(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2012년05월23일
(11) 등록번호 10-1148552
(24) 등록일자 2012년05월15일

(51) 국제특허분류(Int. Cl.)
G06F 17/40 (2006.01) *G06F 17/30* (2006.01)
(21) 출원번호 10-2010-0097229
(22) 출원일자 2010년10월06일
심사청구일자 2010년10월06일
(65) 공개번호 10-2012-0035605
(43) 공개일자 2012년04월16일
(56) 선행기술조사문헌
JP2001075982 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
엔에이치엔(주)
경기도 성남시 분당구 불정로 6, 그린팩토리 (정자동)
(72) 발명자
원태륜
경기도 성남시 분당구 서현로 170, 풍림아이원플러스 C동 1116호 (서현동)
심상옥
경기도 성남시 분당구 탄천로 35, 508동 1404호 (이매동, 아름마을)
(74) 대리인
특허법인무한

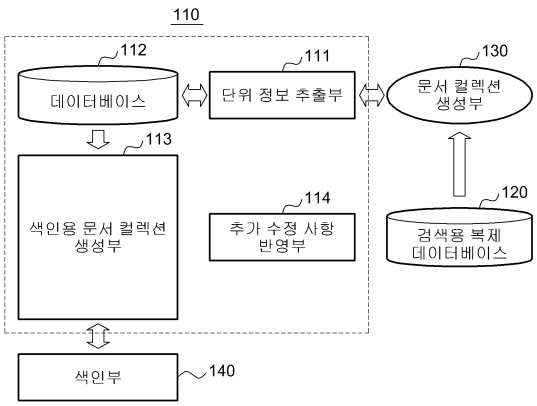
전체 청구항 수 : 총 14 항 심사관 : 손현웅

(54) 발명의 명칭 **수정된 문서의 정보를 이용한 문서 색인 시스템 및 방법**

(57) 요 약

수정된 문서의 정보를 이용한 문서 색인 시스템 및 방법이 개시된다. 문서 색인 시스템은 전체 문서 중 신규 생성된 문서 및 수정된 문서를 포함하는 문서 컬렉션을 수신하여 문서 컬렉션에서 각 문서의 저장 위치에 대한 정보를 적어도 포함하는 단위 정보를 추출하고, 추출된 단위 정보 및 문서 컬렉션을 저장하는 단위 정보 추출부 및 색인용 문서 컬렉션에 대한 생성 요청에 대응하는 단위 정보를 통해 저장된 문서 컬렉션에서 생성 요청에 해당하는 문서를 추출하고, 추출된 문서를 이용하여 색인용 문서 컬렉션을 생성하는 색인용 문서 컬렉션 생성부를 포함한다.

대 표 도 - 도1



특허청구의 범위

청구항 1

검색 색인을 위해 문서를 가공한 데이터인 색인용 문서 컬렉션을 제공하는 문서 색인 시스템에 있어서,

전체 문서 중 신규 생성된 문서 및 수정된 문서를 포함하는 문서 컬렉션을 수신하여 상기 문서 컬렉션에서 각 문서의 저장 위치에 대한 정보를 적어도 포함하는 단위 정보를 추출하고, 상기 추출된 단위 정보 및 상기 문서 컬렉션을 저장하는 단위 정보 추출부; 및

상기 색인용 문서 컬렉션에 대한 생성 요청에 대응하는 단위 정보를 통해 상기 저장된 문서 컬렉션에서 상기 생성 요청에 해당하는 문서를 추출하고, 상기 추출된 문서를 이용하여 상기 색인용 문서 컬렉션을 생성하는 색인용 문서 컬렉션 생성부

를 포함하고,

상기 수정된 문서는 전체 문서 각각에 대해 공통적으로 정의된 모든 항목 중 미리 설정된 항목이 기선정된 기간동안 수정된 문서를 포함하는 문서 색인 시스템.

청구항 2

제1항에 있어서,

상기 단위 정보는 해당 문서가 저장된 경로, 상기 문서 컬렉션에서 시작하는 위치(offset) 및 상기 해당 문서의 길이를 포함하는 문서 색인 시스템.

청구항 3

제1항에 있어서,

상기 단위 정보는 해당 문서의 식별자, 상기 해당 문서가 게재된 사이트의 식별자, 상기 해당 문서의 카테고리, 상기 해당 문서의 타입, 상기 해당 문서가 생성된 시간, 상기 해당 문서가 수정된 시간, 상기 해당 문서가 갱신된 시간, 상기 해당 문서의 상태를 정의한 정보 및 상기 해당 문서의 품질을 정의한 정보 중 적어도 하나를 더 포함하는 문서 색인 시스템.

청구항 4

제1항에 있어서,

상기 생성 요청은 필요한 문서의 식별자, 상기 필요한 문서가 게재된 사이트의 식별자, 상기 필요한 문서의 카테고리, 상기 필요한 문서의 타입, 상기 필요한 문서가 생성된 시간, 상기 필요한 문서가 수정된 시간, 상기 필요한 문서가 갱신된 시간, 상기 필요한 문서의 상태를 정의한 정보 및 상기 필요한 문서의 품질을 정의한 정보 중 적어도 하나를 포함하는 문서 색인 시스템.

청구항 5

검색 색인을 위해 문서를 가공한 데이터인 색인용 문서 컬렉션을 제공하는 문서 색인 시스템에 있어서,

전체 문서 중 신규 생성된 문서 및 수정된 문서를 포함하는 문서 컬렉션을 수신하여 상기 문서 컬렉션에서 각 문서의 저장 위치에 대한 정보를 적어도 포함하는 단위 정보를 추출하고, 상기 추출된 단위 정보 및 상기 문서 컬렉션을 저장하는 단위 정보 추출부; 및

상기 색인용 문서 컬렉션에 대한 생성 요청에 대응하는 단위 정보를 통해 상기 저장된 문서 컬렉션에서 상기 생성 요청에 해당하는 문서를 추출하고, 상기 추출된 문서를 이용하여 상기 색인용 문서 컬렉션을 생성하는 색인용 문서 컬렉션 생성부

를 포함하고,

최근 기간에 생성된 색인용 문서 컨렉션과 기존에 생성된 색인용 문서 컬렉션을 이용하여 상기 전체 문서에 대한 검색이 수행되는 문서 색인 시스템.

청구항 6

검색 색인을 위해 문서를 가공한 데이터인 색인용 문서 컬렉션을 제공하는 문서 색인 시스템에 있어서,

전체 문서 중 신규 생성된 문서 및 수정된 문서를 포함하는 문서 컬렉션을 수신하여 상기 문서 컬렉션에서 각 문서의 저장 위치에 대한 정보를 적어도 포함하는 단위 정보를 추출하고, 상기 추출된 단위 정보 및 상기 문서 컬렉션을 저장하는 단위 정보 추출부; 및

상기 색인용 문서 컬렉션에 대한 생성 요청에 대응하는 단위 정보를 통해 상기 저장된 문서 컬렉션에서 상기 생성 요청에 해당하는 문서를 추출하고, 상기 추출된 문서를 이용하여 상기 색인용 문서 컬렉션을 생성하는 색인용 문서 컬렉션 생성부

를 포함하고,

상기 추출된 단위 정보가 저장되는 데이터베이스

를 더 포함하고,

상기 색인용 문서 컬렉션 생성부는,

상기 생성 요청에 대응하는 단위 정보를 상기 데이터베이스로부터 추출하고, 상기 추출된 단위 정보를 이용하여 상기 문서 컬렉션에서 상기 생성 요청에 해당하는 문서를 추출하여 상기 색인용 문서 컬렉션을 생성하는 문서 색인 시스템.

청구항 7

삭제

청구항 8

제1항에 있어서,

상기 기선정된 기간과는 다른 기간마다 상기 미리 설정된 항목을 제외한 나머지 항목들의 추가 수정 사항을 상기 단위 정보 또는 상기 문서 컬렉션에 반영하는 추가 수정 사항 반영부

를 더 포함하는 문서 색인 시스템.

청구항 9

검색 색인을 위해 문서를 가공한 데이터인 색인용 문서 컬렉션을 제공하는 문서 색인 방법에 있어서,

전체 문서 중 신규 생성된 문서 및 수정된 문서를 포함하는 문서 컬렉션을 수신하는 단계;

상기 문서 컬렉션에서 각 문서의 저장 위치에 대한 정보를 적어도 포함하는 단위 정보를 추출하고, 상기 추출된 단위 정보 및 상기 문서 컬렉션을 저장하는 단계;

상기 색인용 문서 컬렉션에 대한 생성 요청에 대응하는 단위 정보를 조회하는 단계; 및

상기 단위 정보를 통해 상기 문서 컬렉션에서 추출된 문서를 이용하여 상기 색인용 문서 컬렉션을 생성하는 단계

를 포함하고,

상기 수정된 문서는 전체 문서 각각에 대해 공통적으로 정의된 모든 항목 중 미리 설정된 항목이 기선정된 기간동안 수정된 문서를 포함하는 문서 색인 방법.

청구항 10

제9항에 있어서,

상기 단위 정보는 해당 문서가 저장된 경로, 상기 문서 컬렉션 내에서 시작하는 위치 및 상기 해당 문서의 길이를 포함하는 문서 색인 방법.

청구항 11

제9항에 있어서,

상기 단위 정보는 해당 문서의 식별자, 상기 해당 문서가 게재된 사이트의 식별자, 상기 해당 문서의 카테고리, 상기 해당 문서의 타입, 상기 해당 문서가 생성된 시간, 상기 해당 문서가 수정된 시간, 상기 해당 문서가 갱신된 시간, 상기 해당 문서의 상태를 정의한 정보 및 상기 해당 문서의 품질을 정의한 정보 중 적어도 하나를 더 포함하는 문서 색인 방법.

청구항 12

제9항에 있어서,

상기 생성 요청은 필요한 문서의 식별자, 상기 필요한 문서가 게재된 사이트의 식별자, 상기 필요한 문서의 카테고리, 상기 필요한 문서의 타입, 상기 필요한 문서가 생성된 시간, 상기 필요한 문서가 수정된 시간, 상기 필요한 문서가 갱신된 시간, 상기 필요한 문서의 상태를 정의한 정보 및 상기 필요한 문서의 품질을 정의한 정보 중 적어도 하나를 포함하는 문서 색인 방법.

청구항 13

검색 색인을 위해 문서를 가공한 데이터인 색인용 문서 컬렉션을 제공하는 문서 색인 방법에 있어서,

전체 문서 중 신규 생성된 문서 및 수정된 문서를 포함하는 문서 컬렉션을 수신하는 단계;

상기 문서 컬렉션에서 각 문서의 저장 위치에 대한 정보를 적어도 포함하는 단위 정보를 추출하고, 상기 추출된 단위 정보 및 상기 문서 컬렉션을 저장하는 단계;

상기 색인용 문서 컬렉션에 대한 생성 요청에 대응하는 단위 정보를 조회하는 단계; 및

상기 단위 정보를 통해 상기 문서 컬렉션에서 추출된 문서를 이용하여 상기 색인용 문서 컬렉션을 생성하는 단계

를 포함하고,

최근 기간에 생성된 색인용 문서 컨렉션과 기존에 생성된 색인용 문서 컬렉션을 이용하여 상기 전체 문서에 대한 검색이 수행되는 문서 색인 방법.

청구항 14

삭제

청구항 15

제9항에 있어서,

상기 기선정된 기간과는 다른 기간마다 상기 미리 설정된 항목을 제외한 나머지 항목들의 추가 수정 사항을 상기 단위 정보 또는 상기 문서 컬렉션에 반영하는 단계

를 더 포함하는 문서 색인 방법.

청구항 16

제9항 내지 제13항 또는 제15항 중 어느 한 항의 방법을 수행하는 프로그램을 기록한 컴퓨터 판독 가능 기록매체.

명 세 서

기 술 분 야

[0001] 본 발명의 실시예들은 수정된 문서의 정보를 이용한 문서 색인 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 검색 색인 생성에 사용할 수 있도록 가공된 데이터를 검색 문서 컬렉션이라 부르는데, 기존에는 데이터의 변경여부에 상관없이 일정주기로 전체 문서 컬렉션을 새로 생성한다. 즉, 수 억 건이 넘는 문서 중 약 99% 이

상은 변경이 일어나지 않음에도 불구하고, 기존에는 변경된 문서의 구분과 여러 가지 유지보수의 어려움 때문에 문서의 수정 여부와는 상관없이 전체 문서 컬렉션을 재생성한다.

[0003] 본 명세서에서는 효율적으로 문서를 색인할 수 있는 문서 색인 시스템 및 문서 색인 방법이 제안된다.

발명의 내용

해결하려는 과제

[0004] 수정이 발생하지 않은 문서는 기존에 생성한 색인용 문서 컬렉션을 재사용하고, 신규로 생성된 문서와 수정이 발생한 문서에 대해서만 색인용 문서 컬렉션을 새로 생성할 수 있는 문서 색인 시스템 및 방법이 제공된다.

[0005] 신규로 생성된 문서와 수정이 발생한 문서에 대해서만 색인용 문서 컬렉션을 새로 생성하고, 수정이 발생하지 않은 문서에 대해서는 기존에 생성한 색인용 문서 컬렉션을 재사용함으로써, 검색 문서의 색인을 생성하는 시간을 획기적으로 단축하고, 문서들에 대한 데이터를 제공하는 검색용 복제 데이터베이스의 부하도 현저하게 줄일 수 있는 문서 색인 시스템 및 방법이 제공된다.

과제의 해결 수단

[0006] 전체 문서 중 신규 생성된 문서 및 수정된 문서를 포함하는 문서 컬렉션을 문서 컬렉션에서 각 문서의 저장 위치에 대한 정보를 적어도 포함하는 단위 정보를 추출하고, 추출된 단위 정보 및 문서 컬렉션을 저장하는 단위 정보 추출부 및 색인용 문서 컬렉션에 대한 생성 요청에 대응하는 단위 정보를 통해 저장된 문서 컬렉션에서 생성 요청에 해당하는 문서를 추출하고, 추출된 문서를 이용하여 색인용 문서 컬렉션을 생성하는 색인용 문서 컬렉션 생성부를 포함하는 문서 색인 시스템이 제공된다.

[0007] 일측에 따르면, 단위 정보는 해당 문서가 저장된 경로, 문서 컬렉션에서 시작하는 위치(offset) 및 해당 문서의 길이를 포함할 수 있다.

[0008] 다른 측면에서, 단위 정보는 해당 문서의 식별자, 해당 문서가 게재된 사이트의 식별자, 해당 문서의 카테고리, 해당 문서의 타입, 해당 문서가 생성된 시간, 해당 문서가 수정된 시간, 해당 문서가 갱신된 시간, 해당 문서의 상태를 정의한 정보 및 해당 문서의 품질을 정의한 정보 중 적어도 하나를 더 포함할 수 있다.

[0009] 또 다른 측면에서, 생성 요청은 필요한 문서의 식별자, 필요한 문서가 게재된 사이트의 식별자, 필요한 문서의 카테고리, 필요한 문서의 타입, 필요한 문서가 생성된 시간, 필요한 문서가 수정된 시간, 필요한 문서가 갱신된 시간, 필요한 문서의 상태를 정의한 정보 및 필요한 문서의 품질을 정의한 정보 중 적어도 하나를 포함할 수 있다.

[0010] 또 다른 측면에서, 최근 기간에 생성된 색인용 문서 컨렉션과 기존에 생성된 색인용 문서 컬렉션을 이용하여 전체 문서에 대한 검색이 수행될 수 있다.

[0011] 또 다른 측면에서, 문서 색인 시스템은 추출된 단위 정보 및 문서 컬렉션이 저장되는 데이터베이스를 더 포함할 수 있고, 색인용 문서 컬렉션 생성부는 생성 요청에 대응하는 단위 정보를 데이터베이스로부터 추출하고, 추출된 단위 정보를 이용하여 데이터베이스에 저장된 문서 컬렉션에서 생성 요청에 해당하는 문서를 추출하여 색인용 문서 컬렉션을 생성할 수 있다.

[0012] 또 다른 측면에서, 수정된 문서는 전체 문서 각각에 대해 공통적으로 정의된 모든 항목 중 미리 설정된 항목이 기선정된 기간동안 수정된 문서를 포함할 수 있다. 이 경우, 문서 컬렉션 생성부는 기선정된 기간과는 다른 기간마다 미리 설정된 항목을 제외한 나머지 항목들의 추가 수정 사항을 단위 정보 또는 문서 컬렉션에 반영하는 추가 수정 사항 반영부를 더 포함할 수 있다.

[0013] 전체 문서 중 신규 생성된 문서 및 수정된 문서를 포함하는 문서 컬렉션을 수신하는 단계, 문서 컬렉션에서 각 문서의 저장 위치에 대한 정보를 적어도 포함하는 단위 정보를 추출하고, 추출된 단위 정보 및 문서 컬렉션을 저장하는 단계, 색인용 문서 컬렉션에 대한 생성 요청에 대응하는 단위 정보를 조회하는 단계 및 단위 정보를 통해 문서 컬렉션에서 추출된 문서를 이용하여 색인용 문서 컬렉션을 생성하는 단계를 포함하는 문서 색인 방법이 제공된다.

발명의 효과

[0014] 수정이 발생하지 않은 문서는 기존에 생성한 색인용 문서 컬렉션을 재사용하고, 신규로 생성된 문서와 수정이

발생한 문서에 대해서만 색인용 문서 컬렉션을 새로 생성할 수 있다.

[0015] 신규로 생성된 문서와 수정이 발생한 문서에 대해서만 색인용 문서 컬렉션을 새로 생성하고, 수정이 발생하지 않은 문서에 대해서는 기존에 생성한 색인용 문서 컬렉션을 재사용함으로써, 검색 문서의 색인을 생성하는 시간을 획기적으로 단축하고, 문서들에 대한 데이터를 제공하는 검색용 복제 데이터베이스의 부하도 현저하게 줄일 수 있다.

도면의 간단한 설명

[0016] 도 1은 본 발명의 일실시예에 있어서, 문서 색인 시스템을 도시한 블록도이다.

도 2는 본 발명의 일실시예에 있어서, 문서 색인 방법을 도시한 흐름도이다.

도 3은 본 발명의 일실시예에 있어서, 단위 정보의 일례를 나타낸 표이다.

도 4는 본 발명의 일실시예에 있어서, 색인용 문서 컬렉션의 일례를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0017] 이하, 본 발명의 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0018] 본 발명의 실시예들에 다른 문서 색인 시스템 및 문서 색인 방법은 검색 색인을 위해 문서를 가공한 데이터인 색인용 문서 컬렉션을 제공한다. 이때, 문서 색인 시스템 및 문서 색인 방법에서는 전체 문서를 이용하는 것이 아니라 신규 생성된 문서와 수정된 문서를 이용하여 색인용 문서 컬렉션을 생성하고, 수정이 발생하지 않은 문서에 대해서는 기존에 생성한 색인용 문서 컬렉션을 재사용함으로써, 검색 문서의 색인을 생성하는 시간을 획기적으로 단축하고, 문서들에 대한 데이터를 제공하는 검색용 복제 데이터베이스의 부하도 현저하게 줄일 수 있다.

[0019] 도 1은 본 발명의 일실시예에 있어서, 문서 색인 시스템을 도시한 블록도이다. 본 실시예에 따른 문서 색인 시스템(110)은 도 1에 도시된 바와 같이 단위 정보 추출부(111), 데이터베이스(112), 색인용 문서 컬렉션 생성부(113) 및 추가 수정 사항 반영부(114)를 포함할 수 있다. 도 1에 더 나타난 검색용 복제 데이터베이스(120), 문서 컬렉션 생성부(130) 및 색인부(140) 중 적어도 하나는 필요에 따라 문서 색인 시스템(110)에 포함될 수도 있다. 또는 서로 다른 시스템으로서 문서 색인 시스템(110)과 연계하여 동작될 수도 있다.

[0020] 문서의 색인을 위해 각각의 문서들은 미리 정의된 항목들로 구성된 데이터로 가공될 필요가 있다. 이러한 문서의 가공은 문서 색인 시스템(110)의 외부에서 수행되어 가공된 문서가 문서 색인 시스템(110)으로 수신될 수도 있고, 문서 색인 시스템(110)에서 직접 수신된 문서를 가공할 수도 있다. 문서 색인 시스템(110)에서 문서를 가공하는 경우, 문서의 가공은 단위 정보 추출부(111)에서 수신된 문서를 통해 단위 정보를 추출할 때 또는 색인용 문서 컬렉션 생성부(113)에서 색인용 문서 컬렉션을 생성할 때 수행될 수 있다. 가공된 문서에 대해서는 이후 도 4를 통해 더욱 자세히 설명한다.

[0021] 단위 정보 추출부(111)는 전체 문서 중 신규 생성된 문서 및 수정된 문서를 포함하는 문서 컬렉션을 수신하여 문서 컬렉션에서 각 문서의 저장 위치에 대한 정보를 적어도 포함하는 단위 정보를 추출하고, 추출된 단위 정보 및 문서 컬렉션을 저장한다. 이때, 신규 생성된 문서 및 수정된 문서를 포함하는 문서 컬렉션은 단위 정보 추출부(111)에서 발생하는 요청에 따라 문서 컬렉션 생성부(130)가 검색용 복제 데이터베이스(120)를 통해 생성할 수 있다. 예를 들어, 단위 정보 추출부(111)는 1분마다 신규 생성된 문서와 수정된 문서를 포함하는 문서 컬렉션을 문서 컬렉션 생성부(130)로 요청할 수 있고, 문서 컬렉션 생성부(130)는 1분 동안 신규 생성된 문서와 수정된 문서를 검색용 복제 데이터베이스(120)로부터 수신하여 문서 컬렉션을 생성할 수 있다.

[0022] 데이터베이스(112)에는 추출된 단위 정보가 저장된다. 여기서, 문서 컬렉션은 파일의 형태로 파일 시스템에 저장될 수 있고, 기간마다 수신되는 문서들이 해당 파일에 저장될 수 있다. 이 경우, 문서 색인 시스템(110)은 데이터베이스(112)에 저장된 단위 정보를 이용하여 파일 시스템에 저장된 문서 컬렉션의 문서들 중 원하는 문서를 찾을 수 있게 된다.

[0023] 또한, 현재 기간에 수정된 문서 중 적어도 일부의 문서는 이미 이전의 다른 기간에 수정되어 문서 컬렉션에 저장되어 있을 수 있다. 따라서, 문서 색인 시스템(110) 또는 색인용 문서 컬렉션 생성부(113)는 이미 저장된 문서가 존재하는 경우, 새로 수신된 문서를 저장하고, 해당 단위 정보가 새로 저장된 문서의 위치에 대한 정보를 포함하도록 갱신함으로써, 색인용 문서 컬렉션 생성부(113)는 항상 최신의 문서를 추출할 수 있다.

[0024] 색인용 문서 컬렉션 생성부(113)는 색인용 문서 컬렉션에 대한 생성 요청에 대응하는 단위 정보를 통해 상기 저장된 문서 컬렉션에서 상기 생성 요청에 해당하는 문서를 추출하고, 상기 추출된 문서를 이용하여 상기 색인용 문서 컬렉션을 생성한다. 여기서, 단위 정보는 해당 문서가 저장된 파일의 경로, 파일 내에서 시작하는 위치(offset) 및 해당 문서의 길이를 포함하는 것으로, 상술한 바와 같이 데이터베이스(112)에 저장될 수 있다. 이러한 단위 정보는 해당 문서가 어떠한 파일의 어느 위치에 존재하는 가를 나타내는 정보를 포함할 수 있다. 따라서, 색인용 문서 컬렉션 생성부(113)는 색인부(140)를 통해 색인용 문서 컬렉션에 대한 생성 요청이 발생하는 경우, 생성 요청의 조건에 부합하는 문서들의 파일 위치를 포함하는 단위 정보를 데이터베이스(130)에서 조회하고, 조회된 단위 정보를 이용하여 문서 컬렉션에서 해당 문서들을 추출할 수 있다. 이때, 색인용 문서 컬렉션 생성부(113)는 추출한 문서들을 이용하여 색인용 문서 컬렉션을 생성할 수 있다.

[0025] 색인용 문서 컬렉션에 대한 생성 요청은 도 1에 도시된 색인부(140)를 통해 수신될 수 있고, 생성된 색인용 문서 컬렉션 역시 색인부(140)로 제공될 수 있다. 즉, 색인부(140)는 신규 생성되거나 수정된 문서들에 대한 생성된 색인용 문서 컬렉션과 기존에 생성된 색인용 문서 컬렉션을 검색 색인에 이용할 수 있다. 따라서, 전체 문서에 대한 색인용 문서 컬렉션을 새로 생성할 필요 없이 신규 생성되거나 수정된 문서들에 대해서만 색인용 문서 컬렉션을 생성할 수 있다.

[0026] 또한, 단위 정보는 해당 문서의 식별자, 해당 문서가 게재된 사이트의 식별자, 해당 문서의 카테고리, 해당 문서의 타입, 해당 문서가 생성된 시간, 해당 문서가 수정된 시간, 해당 문서가 갱신된 시간, 해당 문서의 상태를 정의한 정보 및 해당 문서의 품질을 정의한 정보 중 적어도 하나를 더 포함할 수 있다. 이러한 정보들은 색인부(140)로부터 수신되는 생성 요청의 조건에 관한 것으로, 생성 요청 역시 필요한 문서의 식별자, 필요한 문서가 게재된 사이트의 식별자, 필요한 문서의 카테고리, 필요한 문서의 타입, 필요한 문서가 생성된 시간, 필요한 문서가 수정된 시간, 필요한 문서가 갱신된 시간, 필요한 문서의 상태를 정의한 정보 및 필요한 문서의 품질을 정의한 정보 중 적어도 하나를 포함할 수 있다. 예를 들어, 색인부(140)가 최근 하루 동안 생성 및 수정된 문서가 필요한 경우, 필요한 문서의 생성 및 수정된 시간에 대한 정보를 생성 요청에 포함시켜 문서 색인 시스템(110)으로 전송할 수 있고, 문서 색인 시스템(110)의 색인용 문서 컬렉션 생성부(113)는 데이터베이스(112)에 저장된 단위 정보를 통해 최근 하루 동안 생성 및 수정된 문서가 문서 컬렉션의 어느 위치에 저장되어 있는지 확인하여 문서 컬렉션에서 해당 문서를 추출할 수 있다. 또 다른 예로, 특정 카테고리나 특정 타입의 문서 또는 특정 사이트에 게시된 문서 등이 필요한 경우, 생성 요청은 필요한 문서의 카테고리나 타입 또는 필요한 문서가 게재된 사이트의 식별자를 포함할 수 있고, 문서 색인 시스템(110)은 단위 정보를 이용하여 생성 요청의 조건에 맞는 문서를 추출할 수 있다.

[0027] 수정된 문서는 전체 문서 각각에 대해 공통적으로 정의된 모든 항목 중 미리 설정된 항목이 기선정된 기간동안 수정된 문서를 포함할 수 있다. 예를 들어 기선정된 기간은 단위 정보 추출부(111)의 문서 컬렉션 요청과 요청 사이의 기간을 포함할 수 있다. 이때, 상술한 바와 같이, 각각의 문서들은 색인을 위한 항목들이 미리 정의될 수 있다. 그러나, 정의된 항목이 많을수록 매 기간마다 수신되는 수정된 문서의 양이 방대해질 수 있다. 예를 들어, 단위 정보 추출부(111)에서 1분 단위로 수정된 문서를 수신할 때, 스크랩 수나 조회 수 등은 매우 빈번하게 변경되기 때문에 이러한 항목들이 수정 항목에 포함되는 경우, 단위 정보 추출부(111)로 수신되는 수정된 문서의 양이 매우 많아질 수 있다. 따라서, 수정 항목은 별도로 설정될 필요가 있다.

[0028] 그러나 이 경우, 수정 항목에 정의되지 못한 항목들은, 시간이 지남에 따라 원본 데이터와의 오차가 점점 누적될 수 있다. 이러한 문제를 해결하기 위해, 문서 색인 시스템(110)은 기선정된 기간과는 다른 기간마다 상기 미리 설정된 항목을 제외한 나머지 항목들의 추가 수정 사항을 상기 단위 정보 또는 상기 문서 컬렉션에 반영하는 추가 수정 사항 반영부(114)를 포함할 수 있다. 즉, 추가 수정 사항 반영부(114)는 기본적으로 기설정된 기간과는 별도의 기간(예를 들어, 1시간이나 하루)마다 상술한 일례의 스크랩 수나 조회수 등 수정 항목으로 설정되지 않은 추가 수정 항목들이 수정된 문서들을 수신할 수 있다. 또한, 추가 수정 사항 반영부(114)는 수신된 문서들을 문서 컬렉션의 해당 문서에 덮어쓰기하여 수정 항목을 반영하거나 또는 단위 정보에 이러한 수정 항목들이 반영되도록 해당 단위 정보를 수정함으로써, 원본 데이터와의 오차가 발생하지 않도록 할 수 있다. 필요에 따라, 수신된 문서들의 덮어쓰기를 통한 문서에도 수정 항목을 반영하고, 해당 단위 정보에도 수정 항목을 반영할 수 있다.

[0029] 도 2는 본 발명의 일실시예에 있어서, 문서 색인 방법을 도시한 흐름도이다. 본 실시예에 따른 문서 색인 방법은 도 1을 통해 설명한 문서 색인 시스템(110)을 통해 수행될 수 있다. 도 2에서는 문서 색인 시스템(110)에서 각각의 단계를 수행하는 과정을 설명함으로써, 문서 색인 방법을 설명한다.

[0030] 문서의 색인을 위해 각각의 문서들은 미리 정의된 항목들로 구성된 데이터로 가공될 필요가 있다. 이러한 문

서의 가공은 문서 색인 시스템(110)의 외부에서 수행되어 가공된 문서가 문서 색인 시스템(110)으로 수신될 수도 있고, 문서 색인 시스템(110)에서 직접 수신된 문서를 가공할 수도 있다. 문서 색인 시스템(110)에서 문서를 가공하는 경우, 문서의 가공은 단위 정보를 추출하는 과정 또는 색인용 문서 컬렉션을 생성하는 과정에서 수행될 수 있다. 가공된 문서에 대해서는 이후 도 4를 통해 더욱 자세히 설명한다.

[0031] 단계(210)에서 문서 색인 시스템(110)은 전체 문서 중 신규 생성된 문서 및 수정된 문서를 포함하는 문서 컬렉션을 수신한다. 이러한 신규 생성된 문서 및 수정된 문서는 문서 색인 시스템(110)에서 발생하는 요청에 따라 문서 컬렉션 생성부(130)에서 검색용 복제 데이터베이스(120)를 통해 생성할 수 있다. 예를 들어, 문서 색인 시스템(110)는 1분마다 신규 생성된 문서와 수정된 문서를 문서 컬렉션 생성부(130)로 요청할 수 있고, 문서 컬렉션 생성부(130)는 1분 동안 신규 생성된 문서와 수정된 문서를 검색용 복제 데이터베이스(120)로부터 수신할 수 있다.

[0032] 단계(220)에서 문서 색인 시스템(110)은 문서 컬렉션에서 각 문서의 저장 위치에 대한 정보를 적어도 포함하는 단위 정보를 추출하고, 추출된 단위 정보 및 문서 컬렉션을 저장한다. 여기서, 단위 정보는 해당 문서가 저장된 파일의 경로, 파일 내에서 시작하는 위치(offset) 및 해당 문서의 길이를 포함할 수 있다.

[0033] 이때, 문서 색인 시스템(110)은 문서 컬렉션을 파일의 형태로 파일 시스템에 저장할 수 있고, 기간마다 수신되는 문서들을 해당 파일에 저장할 수 있다. 또한, 문서 색인 시스템(110)은 단위 정보를 데이터베이스에 저장할 수 있다. 이때, 문서 색인 시스템(110)은 데이터베이스에 저장된 단위 정보를 이용하여 파일 시스템에 저장된 문서 컬렉션의 문서들 중 원하는 문서를 찾을 수 있게 된다.

[0034] 또한, 현재 기간에 수정된 문서 중 적어도 일부의 문서는 이미 이전의 다른 기간에 수정되어 문서 컬렉션에 저장되어 있을 수 있다. 따라서, 문서 색인 시스템(110)은 이미 저장된 문서가 존재하는 경우, 새로 수신된 문서를 저장하고, 해당 단위 정보가 새로 저장된 문서의 위치에 대한 정보를 포함하도록 갱신함으로써, 항상 최신의 문서를 추출할 수 있다.

[0035] 단계(230)에서 문서 색인 시스템(110)은 색인용 문서 컬렉션에 대한 생성 요청에 대응하는 단위 정보를 조회한다. 상술한 바와 같이, 단위 정보는 해당 문서가 어떠한 파일의 어느 위치에 존재하는가를 나타내는 정보를 포함할 수 있다. 따라서, 문서 색인 시스템(110)은 색인용 문서 컬렉션에 대한 생성 요청이 발생하는 경우, 생성 요청의 조건에 부합하는 문서들의 파일 위치를 데이터베이스에서 조회할 수 있다.

[0036] 단계(240)에서 문서 색인 시스템(110)은 단위 정보를 통해 문서 컬렉션에서 추출된 문서를 이용하여 색인용 문서 컬렉션을 생성한다. 즉, 문서 색인 시스템(110)은 단계(230)에서 조회된 단위 정보를 이용하여 문서 컬렉션에서 해당 문서들을 추출할 수 있다.

[0037] 색인용 문서 컬렉션에 대한 생성 요청은 도 1에 도시된 색인부(140)를 통해 수신될 수 있고, 생성된 색인용 문서 컬렉션 역시 색인부(140)로 제공될 수 있다. 즉, 색인부(140)는 신규 생성되거나 수정된 문서들에 대한 생성된 색인용 문서 컬렉션과 기존에 생성된 색인용 문서 컬렉션을 검색 색인에 이용할 수 있다. 따라서, 전체 문서에 대한 색인용 문서 컬렉션을 새로 생성할 필요 없이 신규 생성되거나 수정된 문서들에 대해서만 색인용 문서 컬렉션을 생성할 수 있다.

[0038] 또한, 단위 정보는 해당 문서의 식별자, 해당 문서가 게재된 사이트의 식별자, 해당 문서의 카테고리, 해당 문서의 타입, 해당 문서가 생성된 시간, 해당 문서가 수정된 시간, 해당 문서가 갱신된 시간, 해당 문서의 상태를 정의한 정보 및 해당 문서의 품질을 정의한 정보 중 적어도 하나를 더 포함할 수 있다. 이러한 정보들은 색인부(140)로부터 수신되는 생성 요청의 조건에 관한 것으로, 생성 요청 역시 필요한 문서의 식별자, 필요한 문서가 게재된 사이트의 식별자, 필요한 문서의 카테고리, 필요한 문서의 타입, 필요한 문서가 생성된 시간, 필요한 문서가 수정된 시간, 필요한 문서가 갱신된 시간, 필요한 문서의 상태를 정의한 정보 및 필요한 문서의 품질을 정의한 정보 중 적어도 하나를 포함할 수 있다. 예를 들어, 색인부(140)가 최근 하루 동안 생성 및 수정된 문서가 필요한 경우, 필요한 문서의 생성 및 수정된 시간에 대한 정보를 생성 요청에 포함시켜 문서 색인 시스템(110)으로 전송할 수 있고, 문서 색인 시스템(110)은 단계(230) 및 단계(240)에서 데이터베이스에 저장된 단위 정보를 통해 최근 하루 동안 생성 및 수정된 문서가 문서 컬렉션의 어느 위치에 저장되어 있는지 확인하여 문서 컬렉션에서 해당 문서를 추출할 수 있다.

[0039] 수정된 문서는 전체 문서 각각에 대해 공통적으로 정의된 모든 항목 중 미리 설정된 항목이 기선정된 기간동안 수정된 문서를 포함할 수 있다. 예를 들어 기선정된 기간은 문서 색인 시스템(110)의 문서 컬렉션 요청과 요청 사이의 기간을 포함할 수 있다. 이때, 상술한 바와 같이, 각각의 문서들은 색인을 위한 항목들이 미리

정의될 수 있다. 그러나, 정의된 항목이 많을수록 매 기간마다 수신되는 수정된 문서의 양이 방대해질 수 있다. 예를 들어, 문서 색인 시스템(110)에서 1분 단위로 수정된 문서를 수신할 때, 스크랩 수나 조회 수 등은 매우 빈번하게 변경되기 때문에 이러한 항목들이 수정 항목에 포함되는 경우, 수신되는 수정된 문서의 양이 매우 많아질 수 있다. 따라서, 수정 항목은 별도로 설정될 필요가 있다.

[0040] 그러나 이 경우, 수정 항목에 정의되지 못한 항목들은, 시간이 지남에 따라 원본 데이터와의 오차가 점점 누적될 수 있다. 이러한 문제를 해결하기 위해, 문서 색인 시스템(110)은 기선정된 기간과는 다른 기간마다 상기 미리 설정된 항목을 제외한 나머지 항목들의 추가 수정 사항을 상기 단위 정보 또는 상기 문서 컬렉션에 반영하는 단계를 더 수행할 수 있다. 즉, 문서 색인 시스템(110)은 기본적으로 기설정된 기간과는 별도의 기간(예를 들어, 1시간이나 하루)마다 상술한 일례의 스크랩 수나 조회 수 등 수정 항목으로 설정되지 않은 항목들이 수정된 문서들을 수신할 수 있다. 또한, 문서 색인 시스템(110)은 수신된 문서들을 문서 컬렉션의 해당 문서에 덮어쓰기하여 수정 항목을 반영하거나 또는 단위 정보에 이러한 수정 항목들이 반영되도록 해당 단위 정보를 수정함으로써, 원본 데이터와의 오차가 발생하지 않도록 할 수 있다. 필요에 따라, 수신된 문서들의 덮어쓰기를 통한 문서에도 수정 항목을 반영하고, 해당 단위 정보에도 수정 항목을 반영할 수 있다.

[0041] 도 3은 본 발명의 일실시예에 있어서, 단위 정보의 일례를 나타낸 표이다. 제1 표(310)와 제2 표(320)는 서로 연결된 하나의 표이나 도면의 표현상 두 개의 표로 나누어 표시하였다. 여기서, '포스트 식별자'는 문서의 식별자를, '파일 경로'는 문서가 저장된 파일의 경로를, '블로그 식별자'는 문서가 게재된 사이트의 식별자를, '카테고리 식별자'는 문서의 카테고리를, '블로거 식별자'는 문서가 게재된 사이트 사용자의 식별자를, '문서 타입'은 문서의 타입을, '생성 시간'은 문서가 생성된 시간을, '수정 시간'은 문서가 수정된 시간을, '갱신 시간'은 문서가 문서 컬렉션에 갱신된 시간을 각각 의미할 수 있다. 또한, '오리지널 스코어'는 문서의 상태나 품질을 정의한 정보를 의미할 수 있다.

[0042] 즉, 도 1에 도시된 색인부(140)는 단위 정보에 대응되는 필요한 정보를 생성 요청에 포함시켜 문서 색인 시스템(110)으로 전송함으로써, 문서 색인 시스템(110)으로 하여금 생성 요청의 조건에 해당하는 문서를 추출할 수 있도록 할 수 있다. 예를 들어, 색인부(140)는 '블로그 식별자 132456'을 생성 요청에 포함시켜 전송할 수 있고, 생성 요청을 수신한 문서 색인 시스템(110)은 데이터베이스(112)에서 '블로그 식별자 132456'에 해당하는 문서가 문서 컬렉션의 어느 위치에 저장되어 있는지를 조회하여 해당 문서를 추출할 수 있고, 추출된 문서를 이용하여 색인용 문서 컬렉션을 생성할 수 있다. 생성된 색인용 문서 컬렉션은 색인부(140)로 전송되어 검색 색인에 이용될 수 있다.

[0043] 도 4는 본 발명의 일실시예에 있어서, 색인용 문서 컬렉션의 일례를 나타낸 도면이다. 색인용 문서 컬렉션은 파일의 형태로 데이터를 저장할 수 있고, 복수의 문서로 구성될 수 있다. 이때, 각 문서에 대해서는 색인에 필요한 항목과 사용자 인터페이스 노출을 위해 필요한 항목이 저장될 수 있다. 여기서, 사용자 인터페이스 노출을 위해 필요한 항목은 색인에도 이용될 수 있다. 네모 박스(400)는 색인용 문서 컬렉션에 포함된 하나의 문서에 대한 항목들과 각 항목들에 해당하는 내용들을 도시하고 있다. 예를 들어, 도 1에서 설명한 색인부(140)는 사용자가 검색어를 입력하거나 카테고리를 선택하면, 검색어를 포함하는 문서나 카테고리에 해당하는 문서를 색인용 문서 컬렉션에서 추출하고, 사용자 인터페이스 노출을 위해 필요한 항목의 내용들을 사용자에게 노출시킴으로써, 검색 결과를 제공할 수 있다.

[0044] 도 3과 도 4에서는 블로그의 포스트들을 문서로서 설명하고 있으나, 본 발명의 문서가 포스트로 한정되는 것은 아니다. 즉, 웹 상의 모든 문서에 대해서도 동일하게 적용될 수 있다.

[0045] 이와 같이, 본 발명의 실시예들에 따르면, 수정이 발생하지 않은 문서는 기존에 생성한 색인용 문서 컬렉션을 재사용하고, 신규로 생성된 문서와 수정이 발생한 문서에 대해서만 색인용 문서 컬렉션을 새로 생성할 수 있다. 또한, 신규로 생성된 문서와 수정이 발생한 문서에 대해서만 색인용 문서 컬렉션을 새로 생성하고, 수정이 발생하지 않은 문서에 대해서는 기존에 생성한 색인용 문서 컬렉션을 재사용함으로써, 검색 문서의 색인을 생성하는 시간을 획기적으로 단축하고, 문서들에 대한 데이터를 제공하는 검색용 복제 데이터베이스의 부하도 현저하게 줄일 수 있다.

[0046] 본 발명의 실시 예에 따른 방법들은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다.

[0047] 이상과 같이 본 발명은 비록 한정된 실시예와 도면에 의해 설명되었으나, 본 발명은 상기의 실시예에 한정되는 것은 아니며, 본 발명이 속하는 분야에서 통상의 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형이 가능하다.

[0048] 그러므로, 본 발명의 범위는 설명된 실시예에 국한되어 정해져서는 아니 되며, 후술하는 특허청구범위뿐 아니라 이 특허청구범위와 균등한 것들에 의해 정해져야 한다.

부호의 설명

[0049] 110: 문서 색인 시스템

111: 단위 정보 추출부

112: 데이터베이스

113: 색인용 문서 컬렉션 생성부

114: 추가 수정 사항 반영부

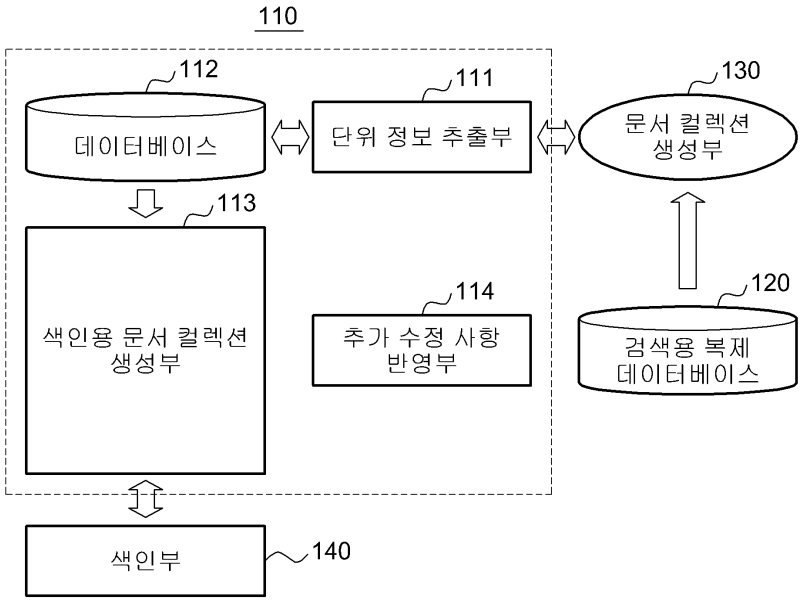
120: 검색용 복제 데이터베이스

130: 문서 컬렉션 생성부

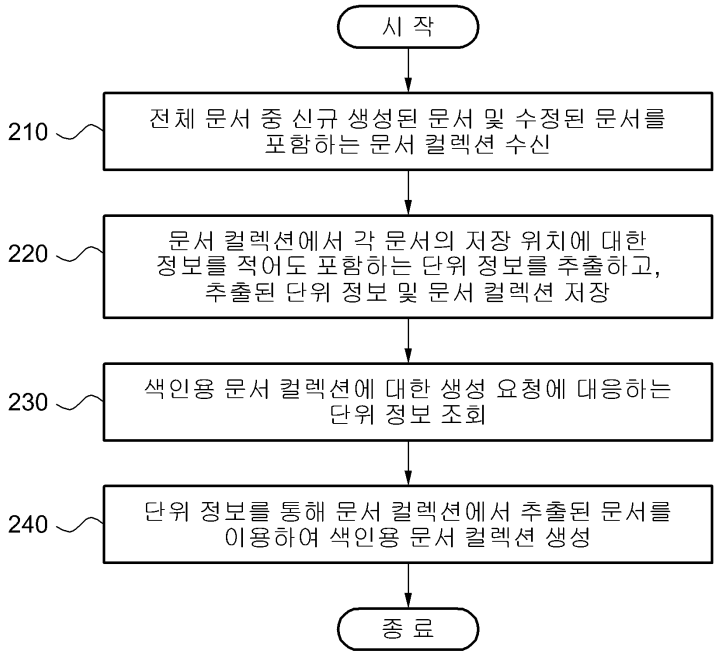
140: 색인부

도면

도면1



도면2



도면3

310

포스트 식별자	파일 경로	블로그 식별자	카테고리 식별자	블로거 식별자	문서 타입
108834	/AA/BB	132456	1	aaa	pri
107268	/AA/BB	159753	13	bbb	pub
...	...	...	...	...	...
108832	/AA/BB	167854	15	ccc	pri

320

생성 시간	수정 시간	갱신 시간	오리지널 스코어
2010...	2010...	2010...	0.404
2010...	2010...	2010...	0.387
...	...	...	...
2010...	2010...	2010...	0.213

도면4

400

<<<postno>>>90009439409
<<<title>>>[본문스크랩] [Emule사용법] 1부, P2P와 당나귀의 간단한 소개
<<<blogno>>>17021260
<<<blogid>>>ehsk830
<<<nickname>>>아우라
<<<blogname>>>도이칠란트으으!!!!!!!!
<<<categoryno>>>18
<<<categoryname>>>외국사이트
<<<addddate>>>20061012000O
<<<scrapcnt>>>0
<<<openmode>>>0
<<<spamscore>>>60
<<<thumbnailurl>>>
<<<score>>>0.603286
<<<bioscore>>>1
<<<gdid>>>90000003_000000000000014F4FB0CB1
<<<review>>>0
<<<feature>>>5167 0 0 0 0 60 1 2 0 0.603286 0 0
<<<domainurl>>>blog.naver.com/ehsk830
<<<contents>>>저는 전문가가 아닙니다. 그저 비교적 숙련된 사용자일 뿐입니다. ^M
그래서, 사용자 입장에서 비교적 쉽게 설명드릴 수 있을 겁니다. ^M

비교적 긴-- 내용이 될듯하여, 몇개로 나누어서 올리려합니다. ^M
1부. P2P와 당나귀의 간단한 소개 (완료) ^M
2부. Emule의 설치 ^M
3부. Emule의 설정 ^M
4부. Emule 사용하기 ^M
5부. 관련 사이트 소개 ^M