(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2016년08월29일
(11) 등록번호 10-1652121
(24) 등록일자 2016년08월23일

(51) 국제특허분류(Int. Cl.)
G06F 17/20 (2006.01) *G06F 17/25* (2006.01)
(21) 출원번호 10-2010-0119604
(22) 출원일자 2010년11월29일
심사청구일자 2015년03월27일
(65) 공개번호 10-2011-0063321
(43) 공개일자 2011년06월10일
(30) 우선권주장
12/629,465 2009년12월02일 미국(US)
(56) 선행기술조사문헌
US06615168 B1*
US20020111792 A1*
US20050108554 A1*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
인터내셔널 비지네스 머신즈 코포레이션
미국 10504 뉴욕주 아몬크 뉴오차드 로드
(72) 발명자
시카, 딥
인도 122001 하리아나 구르가온 섹터 30 내셔널 하이웨이 8 디엘에프 실로케라 인터내셔널 비지네스 머신즈 인디아 엘티디.
(74) 대리인
허정훈

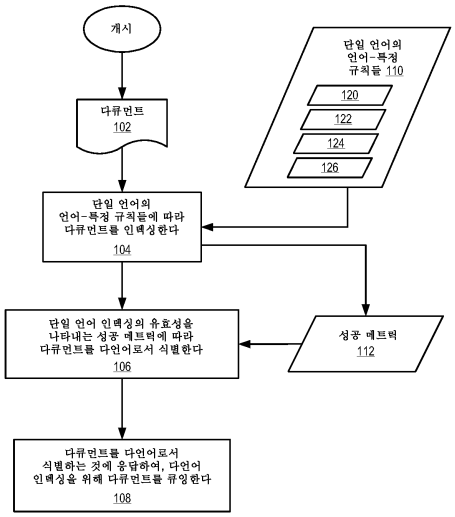
전체 청구항 수 : 총 10 항 심사관 : 홍경아

(54) 발명의 명칭 **다큐먼트 인덱싱**

(57) 요 약

인덱싱될 다큐먼트는 초기에 단일 언어의 언어-특정 규칙들에 따라 인덱싱된다. 성공 메트릭은 단일 언어 인덱싱의 유효성(effectiveness)을 나타내는데 사용된다. 성공 임계 레벨에 도달하지 않으면, 다큐먼트는 다언어로서 식별된다. 다큐먼트를 다언어로서 식별하는 것에 응답해서, 다언어 인덱싱을 위해 다큐먼트가 큐잉된다(queued). 다큐먼트는, 각각 개별적으로 인덱싱되는, 다수의 보다 더 작은 다큐먼트들로 프래그먼트될 수 있다.

대 표 도 - 도1a



명 세 서

청구범위

청구항 1

다큐먼트를 인덱싱하는 컴퓨터 구현 방법에 있어서,

하나 또는 그 이상의 자연어들(one or more natural languages)로 작성된 다큐먼트를 인덱싱하는 단계(indexing) - 상기 인덱싱하는 단계는, 상기 다큐먼트의 콘텐츠에 대하여 하나 또는 그 이상의 자연어들 중 하나의 언어-특정 규칙들(language-specific rules)을 적용하는 단계(applying) 및 상기 언어-특정 규칙들을 따르는 상기 다큐먼트의 콘텐츠로부터 여러 토큰들(a number of tokens)을 생성하는 단계를 포함함 -;

상기 언어-특정 규칙들을 따르는 상기 생성된 여러 토큰들에 기초하여 성공 메트릭(a success metric)을 결정하는 단계(determining);

상기 성공 메트릭을 임계 값(a threshold value)와 비교하는 단계(comparing) - 상기 임계 값은 다언어 다큐먼트(a multi-lingual document)를 지정하는 상기 다큐먼트 내 콘텐츠의 미리 정해진 양을 표시함 -;

상기 비교하는 단계에 기초하여, 상기 다큐먼트를 다언어 다큐먼트로서 식별하는 단계; 및

상기 다큐먼트를 다언어 다큐먼트로서 식별하는 단계에 응답하여, 다언어 인덱싱을 위해 상기 다큐먼트를 큐잉(queuing)하는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 2

제1항에 있어서, 상기 다큐먼트를 복수의 더 작은 다큐먼트들로 프래그먼트하고 상기 더 작은 다큐먼트들을 개별적으로(individually) 따로따로(separately) 인덱싱하는 단계를 더 포함하는 컴퓨터 구현 방법.

청구항 3

제1항에 있어서, 상기 성공 메트릭은 상기 다큐먼트 내 콘텐츠의 백분율을 포함하는, 컴퓨터 구현 방법.

청구항 4

제1항에 있어서, 상기 언어-특정 규칙들은 스테밍(stemming) 규칙들, 음성 태깅 규칙들의 파트, 액센트-민감 규칙들, 동의어 규칙들, 및 정지 단어 규칙들의 그룹 중 하나로부터 선택된 규칙들을 포함하는, 컴퓨터 구현 방법.

청구항 5

제1항에 있어서, 다언어 인덱싱을 위해 상기 다큐먼트를 큐잉하는 상기 단계는 중간 프로세싱을 위해 상기 다큐먼트를 큐잉하는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 6

제1항에 있어서, 다언어 인덱싱을 위해 상기 다큐먼트를 큐잉하는 상기 단계는 피크가 아닌 프로세싱 시간(non-peak processing hours) 동안의 프로세싱을 위해 상기 다큐먼트를 큐잉하는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 7

제1항에 있어서, 상기 다큐먼트를 다언어 다큐먼트로서 식별하는 단계에 응답하여, 상기 다큐먼트의 인덱싱을 역으로 되돌리는(reversing) 단계를 더 포함하는, 컴퓨터 구현 방법.

청구항 8

다큐먼트를 인덱싱하는 컴퓨터 구현 방법을 처리하도록 구성된 컴퓨터 판독가능 프로그램 코드를 저장하는 컴퓨터 판독가능 기억 매체로서, 상기 방법은:

하나 또는 그 이상의 자연어들(one or more natural languages)로 작성된 다큐먼트를 인덱싱하는 단계(indexing) - 상기 인덱싱하는 단계는, 상기 다큐먼트의 콘텐츠에 대하여 하나 또는 그 이상의 자연어들 중 하나의 언어-특정 규칙들(language-specific rules)을 적용하는 단계(applying) 및 상기 언어-특정 규칙들을 따르는 상기 다큐먼트의 콘텐츠로부터 여러 토큰들(a number of tokens)을 생성하는 단계를 포함함 -;

상기 언어-특정 규칙들을 따르는 상기 생성된 여러 토큰들에 기초하여 성공 메트릭(a success metric)을 결정하는 단계(determining);

상기 성공 메트릭을 임계 값(a threshold value)과 비교하는 단계(comparing) - 상기 임계 값은 다언어 다큐먼트(a multi-lingual document)를 지정하는 상기 다큐먼트 내 콘텐츠의 미리 정해진 양을 표시함 -;

상기 비교하는 단계에 기초하여, 상기 다큐먼트를 다언어 다큐먼트로서 식별하는 단계; 및

상기 다큐먼트를 다언어 다큐먼트로서 식별하는 단계에 응답하여, 다언어 인덱싱을 위해 상기 다큐먼트를 큐잉(queuing)하는 단계를 포함하는, 컴퓨터 판독가능 기억 매체.

청구항 9

제8항에 있어서, 상기 방법은 상기 다큐먼트를 복수의 더 작은 다큐먼트들로 프래그먼트하고 상기 더 작은 다큐먼트들을 개별적으로(individually) 따로따로(separately) 인덱싱하는 단계를 더 포함하는, 컴퓨터 판독가능 기억 매체.

청구항 10

다큐먼트를 인덱싱하는 방법을 처리하도록 구성된 시스템에 있어서, 상기 시스템은:

프로세서; 및

상기 프로세서에 동작 가능하게 연결된 컴퓨터 메모리를 포함하고, 상기 방법은:

하나 또는 그 이상의 자연어들(one or more natural languages)로 작성된 다큐먼트를 인덱싱하는 단계(indexing) - 상기 인덱싱하는 단계는 : 상기 다큐먼트의 콘텐츠에 대하여 하나 또는 그 이상의 자연어들 중 하나의 언어-특정 규칙들(language-specific rules)을 적용하는 단계(applying) 및 상기 언어-특정 규칙들을 따르는 상기 다큐먼트의 콘텐츠로부터 여러 토큰들(a number of tokens)을 생성하는 단계를 포함함 -;

상기 언어-특정 규칙들을 따르는 상기 생성된 여러 토큰들에 기초하여 성공 메트릭(a success metric)을 결정하는 단계(determining);

상기 성공 메트릭을 임계 값(a threshold value)과 비교하는 단계(comparing) - 상기 임계 값은 다언어 다큐먼트(a multi-lingual document)를 지정하는 상기 다큐먼트 내 콘텐츠의 미리 정해진 양을 표시함 -;

상기 비교하는 단계에 기초하여, 상기 다큐먼트를 다언어 다큐먼트로서 식별하는 단계; 및

상기 다큐먼트를 다언어 다큐먼트로서 식별하는 단계에 응답하여, 다언어 인덱싱을 위해 상기 다큐먼트를 큐잉(queuing)하는 단계를 포함하는, 시스템.

발명의 설명

배 경 기 술

[0001] 다큐먼트에 대한 액세스는 요즘 급격히 증가했다. 유용한 다큐먼트들의 증가된 수는 원하는 주제에 속한 다큐먼트들을 찾는 것을 더욱 더 어렵게 했다. 이러한 다큐먼트들의 탐색 및 검색은 인덱싱 유용 다큐먼트들에 의해 더 쉽고, 더 빠르고, 더 효율적이게 된다.

[0002] 인덱싱은 적합한 다큐먼트들의 검색을 용이하게 하기 위해 데이터를 저장하는 것을 포함한다. 인덱싱된 다큐먼트들은 이러한 데이터를 획득하도록 파싱 또는 분석될 수 있다. 자연어 프로세싱 또는 다른 언어-민감 분석(other language-sensitive analysis)은 차후 탐색에 유익한 다큐먼트들로부터 정보를 추출할 수 있다. 이러한 프로세스들이 언어 민감 정보를 사용하기 때문에, 프로세싱되는 다큐먼트의 언어는 이러한 분석을 위해 결정되어야만 한다.

발명의 내용

과제의 해결 수단

[0003] 다큐먼트 인덱싱을 위한 방법들, 시스템들, 및 컴퓨터 프로그램 제품들이 본 명세서에 기술된다. 본 발명의 일반적인 실시예들은 단일 언어의 언어-특정 규칙들에 따라 다큐먼트를 인덱싱하는 단계, 단일 언어 인덱싱의 유효성(effectiveness)을 나타내는 성공 메트릭에 따라 다언어(multi-lingual)로서 다큐먼트를 식별하는 단계, 다큐먼트를 다언어로서 식별하는 단계에 응답해서, 다언어 인덱싱을 위해 다큐먼트를 큐잉(queuing)하는 단계를 포함한다. 다른 실시예들은 다큐먼트를 복수의 더 작은 다큐먼트들로 프래그먼트(fragmenting)해서 더 작은 다큐먼트들을 개별적으로 따로따로 인덱싱하는 단계를 포함한다.

[0004] 다른 일반적인 실시예들은 하나 이상의 데이터 프로세싱 시스템을 포함하는 다큐먼트 인덱싱 시스템을 포함한다. 데이터 프로세싱 시스템은 프로세서 및 프로세서에 동작 가능하게 연결된 컴퓨터 메모리를 포함한다. 하나 이상의 시스템들의 컴퓨터 메모리는 그 내부에 상술된 하나 이상의 방법 실시예들을 구현하기 위해 프로세서에서 실행될 컴퓨터 프로그램 명령들을 배치했다. 특히, 컴퓨터 메모리는 그 내부에 배치된 단일 언어의 언어-특정 규칙들에 따라 다큐먼트를 인덱싱하도록 구성된 컴퓨터 판독 가능 프로그램 코드, 단일 언어 인덱싱의 유효성을 나타내는 성공 메트릭에 따라 다큐먼트를 다언어로서 식별하도록 구성된 컴퓨터 판독 가능 프로그램 코드, 다큐먼트를 다언어로서 식별하는 것에 응답해서, 다언어 인덱싱을 위해 다큐먼트를 큐잉하도록 구성된 컴퓨터 판독 가능 프로그램 코드를 가질 수 있다.

[0005] 본 발명의 여타 목적들, 특징들 및 장점들은, 유사한 참조 부호들이 통상 본 발명의 일례의 실시예들의 유사한 부분들을 나타내는 첨부 도면들에 예시된 본 발명의 일례의 실시예들의 이하의 보다 더 상세한 설명으로부터 명백해질 것이다.

도면의 간단한 설명

[0006] 도 1a 내지 도 1c는 본 발명의 실시예들에 따른 방법들을 도시한다.

도 2는 본 발명의 실시예들에 따라 다큐먼트들을 인덱싱하는 컴퓨터의 블록도를 도시한다.

도 3은 본 발명의 실시예들에 따라 다큐먼트들을 인덱싱하는 소프트웨어 아키텍처를 도시한 데이터 흐름도이다.

도 4는 본 발명의 일 실시예에 따라 다큐먼트들을 인덱싱하는 방법을 도시한 플로우챠트이다.

도 5는 본 발명의 실시예들에 따라 다큐먼트들을 인덱싱하는 방법을 도시한 데이터 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0007] 본 발명의 실시예들에 따라 다큐먼트들을 인덱싱하는 예시적인 방법들, 시스템들, 및 디자인 구조들이 첨부 도면을 참조해서 기술된다. 본 명세서에서 사용된 용어는 오직 특정 실시예들을 설명하기 위한 목적으로 본 발명을 제한하려는 의도는 아니다. 본 명세서에서 사용된 바와 같이, 단수형들["하나의(a)", "하나의(an)", "그(the)"]은, 문맥이 명백하게 다른 경우를 나타내지 않는 한, 복수형들 또한 포함한다. 용어들 "포함하다(comprises)" 및/또는 "포함(comprising)"은, 본 명세서에서 사용될 때, 기술된 특징들, 정수들, 단계들, 오퍼레이션들, 소자들, 및/또는 컴포넌트들의 존재를 열거하지만, 하나 이상의 다른 특징들, 정수들, 단계들, 오퍼레이션들, 소자들, 컴포넌트들, 및/또는 그 그룹들의 존재 또는 추가를 배제하지 않음을 또한 알 것이다.

[0008] 이하의 청구항들의 대응 구조들, 컴포넌트들, 동작들, 및 모든 수단들 또는 단계 플러스 기능 소자들의 동등물은 특별히 청구된 다른 청구된 소자들과 협력해서 기능을 실행하는 임의의 구조, 재료 또는 동작을 포함한다. 본 발명의 다양한 실시예들의 기술은 설명 및 기술을 위한 것으로 제시되었지만, 기술된 형태로 본 발명을 속속들이 규명하거나 제한하려는 의도는 아니다. 다수의 변경들 및 변형들은 당업자에게 명백하다. 실시예가 본 발명의 원리들 및 실제 애플리케이션을 가장 잘 설명하기 위해, 또한, 예상되는 특정 용도에 적합한 각종 변경들로 각종 실시예들의 본 발명을 당업자가 이해할 수 있도록 선택 및 기술된 것이다.

[0009] 당업자가 알 수 있는 바와 같이, 본 발명의 양상들은 시스템, 방법 또는 컴퓨터 프로그램 제품으로 구현될 수 있다. 따라서, 본 발명의 양상들은 전체 하드 웨어 구현, 전체 소프트웨어 구현(펌웨어, 상주 소프트웨어, 마이크로-코드 등을 포함함) 또는 모두 일반적으로 본 명세서에서 "회로", "모듈" 또는 "시스템"이라고 할 수 있는 소프트웨어 및 하드웨어 양상들을 조합한 구현의 형태를 취할 수 있다. 또한, 본 발명의 양상들은 그 내부

에 구현된 컴퓨터 판독 가능 프로그램 코드를 갖는 하나 이상의 컴퓨터 판독 가능 매체(들)로 구현된 컴퓨터 프로그램 제품의 형태를 취할 수 있다.

[0010] 하나 이상의 컴퓨터 판독 가능 매체(들)의 임의의 조합이 사용될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터 판독 가능 신호 매체 또는 컴퓨터 판독 가능 기억 매체일 수 있다. 컴퓨터 판독 가능 기억 매체는, 예를 들어, 전자, 자기, 광, 전자기, 적외선, 또는 반도체 시스템, 장치, 또는 디바이스, 또는 상술된 바의 임의의 적합한 조합일 수 있으나, 이들로만 제한되지는 않는다. 컴퓨터 판독 가능 기억 매체의 보다 더 특정적인 일례들(철저하지 않은 리스트)은, 휴대형 컴퓨터 디스켓, 하드 디스크, RAM(random access memory), ROM(read-only memory), EPROM(erasable programmable read-only memory) 또는 플래시 메모리, 휴대형 CD-ROM(compact disc read-only memory), 광 기억 장치, 자기 기억 장치, 또는 상술된 바의 임의의 적합한 조합을 포함한다. 이러한 문맥에서, 컴퓨터 판독 가능 기억 매체는 명령 실행 시스템, 장치, 또는 디바이스에 의해 사용되거나 또는 그와 관련하여 프로그램을 포함 또는 저장할 수 있는 임의의 유형 매체일 수 있다.

[0011] 컴퓨터 판독 가능 기억 매체에서 구현된 프로그램 코드는, 무선, 유선, 광섬유 케이블, RF 등, 또는 상술된 바의 임의의 적합한 조합을 포함하지만 이들로만 제한되지는 않는 임의의 적합한 매체를 사용해서 송신될 수 있다.

[0012] 본 발명의 양상들의 오퍼레이션들을 수행하는 컴퓨터 프로그램 코드는, 자바, 스몰토크, C++ 등의 객체 지향 프로그래밍 언어 및 "C" 프로그래밍 언어 또는 유사한 프로그래밍 언어들 등의 종래의 절차 프로그래밍 언어들을 포함하는 하나 이상의 프로그래밍 언어들의 임의의 조합으로 기록될 수 있다.

[0013] 본 발명의 양상들은 본 발명의 실시예들에 따른 방법들, 장치들(시스템들) 및 컴퓨터 프로그램 제품들의 플로우챠트 및/또는 블록도를 참조해서 후술된다. 플로우챠트 및/또는 블록도의 각 블록, 및 플로우챠트 및/또는 블록도의 블록들의 조합은 컴퓨터 프로그램 명령들에 의해 구현될 수 있음을 알 것이다. 컴퓨터 프로그램 명령들은 범용 컴퓨터의 프로세서, 특별 목적 컴퓨터, 또는 다른 프로그램 가능 데이터 프로세싱 장치에 제공되어서, 컴퓨터의 프로세서 또는 다른 프로그램 가능 데이터 프로세싱 장치를 통해 실행하는 명령들이 플로우챠트 및/또는 블록도 블록 또는 블록들에 기술된 기능/동작을 구현하는 수단을 생성하도록 기계를 생성할 수 있다.

[0014] 이러한 컴퓨터 프로그램 명령들은, 또한, 컴퓨터 판독 가능 매체에 저장된 명령들이 플로우챠트 및/또는 블록도 블록 또는 블록들에 기술된 기능/동작을 구현하는 명령들을 포함하는 제조 물품을 생성하도록, 특정 방식으로 컴퓨터, 다른 프로그램 가능 데이터 프로세싱 장치, 또는 다른 디바이스들이 작용할 수 있게 할 수 있는 컴퓨터 판독 가능 매체에 저장될 수 있다.

[0015] 컴퓨터 프로그램 명령들은, 컴퓨터, 다른 프로그램 가능 장치, 또는 다른 디바이스들에서 실행되는 일련의 동작 단계들이, 컴퓨터 또는 다른 프로그램 가능 장치에서 실행하는 명령들이 플로우챠트 및/또는 블록도 블록 또는 블록들에 기술된 기능/동작을 구현하는 프로세스들을 제공하게 하는 컴퓨터 구현 프로세스를 생성하게 하도록, 컴퓨터, 다른 프로그램 가능 데이터 프로세싱 장치, 또는 다른 디바이스들에 로드될 수 있다.

[0016] 인덱싱 프로세스는 다큐먼트 텍스트를 정보의 탐색 가능한 유닛들로 변환하는 언어-특정 태스크를 포함한다. 도 1a 및 도 1b는 본 발명의 실시예들에 따라 다큐먼트를 인덱싱하는 방법들을 도시한다. 도 1a를 참조하면, 방법은 단일 언어의 언어-특정 규칙들(110)에 따라 다큐먼트(102)를 인덱싱하는 단계[블록(104)], 단일 언어 인덱싱의 유효성을 나타내는 성공 메트릭(112)에 따라 다큐먼트를 다언어로서 식별하는 단계[블록(106)], 다큐먼트를 다언어로서 식별하는 단계에 응답해서, 다언어 인덱싱을 위해 다큐먼트를 지정(designating)하는 단계[블록(108)]를 포함한다.

[0017] 본 명세서에서 사용된 다큐먼트(102)는 인덱싱에 유용한 전자 파일을 나타낸다. 다큐먼트는 다큐먼트에 유용한 임의의 수백개의 상이한 파일 포맷들로 된 임의의 타입의 파일(예를 들어, 스프레드시트, 차트, 프리젠테이션, 및 워드 프로세싱 다큐먼트, 아카이브, 이미지, 텍스트, 웹 페이지 등)일 수 있다. 다큐먼트는 또한, 예를 들어, 압축 파일에 압축 및 통합된 고유 파일 등의 보다 더 큰 다큐먼트 또는 파일의 섹션 또는 일부, 전체로부터 분리된 다큐먼트의 로지컬 섹션(챕터, 서브헤딩 등) 등일 수 있다. 예를 들어, 다큐먼트의 일례는 각종 압축 다큐먼트들로 구성될 수 있는 ZIP 파일 포맷 또는 TAR 파일 포맷으로 된 파일 등의 압축 다큐먼트 아카이브이다. 압축 다큐먼트들 각각은 각종 언어들을 조합한 임의의 언어 또는 다언어 다큐먼트의 단일 언어 다큐먼트일 수 있다.

[0018] 언어-특정 규칙들(110)은 언어 민감 정보를 사용해서 인덱싱을 개선하는 규칙들이다. 언어 민감 정보는 언어 문맥에 따라 다큐먼트(102)로부터 추출될 수 있는 정보이다. 언어 민감 규칙들(110)은 스테밍(stemming)(120),

음성 태깅 파트(122), 액센트-민감 규칙들, 동의어(124), 정지 단어(126), 또는 당업자들이 생각할 수 있는 임의의 다른 언어 종속 규칙들에 속한 규칙들을 포함할 수 있다.

[0019] 언어-특정 규칙들(110)은 단일 언어에 특정된다. 예를 들어, 한 일례에서, 언어-특정 규칙들(110)은 영어에 대한 규칙 세트로 구성될 수 있으며, 다른 일례에서, 언어-특정 규칙들(110)은 프랑스어, 스페인어, 독일어, 일어, 칸나다어, 카슈미르어, 우르두어, 도그리어, 중국어(간체자), 중국어(변체자), 병음, 북경어, 표준 힌디어, 또는 당업자들이 생각할 수 있는 임의의 다른 언어에 대한 규칙 세트로 구성될 수 있다. 언어는 2차 언어(sub-language) 또는 방언을 포함할 수 있다. 일부의 경우에, 언어-특정 규칙들(110)을 포함하는 개별 규칙들 또는 규칙 부집합들은 상이한 언어들에 대한 개별 언어-특정 규칙들(110)에 공통(또는 그들 간에 공유)될 수 있다.

[0020] 단일 언어 인덱싱의 유효성을 나타내는 성공 메트릭(112)은 단일 언어 인덱싱의 성공에 대응하는 임의의 데이터(최소 오버헤드로 획득될 수 있음)일 수 있다. 성공 메트릭(112)은 단일 언어 인덱싱 프로세스로부터 고유하게 생성된 데이터일 수 있다. 대안으로, 성공 메트릭(112)은, 무시할만한 프로세싱 크기(footprint)를 갖는 코드 또는 다른 프로그래밍을 추적 또는 카운팅하는 것을 추가한 후에 단일 언어 인덱싱 프로세스로부터 생성될 수 있다. 도 1b를 참조하면, 성공 메트릭(112)은, 예를 들어, 인덱싱 중에 성공적으로 처리된 다큐먼트의 콘텐츠(130)(예를 들어, 단어, 절, 섹션, 페이지 등)의 백분율을 포함할 수 있다. 성공적으로 처리된 콘텐츠는 토큰이 성공적으로 생성될 수 있는 콘텐츠이거나, 또는 당업자가 생각할 수 있는 성공적인 프로세싱의 임의의 다른 측정치이다.

[0021] 도 1b를 다시 참조하면, 성공 메트릭(112)에 따라 다큐먼트를 다언어로서 식별하는 단계[블록(106)]는 성공적으로 처리된 콘텐츠(130)의 백분율이 임계 백분율 값(132)을 만족시키는 데 실패함을 결정[블록(134)]함으로써 수행될 수 있다. 성공적으로 처리된 단어들의 임시 합계는 보유되어 성공적으로 처리된 단어들의 백분율을 획득하는데 사용될 수 있다. 임계 백분율 값(132)은 다언어로서 표시된 다큐먼트들의 수를 정련(refine)하도록 구성될 수 있다. 예를 들어, 임계 백분율 값(132)은 80%로 설정될 수 있다. 성공적으로 처리된 단어들의 백분율이 임계 백분율 값을 초과하는데 실패하면(138), 방법은 블록(108)으로 진행한다. 성공적으로 처리된 단어들의 백분율이 임계 백분율 값을 초과하면, 다큐먼트는 탐색 준비가 되며, 시스템은 다큐먼트에 대한 다른 프로세싱 리소스들을 사용하지 않는다. 예를 들어, 80% 미만의 단어들이 성공적으로 처리되면, 다큐먼트는 다언어로서 식별된다. 다른 실시예들에서, 추가의 또는 다른 통계 측정치가 다큐먼트를 다언어로 식별하기 위해 계산될 수 있다.

[0022] 도 1b를 다시 참조해서, 성공적으로 처리된 단어들의 백분율이 임계 백분율 값을 초과하는데 실패하면(138), 시스템은 다언어 인덱싱을 위해 다큐먼트(102)를 큐잉한다[블록(108)]. 다언어 인덱싱을 위해 다큐먼트를 큐잉하는 단계[블록(108)]는 다언어 인덱싱을 위한 개별 저장소에 다큐먼트(102)를 저장하는 단계 또는 개별 시스템 또는 저장소에 다큐먼트를 송신하는 단계를 포함할 수 있다. 시스템은, 프로세싱 로드 큐의 경우에서와 같이, 중간 프로세싱을 위해 다큐먼트(102)를 큐잉할 수 있다[블록(140)]. 다른 실시예들에서, 다큐먼트는 차후 인덱싱을 위해 큐잉될 수 있다. 예를 들어, 다언어 인덱싱은 더 프로세싱 집약적이기 때문에, 다큐먼트는 피크가 아닌 프로세싱 시간에 다언어 인덱싱을 위해 큐잉될 수 있다[블록(142)]. 일부 구현들에서, 큐는 FIFO(First-In-First-Out) 데이터 구조일 수 있다. 다른 구현들에서, 큐는 LIFO(Last-In-First-Out) 데이터 구조일 수 있으며, 또는 큐의 다큐먼트들이 각종 우선순위 시스템들을 사용해서 우선 순위에 따라 디큐잉될 수 있다.

[0023] 도 1c는 본 발명의 다른 실시예에 따라 다큐먼트를 인덱싱하는 방법을 도시한다. 도 1c를 참조하면, 방법은 단일 언어의 언어-특정 규칙들에 따른 다큐먼트의 인덱싱의 유효성을 나타내는 성공 메트릭(112)에 따라 다언어로서 다큐먼트(102)를 식별하는 단계[블록(106)], 및 다큐먼트를 다언어로서 식별하는 단계에 응답해서, 다언어 인덱싱을 위해 다큐먼트(102)를 큐잉하는 단계[블록(108)]를 포함한다. 도 1c의 방법은 도 1a와 유사하게 수행되지만, 개별적으로 수행되는 단일 언어 인덱싱 단계가 앞선다. 도 1c의 방법은 단일 언어 인덱싱 단계가 개시된 후에 실행된다.

[0024] 본 발명의 실시예들은 후술되는 컴퓨터 구현 방법들을 포함한다. 일부 실시예들에서, 이러한 방법들은 시스템의 하나의 장치 또는 컴퓨터에서 전체가 수행될 수 있다. 대안으로, 방법들의 일부가, 하나 이상의 LAN(local area network), WAN(wide area network), 유선 또는 셀룰러 폰 네트워크, 인트라넷, 또는 인터넷 등의 네트워크에 의해 연결된 둘 이상의 컴퓨터들에서 수행될 수 있다. 본 명세서에 기술된 방법 요소들의 순서는 요소들이 실행될 수 있는 순서를 제한하는 것이 아니다.

[0025] 도 2는 본 발명의 실시예들에서 사용되는 컴퓨터의 블록도를 도시한다. 컴퓨터(202), 휘발성 RAM(random

access memory)(204) 및 하드 디스크 드라이브, 광 디스크 드라이브, 또는 전기적 소거가능 프로그램가능 판독 전용 메모리 스페이스('EEPROM' 또는 '플래시' 메모리로도 공지됨) 등의 비휘발성 컴퓨터 메모리(250)의 일부 형태 또는 형태들을 포함하는, 컴퓨터 메모리 뿐만 아니라 적어도 하나의 컴퓨터 프로세서(254)를 포함한다. 컴퓨터 메모리는 시스템 버스(240)를 통해 프로세서(254) 및 다른 시스템 컴포넌트들에 연결된다. 따라서, 소프트웨어 모듈들은 컴퓨터 메모리에 저장된 프로그램 명령들이다.

[0026] 오퍼레이팅 시스템(210)은 컴퓨터 메모리에 저장된다. 오퍼레이팅 시스템(210)은 윈도우 오퍼레이팅 시스템, Mac OS X, UNIX, LINUX, 또는 IBM(International Business Machines Corporation)(아몬크, 뉴욕주)의 AIX 등의 임의의 적합한 오퍼레이팅 시스템일 수 있다. 네트워크 스택(212)이 또한 메모리에 저장된다. 네트워크 스택(212)은 네트워크 통신을 용이하게 하는 협동 컴퓨터 네트워킹 프로토콜들의 소프트웨어 구현이다.

[0027] 컴퓨터(202)는 또한 하나 이상의 입력/출력 인터페이스 어댑터들(256)을 포함한다. 입력/출력 인터페이스 어댑터들(256)은 컴퓨터 디스플레이 스크린 등의 출력 디바이스들(272)로의 출력, 및 키보드, 마우스 등의 입력 디바이스들(270)로부터의 사용자 입력을 제어하기 위해 소프트웨어 드라이버들 및 컴퓨터 하드웨어를 통해 사용자 지향 입력/출력을 구현할 수 있다.

[0028] 컴퓨터(202)는 또한 다른 디바이스들(260)과의 데이터 통신을 구현하기 위한 통신 어댑터(252)를 포함한다. 통신 어댑터(252)는, 하나의 컴퓨터가 네트워크를 통해 다른 컴퓨터에 데이터 통신을 송신하는 데이터 통신의 하드웨어 레벨을 구현한다.

[0029] 인덱싱 모듈(206)이 또한 컴퓨터 메모리에 저장된다. 인덱싱 모듈(206)은 본 발명의 실시예들에 따라 다큐먼트를 인덱싱하는 컴퓨터 프로그램 명령들을 포함한다. 인덱싱 모듈(206)은 개별 소프트웨어 계층들 또는 또는 동일한 계층에서 동작하는 하나 이상의 서브-모듈들로서 구현될 수 있다. 오퍼레이팅 시스템(210)으로부터의 개별 모듈로서 도시되었지만, 인덱싱 모듈(206) 또는 하나 이상의 서브-모듈들은 오퍼레이팅 시스템(210)의 파트로서 통합될 수 있다. 각종 실시예들에서, 인덱싱 모듈(206)은 소프트웨어 스택 또는 펌웨어로 구현될 수 있다.

[0030] 도 2에 도시된 컴퓨터(202)는 제한이 아닌 설명을 위해 제공된 것이다. 본 발명의 실시예들은, 로직 및 메모리를 포함하는 임의의 실행 가능한 컴퓨팅 디바이스, 또는 FPGA(field-programmable gate array), ASIC(application-specific integrated circuit) 등으로서 로직이 구현되는 디바이스들을 포함해서, 당업자가 생각할 수 있는, 실행되는 컴퓨터 프로그램 명령들을 포함하는 소프트웨어 모듈들로서 구현될 수 있다.

[0031] 추가적인 설명을 위해, 도 3은 다큐먼트들을 인덱싱하는 일례의 소프트웨어 아키텍처를 설명한 데이터 흐름도를 도시한다. 도 3의 소프트웨어 아키텍처는, 다큐먼트 헤더 파서(304), 다큐먼트 프래그멘터(310), 언어 식별자(318), 인덱서(306), 및 인덱스 저장소(308)를 포함하는 각종 소프트웨어 모듈들을 포함한다. 다큐먼트 헤더 파서(304)는 다큐먼트(302)의 파일 포맷을 식별하고, 다큐먼트(302)의 다큐먼트 헤더를 판독 및 파싱한다. 다큐먼트 헤더 파서(304)는 또한, 언어-특정 정보가 존재하는 경우, 다큐먼트 헤더에 포함된 언어-특정 정보를 판독한다. 다큐먼트 프래그멘터(310)는 다큐먼트 프래그먼트들에 대해 다큐먼트 헤더에 주어진 정보를 기반으로 상이한 프래그먼트들 또는 섹션들로 다큐먼트를 분할한다. 다큐먼트 프래그멘터(310)는 다수의 보다 더 작은 다큐먼트들(312-316)의 집합을 출력으로서 제공한다. 언어 식별자(318)는 다수의 다큐먼트들(312-316)의 집합의 각각의 다큐먼트에 대해 반복되고 각각에 대해 주요 언어를 식별한다. 인덱서(306)는 문자 세트 인코딩, 변환, 토큰화, 및 인덱싱 정보 생성을 실행한다. 인덱서(306)는 또한 언어 정보를 사용해서 특정 언어의 규칙에 따라 언어 민감 인덱싱 정보를 생성한다. 토큰화를 맡고 있는 인덱서(306) 컴포넌트는 성공적으로 처리되지 않은 토큰들을 카운트하도록 구성될 수 있다. 성공적으로 처리되지 않은 토큰들의 카운트는 사실상 추가의 프로세싱 전력을 요구하지 않지만, 성공적으로 처리된 단어들의 백분율을 결정하는데 사용될 수 있다. 인덱스 저장소(308)는 탐색에 사용될 수 있는 인덱싱 정보를 위한 저장소이다. 인덱스 저장소(308)는 포함하는 다큐먼트들[예를 들어, 다큐먼트(302)]에 대한 포인터들 및 파싱된 단어들의 리스트를 포함한다. 인덱스 저장소(308)는 또한 언어 인식 특징들을 지원하기 위해 언어-와이즈(language-wise) 인덱스 데이터를 포함한다.

[0032] 도면들의 플로우챠트 및 블록도들은 본 발명의 각종 실시예들에 따른 시스템들, 방법들 및 컴퓨터 프로그램 제품들의 가능한 구현들의 아키텍처, 기능, 및 오퍼레이션을 도시한다. 이와 관련하여, 플로우챠트 또는 블록도의 각 블록은, 특정 로지컬 기능(들)을 구현하는 하나 이상의 실행가능 명령들을 포함하는 모듈, 세그먼트, 코드의 일부를 나타낼 수 있다. 일부 다른 구현들에서, 블록에 기술된 기능들이 도면들에 기술된 순서와 다르게 발생할 수 있음을 주지해야만 한다. 예를 들어, 포함된 기능에 따라서, 연속해서 도시된 두개의 블록들이 실제로는 거의 동시에 실행될 수 있으며, 블록들이 때때로는 역 순서로 실행될 수 있다. 블록도들 및/또는 플로우

챠트의 각 블록, 및 블록도들 및/또는 플로우챠트의 블록들의 조합들은 특정 기능 또는 동작, 또는 특별 목적 하드웨어 및 컴퓨터 명령들의 조합을 실행하는 특별 목적 하드웨어-기반 시스템들에 의해 구현될 수 있음을 주지해야 한다.

[0033] 추가적인 설명을 위해, 도 4는 본 발명의 일 실시예에 따라 다큐먼트들을 인덱싱하는 방법을 설명한 플로우챠트를 도시한다. 본 명세서에 기술된 방법 요소들의 순서는 요소들이 실행될 수 있는 순서를 제한하는 것이 아니다. 방법은 컴퓨터 시스템(예를 들어, 인덱싱 서버)이 다큐먼트 타입을 식별하고 대응 필터를 로드하는 단계[블록(402)]로 시작된다. 다큐먼트 타입은, 예를 들어, 마이크로소프트 워드 다큐먼트, 마이크로소프트 엑셀 다큐먼트, HTML(Hyper Text Markup Language), PDF(Portable Document Format) 등의 파일 포맷을 포함할 수 있다. 인덱싱에 유용한 수백개의 상이한 파일 포맷들이 있으며, 상기 포맷들 각각은 다큐먼트에 포함된 텍스트에 대한 언어-관련 메타데이터를 포함할 수 있다. HTML 및 XML(Extensible Markup Language) 등의 다큐먼트 포맷들은 텍스트 언어를 식별하기 위해 언어 태그를 포함할 수 있다.

[0034] 다큐먼트 타입에 적합한 다큐먼트 필터를 호출(invoking)한 후에, 시스템은 다큐먼트의 주요 언어를 식별한다[블록(404)]. 다큐먼트 타입을 식별하는 단계는 언어-관련 메타데이터를 검출 및 파싱함으로써 수행될 수 있다. 다큐먼트 헤더 표준들이 존재하지만, 다수의 다큐먼트 파일 포맷들은 이러한 표준을 따르지 않아서 언어 정보를 포함하지 않는다. 언어 관련 데이터가 유용하지 않으면, 언어 식별자(318)는 인덱싱 프로세스들을 사용해서 소스 다큐먼트의 하나의 특정 언어를 예측할 수 있다. 언어 식별자는 텍스트의 샘플링 또는 다수를 기반으로 주요 언어를 예측할 수 있다. 시스템은 주요 언어의 언어 규칙들을 전체 다큐먼트에 적용할 수 있다. 압축된 다큐먼트 아카이브들의 경우, 제1 다큐먼트에 대해 식별된 언어가 인덱싱될 남은 모든 다큐먼트들에 대해 고려될 수 있다. 또는, 한 다큐먼트에 대해, 제1 n개의 바이트들이 샘플링되어 주요 언어를 식별할 수 있다. 일단 주요 언어가 식별되면, 인덱싱은 개시될 수 있다.

[0035] 인덱싱 프로세스는, 스템 인덱스 뿐만 아니라, 인덱싱된 모든 다큐먼트들의 위치 및 다큐먼트들의 텍스트 내에 포함된 모든 단어들의 리스트를 저장하는 콜렉션(collections), 또는 유니버설 인덱스들을 생성한다. 도 4의 방법은, 인덱서(306)가 다큐먼트 메타데이터 및 퓨어-텍스트를 추출하는 단계[블록(406)]로 이어진다. 인덱서는 주요 언어의 언어-특정 규칙들에 따라 다큐먼트를 단어들 또는 토큰들로 분할하고[블록(408)], 다큐먼트의 각 토큰의 발생을 결정해서 단어 인덱스를 생성한다[블록(410)]. 다큐먼트가 처리됨에 따라, 성공적으로 처리된 다큐먼트 콘텐츠(또는, 역으로 성공적으로 처리되지 않은 다큐먼트 콘텐츠)가 추적된다. 단어 인덱스는 탐색에 사용되는 인덱스 데이터를 포함하는 저장소(412)에 저장될 수 있다. 인덱서(306)는 주요 언어의 언어-특정 규칙들에 따라 언어-특정 인덱스 데이터를 생성한다[블록(414)]. 인덱스 데이터는 저장소(412)에 저장될 수 있다.

[0036] 본 방법은 성공적으로 처리된 단어들의 백분율이 임계 백분율 값 보다 작은지를 결정하는 단계[블록(416)]를 포함한다. 임계 백분율 값은, 적합한 언어의 특정 언어 규칙들에 따라 인덱싱할 때 처리될 수 있는 단어들의 최소 또는 명목상의 백분율을 나타내는 구성 가능한 값이다. 성공적으로 처리된 단어들의 백분율이 임계 백분율 값을 초과하면(420), 다큐먼트는 탐색 준비가 되고, 시스템은 다큐먼트에 대한 다른 프로세싱 리소스들을 사용하지 않는다. 성공적으로 처리된 단어들의 백분율이 임계 백분율 값을 초과하지 않으면(418), 시스템은 다큐먼트를 다언어로서 지정한다[블록(422)]. 또한, 시스템은 단일 언어 인덱싱을 원래대로 되돌릴 수 있으며(즉, 역으로 하거나 취소할 수 있으며)[블록(424)], 다언어 프로세싱으로 다큐먼트를 인덱싱한다[블록(426)]. 단일 언어 인덱싱의 원래대로 되돌리기는 저장소(412)로부터 단어 인덱스 및 언어-특정 인덱스 데이터를 제거함으로써 수행될 수 있다.

[0037] 도 5는 본 발명의 일 실시예에 따라 다큐먼트들을 인덱싱하는 방법을 도시한 데이터 흐름도이다. 도 5의 방법은 도 1a와 유사하게 수행되지만, 다큐먼트(102)를 복수의 보다 더 작은 다큐먼트들(504)로 프래그먼트하는 단계[블록(502)]를 더 포함한다. 프래그먼테이션 후에, 인덱서는 해당 다큐먼트에 대한 단일 언어의 언어-특정 규칙들(506)에 따라 보다 더 작은 다큐먼트들(504)의 각각의 다큐먼트를 개별적으로 인덱싱한다[블록(508)]. 시스템은, 그 후, 계속해서, 상세히 상술된 바와 같이, 단일 언어의 언어-특정 규칙들에 따른 다큐먼트의 인덱싱의 유효성을 나타내는 성공 메트릭(112)에 따라 다언어로서 다큐먼트(102)를 식별하고[블록(106)], 다큐먼트를 다언어로서 식별하는 것에 응답해서, 다언어 인덱싱을 위해 다큐먼트(102)를 큐잉한다[블록(108)].

[0038] 본 명세서에 기술된 본 발명의 개념들은 다양하게 변경될 수 있음을 알 것이다. 이러한 변경은 토큰화, 스테밍, 언어 인식, 큐잉 등을 위한 각종 방법들, 시스템들, 및 프로그램들을 포함할 수 있다. 상기 변경들이 첨부된 청구항들 및 그 동등물의 범위 내에 속하는 정도에서, 본 특허에 의해 포함된다.

부호의 설명

[0039] 102 : 다큐먼트

104 : 단일 언어의 언어-특정 규칙들에 따라 다큐먼트를 인덱싱한다

106 : 단일 언어 인덱싱의 유효성을 나타내는 성공 메트릭에 따라 다큐먼트를 다언어로서 식별한다

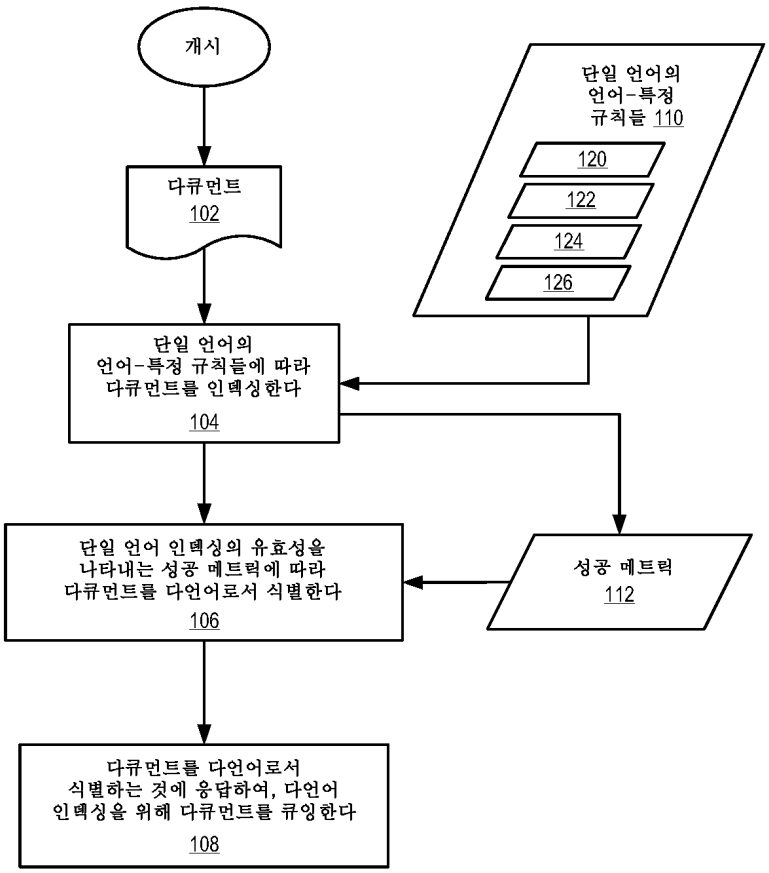
108 : 다큐먼트를 다언어로서 식별하는 것에 응답해서, 다언어 인덱싱을 위해 다큐먼트를 큐잉한다

110 : 단일 언어의 언어-특정 규칙들

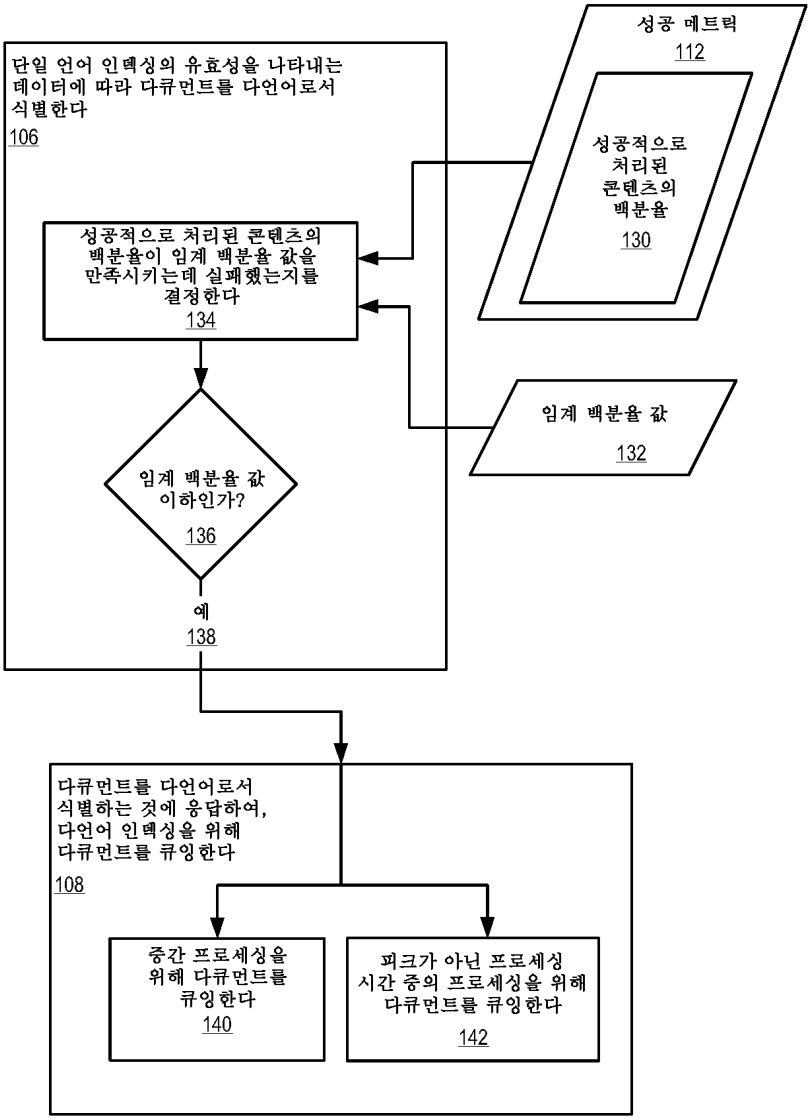
112 : 성공 메트릭

도면

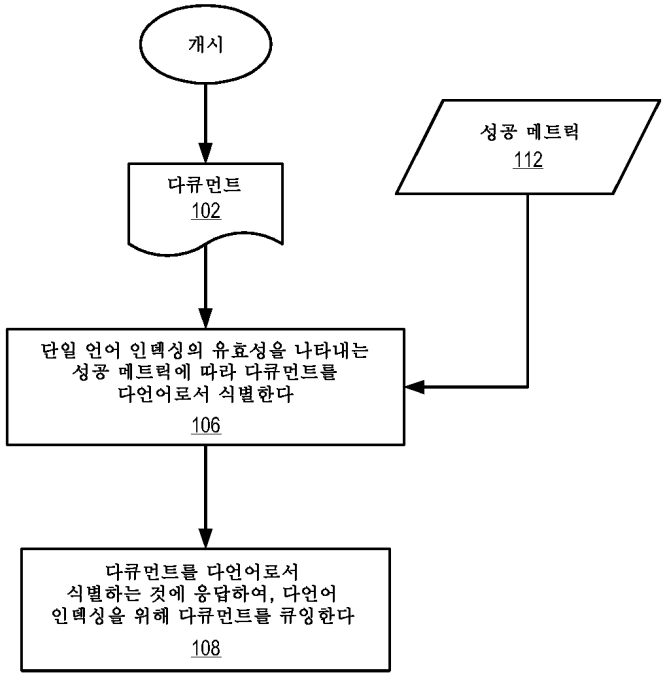
도면1a



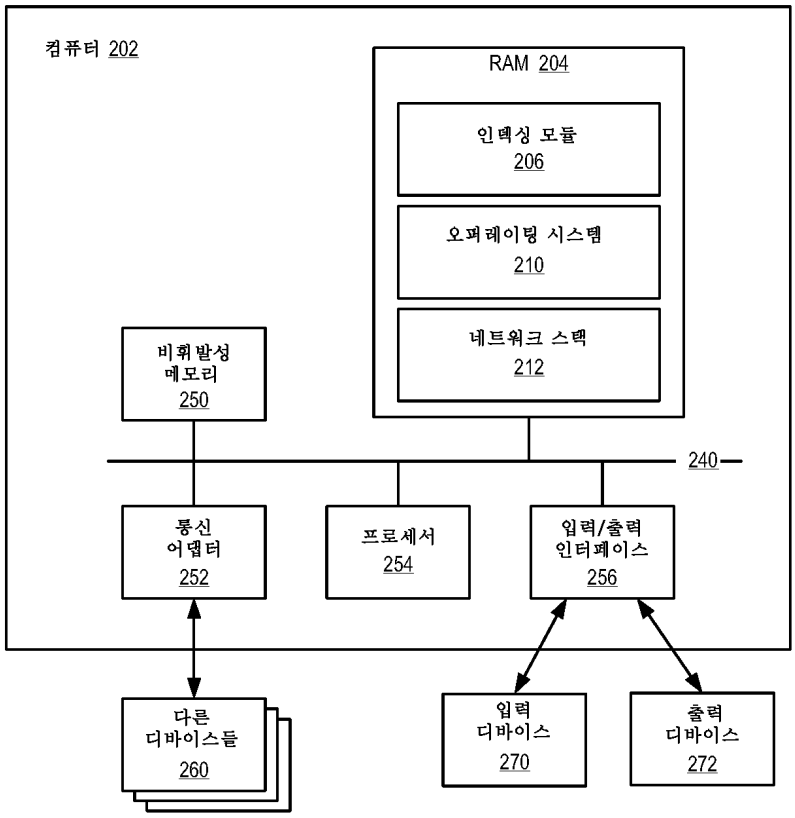
도면1b


단일 언어 인덱싱의 유효성을 나타내는 데이터에 따라 다큐먼트를 다언어로서 식별한다
106
성공 메트릭
112
성공적으로 처리된 콘텐츠의 백분율
130
성공적으로 처리된 콘텐츠의 백분율이 임계 백분율 값을 만족시키는데 실패했는지를 결정한다
134
임계 백분율 값
132
임계 백분율 값 이하인가?
136
예
138
다큐먼트를 다언어로서 식별하는 것에 응답하여, 다언어 인덱싱을 위해 다큐먼트를 큐잉한다
108
중간 프로세싱을 위해 다큐먼트를 큐잉한다
140
피크가 아닌 프로세싱 시간 중의 프로세싱을 위해 다큐먼트를 큐잉한다
142

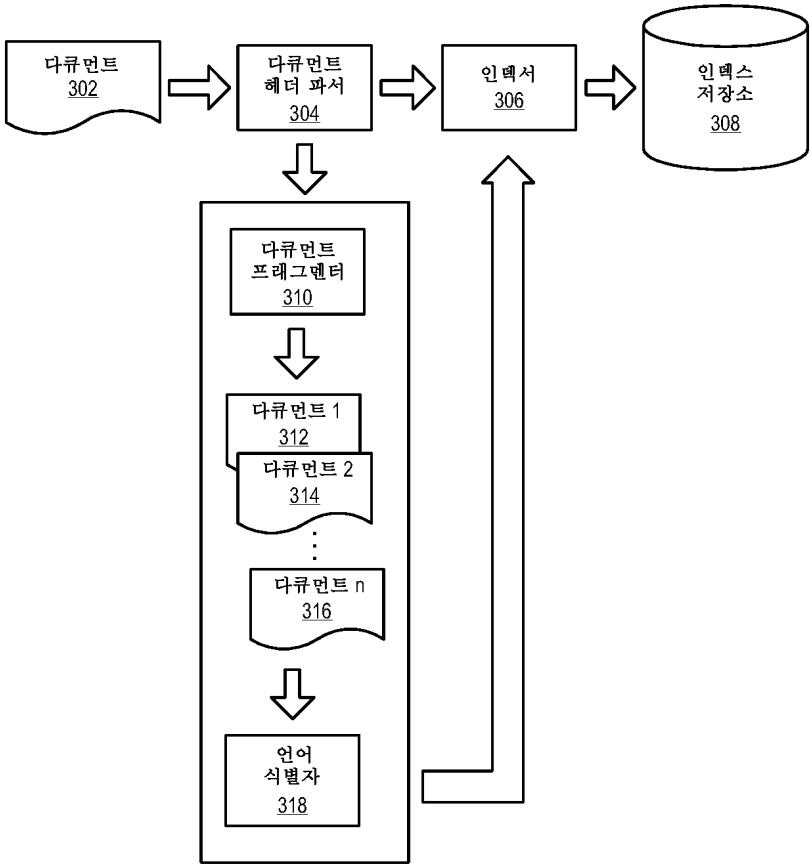
도면1c


개시
다큐먼트
102
성공 메트릭
112
단일 언어 인덱싱의 유효성을 나타내는
성공 메트릭에 따라 다큐먼트를
다언어로서 식별한다
106
다큐먼트를 다언어로서
식별하는 것에 응답하여, 다언어
인덱싱을 위해 다큐먼트를 큐잉한다
108

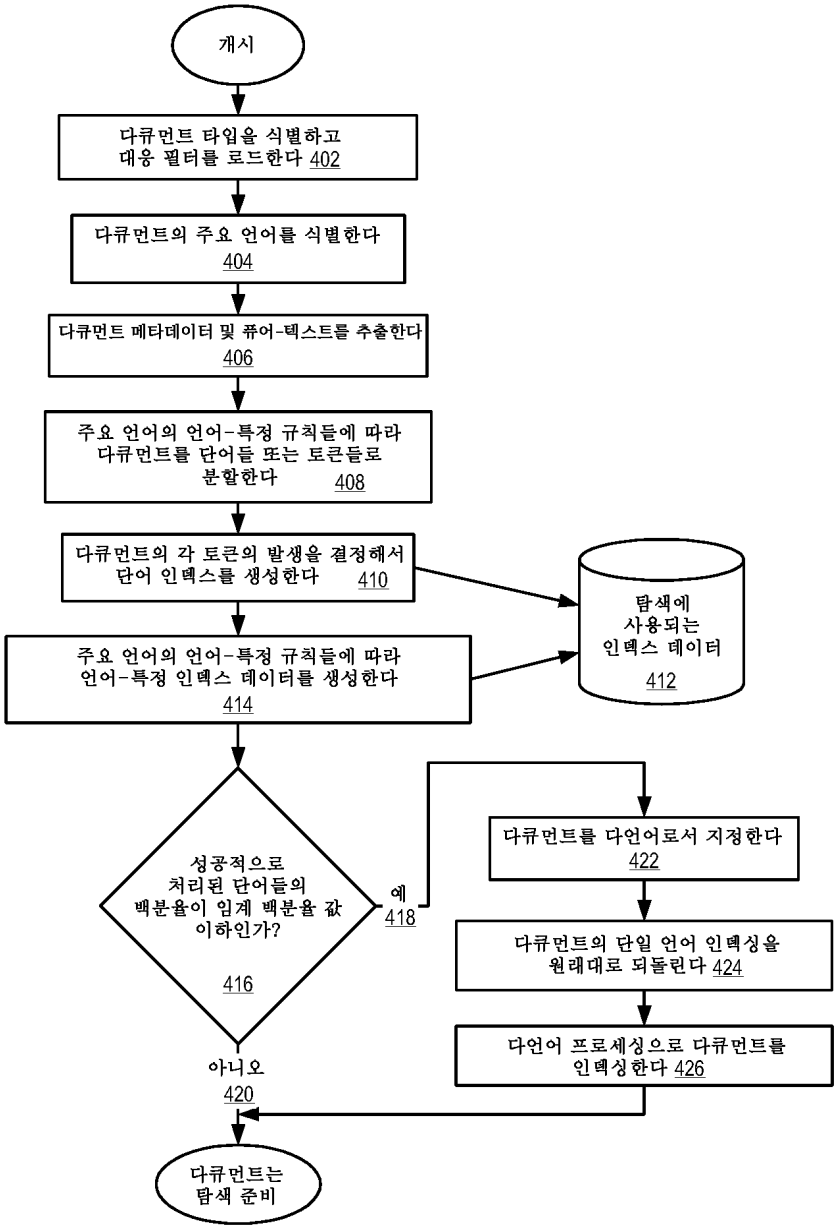
도면2


컴퓨터 202
RAM 204
인덱싱 모듈
206
오퍼레이팅 시스템
210
네트워크 스택
212
비휘발성
메모리
250
240
통신
어댑터
252
프로세서
254
입력/출력
인터페이스
256
다른
디바이스들
260
입력
디바이스
270
출력
디바이스
272

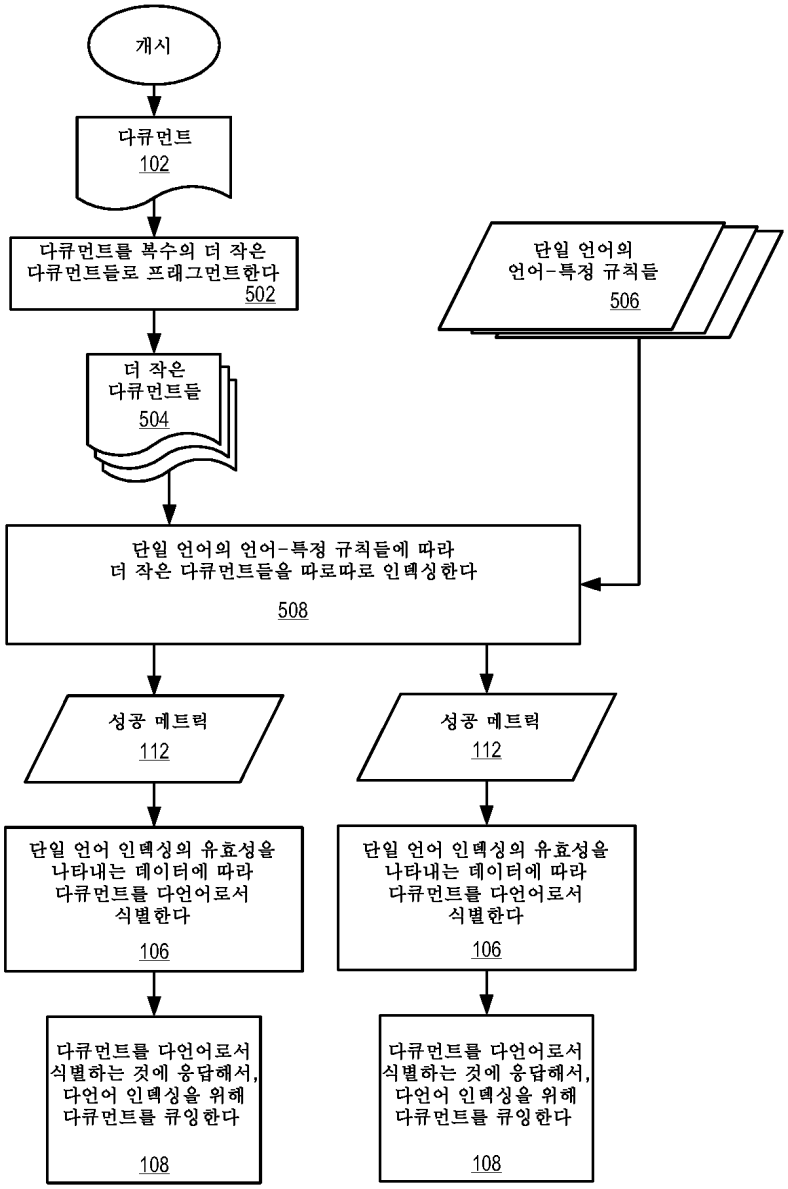
도면3


다큐먼트
302
다큐먼트
헤더 파서
304
인덱서
306
인덱스
저장소
308
다큐먼트
프래그멘터
310
다큐먼트 1
312
다큐먼트 2
314
다큐먼트 n
316
언어
식별자
318

도면4


개시
다큐먼트 타입을 식별하고 대응 필터를 로드한다 402
다큐먼트의 주요 언어를 식별한다 404
다큐먼트 메타데이터 및 퓨어-텍스트를 추출한다 406
주요 언어의 언어-특정 규칙들에 따라 다큐먼트를 단어들 또는 토큰들로 분할한다 408
다큐먼트의 각 토큰의 발생을 결정해서 단어 인덱스를 생성한다 410
탐색에 사용되는 인덱스 데이터 412
주요 언어의 언어-특정 규칙들에 따라 언어-특정 인덱스 데이터를 생성한다 414
성공적으로 처리된 단어들의 백분율이 임계 백분율 값 이하인가? 416
예 418
다큐먼트를 다언어로서 지정한다 422
다큐먼트의 단일 언어 인덱싱을 원래대로 되돌린다 424
다언어 프로세싱으로 다큐먼트를 인덱싱한다 426
아니오 420
다큐먼트는 탐색 준비

도면5


개시
다큐먼트
102
다큐먼트를 복수의 더 작은 다큐먼트들로 프래그먼트한다
502
더 작은 다큐먼트들
504
단일 언어의 언어-특정 규칙들
506
단일 언어의 언어-특정 규칙들에 따라 더 작은 다큐먼트들을 따로따로 인덱싱한다
508
성공 메트릭
112
성공 메트릭
112
단일 언어 인덱싱의 유효성을 나타내는 데이터에 따라 다큐먼트를 다언어로서 식별한다
106
단일 언어 인덱싱의 유효성을 나타내는 데이터에 따라 다큐먼트를 다언어로서 식별한다
106
다큐먼트를 다언어로서 식별하는 것에 응답해서, 다언어 인덱싱을 위해 다큐먼트를 큐잉한다
108
다큐먼트를 다언어로서 식별하는 것에 응답해서, 다언어 인덱싱을 위해 다큐먼트를 큐잉한다
108