



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0157071
(43) 공개일자 2024년10월31일

(51) 국제특허분류(Int. Cl.)
G06V 30/414 (2022.01) G06V 30/148 (2022.01)
G06V 30/18 (2022.01) G06V 30/412 (2022.01)
(52) CPC특허분류
G06V 30/414 (2022.01)
G06V 30/153 (2022.01)
(21) 출원번호 10-2024-7032812
(22) 출원일자(국제) 2023년02월27일
심사청구일자 없음
(85) 번역문제출일자 2024년09월30일
(86) 국제출원번호 PCT/US2023/013970
(87) 국제공개번호 WO 2023/167824
국제공개일자 2023년09월07일
(30) 우선권주장
17/685,328 2022년03월02일 미국(US)

(71) 출원인
알테릭스 인코포레이티드
미국 92618 캘리포니아 어바인 라구나 캐니언 로
드 17200
(72) 발명자
라드 채드 디노
미국 92612 캘리포니아주 어바인 미켈슨 드라이브
3345 스위트 400 알테릭스 인코포레이티드 씨/오
블랙 아담
미국 92612 캘리포니아주 어바인 미켈슨 드라이브
3345 스위트 400 알테릭스 인코포레이티드 씨/오
(74) 대리인
특허법인코리아나

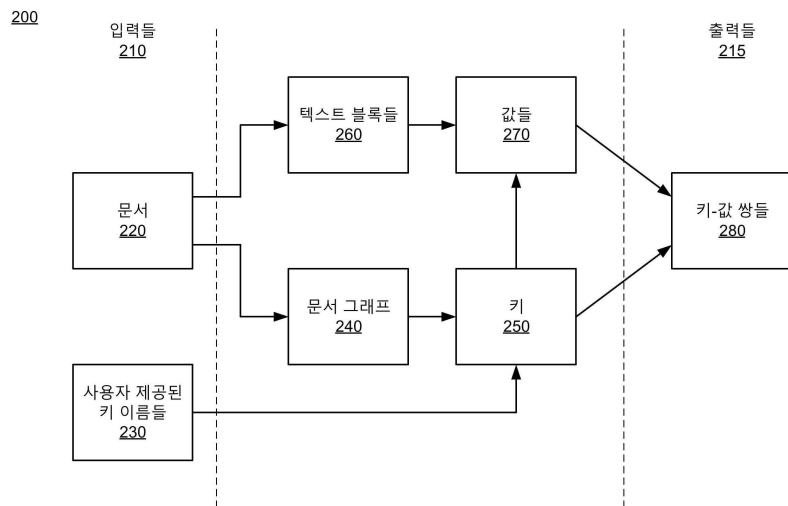
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 자동화된 키-값 쌍 추출

(57) 요약

분석될 문서 및 문서로부터 추출될 키 이름들의 세트가 수신된다. 문서 내에 포함된 문자들의 스트링들의 세트 및 문자의 각각의 스트링에 대한 위치가 식별된다. 또한, 문서에 대한 문서 그래프가 생성된다. 문서 그래프는 노드들의 세트 및 에지들의 세트를 포함한다. 노드들의 세트의 각각의 노드는 문자들의 스트링들의 세트의 문자들의 스트링에 대응한다. 에지들의 세트의 각각의 에지는 둘 이상의 노드들을 함께 연결한다. 또한, 문서 그래프 및 수신된 키 이름들의 세트에 기초하여, 키들의 세트가 식별된다. 또한, 문서로부터 값들의 세트가 추출되고, 식별된 키들의 세트 및 추출된 값들의 세트에 기초하여 키-값 쌍들의 세트가 생성된다.

대표도



(52) CPC특허분류

G06V 30/18 (2023.08)

G06V 30/412 (2022.01)

명세서

청구범위

청구항 1

컴퓨터 구현 방법으로서,

분석될 문서를 수신하는 단계;

상기 문서에서 추출될 키 이름들의 세트를 수신하는 단계;

상기 문서 내에 포함된 복수의 문자들의 스트링들, 및 상기 복수의 문자들의 스트링들의 문자들의 각각의 스트링에 대한 위치를 식별하는 단계;

문서 그래프를 생성하는 단계로서, 상기 문서 그래프는:

노드들의 세트로서, 상기 노드들의 세트의 각각의 노드는 상기 복수의 문자들의 스트링으로부터의 문자들의 스트링에 대응하는, 상기 노드들의 세트, 및

에지들의 세트로서, 상기 에지들의 세트의 각각의 에지는 둘 이상의 노드들을 함께 연결하는, 상기 에지들의 세트

를 포함하는, 상기 문서 그래프를 생성하는 단계;

상기 문서 그래프 및 수신된 상기 키 이름들의 세트에 기초하여 키들의 세트를 식별하는 단계;

상기 문서로부터 값들의 세트를 추출하는 단계; 및

식별된 상기 키들의 세트 및 추출된 상기 값들의 세트에 기초하여 키-값 쌍들의 세트를 생성하는 단계를 포함하는, 컴퓨터 구현 방법.

청구항 2

제1항에 있어서,

상기 복수의 문자들의 스트링들을 식별하는 것은 상기 수신된 문서 상에서 광학 문자 인식(OCR)을 수행하는 것을 포함하는, 컴퓨터 구현 방법.

청구항 3

제1항에 있어서,

상기 문서 그래프를 생성하는 것은:

제2 노드가 제1 노드로부터 임계 거리 내에 있다고 결정하는 것에 응답하여, 상기 에지들의 세트의 에지로 상기 노드들의 세트의 제1 노드를 상기 노드들의 세트의 제2 노드에 연결하는 것을 포함하는, 컴퓨터 구현 방법.

청구항 4

제3항에 있어서,

상기 제1 노드는, 상기 제2 노드가 상기 제1 노드로부터 임계 각도 내에 있다는 결정에 추가로 응답하여 상기 제2 노드에 연결되는, 컴퓨터 구현 방법.

청구항 5

제1항에 있어서,

상기 문서 그래프 및 상기 수신된 키 이름들의 세트에 기초하여 상기 키들의 세트를 식별하는 것은,

키 이름의 제1 토큰에 기초하여 상기 문서 그래프의 상기 노드들의 세트로부터 제1 노드를 식별하는 것으로서, 상기 제1 노드에 대응하는 문자들의 스트링은 상기 키 이름의 제1 토큰에 매칭되는, 상기 제1 노드를 식별하는

것;

상기 키 이름의 제2 토큰에 기초하여 상기 문서 그래프의 노드들의 세트로부터 제2 노드를 식별하는 것으로서, 상기 제2 노드는 상기 문서 그래프의 상기 에지들의 세트의 에지에 의해 상기 제1 노드에 연결되고, 상기 제2 노드에 대응하는 문자들의 스트링은 상기 키 이름의 상기 제2 토큰에 매칭되는, 상기 제2 노드를 식별하는 것을 포함하는, 컴퓨터 구현 방법.

청구항 6

제1항에 있어서,

상기 문서로부터 값들의 세트를 추출하는 것은,

상기 문서를 묘사하여(delining) 묘사된 문서를 생성하는 것;

상기 묘사된 문서에 전경 픽셀들의 세트 및 배경 픽셀들의 세트를 식별하기 위해 상기 묘사된 문서에 RLSA(run length smearing algorithm)를 적용하는 것;

상기 묘사된 문서에서의 상기 전경 픽셀들의 세트에 기초하여 바운딩 박스들의 세트를 생성하는 것; 및

생성된 상기 바운딩 박스들의 세트에 기초하여 상기 값들의 세트를 추출하는 것을 포함하는, 컴퓨터 구현 방법.

청구항 7

제6항에 있어서,

상기 생성된 바운딩 박스들의 세트에 기초하여 상기 값들의 세트를 추출하는 것은,

바운딩 박스들의 세트의 각각의 바운딩 박스에 대해, 바운딩 박스 내에 위치되는 상기 키들의 세트의 키들의 수를 결정하는 것;

상기 키들의 수가 1보다 큰 것에 응답하여, 분할된 바운딩 박스들의 세트를 생성하기 위해 상기 바운딩 박스를 분할하는 것으로서, 각각의 분할된 바운딩 박스는 상기 분할된 바운딩 박스 내에 위치한 상기 키들의 세트 중 하나의 키를 갖는, 상기 바운딩 박스를 분할하는 것을 포함하는, 컴퓨터 구현 방법.

청구항 8

실행가능 컴퓨터 프로그램 명령들을 저장한 비일시적 컴퓨터 판독가능 메모리로서,

상기 명령들은:

분석될 문서를 수신하는 것;

상기 문서로부터 추출될 키 이름들의 세트를 수신하는 것;

상기 문서 내에 포함된 복수의 문자들의 스트링들, 및 상기 복수의 문자들의 스트링들의 문자들의 각각의 스트링에 대한 위치를 식별하는 것;

문서 그래프를 생성하는 것으로서, 상기 문서 그래프는:

노드들의 세트로서, 상기 노드들의 세트의 각각의 노드는 상기 복수의 문자들의 스트링으로부터의 문자들의 스트링에 대응하는, 상기 노드들의 세트, 및

에지들의 세트로서, 상기 에지들의 세트의 각각의 에지는 둘 이상의 노드들을 함께 연결하는, 상기 에지들의 세트

를 포함하는, 상기 문서 그래프를 생성하는 것;

상기 문서 그래프 및 수신된 상기 키 이름들의 세트에 기초하여 키들의 세트를 식별하는 것;

상기 문서로부터 값들의 세트를 추출하는 것; 및

식별된 상기 키들의 세트 및 추출된 상기 값들의 세트에 기초하여 키-값 쌍들의 세트를 생성하는 것

을 포함하는 동작들을 수행하도록 실행가능한, 실행가능 컴퓨터 프로그램 명령들을 저장한 비일시적 컴퓨터 판

독가능 메모리.

청구항 9

제8항에 있어서,

상기 복수의 문자들의 스트링들을 식별하는 것은 상기 수신된 문서 상에서 광학 문자 인식(OCR)을 수행하는 것을 포함하는, 실행가능 컴퓨터 프로그램 명령들을 저장한 비일시적 컴퓨터 판독가능 메모리.

청구항 10

제8항에 있어서,

상기 문서 그래프를 생성하는 것은:

제2 노드가 제1 노드로부터 임계 거리 내에 있다고 결정하는 것에 응답하여, 상기 에지들의 세트의 에지로 상기 노드들의 세트의 제1 노드를 상기 노드들의 세트의 제2 노드에 연결하는 것을 포함하는, 실행가능 컴퓨터 프로그램 명령들을 저장한 비일시적 컴퓨터 판독가능 메모리.

청구항 11

제10항에 있어서,

상기 제1 노드는 상기 제2 노드가 상기 제1 노드로부터 임계 각도 내에 있다는 결정에 추가로 응답하여 상기 제2 노드에 연결되는, 실행가능 컴퓨터 프로그램 명령들을 저장한 비일시적 컴퓨터 판독가능 메모리.

청구항 12

제8항에 있어서,

상기 문서 그래프 및 상기 수신된 키 이름들의 세트에 기초하여 상기 키들의 세트를 식별하는 것은,

키 이름의 제1 토큰에 기초하여 상기 문서 그래프의 상기 노드들의 세트로부터 제1 노드를 식별하는 것으로서, 상기 제1 노드에 대응하는 문자들의 스트링은 상기 키 이름의 상기 제1 토큰에 매칭되는, 상기 제1 노드를 식별하는 것;

상기 키 이름의 제2 토큰에 기초하여 상기 문서 그래프의 상기 노드들의 세트로부터 제2 노드를 식별하는 것으로서, 상기 제2 노드는 상기 문서 그래프의 에지들의 세트의 에지에 의해 상기 제1 노드에 연결되고, 상기 제2 노드에 대응하는 문자들의 스트링은 상기 키 이름의 상기 제2 토큰에 매칭되는, 상기 제2 노드를 식별하는 것을 포함하는, 실행가능 컴퓨터 프로그램 명령들을 저장한 비일시적 컴퓨터 판독가능 메모리.

청구항 13

제8항에 있어서,

상기 문서로부터 값들의 세트를 추출하는 것은,

상기 문서를 묘사하여 묘사된 문서를 생성하는 것;

상기 묘사된 문서에 전경 픽셀들의 세트 및 배경 픽셀들의 세트를 식별하기 위해 상기 묘사된 문서에 RLSA(run length smearing algorithm)를 적용하는 것;

상기 묘사된 문서에서의 전경 픽셀들의 세트에 기초하여 바운딩 박스들의 세트를 생성하는 것; 및

생성된 상기 바운딩 박스들의 세트에 기초하여 상기 값들의 세트를 추출하는 것을 포함하는, 실행가능 컴퓨터 프로그램 명령들을 저장한 비일시적 컴퓨터 판독가능 메모리.

청구항 14

제13항에 있어서,

상기 생성된 바운딩 박스들의 세트에 기초하여 상기 값들의 세트를 추출하는 것은,

바운딩 박스들의 세트의 각각의 바운딩 박스에 대해, 상기 바운딩 박스 내에 위치되는 상기 키들의 세트의 키들

의 수를 결정하는 것;

상기 키들의 수가 1보다 큰 것에 응답하여, 분할된 바운딩 박스들의 세트를 생성하기 위해 상기 바운딩 박스를 분할하는 것으로서, 각각의 분할된 바운딩 박스는 상기 분할된 바운딩 박스 내에 위치한 상기 키들의 세트 중 하나의 키를 갖는, 상기 바운딩 박스를 분할하는 것을 포함하는, 실행가능 컴퓨터 프로그램 명령들을 저장한 비일시적 컴퓨터 판독가능 메모리.

청구항 15

컴퓨터 시스템으로서,

컴퓨터 프로그램 명령들을 실행하기 위한 컴퓨터 프로세서; 및

동작들을 수행하도록 상기 컴퓨터 프로세서에 의해 실행가능한 컴퓨터 프로그램 명령들을 저장한 비일시적 컴퓨터 판독가능 메모리를 포함하고,

상기 동작들은,

분석될 문서를 수신하는 것;

상기 문서로부터 추출될 키 이름들의 세트를 수신하는 것;

상기 문서 내에 포함된 복수의 문자들의 스트링들, 및 상기 복수의 문자들의 스트링들의 문자들의 각각의 스트링에 대한 위치를 식별하는 것;

문서 그래프를 생성하는 것으로서, 상기 문서 그래프는:

노드들의 세트로서, 노드들의 세트의 각각의 노드는 복수의 문자들의 스트링으로부터 문자들의 스트링에 대응하는, 상기 노드들의 세트, 및

에지들의 세트로서, 상기 에지들의 세트의 각각의 에지는 둘 이상의 노드들을 함께 연결하는, 상기 에지들의 세트

를 포함하는, 상기 문서 그래프를 생성하는 것;

상기 문서 그래프 및 수신된 상기 키 이름들의 세트에 기초하여 키들의 세트를 식별하는 단계;

상기 문서로부터 값들의 세트를 추출하는 것; 및

식별된 상기 키들의 세트 및 추출된 상기 값들의 세트에 기초하여 키-값 쌍들의 세트를 생성하는 것을 포함하는, 컴퓨터 시스템.

청구항 16

제15항에 있어서,

상기 복수의 문자들의 스트링들을 식별하는 것은 상기 수신된 문서 상에서 광학 문자 인식(OCR)을 수행하는 것을 포함하는, 컴퓨터 시스템.

청구항 17

제15항에 있어서,

상기 문서 그래프를 생성하는 것은:

제2 노드가 제1 노드로부터 임계 거리 내에 있고 상기 제2 노드가 상기 제1 노드로부터 임계 각도 내에 있다고 결정하는 것에 응답하여, 상기 에지들의 세트의 에지로 상기 노드들의 세트의 제1 노드를 상기 노드들의 세트의 제2 노드에 연결하는 것을 포함하는, 컴퓨터 시스템.

청구항 18

제15항에 있어서,

상기 문서 그래프 및 상기 수신된 키 이름들의 세트에 기초하여 상기 키들의 세트를 식별하는 것은,

키 이름의 제1 토큰에 기초하여 상기 문서 그래프의 상기 노드들의 세트로부터 제1 노드를 식별하는 것으로서, 상기 제1 노드에 대응하는 문자들의 스트링은 상기 키 이름의 제1 토큰에 매칭되는, 상기 제1 노드를 식별하는 것;

상기 키 이름의 제2 토큰에 기초하여 상기 문서 그래프의 상기 노드들의 세트로부터 제2 노드를 식별하는 것으로서, 상기 제2 노드는 상기 문서 그래프의 에지들의 세트의 에지에 의해 상기 제1 노드에 연결되고, 상기 제2 노드에 대응하는 문자들의 스트링은 상기 키 이름의 제2 토큰에 매칭되는, 상기 제2 노드를 식별하는 것을 포함하는, 컴퓨터 시스템.

청구항 19

제15항에 있어서,

상기 문서로부터 값들의 세트를 추출하는 것은,

상기 문서를 묘사하여 묘사된 문서를 생성하는 것;

상기 묘사된 문서에 전경 픽셀들의 세트 및 배경 픽셀들의 세트를 식별하기 위해 상기 묘사된 문서에 RLSA(run length smearing algorithm)를 적용하는 것;

상기 묘사된 문서에서의 전경 픽셀들의 세트에 기초하여 바운딩 박스들의 세트를 생성하는 것; 및

생성된 상기 바운딩 박스들의 세트에 기초하여 상기 값들의 세트를 추출하는 것을 포함하는, 컴퓨터 시스템.

청구항 20

제19항에 있어서,

상기 생성된 바운딩 박스들의 세트에 기초하여 상기 값들의 세트를 추출하는 것은,

바운딩 박스들의 세트의 각각의 바운딩 박스에 대해, 바운딩 박스 내에 위치되는 상기 키들의 세트의 키들의 수를 결정하는 것;

상기 키들의 수가 1보다 큰 것에 응답하여, 분할된 바운딩 박스들의 세트를 생성하기 위해 상기 바운딩 박스를 분할하는 것으로서, 각각의 분할 바운딩 박스는 상기 분할된 바운딩 박스 내에 위치된 상기 키들의 세트 중 하나의 키를 갖는, 상기 바운딩 박스를 분할하는 것을 포함하는, 컴퓨터 시스템.

발명의 설명

기술 분야

[0001] 설명된 실시형태들은 일반적으로 데이터 분석에 관한 것이고, 특히 문서들로부터 키-값 쌍들을 추출하는 것이다.

배경 기술

[0002] 빅 데이터 분석(Big Data Analytics)과 같은 데이터 분석 플랫폼들의 성장은, 비즈니스 가치를 갖는 정보를 추출하기 위해 대량의 데이터를 프로세싱하기 위한 툴(tool)로 데이터 프로세싱을 확장하였다. 예를 들어, 소기업은 외부 데이터 제공자, 내부 데이터 소스(예를 들어, 로컬 컴퓨터 상의 파일), 빅 데이터 스토어, 및 클라우드 기반 데이터(예를 들어, 소셜 미디어 정보)와 같은 다양한 소스들로부터 방대한 양의 데이터를 수집, 프로세싱 및 분석하기 위해 전용 컴퓨팅 및 인적 리소스들을 채용하는 제3자 데이터 분석 환경을 활용할 수도 있다. 유용한 정량적 및 정성적 정보를 추출하는 방식으로, 데이터 분석에 사용되는 바와 같은, 이러한 대규모 데이터 세트들을 프로세싱하는 것은 통상적으로 강력한 컴퓨터 디바이스들 상에서 구현되는 복잡한 소프트웨어 툴들을 필요로 한다.

[0003] 이러한 데이터 분석 플랫폼들에 사용되는 데이터는 문서들의 세트로부터 기원할 수 있다. 대부분의 문서들은 키들과 값들 사이에 연관성들(예를 들어, 세금 양식에 대한 "제1 이름" 필드)을 포함한다. 이러한 키-값 쌍들은 컴퓨터가 추출하기 매우 어려울 수 있다. 예를 들어, 기본 OCR 추출은 복잡하고 힘든 포스트-프로세싱 및 텍스트 마이닝으로 이어질 수 있다. 따라서, 사용자가 추출하고자 하는 키의 이름만을 제공하여 문서들로부터 키-값 쌍들을 추출할 수 있는 툴들이 필요하다.

발명의 내용

과제의 해결 수단

[0004] 위의 문제들 및 다른 문제들은 문서들로부터 키-값 쌍들을 추출하기 위한 컴퓨터 구현 방법, 컴퓨터 시스템 및 컴퓨터 판독가능 저장 매체에 의해 해결된다. 컴퓨터 구현 방법의 일 실시형태는 분석될 문서 및 문서로부터 추출될 키들의 제1 세트를 수신하는 단계를 포함한다. 방법은 문서 내에 포함된 문자들의 스트링들의 세트, 및 문자들의 각각의 스트링에 대한 위치를 식별하는 단계를 더 포함한다. 방법은 문서 그래프를 생성하는 단계를 더 포함한다. 문서 그래프는 노드들의 세트 및 에지들의 세트를 포함한다. 노드들의 세트의 각각의 노드는 문자들의 스트링들의 세트의 문자들의 스트링에 대응한다. 에지들의 세트의 각각의 에지는 둘 이상의 노드들을 함께 연결한다. 방법은 문서 그래프 및 수신된 키들의 제1 세트에 기초하여 키들의 제2 세트를 식별하는 단계를 더 포함한다. 또한, 방법은 문서로부터 값들의 세트를 추출하는 단계를 포함한다. 마지막으로, 방법은 식별된 키들의 제2 세트 및 추출된 값들의 세트에 기초하여 키-값 쌍들의 세트를 생성하는 단계를 포함한다.

[0005] 컴퓨터 시스템의 일 실시형태는 컴퓨터 프로그램 명령들을 실행하기 위한 컴퓨터 프로세서를 포함한다. 그 시스템은 또한, 실행될 때 컴퓨터 프로세서로 하여금 동작들을 수행하게 하는 컴퓨터 프로그램 명령들을 저장하는 비일시적(non-transitory) 컴퓨터 판독가능 저장 매체를 포함한다. 동작들은 분석될 문서 및 문서로부터 추출될 키들의 제1 세트를 수신하는 것을 포함한다. 동작들은 문서 내에 포함된 문자들의 스트링들의 세트, 및 문자들의 각각의 스트링에 대한 위치를 식별하는 것을 더 포함한다. 동작들은 문서 그래프를 생성하는 것을 더 포함한다. 문서 그래프는 노드들의 세트 및 에지들의 세트를 포함한다. 노드들의 세트의 각각의 노드는 문자들의 스트링들의 세트의 문자들의 스트링에 대응한다. 에지들의 세트의 각각의 에지는 둘 이상의 노드들을 함께 연결한다. 동작들은 문서 그래프 및 수신된 키들의 제1 세트에 기초하여 키들의 제2 세트를 식별하는 것을 더 포함한다. 또한, 동작들은 문서로부터 값들의 세트를 추출하는 것을 포함한다. 마지막으로, 동작들은 식별된 키들의 제2 세트 및 추출된 값들의 세트에 기초하여 키-값 쌍들의 세트를 생성하는 것을 포함한다.

[0006] 비일시적 컴퓨터 판독가능 메모리의 일 실시형태는 실행가능 컴퓨터 프로그램 명령들을 저장한다. 그 명령들은 동작들을 수행하도록 실행가능하다. 동작들은 분석될 문서 및 문서로부터 추출될 키들의 제1 세트를 수신하는 것을 포함한다. 동작들은 문서 내에 포함된 문자들의 스트링들의 세트, 및 문자들의 각각의 스트링에 대한 위치를 식별하는 것을 더 포함한다. 동작들은 문서 그래프를 생성하는 단계를 더 포함한다. 문서 그래프는 노드들의 세트 및 에지들의 세트를 포함한다. 노드들의 세트의 각각의 노드는 문자들의 스트링들의 세트의 문자들의 스트링에 대응한다. 에지들의 세트의 각각의 에지는 둘 이상의 노드들을 함께 연결한다. 동작들은 문서 그래프 및 수신된 키들의 제1 세트에 기초하여 키들의 제2 세트를 식별하는 것을 더 포함한다. 또한, 동작들은 문서로부터 값들의 세트를 추출하는 것을 포함한다. 마지막으로, 동작들은 식별된 키들의 제2 세트 및 추출된 값들의 세트에 기초하여 키-값 쌍들의 세트를 생성하는 것을 포함한다.

도면의 간단한 설명

[0007] 도 1은 하나 이상의 실시형태들에 따라, 데이터 분석 시스템을 포함하는 데이터 분석 환경을 나타내는 블록 다이어그램이다.

도 2는 하나 이상의 실시형태들에 따라, 문서로부터 키-값 쌍들을 추출하기 위한 워크플로우를 예시하는 블록 다이어그램이다.

도 3은 하나 이상의 실시형태에 따라 데이터 분석 애플리케이션을 예시하는 블록 다이어그램이다.

도 4a는 하나 이상의 실시형태들에 따라, OCR 모듈의 출력을 예시하는 문서의 마크업 이미지를 예시한다.

도 4b는 하나 이상의 실시형태들에 따라, 문서 그래프의 표현을 예시하는 문서의 마크업 이미지를 예시한다.

도 4c는 하나 이상의 실시형태들에 따라, 키 식별 모듈의 출력을 예시하는 문서의 마크업 이미지를 예시한다.

도 4d는 하나 이상의 실시형태들에 따라, 문서의 이미지에 커널을 적용함으로써 생성된 수평 및 수직 라인들에 대응하는 이미지를 예시한다.

도 4e는 하나 이상의 실시형태들에 따라, 문서의 이미지로부터 문서에서의 수평 및 수직 라인들에 대응되는 이

미지를 감산하여 생성된 묘사된 문서를 예시한다.

도 4f는 하나 이상의 실시형태들에 따라, 값 식별 모듈에 의해 생성된 바운딩 박스들을 예시하는 문서의 마크업 이미지를 예시한다.

도 4g는 하나 이상의 실시형태들에 따라, 식별된 키들 및 값들을 갖는 바운딩 박스들을 예시하는 문서의 마크업 이미지를 예시한다.

도 5는 하나 이상의 실시형태들에 따라, 분석 워크플로우를 사용하여 키-값 쌍들을 추출하기 위한 프로세스를 예시하는 플로우차트이다.

도 6은 하나의 실시형태에 따라, 도 1의 데이터 분석 시스템으로서 사용하기 위한 통상의 컴퓨터 시스템의 기능 블록을 예시하는 하이-레벨 블록 다이어그램이다.

도면들은 다양한 실시형태들을 오직 예시의 목적들을 위해 나타낸다. 당업자는, 본 명세서에서 예시된 구조들 및 방법들의 대안적인 실시형태들이 본 명세서에서 설명된 실시형태들의 원리들로부터 일탈함이 없이 채용될 수도 있음을 다음의 논의로부터 용이하게 인식할 것이다. 다양한 도면들에 있어서 동일한 참조 부호들 및 지정들은 동일한 엘리먼트들을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0008] 도 1은 하나의 실시형태에 따른 데이터 분석 시스템(110)을 포함하는 데이터 분석 환경(100)을 도시하는 블록 다이어그램이다. 환경(100)은 네트워크(130)를 통해 데이터 분석 시스템(110)에 연결된 다수의 데이터 소스들(120)을 더 포함한다. 도시된 환경(100)은 다수의 데이터 소스들(120)에 커플링된 하나의 데이터 분석 시스템(110)만을 포함하지만, 실시형태들은 다수의 데이터 분석 시스템들 및 단일의 데이터 소스를 가질 수 있다.

더욱이, 데이터 소스들(120)이 네트워크를 통해 데이터 분석 시스템(110)에 연결되는 것으로 도시되지만, 하나 이상의 데이터 소스들이 데이터 분석 시스템에 로컬로 저장될 수도 있다. 대안적으로, 사용자들은 (네트워크(130)를 통해 또는 데이터 분석 시스템의 로컬 인터페이스를 통해) 직접 데이터 분석 시스템에 데이터를 제공할 수 있다.

[0009] 데이터 분석 시스템(110)은 대량의 데이터를 프로세싱하기 위해 활용된 컴퓨터-기반 시스템이다. 일부 환경들에서, 데이터는 네트워크(130)를 통해 다수의 데이터 소스들(120)로부터 수집되고, 합쳐지고, 또는 달리 액세스된다. 데이터 분석 시스템(110)은 광범위한 데이터 소스들로부터 데이터 프로세싱 태스크들에서 채용되는 스케일러블(scalable) 소프트웨어 툴들 및 하드웨어 리소스들을 구현할 수 있다. 데이터 프로세싱 태스크(태스크(task) 또는 데이터 분석 태스크라고도 함)는 데이터를 프로세싱하기 위한 컴퓨팅 동작이다. 데이터 프로세싱 태스크는 예를 들어 데이터 액세스, 데이터 준비, 데이터 블렌딩, 데이터 패키징, 데이터 분석, 데이터에 대한 다른 유형들의 동작들, 또는 이들의 일부 조합을 포함한다. 데이터 분석 시스템(110)은 다수의 태스크들을 병렬로 수행할 수 있다. 그것이 이용가능한 컴퓨터 리소스들(예를 들어, 메모리)을 가질 때, 데이터 분석 시스템(110)은 수행될 준비가 된 복수의 태스크들로부터 수행할 태스크를 선택할 수 있다.

[0010] 도 1의 실시형태에서, 데이터 분석 시스템(110)은 메모리(140)와 데이터 분석 애플리케이션(150)을 포함한다. 메모리(140)는 데이터 분석 동작들에서 사용되는 데이터를 저장하고 데이터 저장 영역(들)에 대한 액세스를 제공한다. 일부 실시형태들에서, 메모리(140)는 RAM(Random Access Memory), HDD(Hard Disk Drive), SSD(Solid-State Drive), 다른 형태들의 컴퓨터 메모리, 또는 이들의 일부 조합을 포함한다. 일부 실시형태들에서, 메모리(140)는 비교적 작은 데이터 저장 영역에 대한 비교적 빠른 액세스를 제공하는 프라이머리 메모리(140) 및 비교적 큰 데이터 저장 영역에 대한 비교적 느린 액세스를 제공하는 세컨더리 메모리(150)를 포함한다.

[0011] 데이터 분석 애플리케이션(150)은 도 2에 예시된 워크플로우와 같이 분석 워크플로우들(또한, "워크플로우"로 지칭됨)을 사용함으로써 데이터 분석 태스크들의 실행을 지원하는 소프트웨어 애플리케이션이다. 워크플로우는 데이터 분석 툴들(또한 "툴들"로 지칭됨) 또는 데이터 분석 모듈들(또한 "모듈들"로 지칭됨)의 시퀀스이며, 이를 통하여, 데이터 분석 태스크가 개시(예를 들어, 데이터 파일의 입력)로부터 완료(예를 들어, 태스크의 결과의 출력)까지 통과한다. 시퀀스에서의 각각의 툴은 입력, 출력, 준비, 결합(join), 예측, 공간, 조사, 파싱, 변환 등과 같은 특정 프로세싱 동작 또는 데이터 분석 기능을 수행한다. 워크플로우의 실행은 워크플로우 내의 툴들의 동작들을 포함하고, 태스크의 결과를 나타내는 결과 아이템을 생성한다. 결과 아이템(result item)은, 일 예에서, 태스크의 결과의 시각적 표현을 제공하는 차트이다.

- [0012] 데이터 분석 애플리케이션(150)은 워크플로우들의 생성을 용이하게 하는 환경을 제공한다. 예를 들어, 데이터 분석 애플리케이션(150)은 데이터 분석 애플리케이션(150)의 GUI에 의해 지원되는 시각적 워크플로우 환경을 제공한다. 시각적 워크플로우 환경은 워크플로우를 설계하기 위해 소프트웨어 코드를 제공할 필요성을 없애고 워크플로우에 의해 구현될 복잡한 공식들을 식별할 필요성을 없애는 드래그 앤 드롭 툴들의 세트를 가능하게 한다. 다른 실시형태에서, 워크플로우(200)는 XML(Extensible Markup Language) 문서와 같은 문서에서 형성되고 기술된다.
- [0013] 환경은 또한 도 2에 예시된 워크플로우와 같은 논리 제어 워크플로우들의 생성을 용이하게 할 수 있다. 로직-제어된 워크플로우는 하나 이상의 로직 제어기들과 통합된 툴들의 시퀀스를 포함하는 워크플로우이다. 로직 제어기는 논리 함수를 가지며 입력에 기초하여 툴 또는 모듈을 트리거할지(및/또는 트리거하지 않을지) 여부를 결정하도록 적응된다. 데이터 분석 애플리케이션(150)은 데이터 파일들을 프로세싱하기 위해 툴들의 시퀀스를 사용하고 데이터 파일들의 프로세싱 동안 시퀀스에서 툴들의 동작들을 제어하기 위해 로직 제어기들을 사용한다. 일부 실시형태들에서, 로직 제어기는 워크플로우에서의 적어도 2개의 툴들: 시퀀스에서 제1 툴 및 제1 툴에 후속하는 제2 툴과 연관된다. 데이터 분석 애플리케이션(150)은 데이터 파일의 프로세싱에서 제1 툴의 데이터 분석 기능의 수행 동안 발생하는 이벤트들을 로깅하기 위해 제1 툴 상에 로깅 기능을 적용한다. 제1 툴에 의해 생성된 로그 데이터는 데이터 파일의 프로세싱 동안 제1 툴에 의해 수행된 활동들 및/또는 검출된 이벤트들을 기술한다. 이벤트들은, 예를 들어, 툴의 데이터 분석 기능의 수행 동안 검출된 에러들, 툴의 데이터 분석 기능의 수행 동안 생성된 경고, 툴에 의해 프로세싱된 레코드들의 수, 메타데이터 파일의 출력, 툴의 데이터 분석 기능의 수행의 완료 등일 수 있다. 제1 툴은 또한, 예를 들어, 메타데이터 파일에 분석 데이터를 출력한다. 분석 데이터는 데이터 파일의 프로세싱에서 데이터 분석 기능의 수행으로부터 제1 툴에 의해 생성된다.
- [0014] 데이터 분석 애플리케이션(150)은 제1 툴에 의해 생성된 로그 데이터 및/또는 분석 데이터를 로직 제어기에 대한 입력으로서 사용한다. 예를 들어, 로직 제어기는 로그 데이터를 분석하여 로그 데이터에 의해 설명된 이벤트를 검출하고 검출된 이벤트에 기초하여 제2 툴을 트리거할지 여부를 결정할 수 있다. 일부 실시형태들에서, 로직 제어기는 논리 함수를 갖는 데이터 분석 툴이다. 로직 제어기에서의 툴은 제1 툴에 의해 로깅된 이벤트들을 검출하기 위해 로그에 대한 데이터 분석 기능을 수행한다. 논리 함수는 논리 제어기에서 툴의 출력에 기초하여 커맨드를 출력한다. 제2 툴은 커맨드를 수신하고 커맨드에 따라 데이터를 프로세싱하도록 적응된다.
- [0015] 예를 들어, 제1 툴의 데이터 분석 기능의 성공적인 완료를 검출하는 것에 응답하여, 로직 제어기는 제2 툴이 동작하도록 트리거한다. 일부 실시형태들에서, 로직 제어기는 제1 툴로부터 메타데이터 파일을 수신하고, 제2 툴을 트리거하는 결정에 응답하여 메타데이터 파일을 제2 툴에 전송한다. 로직 제어기로부터 메타데이터 파일을 수신하는 제2 툴은 메타데이터 파일을 프로세싱한다. 로직 제어기는 제1 툴에 의해 생성된 로그에 기초하여 워크플로우에서 다른 툴들을 선택적으로 실행할 수 있다. 제2 툴에 후속하는 툴들(존재하는 경우)은 또한 그들의 데이터 분석 기능들을 수행하고 데이터 파일의 프로세싱을 완료하도록 동작한다. 툴들의 시퀀스에 의해 데이터 파일의 프로세싱의 결과를 나타내는 결과 아이템은 생성된다. 데이터 분석 애플리케이션(150)은 사용자에게 디스플레이하기 위한 결과 아이템을 제공한다.
- [0016] 다른 예로서, 로직 제어기는 제1 툴의 동작 동안 발생한 에러를 검출하는 것에 응답하여 제2 툴의 동작을 트리거하지 않는다. 일부 실시형태들에서, 제2 툴의 동작을 트리거하지 않기로 결정하는 것에 응답하여, 로직 제어기는 제3 툴의 동작을 트리거한다. 제3 툴은 시퀀스에 있지 않을 수 있다. 제3 툴은 보충 결과 아이템을 생성하도록 적응된다. 일 예에서, 보충 결과 아이템은 제1 툴의 동작 동안 발생한 이벤트를 특정하는 메시지가고, 이에 기초하여 로직 제어기는 제2 툴의 데이터 분석 기능을 트리거하지 않기로 결정한다. 제3 툴은 로직 제어기로부터의 이벤트의 정보에 기초하여 메시지를 생성하고, 사용자와 연관된 클라이언트 디바이스에 메시지를 전송할 수 있다. 일부 실시형태들에서, 데이터 분석 애플리케이션(150)은 이벤트를 해결하기 위한 액션을 사용자로부터 수신한다. 데이터 분석 애플리케이션(150)은 사용자의 액션에 기초하여 이벤트를 해결할 수도 있다. 사용자의 액션을 해결하는 것에 응답하여, 데이터 분석 애플리케이션(150)은 데이터 파일의 프로세싱을 완료하기 위해 제2 툴을 트리거하도록 로직 제어기에 명령할 수도 있다.
- [0017] 예들은 예시의 목적으로 제공된다. 로직 제어기는 다른 유형들의 로깅된 이벤트들을 검출할 수 있다. 또한, 로직 제어기는 다른 툴들의 실행을 제어할 수 있다. 예를 들어, 로직 제어기는 제2 툴을 트리거하기로 결정하는 것에 더하여 하나 이상의 다른 툴들을 트리거하기로(또는 트리거하지 않기로) 결정할 수 있다. 하나 이상의 다른 툴들은, 예를 들어, 이들이 툴들의 시퀀스로 배열되는 것과 동일한 순서로 제2 툴에 연결될

수 있다. 이와 유사하게, 로직 제어기는 제2 톨을 트리거하지 않기로 결정하는 것에 더하여 하나 이상의 다른 톨들을 트리거하기로(또는 트리거하지 않기로) 결정할 수 있다. 또한, 로직 제어기는 제1 톨에 더하여 하나 이상의 다른 톨들에 의해 생성된 로그를 입력으로서 수신하고 입력을 사용할 수 있다.

[0018] 데이터 분석 애플리케이션(150)에 의해 생성된 로직-제어된 워크플로우는 사용자들이 데이터 분석 프로세스를 모니터링하고 제어할 수 있게 하기 때문에 유리하다. 이러한 로직-제어 워크플로우를 사용함으로써, 데이터 분석 프로세스는 네거티브 이벤트(즉, 로직 제어기가 제2 톨을 트리거하지 않기로 결정하는 것에 기초한 이벤트) 후에 제1 톨에서 일시 정지될 수 있다. 네거티브 이벤트는 제1 톨에 의해 생성된 데이터로 하여금 추가 프로세싱에 부적절하도록 할 수 있다. 예를 들어, 제1 톨은 제1 톨에 의해 생성된 데이터가 불완전하도록 에러가 발생할 때 데이터 파일을 프로세싱하는 것을 중단할 수 있다. 따라서, 데이터 분석 프로세스의 일시정지는 다른 톨들에 의해 부적절한 데이터를 프로세싱하기 위해 컴퓨팅 리소스들을 낭비하는 것을 회피한다. 또한, 사용자는 워크플로우에 세 번째 톨을 추가하여, 추가 결과 아이템을 통해 네가티브 이벤트를 알릴 수 있다. 이러한 방식으로, 사용자들은 네거티브 이벤트를 해결하고 데이터 분석 프로세스를 재개할 기회를 얻는다. 종래의 데이터 분석 기술들과 비교하여, 로직-제어 워크플로우는 데이터 분석 프로세스들에 대한 사용자 기여들을 용이하게 하고 부적절한 데이터를 프로세싱하는 데 낭비되었을 컴퓨팅 리소스들을 절약한다.

[0019] 일부 경우들에서, 데이터 분석 애플리케이션(150)은 다수의 엔드 사용자들에게 데이터 분석 톨들 및 매크로들에 대한 네트워킹된, 또는 클라우드-기반 액세스를 지원하는 소프트웨어를 제공한다. 일 예로서, 데이터 분석 애플리케이션(150)은 대시보드, 웹 애플리케이션, 모바일 애플리케이션 등의 방식으로 워크플로우들의 형성을 지원한다. 다른 예로서, 데이터 분석 애플리케이션(150)은 모바일 애플리케이션 스토어 또는 다른 클라우드-기반 서비스와 유사한 방식으로 사용자들이 분석들을 공유, 브라우즈 및 소비할 수 있도록 한다. 분석 데이터, 매크로들 및 워크플로우들은 예를 들어 데이터 분석 시스템(110)의 다른 사용자에게 의해 액세스될 수 있는, 더 작은 규모 및 맞춤형(customizable) 분석 애플리케이션(즉, 앱)으로서 패키징되고 실행될 수 있다. 일부 경우들에서, 공개된 분석 앱들에 대한 액세스는 데이터 분석 시스템(110)에 의해 관리될 수 있다. 즉 액세스를 승인 또는 철회하고, 이에 의해 액세스 제어 및 보안 능력들을 제공할 수 있다. 데이터 분석 애플리케이션(150)은 형성, 배치, 공개, 반복, 업데이트 등과 같은 분석 앱들과 연관된 기능들을 수행할 수 있다.

[0020] 추가적으로, 데이터 분석 애플리케이션(150)은 분석 결과들을 액세스, 준비, 블렌드, 분석 및 출력하기 위한 능력과 같은, 데이터 분석에 수반되는 다양한 스테이지들에서 수행되는 기능들을 지원할 수 있다. 일부 경우들에서, 데이터 분석 애플리케이션(150)은 다양한 데이터 소스들에 액세스하여, 예를 들어 데이터의 스트림에서 원시(raw) 데이터를 추출할 수 있다. 데이터 분석 애플리케이션(150)에 의해 수집된 데이터 스트림들은 원시 데이터의 다수의 레코드들을 포함할 수 있고, 여기서, 원시 데이터는 상이한 포맷들 및 구조들의 것이다. 적어도 하나의 데이터 스트림을 수신한 후에, 데이터 분석 애플리케이션(150)은 데이터 스트림들의 레코드들이 데이터 분석 동작들의 입력으로서 사용되도록 허용하는 동작들을 수행할 수 있다. 또한, 예측 분석(예를 들어, 예측 모델링, 클러스터링, 데이터 조사)과 같은, 레코드들의 통계적, 정성적, 또는 정량적 프로세싱에서 수반되는 분석 기능들은 데이터 분석 애플리케이션(150)에 의해 구현될 수 있다.

[0021] 상기 언급된 바와 같이, 데이터 소스들(120)은 데이터 분석 시스템(110)에 전자 데이터를 제공한다. 데이터 소스(120)는 컴퓨터, 데이터베이스, 네트워크 API(애플리케이션 프로그래밍 인터페이스) 또는 클라우드 저장 시스템일 수도 있다. 데이터 소스(120)는 또한 다른 소스로부터 데이터를 추출할 수 있는 컴퓨터 시스템일 수도 있다. 데이터 소스들(120)은 데이터 분석 시스템(110)으로부터 원격일 수도 있고, 네트워크(130)를 통해 데이터를 제공할 수도 있다. 또한, 일부 또는 모든 데이터 소스들(120)은 데이터 분석 시스템에 직접 결합될 수도 있고, 네트워크(130)를 통해 데이터를 통과시키지 않고 데이터를 제공할 수도 있다. 데이터 소스들(120)로부터 제공된 데이터는 통상적으로 데이터 레코드들로 조직되고, 각각의 데이터 레코드는 하나 이상의 값들을 포함한다. 예를 들어, 데이터 소스에 의해 제공된 데이터 레코드는 콤마-분리된 값들의 시리즈들을 포함할 수도 있다. 데이터는 데이터 분석 시스템을 이용하는 기업에 관한 정보를 기술한다. 예를 들어, 데이터 소스(120)로부터의 데이터는 웹사이트들 상에서 액세스가능한 콘텐츠와의 및/또는 소셜 미디어 애플리케이션들과의 컴퓨터-기반 상호작용들(예를 들어, 클릭 트래킹 데이터)을 기술할 수 있다.

[0022] 네트워크(130)는 데이터 분석 시스템(110)과 데이터 소스들(120) 사이의 통신 경로들을 나타낸다. 하나의 실시형태에서, 네트워크(130)는 인터넷이고, 표준 통신 기술들 및/또는 프로토콜들을 이용한다. 네트워크(130)를 통해 교환되는 데이터는 하이퍼텍스트 마크업 언어(HTML), 확장가능 마크업 언어(XML) 등을 포함하는 기술들 및/또는 포맷들을 이용하여 표현될 수 있다. 다른 실시형태에서, 엔티티들은 상기 설명된 것들 대신

에 또는 추가하여 맞춤형 및/또는 전용 데이터 통신 기술들을 이용할 수 있다.

- [0023] 도 2는 하나의 실시형태들에 따른, 데이터를 프로세싱하기 위해 데이터 분석 시스템(110)을 이용하여 수행된 워크플로우(200)를 나타내는 블록 다이어그램이다. 워크플로우(200)는 데이터 분석 시스템(110)의 컴퓨터 디바이스에 의해 실행된다. 그러나, 다른 실시형태들에서, 워크플로우(200)는 데이터 분석 시스템(110)에 네트워크 (예를 들어, 네트워크(130))를 통해 통신가능하게 연결될 수도 있는 다른 컴퓨터 디바이스에 대해 전개된다.
- [0024] 예시된 바와 같이, 도 2의 워크플로우(200)는 입력들의 세트(210)를 수신하는 것, 및 출력들의 세트(215)를 생성 및/또는 제시하는 것을 포함한다. 일부 실시형태들에서, 입력들의 세트는 입력 톨들의 세트를 사용하여 (데이터 분석 시스템(110)의 사용자로부터, 데이터 분석 시스템(110)과 상호작용하는 다른 프로세스에 의해, 저장 디바이스 등으로부터) 수신된다. 더욱이, 출력들의 세트는 브라우징 톨을 통해 데이터 분석 시스템(110)의 사용자에게 제시될 수도 있다.
- [0025] 데이터 분석 톨(110)은 문서(220) 및 사용자 제공된 키 이름들(230)의 세트를 입력들로서 수신한다. 일부 실시형태들에서, 문서는 특정된 문서 포맷(예를 들어, 포터블 문서 포맷(PDF))으로 수신된다. 그 다음, 데이터 분석 톨(110)은 사용자 제공된 키 이름들(230)의 세트에 기초하여 문서들(220)로부터 키-값 쌍들의 세트(280)를 추출한다.
- [0026] 일부 실시형태들에서, 키 값 쌍들의 세트(280)를 추출하기 위해, 데이터 분석 시스템(110)은 수신된 문서(220)를 분석함으로써 문서 그래프(240)를 생성한다. 문서 그래프(240)는 노드들의 세트 및 에지들의 세트를 포함할 수 있다. 노드들의 세트의 각각의 노드는 문서(220)에서 식별된 문자들의 스트링에 대응할 수 있다. 일부 실시형태에서, 각각의 노드는 그의 대응하는 문자들의 스트링 및 문자들의 스트링이 문서(220) 내에 위치되는 연관된 위치를 저장한다. 또한, 에지들의 세트의 각각의 에지는 둘 이상의 노드들을 서로 연결한다.
- [0027] 또한, 생성된 문서 그래프(240) 및 사용자 제공된 키 이름들(230)의 세트에 기초하여, 데이터 분석 시스템(110)은 문서 내의 키투(250)의 세트를 식별한다. 일부 실시형태들에서, 데이터 분석 시스템(110)은 문서 그래프(240) 내에서 사용자 제공된 키 이름들(230) 각각을 위치시키기 위해 문서 그래프(240)를 검색한다. 데이터 분석 톨은 문서 그래프(240)로부터 문서 내의 각각의 키투(250)의 위치를 추가로 추출할 수 있다.
- [0028] 데이터 분석 시스템(110)은 문서(220) 내의 텍스트 블록들(260)의 세트를 식별하고, 텍스트 블록들(260)의 식별된 세트 및 식별된 키투의 세트 (250)에 기초하여 문서로부터 값들(270)의 세트를 추출한다. 그 다음, 데이터 분석 시스템(110)은 식별된 키투(250)의 세트로부터의 키투 및 식별된 값들(270)의 세트로부터의 대응하는 값을 매칭시켜 키투-값 쌍들(280)의 세트를 생성한다.
- [0029] 도 3은 하나 이상의 실시형태에 따라 데이터 분석 애플리케이션(300)을 예시하는 블록 다이어그램이다. 데이터 분석 애플리케이션(300)은 도 1의 데이터 분석 애플리케이션(150)의 일 실시형태이다. 도 3의 실시형태에서, 데이터 분석 애플리케이션(300)은 데이터 수신기 모듈(310), 광학 문자 인식(OCR) 모듈 (320), 그래프 생성 모듈(330), 키투 식별 모듈(340), 값 식별 모듈(350), 및 사용자 인터페이스 모듈(360)을 포함한다. 통상의 기술자는, 다른 실시형태들은 본 명세서에서 기술된 것들과는 상이한 및/또는 다른 컴포넌트들을 가질 수 있고, 기능성들은 상이한 방식으로 컴포넌트들 사이에 분포될 수 있음을 인식할 것이다. 예를 들어, 데이터 분석 애플리케이션(300)의 일부 또는 모든 기능들은 클라이언트 디바이스 또는 제3자 서버에서 수행될 수도 있다.
- [0030] 데이터 수신기 모듈(310)은 데이터 소스들, 예를 들어, 데이터 소스들(120)로부터 데이터 파일들을 수신하고, 데이터 분석 애플리케이션(300)에서의 다른 모듈들에 데이터 파일들을 제공한다. 하나의 실시형태에서, 데이터 수신기 모듈(310)은, 데이터 분석 애플리케이션에서의 모듈들 사이에 인터페이스로서 기능한다. 다른 실시형태에서, 데이터 수신기 모듈(310)은 부존재하고, 데이터 수신기 모듈(310)의 기능은 데이터 분석 애플리케이션의 모듈들의 각각에 의해 수행된다.
- [0031] 일부 실시형태들에서, 데이터 수신기 모듈(310)은 분석될 문서(220) 및 문서를 분석하기 위한 파라미터들의 세트를 수신하도록 구성된다. 문서를 분석하기 위한 파라미터들의 세트는 문서(220)로부터 키투-값 쌍들(280)을 추출하기 위한 사용자 제공 키투 이름들(230)의 세트를 포함할 수 있다. 일부 실시형태들에서, 데이터 수신기 모듈(310)은 (PDF 포맷과 같은) 미리 정의된 포맷을 갖는 문서들을 수신하도록 구성된다. 데이터 수신기 모듈(310)은 문서(220)가 정확한 포맷을 갖는지 여부를 식별하도록 구성될 수 있고, 문서가 미리 정의된 포맷을 갖지 않는 경우 에러를 출력할 수 있다.

- [0032] OCR 모듈(320)은 문서(220)를 분석하고 문서(220)에 포함된 문자들의 스트링들을 식별한다. 일부 실시형태들에서, OCR 모듈(320)은 추가적으로, 식별된 문자들의 스트링 각각에 대한 위치 또는 좌표들을 결정한다. 문자들의 스트링의 위치 또는 좌표는 문서(220) 내의 문자들의 스트링 주위의 바운딩 박스의 중심에 대응할 수 있다. 일부 실시형태들에서, OCR 모듈(320)은 문서의 이미지(또는 문서의 페이지들에 대응하는 하나 이상의 이미지들)를 입력으로서 수신한다. OCR 모듈(320)은 입력 이미지 내에 포함된 문자들의 스트링들을 정확하게 검출할 가능성을 증가시키기 전에 입력 이미지를 프리프로세싱할 수 있다. 예를 들어, OCR 모듈은 (예를 들어, 300 dpi(dots per inch)에서) 미리 결정된 해상도로 이미지를 변환하고, 이미지를 그레이스케일로 변환하고, 적응적 가우시안 임계화, 확장, 침식, 중간 블러 및 단순 임계화를 수행할 수 있다.
- [0033] 도 4a는 하나 이상의 실시형태들에 따라, OCR 모듈의 출력을 예시하는 문서의 마크업 이미지를 예시한다. 도 4a에 도시된 바와 같이, OCR 모듈은 문자들의 스트링들을 포함하는 문서의 영역들을 식별하고, 문자들의 스트링에서 문자들을 인식한다. 도 4a의 예에서, 문자들의 각각의 스트링은 박스에 의해 둘러싸여 있다.
- [0034] 다시 도 3을 참조하면, 그래프 생성 모듈(330)은 OCR 모듈(320)의 출력으로부터 문서(220)에 대한 문서 그래프(240)를 생성한다. 구체적으로, OCR 모듈(320)의 출력에 기초하여 그래프 생성 모듈(330)에 의해 생성된 문서 그래프(240)는 에지들의 세트에 의해 상호연결된 노드들의 세트를 포함한다. 일부 실시형태들에서, 그래프 생성 모듈(330)은 OCR 모듈(320)에 의해 문서로부터 식별된 문자들의 각각의 스트링에 대한 노드를 추가한다. 일부 실시형태에서, 각각의 노드는 그의 대응하는 문자들의 스트링, 및 문자들의 스트링이 문서 내에 위치되는 연관된 위치를 저장한다.
- [0035] 일부 실시형태들에서, 문서 그래프(240) 내의 각각의 노드에 대해, 그래프 생성 모듈(330)은 노드에 연결될 이웃 노드들의 세트를 식별한다. 그래프 생성 모듈(330)은 이웃 노드가 노드로부터 임계 거리 내에 있다면, 이웃 노드와 노드를 연결하기 위해 에지를 추가할 수 있다. 일부 실시형태들에서, 그래프 생성 모듈(330)은 이웃 노드가 노드로부터 임계 거리 내에 있고 노드의 위치 및 이웃 노드의 위치가 서로의 각도 범위 내에 있다면, 노드를 이웃 노드와 연결하기 위한 에지들을 추가한다.
- [0036] 도 4b는 하나 이상의 실시형태들에 따라, 그래프 생성 모듈의 출력을 예시하는 문서의 마크업 이미지를 예시한다. 도 4b의 예에서, 도 4a에 도시된 문서의 이미지에서 OCR 모듈(320)에 의해 식별된 문자들의 각각의 스트링은 문서 그래프의 노드(421)에 할당된다. 또한, 각각의 노드(421)는 에지들(423)에 의해 하나 이상의 노드에 연결된다. 예를 들어, 문자들 "Work"의 스트링에 대응하는 노드(421A)는 에지(423A)에 의해 문자들 "Request"(421B)의 스트링에 대응하는 노드(421B)에 연결된다. 그래프 생성 모듈(330)은 노드(421B)가 노드(421A)로부터 임계 거리 내에 있고 노드(421A)로부터 임계 각도 내에 있다고 결정하는 것에 응답하여 노드(421A)를 노드(421B)에 연결할 수 있다. 유사하게, 노드(421A)는 에지(423B)에 의해 문자들 "Requesting"의 스트링에 대응하는 노드(421D)에 연결된다. 그러나, 노드(421A)는 문자들 "NumberEWR-001-1001"의 스트링에 대응하는 노드(421C)에 연결되지 않는다. 도 4b의 예에서, 노드(421C)가 노드(421A)로부터 임계 각도 내에 있지 않기 때문에 노드(421A)는 노드(421C)에 연결되지 않는다.
- [0037] 다시 도 3을 참조하여 보면, 키 식별 모듈(340)은 문서(220)로부터 키들(250)의 세트를 식별하여 문서로부터 추출될 키-값 쌍들(280)의 세트를 형성한다. 키 식별 모듈(340)은 사용자 제공된 키 이름들(230)의 세트를 수신하고, 문서(220) 내에서 사용자 제공된 키 이름들(230)을 위치시키기 위해 문서 그래프(240)를 트래버스한다.
- [0038] 일부 실시형태들에서, 각각의 사용자 제공된 키 이름에 대해, 키 식별 모듈(340)은 사용자 제공된 키 이름이 다수의 단어들을 포함하는지 여부를 결정한다. 사용자 제공된 키 이름이 다수의 단어들을 포함하면, 키 식별 모듈은 사용자 제공된 키 이름을 단어들의 세트(또는 토큰들의 세트)로 분할한다. 그 다음, 키 식별 모듈(340)은 문서 그래프(240) 내에서 사용자 제공된 키 이름의 제1 단어를 검색하고 그리고 사용자 제공된 키 이름의 후속하는 단어들을 식별하도록 문서 그래프를 트래버스하는 것에 의해 사용자 제공된 키 이름의 단어들의 세트에 대응하는 노드들의 하나 이상의 체인을 식별할 수 있다. 즉, 키 식별 모듈(340)은 OCR 모듈(320)에 의해 식별된 노드들의 각각의 스트링을 사용자 제공된 키 이름의 제1 단어와 비교함으로써 사용자 제공된 키 이름의 제1 단어에 매칭되는 노드들의 세트를 식별할 수 있다. 사용자 제공된 키의 제1 단어에 매칭되는 각각의 식별된 노드에 대해, 키 식별 모듈(340)은 에지를 통해 식별된 노드에 연결된 제2 노드가 사용자 제공 키 이름의 제2 단어에 매칭되는지를 결정한다. 식별된 노드가 사용자 제공된 키 이름의 제2 단어에 매칭하는 제2 노드에 연결되지 않으면, 키 식별 모듈(340)은 식별된 노드가 사용자 제공된 키 이름에 대응하지 않는다고 결정한다. 대안적으로, 식별된 노드가 사용자 제공된 키 이름의 제2 단어에 매칭하는 제2 노드를 가지면, 키 식별 모듈(340)은 식별된 노드를 사용자 제공된 키 이름에 대응하는 후보 키로서 유지한다. 사용자 제공된 키

이름이 추가적인 단어들을 포함하면, 키 식별 모듈(340)은 사용자 제공된 키 이름의 모든 단어가 찾아질 때까지 문서 그래프를 계속 트래버스한다. 예를 들어, 사용자 제공된 키 이름이 제3 단어를 포함하면, 키 식별 모듈(340)은 에지를 통해 제2 노드에 연결된 제3 노드가 사용자 제공된 키 이름의 제3 단어에 매칭하는지 등을 결정한다.

[0039] 일부 실시형태들에서, 노드들의 체인(즉, 에지들의 세트를 통해 연결된 노드들의 세트)이 사용자 제공된 키 이름에 매칭하면, 키 식별 모듈(340)은 노드들의 체인에서의 노드들을 단일 노드로 병합할 수 있다. 키 식별 모듈(340)은 병합된 노드에 대한 새로운 위치를 추가적으로 결정할 수 있다. 그 다음, 키 식별 모듈(340)은 노드들의 체인에서의 노드들을 병합된 노드로 대체할 수 있다.

[0040] 일부 실시형태들에서, 키 식별 모듈(340)은 단어가 문서 그래프에서의 노드에 매칭되는지를 결정하기 위해 퍼지 매칭을 사용한다. 예를 들어, 키 식별은 단어를 노드에 대응하는 문자들의 스트링과 비교할 수 있고, 비교에 기초하여 LR(Levenshtein ratio)을 결정할 수 있다. LR이 임계 값(예를 들어, 90%)을 초과하면, 키 식별 모듈(340)은 단어가 노드에 대응하는 문자들의 스트링과 일치한다고 결정한다. 대안적으로, 키 식별 모듈(340)은 고려중인 단어에 대한 최고 LR 값들을 갖는 상위 설정 수의 노드들(예를 들어, 상위 10개의 노드들)을 선택한다.

[0041] 도 4c는 하나 이상의 실시형태들에 따라, OCR 모듈의 출력을 예시하는 문서의 마크업 이미지를 예시한다. 구체적으로, 도 4c는 키 식별 모듈에 의해 식별된 키들의 세트를 예시한다. 키 식별 모듈에 의해 식별된 키들의 각각은 박스에 의해 둘러싸인다.

[0042] 다시 도 3을 참조하여 보면, 값 식별 모듈(350)은 문서(220)로부터 값들(270)의 세트를 식별하여 문서로부터 추출된 키-값 쌍들(280)의 세트를 형성한다. 값 식별 모듈(350)은 문서(220)로부터 텍스트 블록들(260)(즉, 텍스트의 블록들)을 식별하고, 텍스트 블록들(260)을 키 식별 모듈(340)에 의해 식별된 키들(250)에 매칭시킨다. 그 후, 값 식별 모듈(350)은 텍스트 블록들(260)로부터 값들을 추출한다.

[0043] 일부 실시형태들에서, 값 식별 모듈(350)은 문서를 구분한다. 예를 들어, 값 식별 모듈(350)은 문서의 이미지로부터(또는 문서의 각각의 페이지에 대응하는 이미지들로부터) 수평 및 수직 라인들을 제거한다. 문서를 묘사하기 위해, 문서의 이미지에 커널을 적용하여 수평 및 수직 라인들에 대응하는 이미지를 생성한다. 일부 실시형태들에서, 커널은 문서의 크기에 의해 결정된다. 또한, 일부 실시형태들에서, 제1 커널은 수평 라인들에 대응하는 제1 이미지를 생성하기 위해 적용되고, 제2 커널은 수직 라인들에 대응하는 제2 이미지를 생성하기 위해 적용된다. 수평 및 수직 라인들에 대응하는 생성된 이미지(또는 이미지들)는 문서의 이미지로부터 감산되어 문서의 묘사된 이미지를 생성한다.

[0044] 도 4d는 하나 이상의 실시형태들에 따라, 문서의 이미지에 커널을 적용함으로써 생성된 수평 및 수직 라인들에 대응하는 이미지를 예시한다. 도 4e는 하나 이상의 실시형태들에 따라, 문서의 이미지로부터 문서에서의 수평 및 수직 라인들에 대응되는 이미지를 감산하여 생성된 묘사된 문서를 예시한다.

[0045] 이어서, 문서의 묘사된 이미지는 바운딩 박스들의 세트를 생성하도록 프로세싱된다. 일부 실시형태들에서, 바운딩 박스들을 생성하기 위해 RLSA(Run Length Smearing Algorithm)가 묘사된 이미지에 적용된다. 일부 실시형태들에서, RLSA는 전경 픽셀들(예를 들어, 채워진 또는 흑색 픽셀들)의 세트 및 배경 픽셀들(예를 들어, 블랭크 또는 백색 픽셀들)의 세트를 식별한다. 그 다음, RLSA는 전경 픽셀들에 대한 배경 픽셀의 근접성에 기초하여 배경 픽셀들을 전경 픽셀들로 변환한다. 일부 실시형태들에서, 배경 픽셀이 묘사된 이미지의 수평 축 상의 양 방향들에서 전경 픽셀들로부터 설정된 수의 픽셀들 내에 있다면, RLSA는 배경 픽셀들을 전경 픽셀들로 변환한다.

[0046] 일부 실시형태들에서, RLSA는 임계 한계 값보다 낮은 수의 인접한 배경 픽셀들을 갖는 배경 픽셀들을 변환한다. 일부 실시형태들에서, RLSA는 문서의 배경 픽셀들을 0들로서 표현하고, 문서의 전경 픽셀들을 1들로서 표현한다. 그런 다음, 문서의 표현에서 각각의 0에 대해, RLSA는 0의 인접한 수가 임계 한계 값보다 작다면 0을 1로 대체한다. 예를 들어, 임계 한계 값이 3이면, 1000111000001의 표현을 갖는 문서의 일부가 1111111000001의 표현으로 변환된다. 일부 실시형태들에서, RLSA는 수평 방향으로 적용된다. 다른 실시형태들에서, RLSA는 수평 방향 및 수직 방향 둘 모두에 적용된다.

[0047] RLSA를 구동한 후, 값 식별 모듈(350)은 윤곽 검출 알고리즘을 실행하여 함께 연관되는 텍스트 블록들 주위에 바운딩 박스들을 생성한다. 윤곽 검출 알고리즘은 RLSA가 적용된 후에 전경 픽셀들로서 분류된 픽셀들에 대한 바운딩 박스들을 식별한다. 일부 실시형태들에서, 값 식별 모듈은 인접한 박스들을 함께 병합하기 위해

바운딩 박스들에 대해 윤곽 검출 알고리즘을 실행한다. 예를 들어, 값 식별 모듈(350)은 전경 픽셀들로서 바운딩 박스에 의해 둘러싸인 모든 픽셀들을 변환하고 윤곽 검출 알고리즘을 재적용하여, 제1 윤곽 검출 검출이 적용된 후에 전경 픽셀들로서 분류된 픽셀들에 대한 바운딩 박스들을 검출할 수 있다.

[0048] 도 4f는 하나 이상의 실시형태들에 따라, 값 식별 모듈에 의해 생성된 바운딩 박스들을 예시하는 문서의 마크업 이미지를 예시한다. 구체적으로, 도 4f는 RLSA에 의해 식별된 바와 같은 전경 픽셀들의 세트 및 바운딩 박스들의 세트를 예시하며, 각각은 전경 픽셀들의 하나의 그룹을 둘러싼다.

[0049] 그 후, 바운딩 박스를 사용하여, 값 식별 모듈(350)은 키 식별 모듈(340)에 의해 식별된 각각의 키에 대한 하나 이상의 값을 식별한다. 일부 실시형태들에서, 키 식별 모듈(340)에 의해 식별된 각각의 키에 대해, 값 식별 모듈(350)은 키의 위치에 기초하여 바운딩 박스를 식별하고 식별된 바운딩 박스에 의해 둘러싸인 텍스트를 추출함으로써 하나의 값을 식별한다.

[0050] 일부 실시형태들에서, 주어진 키에 대해, 값 식별 모듈(350)은 키를 둘러싸는 바운딩 박스를 식별한다. 대안적으로, 값 식별 모듈(350)은 키에 가장 가까운 바운딩 박스를 식별한다. 일부 실시형태들에서, 값 식별 모듈(350)은 키를 연관시킬 바운딩 박스를 식별하기 위해 문서 그래프를 트레이싱한다. 일부 실시형태들에서, 값 식별 모듈(350)은 임의의 주어진 바운딩 박스가 다수의 키들에 가장 가까운 바운딩 박스인지 여부를 결정한다. 값 식별 모듈(350)이 바운딩 박스가 다수의 키들에 가장 가까운 바운딩 박스라고 결정하면, 값 식별 모듈(350)은 바운딩 박스를 분할하고 각각의 키를 바운딩 박스의 일부에 연관시킨다.

[0051] 대안적으로, 일부 실시형태들에서, 값 식별 모듈(350)은, 키와 연관될 바운딩 박스들을 식별하기 위해 각각의 키를 프로세싱하는 대신에, 값 식별 모듈(350)은 바운딩 박스를 하나 이상의 키와 연관시킬지 여부를 식별하기 위해 각각의 바운딩 박스를 프로세싱한다. 예를 들어, 각각의 바운딩 박스에 대해, 값 식별 모듈(350)은 바운딩 박스에 의해 둘러싸인 키들을 식별할 수 있다. 이 실시형태에서, 값 식별 모듈(350)은 3개의 상이한 시나리오들을 고려할 수 있다. 첫째, 바운딩 박스는 하나의 키 및 연관된 값을 포함할 수 있다. 둘째, 바운딩 박스는 다수의 키들 및 연관된 값들을 포함할 수 있다. 셋째, 바운딩 박스는 값을 포함하고 키를 포함하지 않을 수 있다.

[0052] 값 식별 모듈(350)이 바운딩 박스가 하나의 키를 포함한다고 결정하면, 값 식별 모듈(350)은 바운딩 박스로부터 값을 추출하고, 키 및 추출된 값에 기초하여 키-값 쌍을 생성한다. 예를 들어, 바운딩 박스에 의해 둘러싸인 텍스트로부터 키의 텍스트를 제거함으로써 값이 추출될 수 있다.

[0053] 또한, 값 식별 모듈(350)이 바운딩 박스가 다수의 키들을 포함한다고 결정하면, 값 식별 모듈(350)은 바운딩 박스에 의해 둘러싸인 키들의 수에 기초하여 바운딩 박스를 분할한다. 일부 실시형태들에서, 값 식별 모듈(350)은 키들과 값들 사이의 바운더리들을 최적화하려고 시도함으로써 바운딩 박스를 분할한다. 예를 들어, 바운딩 박스가 열에 적층된 n개의 키들을 포함하면, 값 식별 모듈(350)은 바운딩 박스를 수평으로 n개의 더 작은 바운딩 박스들로 분할하고, 각각의 바운딩 박스는 하나의 키 및 하나의 값을 포함한다. 그 다음, 값 식별 모듈(350)은 바운딩 박스가 하나의 키 및 연관된 값을 포함하는 경우에 대응하는 단계들을 다수의 키들을 포함하는 바운딩 박스를 분할함으로써 획득된 바운딩 박스들 각각에 적용한다.

[0054] 추가적으로, 값 식별 모듈(350)이 바운딩 박스가 키를 포함하지 않는다고 결정하면, 값 식별 모듈은 바운딩 박스를 키에 연관시키기 위해 문서 그래프를 사용한다.

[0055] 도 4g는 하나 이상의 실시형태들에 따라, 식별된 키들 및 값들을 갖는 바운딩 박스들을 예시하는 문서의 마크업 이미지를 예시한다. 도 4g의 예에서, 문서는 하나의 키 및 하나의 값을 포함하는 바운딩 박스들(471)의 세트를 포함한다. 예를 들어, 바운딩 박스(471A)는 키 "Type" 및 값 "Change"를 포함한다. 바운딩 박스(471A)를 프로세싱할 때, 값 식별 모듈은 하나의 키만이 바운딩 박스(471A)에 의해 둘러싸여 있다고 결정하고 바운딩 박스로부터 값을 추출하여 바운딩 박스 내부에 위치한 키에 대한 키-값 쌍을 생성한다. 예를 들어, 값 식별 모듈은 바운딩 박스(471A) 내부에 위치한 문자들의 스트링들로부터 키를 제거함으로써 바운딩 박스로부터 값을 추출한다. 또한, 도 4g의 예에서, 문서는 하나의 키 및 하나의 값을 포함하는 바운딩 박스들(473)의 세트를 포함한다. 예를 들어, 바운딩 박스(473A)는 "CARBURETOR" 값을 포함하지만 키를 포함하지 않는다. 여기서, 값 식별 모듈은 값에 가장 가깝게 위치되고 식별된 키와 바운딩 박스에 의해 바운딩된 값과 연관된 키를 식별할 수 있다. 이 경우, 값 식별 모듈은 키 "Name"이 바운딩 박스(473A)에 가장 가깝다는 것을 식별하고, 키 "Name"을 값 "CARBURETOR"와 연관시킴으로써 키-값 쌍을 생성한다.

[0056] 도 5는 하나 이상의 실시형태들에 따라, 분석 워크플로우를 사용하여 키-값 쌍들을 추출하기 위한 프로세스를

예시하는 플로우차트이다. 일부 실시형태들에서, 프로세스는 데이터 분석 애플리케이션(150)에 의해 수행되지만, 프로세스에서의 동작들의 일부 또는 전부는 다른 실시형태들에서 다른 엔티티들에 의해 수행될 수도 있다. 일부 실시형태들에서, 프로세스에서의 동작들은 상이한 순서로 수행되고, 상이한 및/또는 추가적인 단계들을 포함할 수 있다.

- [0057] 일부 실시형태들에서, 데이터 분석 애플리케이션(150)은 분석 워크플로우에 액세스한다(510). 분석 워크플로우는 톨들의 시퀀스를 포함한다. 시퀀스에서의 각각의 톨은 입력, 출력, 준비, 결합, 예측, 공간, 조사, 파싱, 변환 등과 같은 데이터 분석 기능을 수행하도록 적응된다. 분석 워크플로우는 로직 제어 워크플로우이고, 하나 이상의 로직 제어기들을 더 포함한다.
- [0058] 데이터 분석 애플리케이션(150)의 데이터 수신 모듈(310)은 분석될 문서를 수신한다(510). 분석될 문서는 데이터 파일로서 수신될 수 있다(예를 들어, 데이터 분석 애플리케이션의 사용자에게 의해 업로드된다). 데이터 파일은 지원되는 문서 포맷들의 세트(예를 들어, PDF 포맷) 중 하나로서 수신될 수 있다. 또한, 데이터 분석 애플리케이션(150)의 데이터 수신 모듈(310)은 문서로부터 추출될 사용자 제공된 키 이름들의 세트를 수신한다(515). 일부 실시형태들에서, 키들은 데이터 분석 애플리케이션의 사용자 인터페이스 모듈(360)에 의해 제공된 그래픽 사용자 인터페이스를 통해 데이터 분석 애플리케이션(150)의 사용자로부터 수신된다.
- [0059] 데이터 분석 애플리케이션(150)은 문서를 OCR 모듈(320)에 제공한다. 데이터 분석 애플리케이션(150)의 OCR 모듈(320)은 문서 내에서 문자들의 스트링들(또는 텍스트 스트링들)의 세트를 식별한다(520). 또한, OCR 모듈(320)은 문자의 각각의 식별된 스트링의 위치를 식별할 수 있다. 일부 실시형태들에서, OCR 모듈은 문서에 대응하는 이미지로부터 문자들의 스트링들의 세트를 추출한다. 또한, OCR 모듈(320)은 문서로부터 문자들의 스트링들의 세트를 추출하기 전에 문서 또는 문서의 이미지에 대한 프리-프로세싱을 수행할 수 있다.
- [0060] 데이터 분석 애플리케이션(150)은 OCR 모듈(320)의 출력을 그래프 생성 모듈(330)에 제공할 수도 있다. 데이터 분석 애플리케이션(150)의 그래프 생성 모듈(330)은 OCR 모듈(320)에 의해 추출된 문자들의 스트링들의 세트에 기초하여 문서에 대한 문서 그래프를 생성한다(525). 문서 그래프는 노드들의 세트(각각 OCR 모듈(320)에 의해 식별된 문자들의 스트링들의 세트의 문자들의 스트링에 대응함), 및 에지들의 세트를 포함할 수 있다(각각의 에지는 둘 이상의 노드들을 함께 연결한다).
- [0061] 그 다음, 데이터 분석 애플리케이션(150)은 그래프 생성 모듈(330)에 의해 생성된 문서 그래프 및 사용자 제공 키 이름을 키 식별 모듈(340)에 제공할 수 있다. 데이터 분석 애플리케이션(150)의 키 식별 모듈(340)은 문서에서 키들의 세트를 식별한다(530). 키 식별 모듈(340)은 각각의 사용자 제공 키 이름을 문서 그래프 내의 하나 이상의 노드 또는 노드 체인(즉, 에지 세트를 통해 연결된 노드들의 세트)에 매칭시킬 수 있다.
- [0062] 그 다음, 데이터 분석 애플리케이션(150)은 키 식별 모듈(340)에 의해 식별된 문서 및 키들을 값 식별 모듈(350)에 제공할 수 있다. 일부 실시형태들에서, 데이터 분석 애플리케이션(150)은 문서 그래프를 값 식별 모듈(350)에 추가로 제공한다. 데이터 분석 애플리케이션(150)의 값 식별 모듈(350)은 문서 내에서 텍스트의 블록들의 세트를 식별한다(535). 특히, 값 식별 모듈(350)은 바운딩 박스들의 세트를 생성할 수 있으며, 바운딩 박스들 각각은 문서 내의 문자들의 하나 이상의 스트링들을 둘러싼다. 바운딩 박스들을 생성하기 위해, 값 식별 모듈(350)은 문서를 묘사하고, RLSA를 적용하고, 윤곽 검출 알고리즘을 적용할 수 있다.
- [0063] 또한, 데이터 분석 애플리케이션(150)의 값 식별 모듈(350)은 텍스트의 식별된 블록들로부터 각각의 키에 대한 값들을 추출한다(540). 일부 실시형태들에서, 값들을 추출하기 위해, 값 식별 모듈(350)은 식별된 바운딩 박스들 각각을 프로세싱하고, 하나 이상의 키들이 바운딩 박스들 각각의 내부에 위치하는지를 결정한다. 하나보다 많은 키가 바운딩 박스 내부에 위치되면, 값 식별 모듈(350)은 결과적인 바운딩 박스들 각각의 내부에 하나의 키만이 위치되도록 바운딩 박스를 분할한다. 하나의 키만을 포함하는 각각의 바운딩 박스에 대해, 값 식별 모듈(350)은 바운딩 박스에 의해 둘러싸인 텍스트로부터 바운딩 박스 내부에 위치한 키와 연관될 값을 추출한다. 또한, 키가 바운딩 박스 내에 위치되지 않으면, 값 식별 모듈은 바운딩 박스를 키와 연관시킬지 여부를 결정한다.
- [0064] 마지막으로, 키 식별 모듈(340)에 의해 식별된 키들 및 값 식별 모듈(350)에 의해 식별된 값들에 기초하여, 키-값 쌍들의 세트가 생성된다(545). 일부 실시형태들에서, 데이터 분석 애플리케이션(150)은 생성된 키-값 쌍들을 데이터 분석 애플리케이션의 사용자에게 출력한다. 대안적으로, 데이터 분석 애플리케이션(150)은 생성된 키-값 쌍들을 데이터 분석 애플리케이션에 연결된 파일 또는 데이터베이스에 저장할 수도 있다.
- [0065] 도 6은 하나 이상의 실시형태들에 따라, 도 1의 머신 러닝 서버(110)로서 사용하기 위한 통상적인 컴퓨터 시스

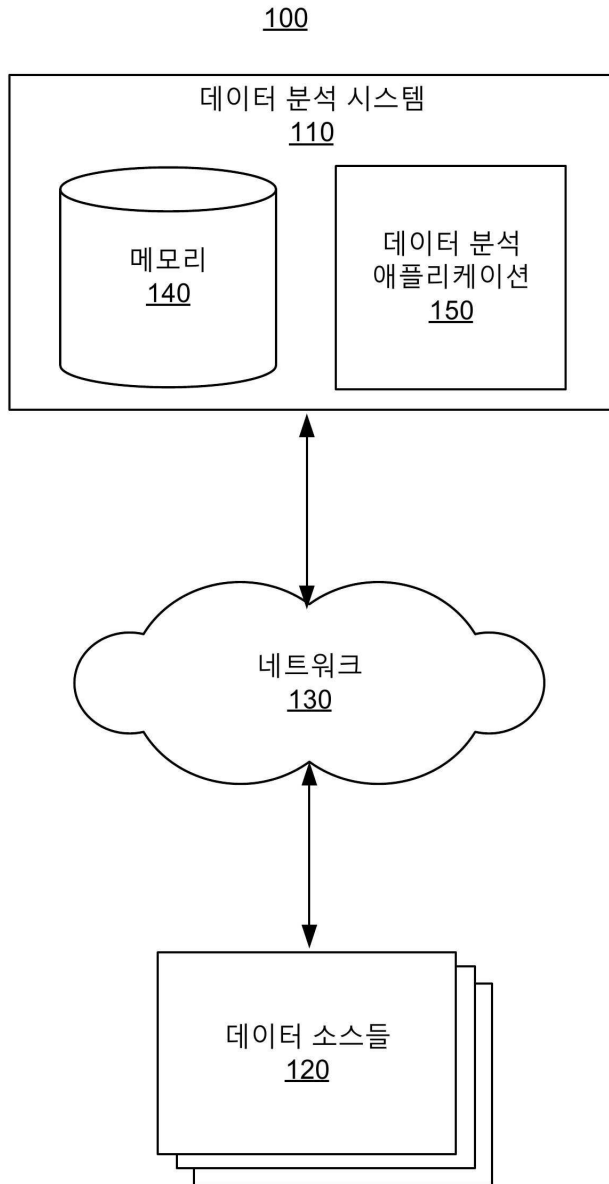
템(600)의 기능적 뷰를 예시하는 하이-레벨 블록 다이어그램이다.

- [0066] 예시된 컴퓨터 시스템은 칩셋(604)에 커플링된 적어도 하나의 프로세서(602)를 포함한다. 프로세서(602)는 동일한 다이 상에 다수의 프로세서 코어들을 포함할 수 있다. 칩셋(604)는 메모리 제어기 허브(620) 및 입력/출력(I/O) 제어기 허브(622)를 포함한다. 메모리(606) 및 그래픽스 어댑터(612)는 메모리 제어기 허브(620)에 커플링되고, 디스플레이(618)는 그래픽스 어댑터(612)에 커플링된다. 저장 디바이스(608), 키보드(610), 포인팅 디바이스(614), 및 네트워크 어댑터(616)가 I/O 제어기 허브(622)에 커플링될 수도 있다. 일부 다른 실시형태들에서, 컴퓨터 시스템(600)은 추가적인, 더 적은, 또는 상이한 컴포넌트들을 가질 수도 있고, 그 컴포넌트들은 상이하게 커플링될 수도 있다. 예를 들어, 컴퓨터 시스템(600)의 실시형태들은 디스플레이들 및/또는 키보드들이 없을 수도 있다. 또한, 컴퓨터 시스템(600)은 일부 실시형태들에서 랙-마운트드 블레이드 서버로서 또는 클라우드 서버 인스턴스로서 인스턴스화될 수도 있다.
- [0067] 메모리(606)는, 프로세서(602)에 의해 사용되는 명령들 및 데이터를 보유한다. 일부 실시형태들에서, 메모리(606)는 랜덤-액세스 메모리이다. 저장 디바이스(608)는 비일시적 컴퓨터 판독가능 저장 매체이다. 저장 디바이스(608)는 HDD, SSD, 또는 다른 타입들의 비일시적 컴퓨터 판독가능 저장 매체일 수 있다. 머신러닝 서버(110)에 의해 프로세싱 및 분석된 데이터는 메모리(606) 및/또는 저장 디바이스(608)에 저장될 수 있다.
- [0068] 포인팅 디바이스(614)는 마우스, 트랙 볼, 또는 다른 타입의 포인팅 디바이스일 수도 있고, 데이터를 컴퓨터 시스템(600)에 입력하기 위해 키보드(610)와 조합하여 사용된다. 그래픽스 어댑터(612)는 디스플레이(618)상에 이미지들 및 다른 정보를 디스플레이한다. 일부 실시형태들에서, 디스플레이(618)는, 사용자 입력 및 선택들을 수신하기 위한 터치 스크린 능력을 포함한다. 네트워크 어댑터(616)는 컴퓨터 시스템(600)을 네트워크(160)에 커플링시킨다.
- [0069] 컴퓨터 시스템(600)은, 본 명세서에서 설명된 기능성을 제공하기 위한 컴퓨터 모듈들을 실행하도록 적응된다. 본 명세서에서 사용된 바와 같이, 용어 "모듈"은, 명시된 기능성을 제공하기 위한 컴퓨터 프로그램 명령 및 다른 로직을 지칭한다. 모듈은 하드웨어, 펌웨어, 및/또는 소프트웨어에서 실현될 수 있다. 모듈은 하나 이상의 프로세스를 포함할 수 있고, 및/또는 프로세스의 부분으로만 제공될 수 있다. 모듈은 통상적으로, 저장 디바이스(608)상에 저장되고, 메모리(606)로 로딩되며, 프로세서(602)에 의해 실행된다.
- [0070] 컴포넌트들의 특정 명명, 용어들의 대문자화, 속성들, 데이터 구조들, 또는 임의의 다른 프로그래밍 또는 구조적 양태는 의무적이거나 중요하지 않으며, 설명된 실시형태들을 구현하는 메커니즘들은 상이한 명칭들, 포맷들, 또는 프로토콜들을 가질 수도 있다. 추가로, 시스템들은, 설명된 바와 같이 하드웨어와 소프트웨어의 조합을 통해, 또는 완전히 하드웨어 엘리먼트들로 구현될 수도 있다. 또한, 본 명세서에서 설명된 다양한 시스템 컴포넌트들 사이의 기능성의 특정 분할은 단지 예시적일 뿐이고 의무적인 것은 아니며; 단일 시스템 컴포넌트에 의해 수행된 기능들은 대신에 다수의 컴포넌트들에 의해 수행될 수도 있고, 다수의 컴포넌트들에 의해 수행된 기능들은 대신에 단일 컴포넌트에 의해 수행될 수도 있다.
- [0071] 상기 설명의 일부 부분들은 정보에 대한 동작들의 알고리즘들 및 심볼 표현들의 관점에서 피쳐들을 제시한다. 이들 알고리즘 설명들 및 표현들은 데이터 프로세싱 기술들에서의 당업자에 의해 그 작업의 실체를 다른 당업자에게 가장 효과적으로 전달하기 위해 사용되는 수단들이다. 이들 동작들은, 기능적으로 또는 논리적으로 설명되지만, 컴퓨터 프로그램들에 의해 구현되는 것으로 이해된다. 또한, 동작들의 이들 배열들을 일반성의 손실 없이, 모듈들로서 또는 기능적 명칭들에 의해 지칭하는 것이 때로는 편리한 것으로 증명되었다.
- [0072] 상기 논의로부터 명백한 바와 같이 특별히 달리 언급되지 않으면, 설명 전반에 걸쳐, "프로세싱하는 것" 또는 "컴퓨팅하는 것" 또는 "계산하는 것" 또는 "결정하는 것" 또는 "디스플레이하는 것" 등과 같은 용어들을 활용하는 논의들은, 컴퓨터 시스템 메모리들 또는 레지스터들 또는 다른 그러한 정보 저장, 송신 또는 디스플레이 디바이스들 내의 물리적 (전자적) 양들로서 표현된 데이터를 조작하고 변환하는, 컴퓨터 시스템 또는 유사한 전자 컴퓨팅 디바이스의 액션 및 프로세스들을 지칭함이 인식된다.
- [0073] 본원에 기술된 특정 실시형태들은 알고리즘의 형태로 기술된 프로세스 단계들 및 명령들을 포함한다. 실시형태들의 프로세스 단계들 및 명령들은 소프트웨어, 펌웨어 또는 하드웨어로 구현될 수 있으며, 소프트웨어로 구현될 경우, 실시간 네트워크 오퍼레이팅 시스템들에 의해 사용된 상이한 플랫폼들 상에 상주하고 그 상이한 플랫폼들로부터 동작되도록 다운로드될 수 있음을 유의해야 한다.

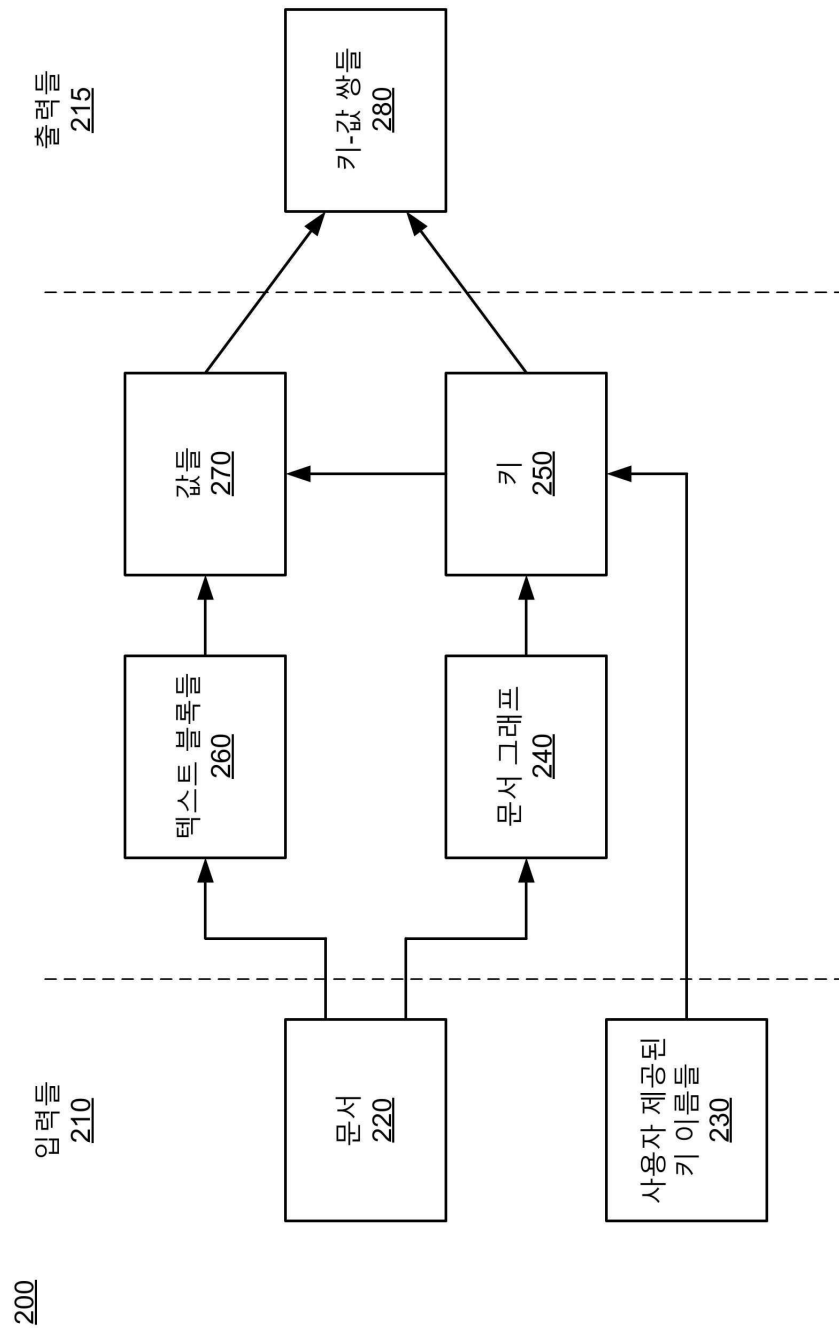
[0074] 마지막으로, 명세서에서 사용된 언어는 주로 관독성 및 교시적 목적들을 위해 선택되었고, 발명의 청구물을 기술하거나 제한하도록 선택되지 않았을 수도 있음을 유의해야 한다. 따라서, 실시형태들의 개시는 예시적인 것으로 의도되며, 제한하는 것으로 의도되지 않는다.

도면

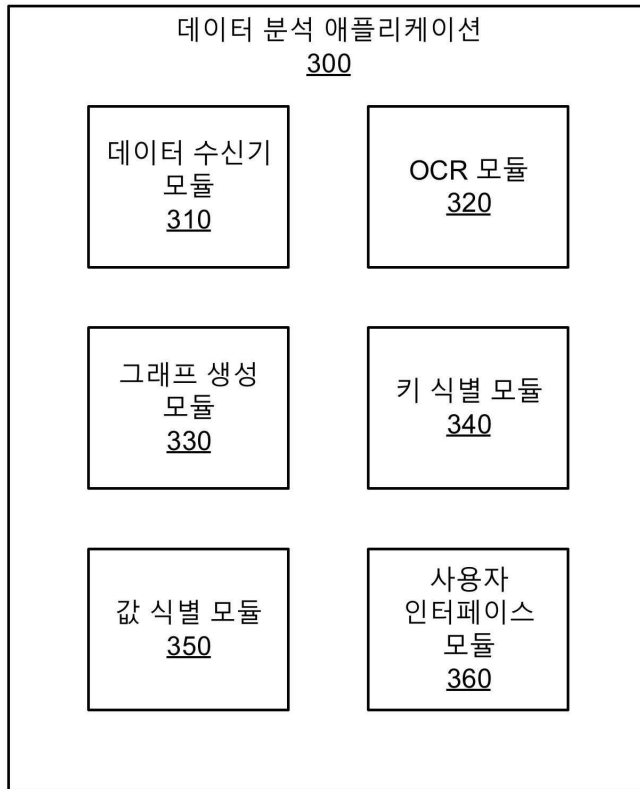
도면1



도면2



도면3



도면4a

Work Request

WLB

Weber und Less AG

Number: EWR-001-1001

Revision: 1

Type: Change

Date: 15.12.2011

Description: We need to reduce the weight of the Carburetor by 10%

Urgency: processing required

Change kind: Functional improvement

Interchangeability: Upward only

Requesting person

Name: Hannan Kaller

Company: Weber und Less AG

Mail address: Hannan.Kaller@weberundless.com

Phone Number: +49-6-143-71462-223

Person in charge

Name: Alexander Koch

Company: Weber und Less AG

Mail address: Alexander.Koch@weberundless.com

Phone Number: +49-6-143-71462-02

Released

Rejected

Signature:

Affected Objects

Type: Number

Item: 10035

Number: 10035

Name: CARBURETOR

Version: 1

Revision: 0

도면4b

421A 421B 423A

Work Request



421C	Number/EWA: 001-0001	Revised: 1	Type: Change	Date: 15.12.2011
423B	Description: need to revise the work order			
	Urgency: processing required			
	Change kind: optional improvement			
421D	Interchangeability: upward only			
	Company: Webber und Less AG			
	Phone Number: 49-0143-71462-02			
Requesting person				
Name: Hannah Kesh				
Mail address: hannah.kesh@webberundless.com				
Person in charge				
Name: Alexander Kesh				
Mail address: Alexander.Kesh@webberundless.com				
Phone Number: 49-0143-71462-02				
Comments:				
Released: ☐ Rejected: ☐				
Signature:				
Affected Objects				
Type		Number		Revision
		100046		1
Item		CARBURATOR		0

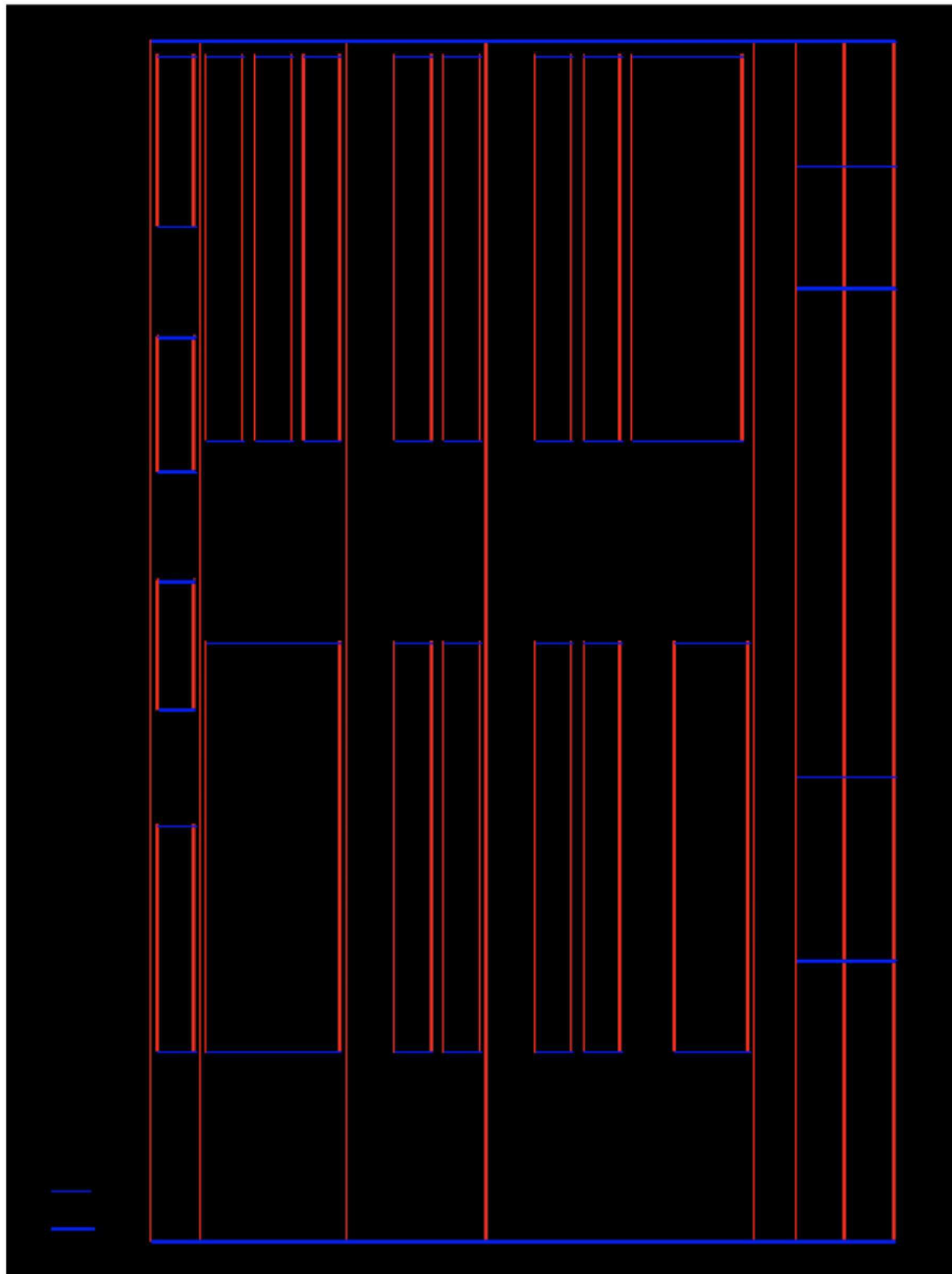
Work Request



Weber und Less AG

Number EWR-001-100		Revision: 1	Type Change	Date: 15.12.2011
Description: We need to reduce the weight of the Carburetor by 10%			Urgency: processing required	
			Change kind: Functional improvement	
			Interchangeability: Upward only	
Requesting person				
Name: Hannah Keller			Company: Weber und Less AG	
Mail address: Hannah.Keller@weberundless.com			Phone Number: +49-6143-71462-223	
Person in charge				
Name: Alexander Koch			Company: Weber und Less AG	
Mail address: Alexander.Koch@weberundless.com			Phone Number: +49-6143-71462-02	
Comments: Released ☐ Rejected ☐			Signature:	
Affected Objects				
Type	Number	Name	Version	Revision
Item	10035	CARBURETOR	1	0

도면4d



도면4e



Work Request

NumberEWR-001-1001

Revision: 1

Type: Change

Date: 15.12.2011

Description: We need to reduce the weight of the Carburetor by 10%

Urgency: processing required

Change kind: Functional improvement

Interchangeability: Upward only

Requesting person

Name: Hannah Keller

Company: Weber und Less AG

Mail address: Hannah.Keller@weberundless.com

Phone Number: +49-6143-71462-223

Person in charge

Name: Alexander Koch

Company: Weber und Less AG

Mail address: Alexander.Koch@weberundless.com

Phone Number: +49-6143-71462-02

Comments:

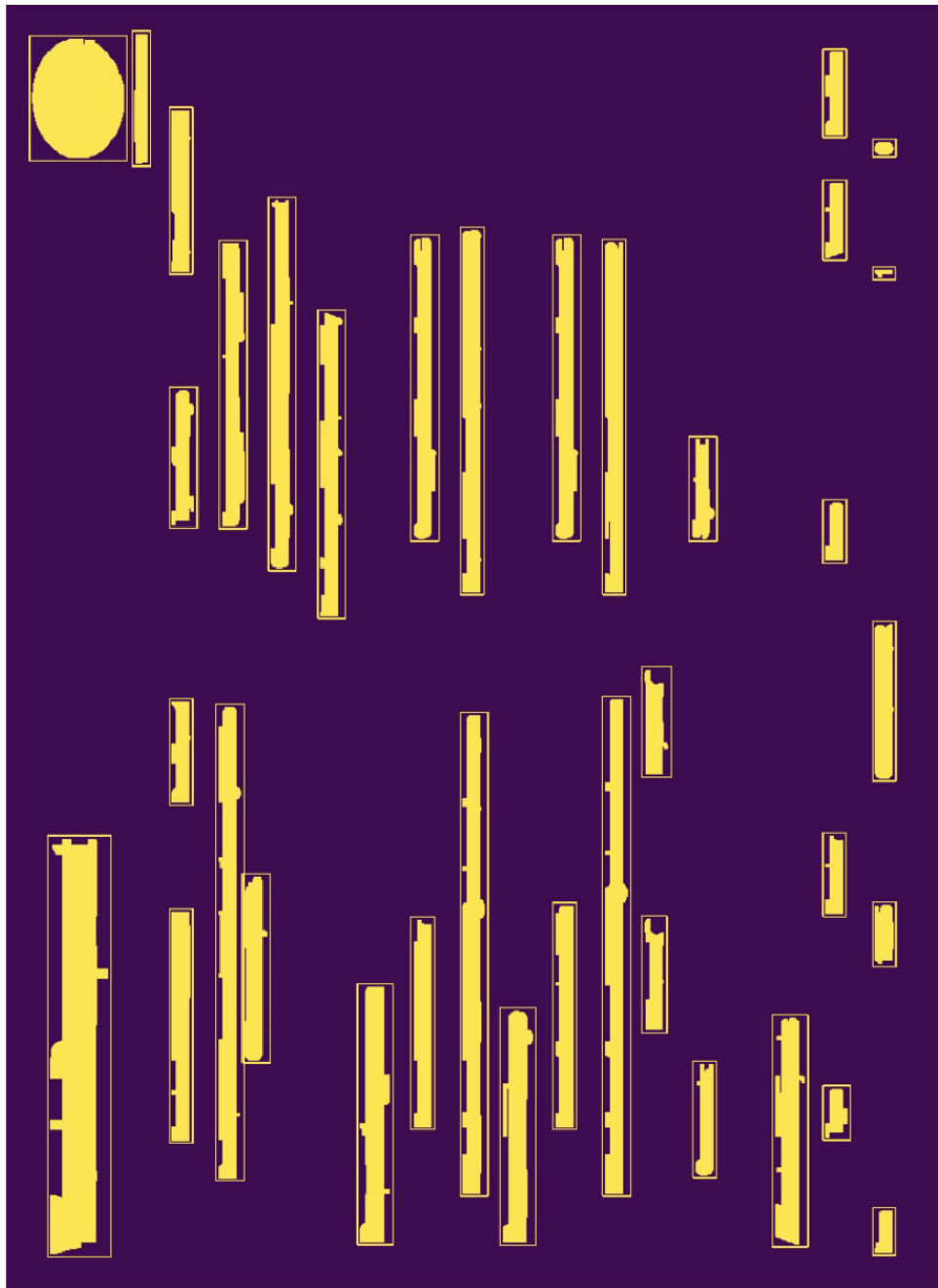
Signature:

Released ☐ Rejected ☒

Affected Objects

Item	Type	Number	Name	Version	Revision
		10035	CARBURETOR	1	0

도면4f



도면4g

Work Request

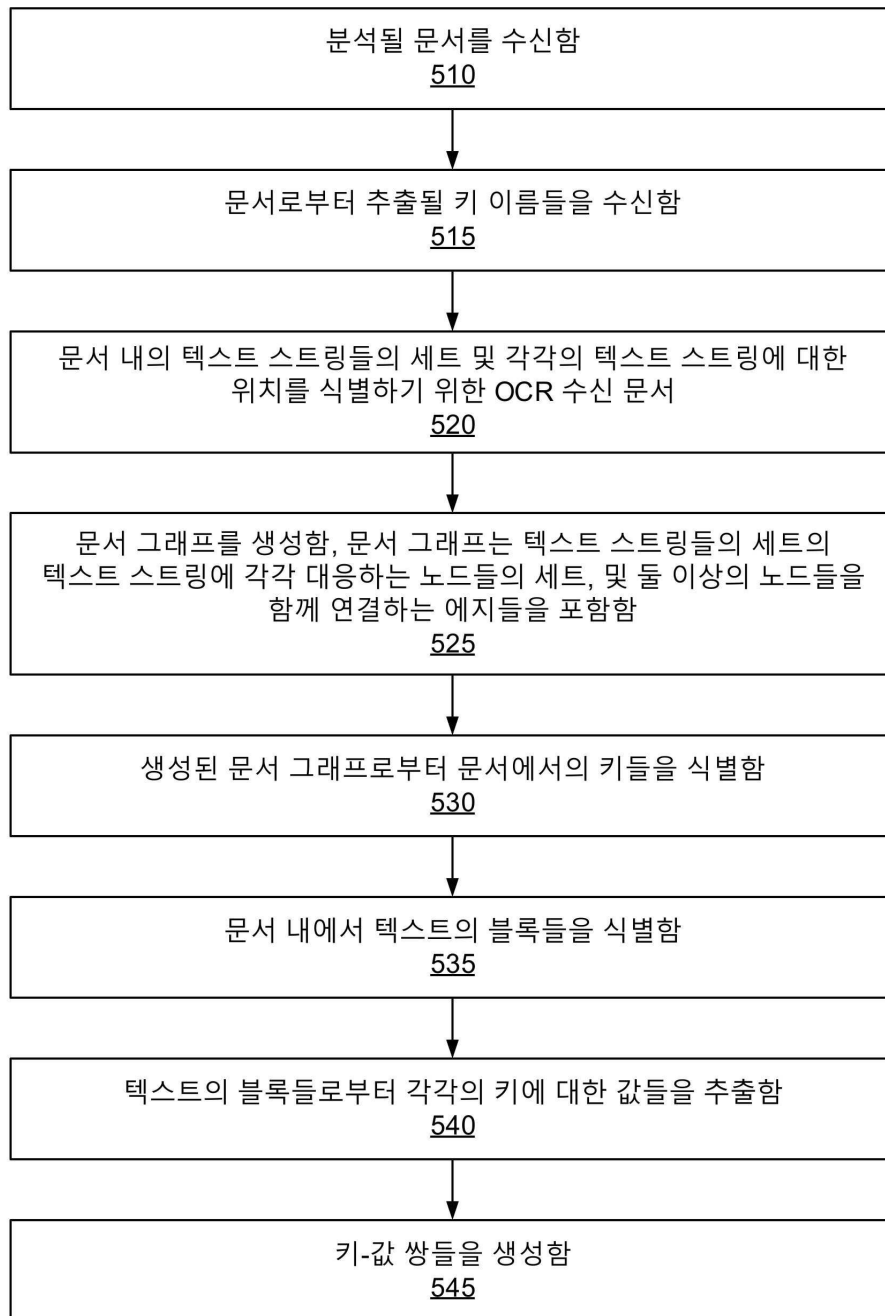


471A

471B

Number: EWR-001-1001	Revision: 1	Type: Change	Date: 15.12.2011
Description: We need to reduce the weight of the Carburetor by 10%		Urgency: processing required Change kind: Functional improvement Interchangeability: Upward only	
Requesting person			
Name: Hannah Keller	Company: Weber und Less AG		
Mail address: Hannah.Keller@weberundless.com	Phone Number: +49-6143-71462-223		
Person in charge			
Name: Alexander Koch	Company: Weber und Less AG		
Mail address: Alexander.Koch@weberundless.com	Phone Number: +49-6143-71462-02		
Comments:	Signature:		
Affected Objects			
Item	Type	Number	Name
10035		473B	CARBURETOR 473A
		1	0
		Version	Revision

도면5



도면6

