 유지된 모든 페이지들(i)에 대해 비율들(M_i/P_i) 발견.

[0086] D. $M_i/P_i<\eta$을 페이지들(i)로부터 삭제.

[0087] E. 유지된 페이지들이 스팸 구성.

[0088] 실험시, 검출된 스팸 페이지들은 실제로 대부분의 경우들에서, 스팸(사람 판단에 의함)인 것으로 확인된다. 이것은 부정적인 잠재 비율이 이들 기술들을 사용하여 낮춰질 수 있다는 것을 의미한다.

[0089] 시드 세트

[0090] 상기된 처리는 소위 시드 세트와 연관된 특정한 텔레포테이션 분배를 가진 신뢰순위, 즉 방정식 1의 솔루션에 따른다. 시드 세트는 스팸없음으로 알려진 k 고품질 페이지들의 세트이다. 본 발명의 실시예들의 측면은 신뢰가치(즉, 스팸없음) 페이지들 또는 사이트들에서 적당한 시드의 발견에 관한 것이다. 신뢰성 있는 웹 페이지들의 시드 세트를 식별하는 한가지 방법은 인간 편집 판단을 바탕으로 특정한 웹 페이지들을 추천하는 것이다. 그러나, 인간 평가는 값비싸고 시간 소비적이다. 실행 가능한 대안으로서 시드 세트를 수동으로 선택하는 것의 옵션을 유지하면서, 시드 세트를 반자동으로 구성하는 다른 기술은 하기된다.

[0091] 시드 선택 처리는 시드 페이지들이 두 개의 중요한 특징들을 가져야 한다는 의견에 따른다, 즉 1) 다수의 다른 페이지들은 시드 페이지들로부터 시작하여 마주하는 웹 페이지들상 아웃링크들을 반복적으로 따라야 도달할 수 있어야 하고, 및 2) 시드 페이지들은 매우 고품질이어서, 스팸없음에서 스팸으로 링크를 조우할 기회가 최소화되어야 한다.

[0092] 제 1 특징을 보장하기 위하여, 모든 페이지들의 순위(즉, 페이지 인덱스 페이지들)는 형성된다. 이를 위해, 방정식 7에 의해 도시된 다음 선형 시스템은 사용된다.

[0093] $y = cU^{T}y + (1-c)v$ (방정식 7)

[0094] 이 시스템에서,

[0095] - U는 만약 링크 j→i이면 엘리먼트들이 $U_{ij} = 1/\mathrm{indeg}(i)$인 리버스 전이이거나, 그렇지 않으면 영이다. 여기서 indeg(i)는 매트릭스 U 확률을 형성하기 위하여 표준화 인자로서 사용하는 페이지(i)에 대한 인링크들의 수이고,

[0096] - c는 일반적으로 범위(0.7-0.9)에서 선택된 텔레포테이션 상수이고,

[0097] - $y = (y_i)$는 권한 벡터이고, 여기서 인덱스 i는 모두 n 페이지들에서 운용되고, i = 1:n,

[0098] - 가능성 분배라 가정된 $v = v(_i)$는 텔레포테이션 벡터이고, $0 \le v_i \le 1$, $v_1 + \ldots + v_n = 1$.

[0099] 방정식 7이 정상적인 변이 매트릭스(T) 대신 리버스 전이 매트릭스(U)를 사용하는 것을 제외하고, 방정식 7에 의해 기술된 시스템이 방정식 1과 유사한 것이 주의된다. 리버스 전이 매트릭스는 리버스된 링크들의 방향성을

가진 웹 그래프에 해당한다. 이를 위하여, 균일한 텔레포테이션을 가진 방정식 7에 대한 솔루션은 인버스 페이지순위라 한다. 인버스 페이지순위는 얼마나 많은 웹이 페이지상 아웃링크들을 따라 하나의 페이지로부터 도달될 수 있는 가의 측정치이다.

[0100] 시드 페이지들의 제 2 특징을 보장하기 위하여, 가장 높은 인버스 페이지순위를 가진 페이지들은 인간 에디터에 의해 추가로 처리된다. 인간 에디터는 후보자들(인버스 페이지순위에 의해 측정된 바와 같은 높은 커버리지를 제공하는 페이지들)이 실제로 고품질 스팸없음 페이지들인 것을 선택한다. 인간 에디터에 의해 선택된 페이지들은 상기된 바와 같이 시드 세트에 포함되고 신뢰순위 계산시 사용된다.

[0101] 예시적인 시드 세트 구성 처리는 다음과 같이 요약된다:

[0102] A. 모든 페이지들(호스트들, 등등)에 대해, i가 방정식 7에 따라 인버스 페이지순위(y_i)를 발견

[0103] B. y_i의 감소 순위에서 페이지들(i) 분류 및 분류된 리스트의 상위 순위를 유지하거나, 그렇지 않으면 가장 높은 순위 페이지들의 세트를 식별 및 유지

[0104] C. 리스트에 유지된 페이지들의 품질을 평가하기 위하여 인간 에디터(들) 사용

[0105] D. 에디터(들)에 의해 적당하지 않은 것을 인식되는 리스트 페이지들 삭제

[0106] E. 유지된 페이지들이 시드 세트 구성.

[0107] 실험은 페이지순위 및 신뢰순위로부터 유도된 양 평가를 바탕으로 결과적인 시드 세트가 신뢰순위 계산 및 스팸 검출에 적당하다는 것을 나타낸다.

[0108] 여기에 기술된 실시예들은 월드 와이드 웹(또는 상기 월드 와이드 웹의 서브세트)이 검색 코퍼스로서 사용되는 경우에 특정한 웹 사이트들, 링크들 및 다른 기술을 인용할 수 있다. 여기에 기술된 시스템들 및 처리들이 다른 검색 코퍼스(전자 데이터베이스 또는 도큐먼트 저장소 같은)에 사용하기 위하여 제공될 수 있고 그 결과들이 콘텐트뿐 아니라 콘텐트가 발견될 수 있는 위치들에 대한 링크들 또는 인용들인 것이 이해되어야 한다.

[0109] 따라서, 비록 본 발명이 특정 실시예들과 관련하여 기술되었지만, 본 발명이 다음 청구항들의 범위내에서 모든 변형들 및 등가물들을 커버하는 것으로 의도된다는 것이 인식된다.

도면의 간단한 설명

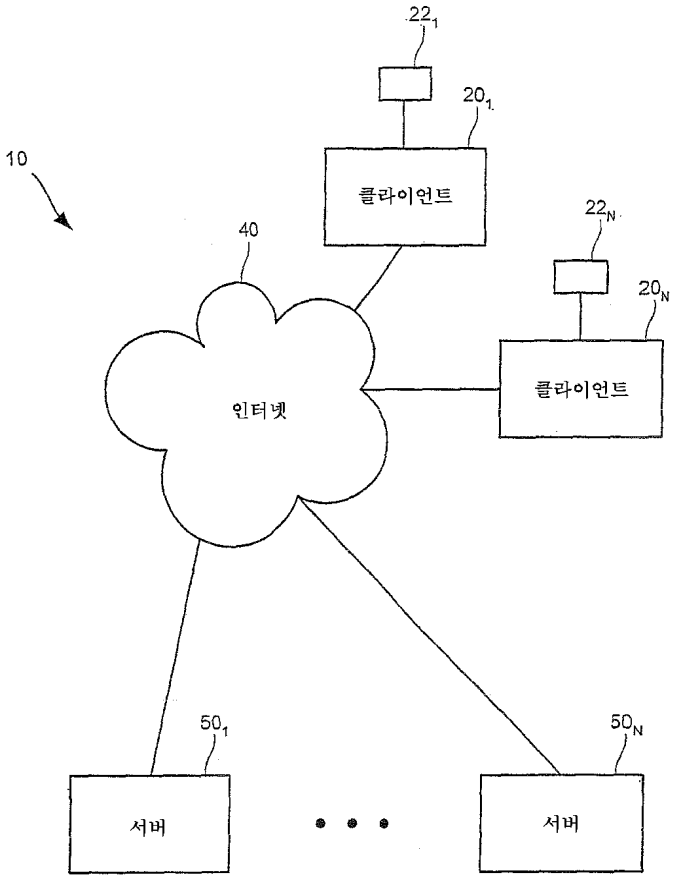
[0014] 도 1은 본 발명의 실시예들을 실행하기 위하여 사용될 수 있는 정보 검색 및 통신 네트워크의 예시적인 블록도이다.

[0015] 도 2는 본 발명의 실시예에 따른 정보 검색 및 통신 네트워크의 예시적인 블록도이다.

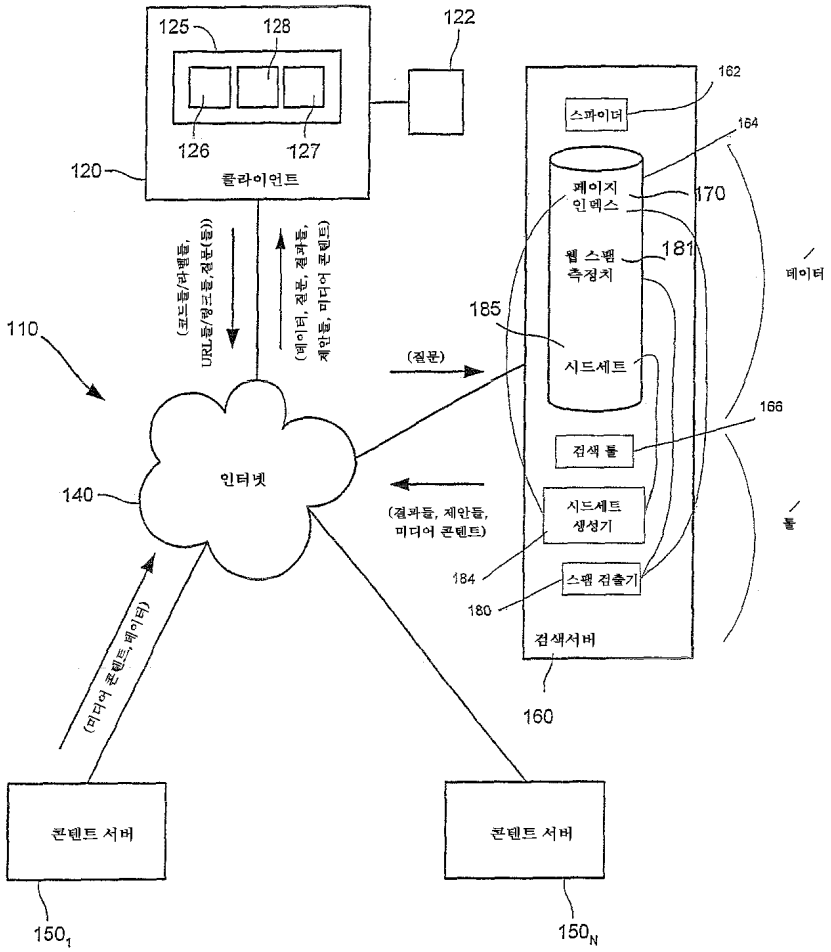
[0016] 도 3A-B는 간단한 스팸 팜들의 예시적인 도면들이다.

도면

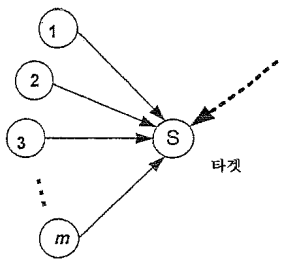
도면1



도면2

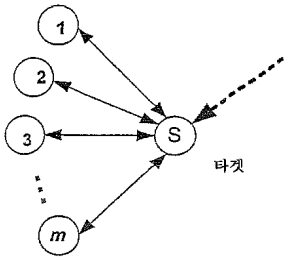


도면3A



간단한 스팸팜

도면3B



백링크들을 가진 간단한 스팸 팜