(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년08월22일
(11) 등록번호 10-1770271
(24) 등록일자 2017년08월16일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01)
(52) CPC특허분류
G06F 17/30613 (2013.01)
G06F 17/30616 (2013.01)
(21) 출원번호 10-2015-0113619
(22) 출원일자 2015년08월12일
심사청구일자 2015년08월12일
(65) 공개번호 10-2017-0019603
(43) 공개일자 2017년02월22일
(56) 선행기술조사문헌
US20140136513 A1*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
주식회사 이디엄
서울특별시 마포구 새창로 7, 1601호(도화동, 에스엔유장학빌딩)
(72) 발명자
양봉열
서울특별시 마포구
황원근
서울특별시 마포구 새창로 7, 1601호(도화동, 에스엔유장학빌딩)
(뒷면에 계속)
(74) 대리인
백도현

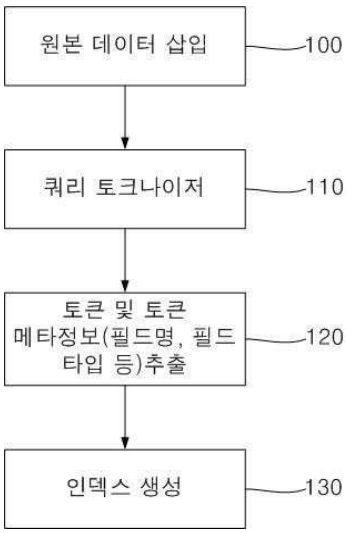
전체 청구항 수 : 총 2 항 심사관 : 이복현

(54) 발명의 명칭 **메시지의 필드 인덱싱 방법**

(57) 요 약

본 발명은 컴퓨터가 수행하는 메시지 인덱싱 방법에 관한 것으로서, 메시지에 대해서 쿼리 토크나이저 모듈을 실행시키는 제1 단계와, 상기 제1 단계에서 토크나이징된 정보에 대해서 인덱싱을 수행하는 제2 단계를 포함한다

대 표 도 - 도1



(72) 발명자

김상국

서울특별시 마포구 새창로 7, 1601호(도화동, 에스엔유장학빌딩)

성준경

서울특별시 마포구 새창로 7, 1601호(도화동, 에스엔유장학빌딩)

명 세 서

청구범위

청구항 1

컴퓨터가 수행하는 메시지 인덱싱 방법에 있어서,

메시지에 대해서 쿼리 토크나이저 모듈을 실행시키는 제1 단계와,

상기 제1 단계에서 토크나이징된 정보에 대해서 인덱싱을 수행하는 제2 단계를 포함하며,

상기 제1 단계는, 쿼리 토크나이저 모듈이 쿼리에 기반을 두어 메시지를 토크나이징하는 단계이며, 쿼리는 메시지의 지정된 필드의 필드 타입을 정의하는 구문을 포함하는,

메시지 인덱싱 방법.

청구항 2

청구항 1에 있어서,

상기 제1 단계에서 상기 쿼리 토크나이저 모듈을 실행할 때의 쿼리는 메시지의 일부 필드를 지정하는 구문을 포함하고,

상기 제2 단계는 상기 지정된 일부 필드에 대해서만 수행되는 단계인,

메시지 인덱싱 방법.

청구항 3

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 메시지의 필드 인덱싱 방법에 관한 것으로서 좀 더 구체적으로는 부하가 적고 고속의 검색이 가능하도록 하는 인덱싱 방법에 관한 것이다.

배 경 기 술

[0003] 인터넷이 발달하고 인터넷 등을 이용한 온라인 서비스가 확산될수록 온라인상에서 서비스를 제공하는 서버의 동작 과정에서 발생하는 작업 내력 예를 들어 웹 로그, 방화벽 로그, 거래 로그 등의 데이터의 양이 방대해지고 있다. 일반적으로 로그에는 복수 개의 로그 라인이 기록되며, 각 로그라인에는 방문자의 접속 인터넷 프로토콜(IP), 방문 시간 정보, 방문 웹 페이지 정보, 방문 상태 등이 기록될 수 있다.

[0004] 이러한 로그의 분석을 위해서 특정 단어나 문자열을 검색해야 하는 경우가 많은데 로그 데이터의 관리 내지 검색을 위한 인덱싱 과정의 중요성이 대두되고 있다.

[0005] 본 출원 발명의 발명자가 특허권자로서 보유하고 있는 특허 제1112568호에는 방대한 데이터의 효율적인 검색을 위한 로그 인덱싱 방법이 개시되어 있다. 이 특허에 개시된 방법에 의하면 로그를 데이터베이스화하는 정규화 과정없이 인덱싱을 통해서 신뢰성 있고 빠른 로그 검색이 가능하게 된다. 이 특허의 내용 전체는 본 명세서의 일부로서 본 명세서에 반영되며 본 명세서에서 명시적인 언급이 없더라도 본 발명의 설명을 위해서 사용될 수 있다. 그러나 이 특허의 내용이 본 발명의 권리범위를 제한하는 것으로 이용되어서는 아니되며, 본 발명의 이해를 돕기 위한 용도로만 이용되어야 한다.

발명의 내용

해결하려는 과제

[0007] 본 발명은 전술한 특허의 로그 인덱싱 방법보다 더 빠르고 부하가 적은 인덱싱 방법을 제공하는 것을 목적으로 한다.

과제의 해결 수단

[0009] 본 발명은 컴퓨터가 수행하는 메시지 인덱싱 방법에 관한 것으로서, 메시지에 대해서 쿼리 토크나이저 모듈을 실행시키는 제1 단계와, 상기 제1 단계에서 토크나이징된 정보에 대해서 인덱싱을 수행하는 제2 단계를 포함한다.

[0010] 상기 제1 단계에서 상기 쿼리 토크나이저 모듈을 실행할 때의 쿼리는 메시지의 일부 필드를 지정하는 구문을 포함하고, 상기 제2 단계는 상기 지정된 일부 필드에 대해서만 수행되는 단계인 것이 바람직하다.

[0011] 상기 제1 단계에서 상기 쿼리 토크나이저 모듈을 실행할 때의 쿼리는 메시지의 지정된 필드의 필드 타입을 정의하는 구문도 포함하는 것이 바람직하다.

발명의 효과

[0013] 본 발명에 의하면, 고속 인덱싱이 가능해지며, 입력/출력 부하가 감소하며, 검색에 있어서도 범위 및 대소 비교가 빠르게 수행될 수 있는 효과가 제공된다.

도면의 간단한 설명

[0015] 도 1은 본 발명에 의한 인덱싱 방법의 흐름도.

도 2는 본 발명에 의한 인덱싱 방법에 의해 인덱싱된 데이터의 검색 방법의 흐름도.

발명을 실시하기 위한 구체적인 내용

[0016] 이하에서는 첨부 도면을 참조하여 본 발명에 대해서 자세하게 설명한다. 본 발명에 의한 인덱싱 방법은 컴퓨터에 의해서 수행되며, 본 명세서에서 컴퓨터라 함은 전자적 연산을 수행하고 프로그램에 의해 작동 가능한 전자기기를 망라하는 것으로 정의된다. 예를 들어, 개인용 컴퓨터(PC) 뿐만 아니라 서버 컴퓨터 또는 본 발며에 의한 데이터 처리에 적합하다면 모바일 기기 등도 포함될 수 있다.

[0017] 도 1에는 본 발명에 의한 인덱싱 방법의 흐름도가 도시되어 있다. 먼저 인덱싱할 원본 데이터를 삽입한다(100). 원본 데이터의 종류는 웹 로그 데이터, 방화벽 로그 데이터, 금융거래 로그 데이터 등이 될 수 있으며, 본 발명에 의한 인덱싱 방법이 적용 가능한 데이터라면 그 종류를 불문하고 모두 포함될 수 있으며, 본 명세서에서는 "메시지"라는 용어로도 혼용되어 사용된다.

[0018] 원본 데이터가 삽입되면 원본 데이터(메시지)에 대해서 쿼리 토크나이저 모듈이 실행된다(110). 쿼리 토크나이저 모듈은 종래의 토크나이저와 달리 쿼리에 기반을 두어 토크나이징을 하는, 범용의 하드웨어와 그 기능을 수행하는 소프트웨어의 논리적 결합으로서 정의된다.

[0019] 쿼리 토크나이저 모듈의 실행에 의해서 쿼리에 기반하여 원본 데이터 중 일부 필드가 지정되어 토크나이징될 수 있다.

[0020] 토크나이저 모듈의 실행에 의해서 토큰 및 토큰 메타 정보가 추출된다(120). 토큰 메타 정보로는 필드명, 필드 타입(int, string, long, ip) 등이 있을 수 있다.

[0021] 다음으로 이렇게 토크나이징된 정보에 대해서 인덱싱하여 인덱스를 생성한다(130). 구체적인 인덱싱 방법으로는 여러가지가 있으며, 예를 들어 전술한 본 발명자의 특허에 개시되어 있는 인덱싱 방법을 적용할 수 있다. 그러나 본 발명에 의한 쿼리 토크나이저 모듈의 실행에 의해서 토크나이징된 후의 데이터에 대해서 인덱싱하는 구체적인 방법에 본 발명의 권리범위가 제한되는 것은 아니며, 공개되어 있는 다양한 인덱싱 방법이 적용될 수 있다.

[0022] 이와 같이 쿼리 토크나이저 모듈을 실행시켜서 원하는 필드에 대해서 필드 타입을 지정해서 인덱싱을 하게 되면

종래의 인덱싱 방법에 비해서 고속으로 인덱싱을 수행할 수 있는 장점이 있다. 본 발명에 의한 인덱싱 방법에 따르면, 데이터의 특정 (원하는) 필드 예를 들어, 포트 정보, 파일 크기, 파일 이름 등에 대해서만 인덱싱을 하기 때문에 불필요한 필드까지 전체를 인덱싱하는 종래의 인덱싱 방법에 비해서 I/O가 감소하여 매우 빠른 검색이 가능해진다. 또한, 쿼리 토크나이저 모듈의 실행에 의해서 인덱싱 전에 필드의 타입을 지정할 수 있기 때문에 문자열 형태로 인덱싱하는 종래의 방법에 비하여 적은 용량으로 인덱싱을 할 수 있어, 검색시 I/O 부하가 감소하여 소요 시간이 감소한다. 예를 들어 IP 주소를 문자열로 저장하면 최대 15바이트가 되지만, 필드 타입을 ip 타입으로 저장하면 4바이트로 감소하여, 데이터를 읽는데 소요되는 시간이 줄어든다. 또한, 인덱싱 후에 필드 타입에 대한 정보(string, int, ip)가 있기 때문에 해당 타입에 맞는 연산을 수행할 수 있게 된다. 따라서 범위를 검색하거나 대소 비교를 할 때에 매우 유용하며, 또한 기존의 문자열에 대한 OR 조건식 대신 해당 타입에 맞는 연산을 직접 수행하여 고속으로 범위 검색이 가능해진다.

[0023] 도 2에는 본 발명에 의한 인덱싱 방법에 의해 인덱싱되어 있는 정보를 검색하는 흐름도가 도시되어 있다.

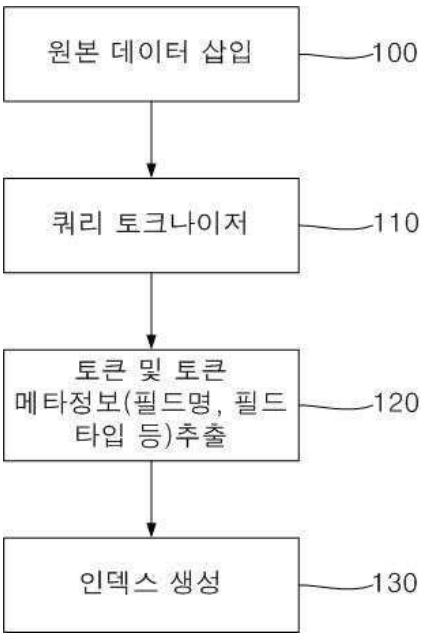
[0024] 먼저 세그먼트의 헤더를 참조하여 검색 기간의 해당 여부를 확인하고 필드 엔트리를 참조하여, 검색하려는 해당 필드의 존재 여부를 확인한다(200). 다음으로 표현식에 해당하는 포스팅 그룹(해당 토큰이 포함되는 로그 식별정보 목록)을 추출한다(210). 그리고 포스팅에 대해서 불린(Boolean) 연산을 수행한다(220). 예를 들어 다수 검색어를 검색할 때에 수행되는 것이 바람직하다.

[0025] 최종적으로 검색된 로그 식별정보(로그 ID) 목록에 기초하여 원본 로그를 추출한다(230).

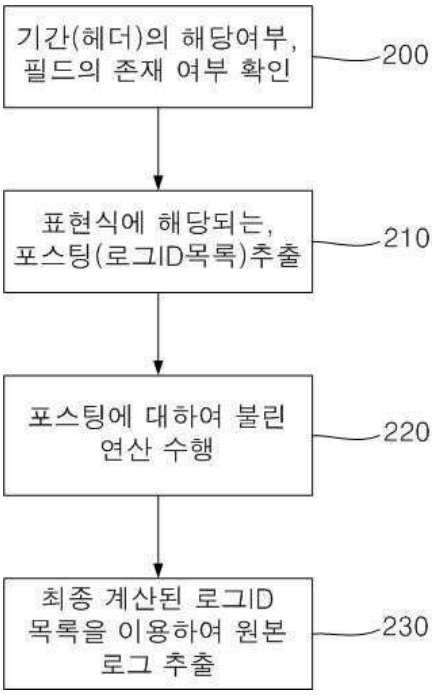
[0026] 이상 첨부 도면을 참고하여 본 발명에 대해서 설명하였지만 본 발명의 권리범위는 후술하는 특허청구범위에 의해 결정되며 전술한 실시예 및/또는 도면에 제한되는 것으로 해석되어서는 아니된다. 그리고 특허청구범위에 기재된 발명의, 당업자에게 자명한 개량, 변경 및 수정도 본 발명의 권리범위에 포함된다는 점이 명백하게 이해되어야 한다.

도면

도면1



도면2


기간(헤더)의 해당여부, 필드의 존재 여부 확인
200
표현식에 해당되는, 포스팅(로그ID목록)추출
210
포스팅에 대하여 불린 연산 수행
220
최종 계산된 로그ID 목록을 이용하여 원본 로그 추출
230