(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2015-0030405
(43) 공개일자 2015년03월20일

(51) 국제특허분류(Int. Cl.)
G06F 17/00 (2006.01) *G06F 17/30* (2006.01)
(21) 출원번호 10-2013-0109597
(22) 출원일자 2013년09월12일
심사청구일자 2013년09월12일

(71) 출원인
이화여자대학교 산학협력단
서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)
(72) 발명자
임혜숙
서울 강남구 남부순환로 2803, 108동 1102호 (도곡동, 삼성래미안아파트)
이나라
서울 서초구 잠원로 213-10, 6동 803호 (잠원동, 한강아파트)
이정원
서울 광진구 용마산로1길 78, 1동 603호 (중곡동, 대영아파트)
(74) 대리인
김인철

전체 청구항 수 : 총 16 항

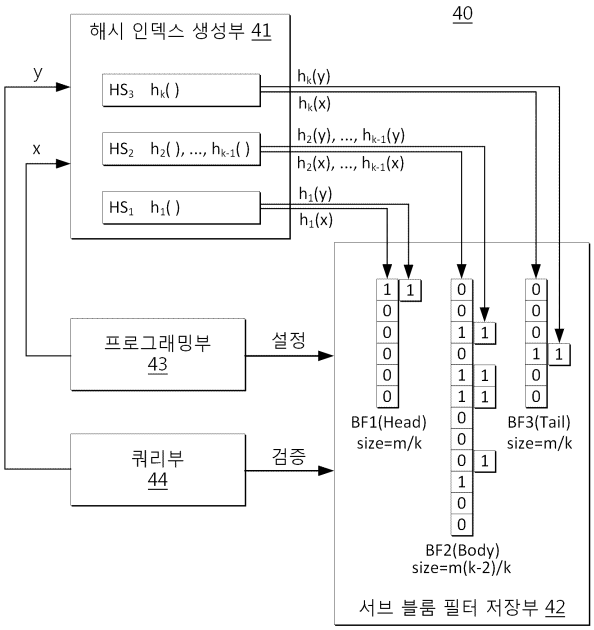
(54) 발명의 명칭 **거짓 양성 성능을 개선할 수 있는 블룸 필터 처리 장치 및 블룸 필터 처리 방법**

(57) 요 약

블룸 필터 처리 장치는 k 개의 해시 함수들을 j 개(j는 2부터 k까지의 정수)의 해시 함수 서브그룹으로 분할하여, 입력되는 원소 또는 데이터에 관하여, j 개의 해시 인덱스 서브그룹으로 분할된 k 개의 해시 인덱스들을 생성하는 해시 인덱스 생성부, 총 m 개의 비트 포지션들을 가지는 j 개의 서브 블룸 필터들을 보유하며, j 개

(뒷면에 계속)

대 표 도 - 도4



의 해시 함수 서브그룹들 중 i 번째(i는 1부터 j까지의 정수) 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 i 번째 서브 블룸 필터에 연결하는 서브 블룸 필터 저장부, 프로그래밍 시에 원소에 관하여 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 서브 블룸 필터 저장부의 i 번째 서브 블룸 필터에 설정하는 프로그래밍부 및 멤버쉽 쿼리 시에 입력 데이터에 관하여 생성된 i 번째 해시 인덱스 서브그룹의 해시 인덱스들에 따라 i 번째 서브 블룸 필터로부터 각각 출력되는 비트열에 기초하여, 입력 데이터의 멤버쉽을 검증하는 쿼리부를 포함할 수 있다.

이 발명을 지원한 국가연구개발사업

과제고유번호	1345192307
부처명	교육과학기술부
연구관리전문기관	한국연구재단
연구사업명	기초연구사업(학술진흥)-중견연구자지원사업-핵심연구(개인)
연구과제명	차세대인터넷을위한고속패킷전달기술에관한연구
기 여 율	50/100
주관기관	이화여자대학교
연구기간	2012.05.01 ~ 2015.04.30이 발명을 지원한 국가연구개발사업
과제고유번호	1345196793
부처명	교육과학기술부
연구관리전문기관	한국연구재단
연구사업명	교육・인력양성사업-2단계연구중심대학(BK21)육성사업-핵심분야
연구과제명	차세대정보통신시스템핵심기술연구
기 여 율	50/100
주관기관	이화여자대학교
연구기간	2006.03.01 ~ 2013.02.28

특허청구의 범위

청구항 1

k 개의 해시 함수들을 j 개(j는 2부터 k까지의 정수)의 해시 함수 서브그룹으로 분할하여, 입력되는 원소 또는 데이터에 관하여, j 개의 해시 인덱스 서브그룹으로 분할된 k 개의 해시 인덱스들을 생성하는 해시 인덱스 생성부;

총 m 개의 비트 포지션들을 가지는 j 개의 서브 블룸 필터들을 보유하며, j 개의 해시 함수 서브그룹들 중 i 번째(i는 1부터 j까지의 정수) 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 i 번째 서브 블룸 필터에 연결하는 서브 블룸 필터 저장부;

프로그래밍 시에 원소를 상기 해시 인덱스 생성부에 입력하고, 상기 원소에 관하여 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 상기 서브 블룸 필터 저장부의 i 번째 서브 블룸 필터에 설정하는 프로그래밍부; 및

멤버쉽 쿼리 시에 입력 데이터를 상기 해시 인덱스 생성부에 입력하고, 상기 입력 데이터에 관하여 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 i 번째 해시 인덱스 서브그룹의 해시 인덱스들이 생성되면, 생성된 i 번째 해시 인덱스 서브그룹의 해시 인덱스들에 따라 상기 서브 블룸 필터 저장부의 i 번째 서브 블룸 필터로부터 각각 출력되는 비트열에 기초하여, 상기 입력 데이터의 멤버쉽을 검증하는 쿼리부를 포함하는 블룸 필터 처리 장치.

청구항 2

청구항 1에 있어서, 상기 서브 블룸 필터들 중 적어도 하나의 서브 블룸 필터의 크기는 $\left[m/k\right]$이고, [X]는 정수 함수(greatest integer function)로서, 실수 X보다 크지 않으면서 가장 큰 정수인 것을 특징으로 하는 블룸 필터 처리 장치.

청구항 3

청구항 1에 있어서, 상기 서브 블룸 필터들의 수 j가 k보다 작은 경우에, 상기 j 개의 서브 블룸 필터들 중 j-1 개의 서브 블룸 필터들의 크기는 $\left[m/k\right]$인 정수 또는 $\left[m/k\right]$의 정수배인 정수로 결정되고, 나머지 하나의 서브 블룸 필터의 크기는 총 비트 포지션의 크기 m에서 j-1 개의 서브 블룸 필터들의 크기를 뺀 값으로 결정되는 것을 특징으로 하는 블룸 필터 처리 장치.

청구항 4

청구항 1에 있어서, 상기 서브 블룸 필터들의 수 j가 k와 같은 경우에, 상기 j 개의 서브 블룸 필터들 중 j-1 개의 서브 블룸 필터들의 크기는 $\left[m/k\right]$인 정수로 결정되고, 나머지 하나의 서브 블룸 필터의 크기는 총 비트 포지션의 크기 m에서 j-1 개의 서브 블룸 필터들의 크기를 뺀 값으로 결정되는 것을 특징으로 하는 블룸 필터 처리 장치.

청구항 5

청구항 1에 있어서, 상기 j 개의 서브 블룸 필터들 중 적어도 하나의 서브 블룸 필터의 크기는 $\left[m/k\right]$인 정수로 결정되고, 여기서 $\lceil X \rceil$는 천정 함수(ceiling function)로서, 실수 X보다 작지 않으면서 가장 작은 정수인 것을 특징으로 하는 블룸 필터 처리 장치.

청구항 6

청구항 1에 있어서, 상기 j 개의 서브 블룸 필터들 중 i 번째 서브 블룸 필터의 크기는 j 개의 해시 함수 서브

그룹들 중 i 번째 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들이 반드시 i 번째 서브 블룸 필터 내에 설정될 수 있는 조건을 만족하도록 결정되는 것을 특징으로 하는 블룸 필터 처리 장치.

청구항 7

청구항 1에 있어서, 상기 쿼리부는 입력 데이터에 대해 상기 해시 인덱스 서브그룹들의 각각에 상응하는 상기 서브 블룸 필터들의 각각에서 모두 비트 1이 출력되어야 입력 데이터에 대해 양성으로 판정하고, 만약 어느 한 서브 블룸 필터에서 비트 0이 나오면, 음성으로 판정하도록 동작하는 것을 특징으로 하는 블룸 필터 처리 장치.

청구항 8

컴퓨터를 청구항 1 내지 청구항 7 중 어느 한 청구항에 따른 블룸 필터 처리 장치로 작동시키도록 구현된 프로그램이 기록된 컴퓨터로 독출 가능한 기록 매체.

청구항 9

입력되는 원소 또는 데이터에 관하여, j 개의 해시 인덱스 서브그룹으로 분할된 k 개의 해시 인덱스들을 생성하도록 j 개(j는 2부터 k까지의 정수)의 해시 함수 서브그룹으로 분할되는 k 개의 해시 함수들을 제공하는 단계;

상기 j 개의 해시 함수 서브그룹들 중 i 번째(i는 1부터 j까지의 정수) 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들이 i 번째 서브 블룸 필터에 연결되도록, 총 m 개의 비트 포지션들을 가지는 j 개의 서브 블룸 필터들을 제공하는 단계;

프로그래밍 시에, 원소에 관하여 상기 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 생성하고, 생성된 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 i 번째 서브 블룸 필터에 설정하는 단계; 및

멤버쉽 쿼리 시에 입력 데이터에 관하여 상기 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 생성하고, 생성된 i 번째 해시 인덱스 서브그룹의 해시 인덱스들 따라 i 번째 서브 블룸 필터로부터 각각 출력되는 비트열에 기초하여, 상기 입력 데이터의 멤버쉽을 검증하는 단계를 포함하는 블룸 필터 처리 방법.

청구항 10

청구항 9에 있어서, 상기 서브 블룸 필터들 중 적어도 하나의 서브 블룸 필터의 크기는 $\lceil m/k \rceil$이고, [X]는 정수함수로서, 실수 X보다 크지 않으면서 가장 큰 정수인 것을 특징으로 하는 블룸 필터 처리 방법.

청구항 11

청구항 9에 있어서, 상기 서브 블룸 필터들의 수 j가 k보다 작은 경우에, 상기 j 개의 서브 블룸 필터들 중 j-1 개의 서브 블룸 필터들의 크기는 $\lceil m/k \rceil$인 정수 또는 $\lceil m/k \rceil$의 정수배인 정수로 결정되고, 나머지 하나의 서브 블룸 필터의 크기는 총 비트 포지션의 크기 m에서 j-1 개의 서브 블룸 필터들의 크기를 뺀 값으로 결정되는 것을 특징으로 하는 블룸 필터 처리 방법.

청구항 12

청구항 9에 있어서, 상기 서브 블룸 필터들의 수 j가 k와 같은 경우에, 상기 j 개의 서브 블룸 필터들 중 j-1 개의 서브 블룸 필터들의 크기는 $\lceil m/k \rceil$인 정수로 결정되고, 나머지 하나의 서브 블룸 필터의 크기는 총 비트 포지션의 크기 m에서 j-1 개의 서브 블룸 필터들의 크기를 뺀 값으로 결정되는 것을 특징으로 하는 블룸 필터 처리 방법.

청구항 13

청구항 9에 있어서, 상기 j 개의 서브 블룸 필터들 중 적어도 하나의 서브 블룸 필터의 크기는 $\lceil m/k \rceil$인 정수로

결정되고, 여기서 $[X]$는 천정 함수로서, 실수 X보다 작지 않으면서 가장 작은 정수인 것을 특징으로 하는 블룸 필터 처리 방법.

청구항 14

청구항 9에 있어서, 상기 j 개의 서브 블룸 필터들 중 i 번째 서브 블룸 필터의 크기는 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들이 반드시 i 번째 서브 블룸 필터 내에 설정될 수 있는 조건을 만족하도록 결정되는 것을 특징으로 하는 블룸 필터 처리 방법.

청구항 15

청구항 9에 있어서, 상기 입력 데이터의 멤버쉽을 검증하는 단계는,

상기 입력 데이터에 대해 상기 해시 인덱스 서브그룹들의 각각에 상응하는 상기 서브 블룸 필터들의 각각에서 모두 비트 1이 출력되어야 입력 데이터에 대해 양성으로 판정하고, 만약 어느 한 서브 블룸 필터에서 비트 0이 나오면, 음성으로 판정하는 단계를 포함하는 것을 특징으로 하는 블룸 필터 처리 방법.

청구항 16

컴퓨터에서 청구항 9 내지 청구항 15 중 어느 한 청구항에 따른 블룸 필터 처리 방법을 구동하도록 구현된 프로그램이 기록된 컴퓨터로 독출 가능한 기록 매체.

명 세 서

기 술 분 야

[0001] 본 발명은 블룸 필터에 관한 것으로, 더욱 상세하게는, 거짓 양성 성능을 개선할 수 있는 블룸 필터에 관한 것이다.

배 경 기 술

[0002] 블룸 필터(Bloom filter)는 1970년 B. Bloom에 의해 고안되었으며, 원소가 집합에 속하는지 여부를 검사할 수 있는 확률적 자료 구조이다.

[0003] 블룸 필터는 m 비트 크기의 비트열로 표현될 수 있는데, m 비트의 비트열 중에서 각각의 원소에 관하여 k 가지의 서로 다른 해시 함수(hash function)에 의해 도출된 원소의 해시 값에 대응하는 비트들이 1로 설정되고 나머지 비트들은 0으로 설정된다.

[0004] 외부에서 어떤 데이터가 집합에 포함되는지 여부를 검사(query)할 경우에, 검사하려는 데이터의 k 가지 해시 함수에 의한 해시 값들이 가지는 인덱스(index)가 얻어지는데, 블룸 필터 비트열 중에서 그러한 인덱스들에 해당하는 위치의 비트들이 모두 1이면 검사하려는 데이터가 집합에 속한다고 양성으로(positive) 판정하고, 그러한 비트들 중에 하나라도 0이면 집합에 속하지 않는다고 음성으로(negative) 판정한다.

[0005] 이때, 어떤 집합을 그보다 더 적은 크기의 다른 집합에 매핑하는 해시 함수의 필연적인 불완전성에 따라, 집합에 속하지 않는 어떤 데이터의 해시 값들이 집합에 속하는 원소의 해시 값들과 동일하게 도출될 수 있다. 블룸 필터는 전적으로 해시 값에 의존하므로, 만약 해시 값이 동일하다면 집합에 속하지 않는 데이터와 집합에 속하는 원소는 해시 함수로는 구별할 수 없고, 집합에 속하지 않는 데이터에 대해 집합에 속한다고 판정될 수 있다.

[0006] 따라서 블룸 필터에 의해 어떤 원소가 집합에 속한다고 양성 판정되었지만 실제로는 원소가 집합에 속하지 않는 거짓 양성의 가능성이 문제된다.

[0007] 원소의 개수가 많아지면 거짓 양성 확률이 크게 증가하며, 일정한 거짓 양성 확률을 달성하려면 원소의 개수에 비례하여 블룸 필터의 크기를 늘려야 한다는 점이 수학적으로 증명되어 있다. 그러나 블룸 필터의 크기를 늘리면 그만큼 스토리지과 메모리를 차지하므로 블룸 필터의 크기를 무작정 늘릴 수도 없다.

[0008] 따라서, 효율과 성능을 유지하면서 거짓 양성을 줄일 수 있는 블룸 필터 구조가 요구된다.

[0009] 한편, 블룸 필터에 의해 어떤 원소가 집합에 속하지 않는다고 음성 판정될 경우에 실제로 원소가 집합에 속할

수도 있는 거짓 음성의 가능성은 전혀 없다.

[0010] 이렇듯, 블룸 필터는 설계에 따라 작은 메모리 공간을 소요하고 판정 속도가 빠르며, 거짓 음성이 불가능하여, 예를 들어 대규모 데이터베이스를 구성하는 여러 개의 소규모 데이터베이스들 중에서 적어도 특정 원소가 속하지 않는, 그래서 검색할 필요가 없는 소규모 데이터베이스들을 정확하고 빠르게 가려낼 수 있다.

[0011] 이러한 장점에 따라 블룸 필터는 다양한 데이터베이스 어플리케이션들과 대규모 스토리지 어플리케이션에 널리 사용되고 있으며, 라우팅 테이블(routing table) 검색, 온라인 트래픽 제어, 침입 탐지 시스템 등과 같이 일견 관련이 없어 보이는 다양한 네트워크 분야에서도 점점 더 자주 응용되고 있다.

발명의 내용

해결하려는 과제

[0012] 본 발명이 해결하고자 하는 과제는 복수의 서브 블룸 필터들을 이용하여 거짓 양성을 줄일 수 있는 블룸 필터 처리 장치 및 블룸 필터 처리 방법을 제공하는 데에 있다.

과제의 해결 수단

[0013] 본 발명의 일 측면에 따른 블룸 필터 처리 장치는,

[0014] k 개의 해시 함수들을 j 개(j는 2부터 k까지의 정수)의 해시 함수 서브그룹으로 분할하여, 입력되는 원소 또는 데이터에 관하여, j 개의 해시 인덱스 서브그룹으로 분할된 k 개의 해시 인덱스들을 생성하는 해시 인덱스 생성부;

[0015] 총 m 개의 비트 포지션들을 가지는 j 개의 서브 블룸 필터들을 보유하며, j 개의 해시 함수 서브그룹들 중 i 번째(i는 1부터 j까지의 정수) 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 i 번째 서브 블룸 필터에 연결하는 서브 블룸 필터 저장부;

[0016] 프로그래밍 시에 원소를 상기 해시 인덱스 생성부에 입력하고, 상기 원소에 관하여 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 상기 서브 블룸 필터 저장부의 i 번째 서브 블룸 필터에 설정하는 프로그래밍부; 및

[0017] 멤버쉽 쿼리 시에 입력 데이터를 상기 해시 인덱스 생성부에 입력하고, 상기 입력 데이터에 관하여 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 i 번째 해시 인덱스 서브그룹의 해시 인덱스들이 생성되면, 생성된 i 번째 해시 인덱스 서브그룹의 해시 인덱스들에 따라 상기 서브 블룸 필터 저장부의 i 번째 서브 블룸 필터로부터 각각 출력되는 비트열에 기초하여, 상기 입력 데이터의 멤버쉽을 검증하는 쿼리부를 포함할 수 있다.

[0018] 본 발명의 다른 측면에 따른 블룸 필터 처리 방법은,

[0019] 입력되는 원소 또는 데이터에 관하여, j 개의 해시 인덱스 서브그룹으로 분할된 k 개의 해시 인덱스들을 생성하도록 j 개(j는 2부터 k까지의 정수)의 해시 함수 서브그룹으로 분할되는 k 개의 해시 함수들을 제공하는 단계;

[0020] 상기 j 개의 해시 함수 서브그룹들 중 i 번째(i는 1부터 j까지의 정수) 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들이 i 번째 서브 블룸 필터에 연결되도록, 총 m 개의 비트 포지션들을 가지는 j 개의 서브 블룸 필터들을 제공하는 단계;

[0021] 프로그래밍 시에, 원소에 관하여 상기 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 생성하고, 생성된 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 i 번째 서브 블룸 필터에 설정하는 단계; 및

[0022] 멤버쉽 쿼리 시에 입력 데이터에 관하여 상기 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 생성하고, 생성된 i 번째 해시 인덱스 서브그룹의 해시 인덱스들 따라 i 번째 서브 블룸 필터로부터 각각 출력되는 비트열에 기초하여, 상기 입력 데이터의 멤버쉽을 검증하는 단계를 포함할 수 있다.

발명의 효과

[0023] 본 발명의 블룸 필터 처리 장치 및 블룸 필터 처리 방법에 따르면, 집합의 규모가 커질 때에도 전체적으로 동일

한 크기의 메모리를 이용하지만, 거짓 양성의 빈도를 훨씬 줄일 수 있다.

도면의 간단한 설명

[0024] 도 1은 본 발명의 일 실시예에 따른 서브 블룸 필터들을 가지는 블룸 필터의 프로그래밍을 예시한 개념도이다.

도 2는 본 발명의 일 실시예에 따른 서브 블룸 필터들을 가지는 블룸 필터에 대한 멤버쉽 쿼리를 예시한 개념도이다.

도 3은 본 발명의 일 실시예에 따른 서브 블룸 필터들을 이용한 블룸 필터 처리 방법에 따라 복수 개의 서브 블룸 필터들을 사용할 경우에 나타난 거짓 양성의 빈도를 예시한 표이다.

도 4는 본 발명의 일 실시예에 따른 서브 블룸 필터들을 이용한 블룸 필터 처리 장치를 예시한 블록도이다.

도 5는 본 발명의 일 실시예에 따른 서브 블룸 필터들을 이용한 블룸 필터 처리 방법을 예시한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0025] 본문에 개시되어 있는 본 발명의 실시예들에 대해서, 특정한 구조적 내지 기능적 설명들은 단지 본 발명의 실시예를 설명하기 위한 목적으로 예시된 것으로, 본 발명의 실시예들은 다양한 형태로 실시될 수 있으며 본문에 설명된 실시예들에 한정되는 것으로 해석되어서는 아니 된다.

[0026] 이하, 첨부한 도면들을 참조하여, 본 발명의 바람직한 실시예를 보다 상세하게 설명하고자 한다. 도면상의 동일한 구성요소에 대해서는 동일한 참조부호를 사용하고 동일한 구성요소에 대해서 중복된 설명은 생략한다.

[0027] 아래 설명에서, 맥락에 따라 달리 해석하도록 명시되지 않은 한, 원소는 이미 집합에 속하거나 현재 추가될 대상물을 지칭하고, 데이터는 집합에 속하는지 여부를 검사하기 위한 대상물을 가리킨다.

[0028] 블룸 필터는 일반적으로 크기 m의 비트 벡터(bit vector)와 k 개의 해시 함수들로 구성된다.

[0029] 해시 함수(hash function)는 가변하는 길이의 데이터를 어떤 고정된 길이의 데이터에 상사(mapping)할 수 있는 여하한 종류의 알고리즘을 의미한다. 해시 함수의 연산 결과, 즉 매핑된 고정 길이의 데이터는 해시 값(hash values), 해시 코드, 해시 섬(sum), 체크섬(checksum), 또는 간단히 해시 등으로 불리고, 특히 블룸 필터에서는 해시 인덱스(hash index), 인덱스 또는 색인(index)이라고도 불린다.

[0030] 블룸 필터에서 각각의 해시 함수는 다양한 길이의 원소 또는 데이터를 입력받아 고정된 길이의 해시 인덱스를 출력한다. 해시 인덱스가 가지는 값은 가변하는 길이의 원소 또는 데이터에 대해 고정된 길이의 비트 벡터 내의 비트 포지션들 중 이진 비트값 1로 설정하기 위한 비트 포지션을 의미한다.

[0031] 해시 함수에 의해 각각의 포지션이 선택될 확률이 균일하고 랜덤하여야 블룸 필터에 적합한 해시 함수라고 할 수 있고, 다양한 상사 함수들이 해시 함수로서 이용될 수 있음이 알려져 있지만, 대표적으로 예를 들어 CRC(Cyclic Redundancy Check) 알고리즘이 이용될 수 있다.

[0032] 통상적으로 CRC 알고리즘은 다음과 같이, a 비트의 주어진 정보가 있을때, 이를 b 비트 만큼 자리를 올리고 미리 약속한 b 비트의 키 값으로 나누면 r 비트의 나머지가 남게 되는데, 정보의 송신측에서는 원래의 정보 비트를 b 비트 자리 올린 것에 r 비트의 나머지를 첨부하여 a+r 비트 크기의 데이타를 만들어 보내고, 정보의 수신측에서는 수신된 a+r 비트 크기의 데이타를 b 비트의 키 값으로 나누어 보고 나머지가 정확히 0이 되는지를 검사함으로써, 수신된 데이터가 송신된 정보와 동일함을 입증하는, 즉 전송 데이터의 무결성을 입증하는 알고리즘이다. b가 16 비트이면 CRC-16, 32 비트이면 CRC-32라 한다.

[0033] 블룸 필터는 CRC 알고리즘에서 얻어진 r 비트의 잉여 비트들의 전부 또는 일부분 또는 추가적인 조합으로 얻은 이진 비트열에 상응하는 십진수를 해시 인덱스로 이용할 수 있다.

[0034] 어떤 n 개의 원소들을 가진 집합이 주어졌을 때, 집합에 속하는 원소들의 각각을 k 개의 해시 함수들에 입력하면, k 개의 해시 함수들에 의해 크기 m의 비트 벡터 내에 k 개의 해시 색인 위치가 선택되고, 선택된 해시 색인 위치들에 상응하는 비트 포지션들이 각각 1로 설정된다. 이를 블룸 필터의 프로그래밍이라 한다.

[0035] 원소 x를 k 개의 해시 함수들 $h_i()$, $1 \le i \le k$을 이용하여 블룸 필터 BF의 비트 벡터 Vector[]에 프로그래밍하는 알고리즘 BFProgram(x)을 의사 코드(pseudo code)로 표현하면 다음과 같다.

[0036] BFProgram (x)

[0037] 1) for (i= 1 to k)

[0038] 2) Vector[$h_i(x)$] ← 1

[0039] $h_i(x)$는 i 번째 해시 함수 $h_i()$에 원소 x가 입력되어 연산된 i 번째 해싱 값 또는 해싱 색인이다.

[0040] 다시 말해, 일반적으로 블룸 필터의 프로그래밍은 매번 원소가 입력될 때마다 원소 x에 대해 비트 벡터 내의 m 개의 비트 포지션 중에 k 개의 해시 함수에 의해 선택되는 k 개의 비트 포지션을 1로 반복적으로 설정하는 과정이다.

[0041] 블룸 필터에 집합 내의 모든 원소들에 대해 프로그래밍을 한 이후에는, 쿼리(query)를 통해, 입력된 데이터가 이전에 프로그래밍된 원소인지 여부를 판정할 수 있다.

[0042] 쿼리 과정도 실질적으로 프로그래밍 과정과 다르지 않다. 입력된 데이터에 대해 k 개의 해시 함수들에 의해 k 개의 해시 색인 위치가 선택되면, k 개의 해시 색인 위치들에 상응하는 비트 벡터 내에 비트 포지션들이 모두 1로 설정되어 있다면, 입력된 데이터가 이전에 프로그래밍된 원소일 확률이 매우 높고, 따라서 양성(positive)으로, 다시 말해 데이터가 집합에 속한다고 판정한다.

[0043] 그렇지 않고, k 개의 해시 색인 위치들에 상응하는 비트 벡터 내에 비트 포지션들 중 어느 하나라도 0으로 설정되어 있다면, 입력된 데이터는 이전에 프로그래밍된 적이 없는 데이터이며, 따라서 음성(negative)으로, 다시 말해 데이터가 집합에 속하지 않는다고 판정한다.

[0044] 입력 데이터 y를 쿼리하는 알고리즘 BFQuery(y)을 의사 코드로 표현하면 다음과 같다.

[0045] BFQuery (y)

[0046] 1) for (i = 1 to k)

[0047] 2) if (Vector[$h_i(y)$] = 0) return negative

[0048] 3) return positive

[0049] 여기서 프로그래밍과 마찬가지로 k 개의 해시 함수들 $h_i()$, $1 \le i \le k$을 이용하여 블룸 필터 BF의 비트 벡터 Vector[]를 참조하며, 한 해시 인덱스라도 0이 있으면 음성(negative)을 반환하고, 그렇지 않고 모두 1인 경우에만 양성(positive)을 반환한다.

[0050] 블룸 필터가 해시 함수를 사용하기 때문에, 해시 함수가 서로 다른 입력 데이터에 대한 해시 값의 충돌(collision)을 피할 수 없는 한, 서로 다른 x, y 입력 데이터가 완전히 동일한 해시 인덱스들을 가질 가능성도 피할 수 없다. 데이터 x가 사전에 원소로서 프로그래밍되어 있을 때에, 다른 데이터 y가 입력된다면, 데이터 y에 의한 해시 인덱스에 해당하는 비트 포지션들은 모두 데이터 x의 프로그래밍에 의해 1로 설정되어 있을 것이므로, 블룸 필터는 데이터 y가 집합의 원소라고 판정한다. 이러한 잘못된 멤버쉽 판정을 거짓 양성 판정이라고 한다.

[0051] 주어진 집합의 크기 n에 대해, 해싱 함수의 개수 k와 블룸 필터 비트 벡터의 크기 m을 늘리면 거짓 양성의 확률을 줄일 수 있다. 그러나 해싱 함수의 개수는 연산 속도와 복잡성에 관련이 있고, 비트 벡터의 크기는 메모리 사용량에 관련이 있으며, 이러한 요인들이 복잡하게 블룸 필터의 전체적인 성능에 영향을 미친다.

[0052] 해시 함수의 개수 k = (m/n)*ln2일 때에 거짓 양성의 확률을 최소화할 수 있다고 알려져 있다.

[0053] 예시적으로, 원소의 개수는 n이고, 이를 위해 블룸필터의 비트 벡터를 원소 엔트리(element entry)마다 32 비트 꼴로 부가되도록 설계한다면, 전체 블룸 필터의 비트 벡터의 크기 m = 32n이다. 이때 거짓 양성의 확률을 최소화하는 해시 함수의 개수 k는 (m/n)*ln2 = 32ln2 ≒ 22이다.

[0054] 한편, 본 출원의 발명자는 기존의 단일 블룸 필터 구조에서 k 개의 해싱 함수들을 가지고 생성된 k 개의 해싱

인덱스들 및 거짓 양성 판정 현상을 관찰하면서, 해싱 인덱스들 사이에 순서에 관한 정보가 활용되지 않고 있음을 발견하였다.

[0055] 예를 들어, 원소 x에 대해 두 개의 해싱 함수 $h_1()$과 $h_2()$를 사용하여 프로그래밍하면 해싱 인덱스 $h_1(x)$과 $h_2(x)$을 얻을 수 있다.

[0056] 이때, 입력 데이터 y에 대해, 만약 해싱 인덱스 $h_1(y)=h_1(x)$이고 또한 $h_2(y)=h_2(x)$라면 두 개의 해싱 함수 $h_1()$과 $h_2()$를 이용하여서는 입력 데이터 y를 원소 x와 결코 구별할 수 없으므로 거짓 양성 판정을 피할 수 없다.

[0057] 나아가, 입력 데이터 y에 대해, 만약 해싱 인덱스 $h_1(y)=h_2(x)$이고 또한 $h_2(y)=h_1(x)$라면, 종래의 단일 블룸 필터 구조에서는 단순히 해당 비트 포지션의 비트가 1인지 여부만 검사하므로, 마찬가지로 입력 데이터 y를 원소 x와 구별할 가능성이 없어 거짓 양성 판정을 피할 수 없을 것이다.

[0058] 하지만, 만약 해싱 인덱스 $h_1(y)=h_2(x)$이고 또한 $h_2(y)=h_1(x)$이더라도, $h_1(y)\neq h_1(x)$이거나 또는 $h_2(y)\neq h_2(x)$라면, 후자의 관계를 식별할 수만 있다면 거짓 양성 판정을 피할 가능성이 있다.

[0059] 다시 말해, 만약 해싱 함수와 해싱 인덱스들의 상관 관계에 관한 정보를 프로그래밍 과정이나 쿼리 과정 중에 유지할 수 있다면, 다시 말해, $h_1(y)\neq h_1(x)$이거나 또는 $h_2(y)\neq h_2(x)$임을 판단할 수 있다면, 데이터 y는 원소 x와 명확히 구별할 수 있을 것이고 데이터 y의 거짓 양성 판정을 피할 가능성이 있다.

[0060] 본 출원의 발명자는 이러한 가능성에 따라 서브 블룸 필터들을 가지는 블룸 필터를 아래와 같이 제안한다.

[0061] 도 1은 본 발명의 일 실시예에 따른 서브 블룸 필터들을 가지는 블룸 필터의 프로그래밍을 예시한 개념도이다.

[0062] 도 1을 참조하면, 블룸 필터는 예를 들어 헤드(head), 바디(body) 및 테일(tail)로 지칭되는 3 개의 서브 블룸 필터들(BF_1, BF_2, BF_3)과 3 개의 해시 함수 서브그룹들(HS_1, HS_2, HS_3)로 구성된다.

[0063] 여기서, 예시적으로 3 개의 서브 블룸 필터들(BF_1, BF_2, BF_3)의 각각의 크기는 m/k, m(k-2)/k 및 m/k이며, 서브 블룸 필터들의 크기의 총합은 m이다. 3 개 해시 함수 서브그룹들(HS_1, HS_2, HS_3)은 각각 첫 해시 함수 $h_1()$, 다음 복수의 해시 함수들 $h_2()$ 내지 $h_{k-1}()$ 및 마지막 해시 함수 $h_k()$를 가지는 것으로 예시되며, 해시 함수들 및 해시 인덱스들의 총 수는 k 개이다.

[0064] 원소 x의 프로그래밍 시에, 예시적으로 첫 해시 인덱스 $h_1(x)$는 헤드 서브 블룸 필터 BF_1에 설정되고, 다음 해시 인덱스 $h_2(x)$ 내지 $h_{k-1}(x)$는 각각 바디 서브 블룸 필터 BF_2에 설정되며, 마지막 해시 인덱스 $h_k(x)$는 테일 서브 블룸 필터 BF_3에 설정된다.

[0065] 예시적으로 원소 x에 대해, 첫 해시 인덱스 $h_1(x)=1$은 헤드 서브 블룸 필터 BF_1의 비트 포지션 1에 설정되고, 다음 해시 인덱스 $h_2(x)$ 내지 $h_{k-1}(x)=3, 5, 6, 10$은 각각 바디 서브 블룸 필터 BF_2의 비트 포지션들 3, 5, 6, 10에 설정되며, 마지막 해시 인덱스 $h_k(x)=4$는 테일 서브 블룸 필터 BF_3의 비트 포지션 4에 설정된다.

[0066] 따라서, 해시 인덱스들의 순서 정보가 프로그래밍 과정 중에 전부 또는 부분적으로 유지될 수 있다.

[0067] 도 2는 본 발명의 일 실시예에 따른 서브 블룸 필터들을 가지는 블룸 필터에 대한 멤버쉽 쿼리를 예시한 개념도이다.

[0068] 도 2를 참조하면, 블룸 필터는 도 1과 마찬가지로 헤드, 바디 및 테일로 지칭되는 3 개의 서브 블룸 필터들(BF_1, BF_2, BF_3)과 3 개의 해시 함수 서브그룹들(HS_1, HS_2, HS_3)로 구성되어 있다.

[0069] 여기서, 예시적으로 3 개의 서브 블룸 필터들(BF_1, BF_2, BF_3)의 각각의 크기는 m/k, m(k-2)/k 및 m/k이며, 서브 블룸 필터들의 크기의 총합은 m이다. 3 개 해시 함수 서브그룹들(HS_1, HS_2, HS_3)은 각각 첫 해시 함수 $h_1()$, 다음 복수의 해시 함수들 $h_2()$ 내지 $h_{k-1}()$ 및 마지막 해시 함수 $h_k()$를 가지는 것으로 예시되며, 해시 함수들 및 해시 인덱스들의 총 수는 k 개이다.

[0070] 입력 데이터 y의 멤버쉽 쿼리가 있으면, 첫 해시 인덱스 $h_1(y)$는 헤드 서브 블룸 필터 BF_1에서 검증되고, 다음

해시 인덱스 $h_2(y)$ 내지 $h_{k-1}(y)$는 바디 서브 블룸 필터 BF_2에서 각각 검증되며, 마지막 해시 인덱스 $h_k(y)$는 테일 서브 블룸 필터 BF_3에서 검증된다.

[0071] 예시적으로 입력 데이터 y에 대해, 첫 해시 인덱스 $h_1(y)=1$은 헤드 서브 블룸 필터 BF_1의 비트 포지션 1에서 검증되고, 다음 해시 인덱스 $h_2(y)$ 내지 $h_{k-1}(y)=3$, 5, 6, 9은 각각 바디 서브 블룸 필터 BF_2의 비트 포지션들 3, 5, 6, 10에서 검증되며, 마지막 해시 인덱스 $h_k(y)=4$는 테일 서브 블룸 필터 BF_3의 비트 포지션 4에서 검증된다.

[0072] 따라서, 해시 인덱스들의 순서 정보가 쿼리 과정 중에서도 전부 또는 부분적으로 유지될 수 있다.

[0073] 입력 데이터 y의 멤버쉽 쿼리는 만약 한 서브 블룸 필터에서라도 음성 판정이 나오면 최종 음성 판정되며, 모든 서브 블룸 필터에서 양성 판정이 나와야 최종 양성 판정된다.

[0074] 이러한 구성을 좀더 일반화시키면 다음과 같다.

[0075] 본 발명의 블룸 필터는 총 m 개의 비트 포지션을 가지는 j(j는 2부터 k까지의 정수) 개의 서브 블룸 필터들 및 k 개의 해시 함수들로 구성된다.

[0076] j 개의 서브 블룸 필터들 중 i 번째(i는 1부터 j까지의 정수) 서브 블룸 필터의 크기 m_i를 모두 합하면 m 이다.

[0077] 원소의 개수를 n이라 하면, k = (m/n)*ln2에 가장 가까운 정수일 때 거짓 양성 확률이 가장 낮은 최적값이라고 알려져 있다.

[0078] 바람직하게는, 서브 블룸 필터들 중 적어도 하나의 서브 블룸 필터의 크기는 $\lceil m/k \rceil$이다. [X]는 가우스 기호 또는 최대 정수 함수(greatest integer function)로서, 실수 X보다 크지 않으면서 가장 큰 정수이다.

[0079] 바람직하게는, j = 2이고, 서브 블룸 필터들 중 하나의 크기는 $\lceil m/k \rceil$인 정수이다. 나머지 하나의 크기는 산술적으로, 다시 말해 $m - \lceil m/k \rceil$으로 결정된다.

[0080] 바람직하게는, j = 3이고, 서브 블룸 필터들 중 둘의 크기는 $\lceil m/k \rceil$인 정수이다. 나머지 하나의 크기는 산술적으로 결정될 수 있다.

[0081] 예를 들어, 간단히 m이 k의 배수라면, 첫 번째(헤드)와 세 번째(테일) 서브 블룸 필터의 크기는 각각 m/k이고, 두 번째(바디) 서브 블룸 필터의 크기는 m(k-2)/k일 수 있다.

[0082] 바람직하게는 k보다 작은 j 개의 서브 블룸 필터들 중 j-1 개의 서브 블룸 필터들의 크기는 $\lceil m/k \rceil$인 정수 또는 $\lceil m/k \rceil$의 정수배인 정수이고, 나머지 하나의 크기는 산술적으로, 다시 말해 총 비트 포지션의 크기 m에서 j-1 개의 서브 블룸 필터들의 크기를 뺀 값으로서 결정될 수 있다.

[0083] 만약 j = k라면, 서브 블룸 필터들의 크기는 $\lceil m/k \rceil$인 정수로 결정될 수 있다.

[0084] 다른 실시예에서, 서브 블룸 필터들 중 적어도 하나의 서브 블룸 필터의 크기는 $\lceil m/k \rceil$인 정수일 수 있다. $\lfloor X \rfloor$는 천정 함수(ceiling function)로서, 실수 X보다 작지 않으면서 가장 작은 정수이다.

[0085] 실시예에 따라, j 개의 서브 블룸 필터들 중 i 번째 서브 블룸 필터의 크기는 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들이 반드시 i 번째 서브 블룸 필터 내에 설정될 수 있는 조건 하에서 결정될 수 있다.

[0086] 다시 말해, 예를 들어, i 번째 해시 인덱스 서브 그룹의 최대 해시 인덱스가 512라면, i 번째 서브 블룸 필터의 크기는 어떠한 해시 인덱스라도 서브 블룸 필터 내에서 설정될 수 있도록 적어도 512보다 큰 크기로 결정될 수

있다.

[0087] k 개의 해시 함수들은 j 개의 해시 함수 서브그룹으로 분할되고, 그에 따라 k 개의 해시 인덱스들도 j 개의 해시 인덱스 서브그룹으로 분할된다.

[0088] 이때, 프로그래밍 시에는 원소 x에 관하여 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들은 i 번째 서브 블룸 필터에 설정된다.

[0089] 바람직하게는, j 개의 해시 함수 서브그룹들 또는 해시 인덱스 서브그룹들의 각각에 속하는 해시 함수들 또는 해시 인덱스들의 개수는 j 개의 서브 블룸 필터들의 각각의 길이에 상응하여 결정될 수 있고, 그 역의 관계도 성립할 수 있다.

[0090] 예를 들어, 총 10 개의 해시 함수들을 사용하는 블룸 필터를 5 개의 서브 블룸 필터들로 나누고 서브 블룸 필터들의 각각의 길이가 m/k, 2m/k, 2m/k, 4m/k 및 m/k라면, 10 개의 해시 함수들 또는 해시 인덱스들로써 구성되는 5 개의 해시 함수 서브그룹들 또는 해시 인덱스 서브그룹들의 각각에 속하는 해시 함수들 또는 해시 인덱스들의 개수는 각각 1 개, 2 개, 2 개, 4 개 및 1 개일 수 있다.

[0091] 쿼리 시에는 입력 데이터 y에 관하여 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들은 i 번째 서브 블룸 필터에서 검증된다.

[0092] 이렇게 하여, 해시 인덱스들의 순서 정보가 프로그래밍 과정 및 쿼리 과정 중에 전부 또는 부분적으로 유지될 수 있다.

[0093] 각각의 해시 인덱스 서브그룹들에 상응하는 각각의 서브 블룸 필터들에서 모두 양성 판정이 나와야 최종적으로 양성 판정을 내릴 수 있다. 어느 하나의 서브 블룸 필터에서만 음성 판정이 나오고 나머지 서브 블룸 필터들에서 양성 판정이 나와도 최종적으로 음성 판정이 내려진다. 따라서, 거짓 양성의 가능성은 대단히 감소한다.

[0094] 도 3은 본 발명의 일 실시예에 따른 서브 블룸 필터들을 이용한 블룸 필터 처리 방법에 따라 복수 개의 서브 블룸 필터들을 사용할 경우에 나타난 거짓 양성의 빈도를 예시한 표이다.

[0095] 거짓 양성 검출 실험은, 각각의 주어진 집합의 크기 및 블룸 필터의 전체 비트 벡터 크기 조건 하에서, 하나부터 다섯 개의 서브 블룸 필터를 이용하여 프로그래밍하고, 프로그래밍된 서브 블룸 필터들에 대해, 집합에 속하지 않도록 랜덤하게 생성되는 입력 데이터를 집합 크기의 세 배만큼 생성하여 멤버쉽 쿼리를 수행하도록 진행되었다.

[0096] 입력 데이터는 집합에 속하지 않도록 생성되므로 모두 음성으로 판정되어야 정상이고, 만약 그러한 입력 데이터의 멤버쉽 쿼리가 양성으로 판정된다면 이는 반드시 거짓 양성일 수 밖에 없다.

[0097] 도 3에 따르면, 하나의 서브 블룸 필터를 이용하는 경우(1BF)는 실질적으로 기존의 블룸 필터 구조와 다르지 않는데, 집합의 크기가 상대적으로 작은 10K, 즉 1만 개이고 블룸 필터의 크기, 즉 메모리 점유량이 40 KB일 때에, 거짓 양성의 빈도가 없거나 매우 낮다. 하지만 집합의 크기가 점점 커지면서 거짓 양성의 빈도가 급증하는데, 예를 들어 집합의 크기가 400K, 즉 40만 개이고 블룸 필터의 크기가 1600 KB일 때, 120만 회의 비 멤버 입력 데이터의 멤버쉽 쿼리에서 483 회의 거짓 양성이 일어났다.

[0098] 다시 말해, 기존의 블룸 필터 구조를 그대로 이용한다면, 쿼리 1만 회에 4 회 이상의 거짓 양성이 일어나는 꼴이므로, 응용 분야에 따라서는 용인될 수 없는 수준일 수도 있는, 적지 않은 거짓 양성이 일어남을 의미한다.

[0099] 두 개의 서브 블룸 필터들을 이용하는 경우(2BF)는, 1BF와 동일한 집합 크기, 동일한 블룸 필터 메모리 점유량, 동일한 회수의 비 멤버 입력 데이터의 멤버쉽 쿼리 조건 하에서, 거짓 양성의 횟수가 1BF에 비해 많게는 1/3 가량으로 줄어들었다. 메모리와 속도 등 모든 외형적 조건이 동일함에도 불구하고 거짓 양성 성능이 기존의 블룸 필터에 비해 두 배나 향상된 것이다.

[0100] 나아가, 세 개의 서브 블룸 필터들을 이용하는 경우(3BF)는, 1BF와 동일한 집합 크기, 동일한 블룸 필터 메모리 점유량, 동일한 회수의 비 멤버 입력 데이터의 멤버쉽 쿼리 조건 하에서, 거짓 양성의 횟수가 1BF에 비해 많게는 15% 정도까지 줄어들었다.

[0101] 네 개의 서브 블룸 필터들을 이용하는 경우(4BF)는, 1BF와 동일한 집합 크기, 동일한 블룸 필터 메모리 점유량, 동일한 회수의 비 멤버 입력 데이터의 멤버쉽 쿼리 조건 하에서, 거짓 양성의 횟수가 1BF에 비해 많게는 5% 가량으로 줄어들었다.

[0102] 다섯 개의 서브 블룸 필터들을 이용하는 경우(5BF)는, 1BF와 동일한 집합 크기, 동일한 블룸 필터 메모리 점유량, 동일한 회수의 비 멤버 입력 데이터의 멤버쉽 쿼리 조건 하에서, 거짓 양성의 횟수가 1BF에 비해 많게는 1.2% 가량으로 줄어들었다.

[0103] 이는 쿼리 20만 회에 1회 수준의 거짓 양성이 일어나는 꼴이므로, 용인 범위(tolerance)가 훨씬 엄격한 응용 분야에도 적용될 수 있다.

[0104] 이를 통해 본 발명의 향상된 거짓 양성 성능을 실증할 수 있었다.

[0105] 도 4는 본 발명의 일 실시예에 따른 서브 블룸 필터들을 이용한 블룸 필터 처리 장치를 예시한 블록도이다.

[0106] 도 4를 참조하면, 블룸 필터 처리 장치(40)는 해시 인덱스 생성부(41), 서브 블룸 필터 저장부(42), 프로그래밍부(43) 및 쿼리부(44)를 포함할 수 있다.

[0107] 해시 인덱스 생성부(41)는 k 개의 해시 함수들을 j 개(j는 2부터 k까지의 정수)의 해시 함수 서브그룹으로 분할하여, 입력되는 원소 또는 데이터에 관하여, j 개의 해시 인덱스 서브그룹으로 분할된 k 개의 해시 인덱스들을 생성한다.

[0108] 서브 블룸 필터 저장부(42)는 총 m 개의 비트 포지션들을 가지는 j 개의 서브 블룸 필터들을 보유하며, j 개의 해시 함수 서브그룹들 중 i 번째(i는 1부터 j까지의 정수) 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 i 번째 서브 블룸 필터에 연결한다.

[0109] j 개의 서브 블룸 필터들 중 i 번째(i는 1부터 j까지의 정수) 서브 블룸 필터의 크기 m_i를 모두 합하면 m 이다.

[0110] 바람직하게는, 서브 블룸 필터들 중 적어도 하나의 서브 블룸 필터의 크기는 $\lceil m/k \rceil$이다.

[0111] 바람직하게는, j = 2이고, 서브 블룸 필터들 중 하나의 크기는 $\lceil m/k \rceil$인 정수이다. 나머지 하나의 크기는 산술적으로, 다시 말해 $m - \lceil m/k \rceil$으로 결정된다.

[0112] 바람직하게는, j = 3이고, 서브 블룸 필터들 중 둘의 크기는 $\lceil m/k \rceil$인 정수이다. 나머지 하나의 크기는 산술적으로 결정될 수 있다.

[0113] 바람직하게는 k보다 작은 j 개의 서브 블룸 필터들 중 j-1 개의 서브 블룸 필터들의 크기는 $\lceil m/k \rceil$인 정수 또는 $\lceil m/k \rceil$의 정수배인 정수이고, 나머지 하나의 크기는 산술적으로 결정될 수 있다.

[0114] 만약 j = k라면, 서브 블룸 필터들의 크기는 $\lceil m/k \rceil$인 정수로 결정될 수 있다.

[0115] 다른 실시예에서, 서브 블룸 필터들 중 적어도 하나의 서브 블룸 필터의 크기는 $\lceil m/k \rceil$인 정수일 수 있다.

[0116] 실시예에 따라, j 개의 서브 블룸 필터들 중 i 번째 서브 블룸 필터의 크기는 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들이 반드시 i 번째 서브 블룸 필터 내에 설정될 수 있는 조건 하에서 결정될 수 있다.

[0117] 프로그래밍부(43)는, 프로그래밍 시에 원소 x를 해시 인덱스 생성부(41)에 입력하고, 원소 x에 관하여 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 서브 블룸 필터 저장부(42)의 i 번째 서브 블룸 필터에 설정하도록 동작한다.

[0118] 쿼리부(44)는, 멤버쉽 쿼리 시에 입력 데이터 y를 해시 인덱스 생성부(41)에 입력하고, 입력 데이터 y에 관하여 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들에 따라 서브 블룸 필터 저장부(42)의 i 번째 서브 블룸 필터로부터 각각 출력되는 비트열에 기초하여, 입력 데이터 y의 멤버쉽을 검증한다.

[0119] 이렇게 하여, 해시 인덱스들의 순서 정보가 프로그래밍 과정 및 쿼리 과정 중에 전부 또는 부분적으로 유지될

수 있다.

[0120] 쿼리부(44)는 각각의 해시 인덱스 서브그룹들에 상응하는 각각의 서브 블룸 필터들에서 모두 비트 1이 출력되어야 최종적으로 양성 판정을 내릴 수 있다. 쿼리부(44)는 어느 하나라도 서브 블룸 필터에서 0이 나오면, 나머지 서브 블룸 필터들에서 모두 1이 나오더라도 최종적으로 음성 판정을 내린다.

[0121] 도 5는 본 발명의 일 실시예에 따른 서브 블룸 필터들을 이용한 블룸 필터 처리 방법을 예시한 순서도이다.

[0122] 먼저 단계(S51)에서, 입력되는 원소 또는 데이터에 관하여, j 개의 해시 인덱스 서브그룹으로 분할된 k 개의 해시 인덱스들을 생성하도록 j 개(j는 2부터 k까지의 정수)의 해시 함수 서브그룹으로 분할되는 k 개의 해시 함수들을 제공한다.

[0123] 다음으로, 단계(S52)에서, j 개의 해시 함수 서브그룹들 중 i 번째(i는 1부터 j까지의 정수) 해시 함수 서브그룹에 의해 생성되는 i 번째 해시 인덱스 서브그룹의 해시 인덱스들이 i 번째 서브 블룸 필터에 연결되도록, 총 m 개의 비트 포지션들을 가지는 j 개의 서브 블룸 필터들을 제공한다.

[0124] 단계(S53)에서, 프로그래밍 시에, 원소 x에 관하여 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 생성하고, 생성된 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 i 번째 서브 블룸 필터에 설정한다.

[0125] 단계(S54)에서, 멤버쉽 쿼리 시에 입력 데이터 y에 관하여 j 개의 해시 함수 서브그룹들 중 i 번째 해시 함수 서브그룹에 의해 i 번째 해시 인덱스 서브그룹의 해시 인덱스들을 생성하고, 생성된 i 번째 해시 인덱스 서브그룹의 해시 인덱스들 따라 i 번째 서브 블룸 필터로부터 각각 출력되는 비트열에 기초하여, 입력 데이터 y의 멤버쉽을 검증한다.

[0126] 구체적으로, 단계(S54)에서는, 입력 데이터 y에 대해 해시 인덱스 서브그룹들의 각각에 상응하는 서브 블룸 필터들의 각각에서 모두 비트 1이 출력되어야 입력 데이터 y에 대해 양성으로 판정하고, 만약 어느 한 서브 블룸 필터라도 비트 0이 나오면, 음성으로 판정할 수 있다.

[0127] 이상과 같이 본 발명은 비록 한정된 실시예와 도면에 의해 설명되었으나, 본 발명이 상기의 실시예에 한정되는 것은 아니며, 이는 본 발명이 속하는 분야에서 통상의 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형이 가능하다. 따라서, 본 발명의 사상은 아래에 기재된 특허청구범위에 의해서만 파악되어야 하고, 이와 균등하거나 또는 등가적인 변형 모두는 본 발명 사상의 범주에 속한다 할 것이다.

[0128] 또한, 본 발명에 따른 장치는 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 기록매체의 예로는 ROM, RAM, 광학 디스크, 자기 테이프, 플로피 디스크, 하드 디스크, 비휘발성 메모리 등이 있으며, 또한 캐리어 웨이브(예를 들어 인터넷을 통한 전송)의 형태로 구현되는 것도 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

부호의 설명

[0129]
40 블룸 필터 처리 장치

41 해시 인덱스 생성부

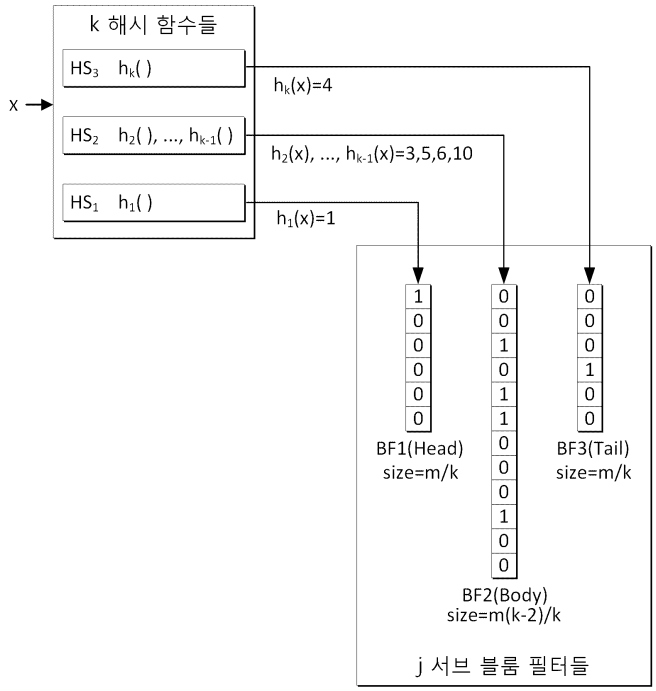
42 서브 블룸 필터 저장부

43 프로그래밍부

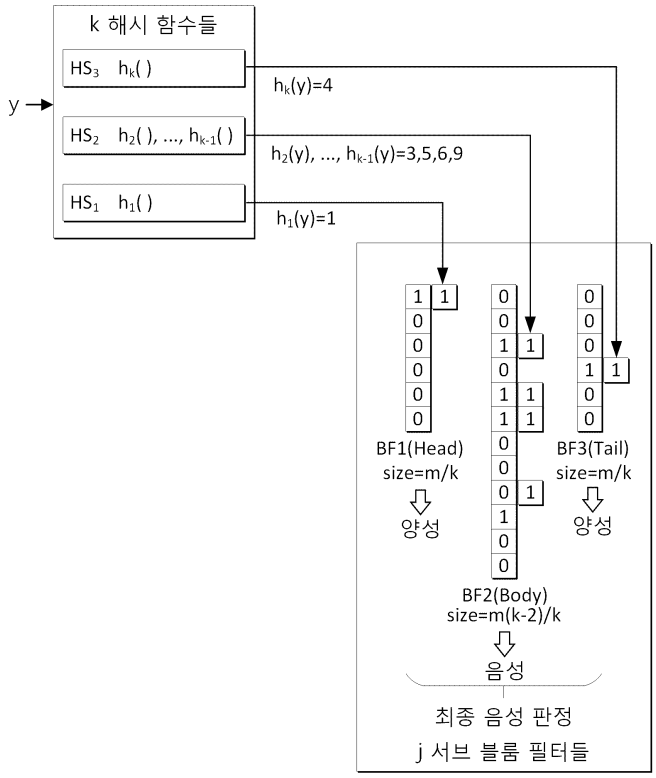
44 쿼리부

도면

도면1


k 해시 함수들
HS3 hk( )
HS2 h2( ), ..., hk-1( )
HS1 h1( )
x
hk(x)=4
h2(x), ..., hk-1(x)=3,5,6,10
h1(x)=1
BF1(Head)
size=m/k
BF2(Body)
size=m(k-2)/k
BF3(Tail)
size=m/k
j 서브 블룸 필터들

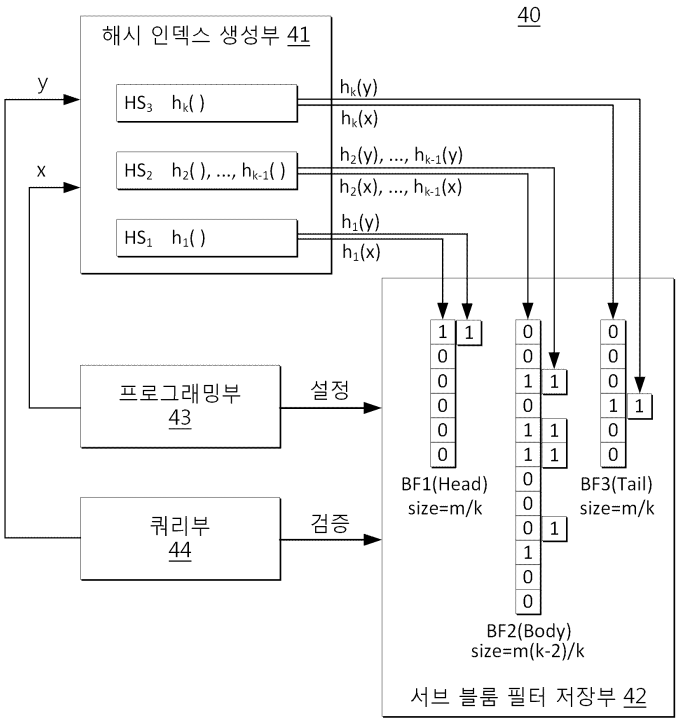
도면2


k 해시 함수들
HS3 hk( )
HS2 h2( ), ..., hk-1( )
HS1 h1( )
y
hk(y)=4
h2(y), ..., hk-1(y)=3,5,6,9
h1(y)=1
BF1(Head)
size=m/k
양성
BF2(Body)
size=m(k-2)/k
음성
BF3(Tail)
size=m/k
양성
최종 음성 판정
j 서브 블룸 필터들

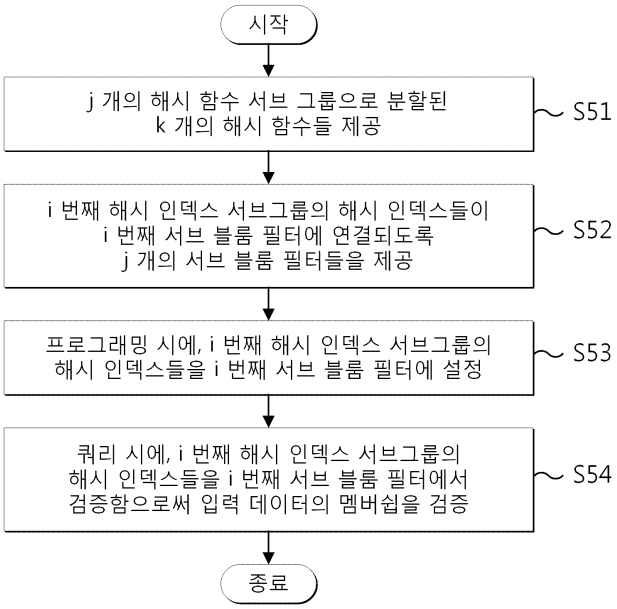
도면3

n	*m* (KB)	거짓양성의 개수				
		1-BF	2-BF	3-BF	4-BF	5-BF
10,000	40	0	0	0	0	0
50,000	200	4	5	1	1	1
100,000	400	30	23	8	1	2
200,000	800	123	71	29	8	1
300,000	1200	240	141	90	32	11
400,000	1600	483	163	74	23	6

도면4



도면5


시작
j 개의 해시 함수 서브 그룹으로 분할된
k 개의 해시 함수들 제공
S51
i 번째 해시 인덱스 서브그룹의 해시 인덱스들이
i 번째 서브 블룸 필터에 연결되도록
j 개의 서브 블룸 필터들을 제공
S52
프로그래밍 시에, i 번째 해시 인덱스 서브그룹의
해시 인덱스들을 i 번째 서브 블룸 필터에 설정
S53
쿼리 시에, i 번째 해시 인덱스 서브그룹의
해시 인덱스들을 i 번째 서브 블룸 필터에서
검증함으로써 입력 데이터의 멤버쉽을 검증
S54
종료