



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0061110
(43) 공개일자 2023년05월08일

- | | |
|--|--|
| <p>(51) 국제특허분류(Int. Cl.)
 <i>G06Q 50/16</i> (2012.01) <i>G06F 16/583</i> (2019.01)
 <i>G06F 18/00</i> (2023.01) <i>G06F 40/151</i> (2020.01)
 <i>G06F 40/237</i> (2020.01)</p> <p>(52) CPC특허분류
 <i>G06Q 50/16</i> (2013.01)
 <i>G06F 16/5846</i> (2019.01)</p> <p>(21) 출원번호 10-2021-0146022
 (22) 출원일자 2021년10월28일
 심사청구일자 2021년10월28일</p> | <p>(71) 출원인
 고려대학교 산학협력단
 서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)</p> <p>(72) 발명자
 최진영
 서울특별시 노원구 동일로248길 16, 506동 1003호 (수락파크빌아파트)</p> <p>김현조
 서울특별시 성북구 안암로 145, 산학관 312-2</p> <p>(74) 대리인
 권성현, 유광철, 백두진, 강일신, 김정연</p> |
|--|--|

전체 청구항 수 : 총 9 항

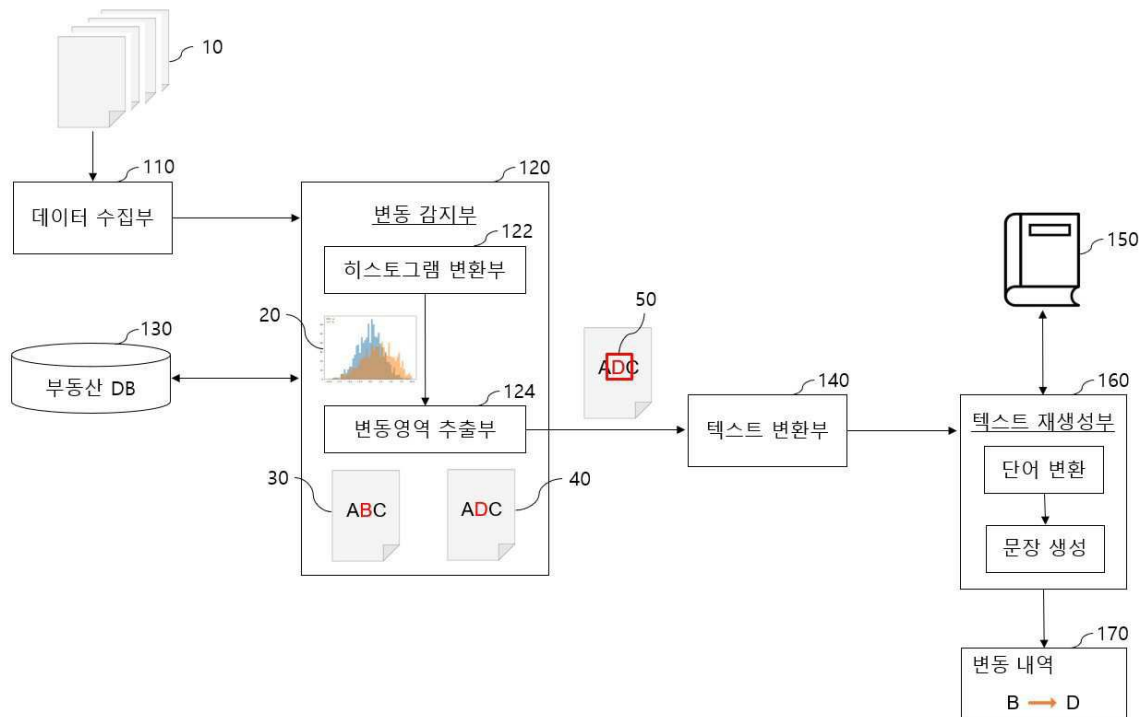
(54) 발명의 명칭 **부동산 문서의 메타 데이터 수집 및 문서 이미지 히스토그램을 활용한 부동산 문서 추적 항목 변동 감지 시스템 및 방법**

(57) 요약

부동산 문서의 메타 데이터 수집 및 문서 이미지 히스토그램을 활용하여 부동산 문서의 변동을 추적하고 변동 항목을 감지하는 부동산 문서 추적 항목 변동 감지 시스템 및 방법이 개시된다. 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 시스템은 부동산 매물 관련 웹사이트에서 부동산 매물에 관련된 등기부등본,

(뒷면에 계속)

대표도 - 도1



토지대장, 및 건축물 관리 대장 중의 적어도 하나를 포함하는 PDF 파일 또는 이미지 형식 파일의 부동산 관련 문서를 정기적으로 수집하도록 구성되는 데이터 수집부; 부동산 데이터베이스에서 상기 부동산 매물과 관련된 부동산 데이터를 독출하고, 상기 부동산 데이터와 상기 부동산 관련 문서를 각각 히스토그램 변환하여 대조함으로써 변동 부분을 추출하도록 구성되는 변동 감지부; 및 상기 부동산 관련 문서 중 상기 변동 부분에서 텍스트 영역을 광학 문자 인식을 기반으로 텍스트화하여 부동산 메타 데이터를 텍스트로 변환하여 추출하도록 구성되는 텍스트 변환부를 포함할 수 있다.

(52) CPC특허분류

G06F 40/151 (2020.01)

G06F 40/237 (2022.01)

G06V 30/413 (2022.01)

명세서

청구범위

청구항 1

부동산 매물 관련 웹사이트에서 부동산 매물에 관련된 등기부등본, 토지대장, 및 건축물 관리 대장 중의 적어도 하나를 포함하는 PDF 파일 또는 이미지 형식 파일의 부동산 관련 문서를 정기적으로 수집하도록 구성되는 데이터 수집부;

부동산 데이터베이스에서 상기 부동산 매물과 관련된 부동산 데이터를 독출하고, 상기 부동산 데이터와 상기 부동산 관련 문서를 각각 히스토그램 변환하여 대조함으로써 변동 부분을 추출하도록 구성되는 변동 감지부; 및

상기 부동산 관련 문서 중 상기 변동 부분에서 이미지 영역을 광학 문자 인식을 기반으로 텍스트화하여 부동산 메타 데이터를 텍스트로 변환하여 추출하도록 구성되는 텍스트 변환부를 포함하는 부동산 문서 추적 항목 변동 감지 시스템.

청구항 2

청구항 1에 있어서,

부동산에 관련된 다양한 부동산 전문 용어와, 각 부동산 전문 용어의 유사어로서 상기 각 부동산 전문 용어 보다 낮은 난이도를 가지는 일상 부동산 용어 간의 대응 관계를 나타내도록 구축되는 부동산 용어 사전; 및

상기 부동산 용어 사전을 기반으로, 상기 부동산 메타 데이터의 텍스트 중에서 추출된 부동산 전문 용어를 부동산 일상 용어로 변환하여 출력하도록 구성되는 텍스트 재생성부를 더 포함하는 부동산 문서 추적 항목 변동 감지 시스템.

청구항 3

청구항 2에 있어서,

상기 텍스트 재생성부는:

상기 부동산 용어 사전을 기반으로 변환된 부동산 일상 용어들의 리스트와 기 정의된 문장들 간의 유사도를 분석하고;

상기 유사도를 기준으로 상기 기 정의된 문장들 중에서 상기 부동산 일상 용어들과 관련된 복수개의 후보 문장을 추출하고;

추출된 상기 복수개의 후보 문장에 대해 각각 문장 간결도 및 문장 가독성을 분석하고; 그리고

상기 문장 간결도 및 상기 문장 가독성을 기반으로 단일 문장을 결정하여 출력하도록 구성되는 부동산 문서 추적 항목 변동 감지 시스템.

청구항 4

청구항 1에 있어서,

상기 변동 감지부는:

상기 부동산 데이터 및 상기 부동산 관련 문서의 이미지를 그레이스케일로 변환하여 임계치를 기준으로 노이즈를 제거하고;

노이즈가 제거된 상기 부동산 데이터 및 상기 부동산 관련 문서에 대해 각각 형태학 연산을 통해 표의 프레임을 추출하고;

부동산 데이터 및 부동산 관련 문서에 대해 각각 설정된 커널 조건을 충족하는 데이터 값을 남기고 각 표의 셀에 대해 레이블링을 수행하고;

레이블링된 부동산 데이터 및 부동산 관련 문서의 영역을 매칭하여 각 영역 별로 히스토그램으로 변환하여 일치

여부를 비교함으로써 변동 여부를 판단하고;

상기 부동산 데이터 및 상기 부동산 관련 문서 중 일치하지 않는 영역인 경우, 변동 사항이 있는 영역만 추출하여 상기 텍스트 변환부로 전달하고; 그리고

상기 부동산 데이터 및 상기 부동산 관련 문서의 일치 영역은 텍스트 변환부로 전달하는 프로세스를 중단하는 부동산 문서 추적 항목 변동 감지 시스템.

청구항 5

데이터 수집부에 의해, 부동산 매물 관련 웹사이트에서 부동산 매물에 관련된 등기부등본, 토지대장, 및 건축물 관리 대장 중의 적어도 하나를 포함하는 PDF 파일 또는 이미지 형식 파일의 부동산 관련 문서를 정기적으로 수집하는 단계;

변동 감지부에 의해, 부동산 데이터베이스에서 상기 부동산 매물과 관련된 부동산 데이터를 독출하고, 상기 부동산 데이터와 상기 부동산 관련 문서를 각각 히스토그램 변환하여 대조함으로써 변동 부분을 추출하는 단계; 및

텍스트 변환부에 의해, 상기 부동산 관련 문서 중 상기 변동 부분에서 이미지 영역을 광학 문자 인식을 기반으로 텍스트화하여 부동산 메타 데이터를 텍스트로 변환하여 추출하는 단계를 포함하는 부동산 문서 추적 항목 변동 감지 방법.

청구항 6

청구항 5에 있어서,

부동산에 관련된 다양한 부동산 전문 용어와, 각 부동산 전문 용어의 유사어로서 상기 각 부동산 전문 용어 보다 낮은 난이도를 가지는 일상 부동산 용어 간의 대응 관계를 나타내는 부동산 용어 사전을 구축하는 단계; 및

텍스트 재생성부에 의해, 상기 부동산 용어 사전을 기반으로, 상기 부동산 메타 데이터의 텍스트 중에서 추출된 부동산 전문 용어를 부동산 일상 용어로 변환하여 출력하는 단계를 더 포함하는 부동산 문서 추적 항목 변동 감지 방법.

청구항 7

청구항 6에 있어서,

상기 부동산 일상 용어로 변환하여 출력하는 단계는:

상기 부동산 용어 사전을 기반으로 변환된 부동산 일상 용어들의 리스트와 기 정의된 문장들 간의 유사도를 분석하는 단계;

상기 유사도를 기준으로 상기 기 정의된 문장들 중에서 상기 부동산 일상 용어들과 관련된 복수개의 후보 문장을 추출하는 단계;

추출된 상기 복수개의 후보 문장에 대해 각각 문장 간결도 및 문장 가독성을 분석하는 단계; 및

상기 문장 간결도 및 상기 문장 가독성을 기반으로 단일 문장을 결정하여 출력하는 단계를 포함하는 부동산 문서 추적 항목 변동 감지 방법.

청구항 8

청구항 5에 있어서,

상기 변동 부분을 추출하는 단계는:

상기 부동산 데이터 및 상기 부동산 관련 문서의 이미지를 그레이스케일로 변환하여 임계치를 기준으로 노이즈를 제거하는 단계;

노이즈가 제거된 상기 부동산 데이터 및 상기 부동산 관련 문서에 대해 각각 형태학 연산을 통해 표의 프레임을 추출하는 단계;

부동산 데이터 및 부동산 관련 문서에 대해 각각 설정된 커널 조건을 충족하는 데이터 값을 남기고 각 표의 셀

에 대해 레이블링을 수행하는 단계;

레이블링된 부동산 데이터 및 부동산 관련 문서의 영역을 매칭하여 각 영역 별로 히스토그램으로 변환하여 일치 여부를 판단하여 변동 여부를 판단하는 단계;

상기 부동산 데이터 및 상기 부동산 관련 문서 중 일치하지 않는 영역인 경우, 변동 사항이 있는 영역만 추출하여 상기 텍스트 변환부로 전달하는 단계; 및

상기 부동산 데이터 및 상기 부동산 관련 문서의 일치 영역은 텍스트 변환부로 전달하는 프로세스를 중단하는 단계를 포함하는 부동산 문서 추적 항목 변동 감지 방법.

청구항 9

청구항 5 내지 청구항 8 중 어느 한 항의 부동산 문서 추적 항목 변동 감지 방법을 실행시키도록 컴퓨터로 판독 가능한 기록 매체에 기록된 컴퓨터 프로그램.

발명의 설명

기술 분야

[0001] 본 발명은 부동산 문서 추적 항목 변동 감지 시스템 및 방법에 관한 것으로, 보다 상세하게는 부동산 문서의 메타 데이터 수집 및 문서 이미지 히스토그램을 활용한 부동산 문서 추적 항목 변동 감지 시스템 및 방법에 관한 것이다.

배경 기술

[0002] 등기부등본, 토지대장, 건축물 관리 대장과 같은 부동산 관련 문서는 부동산 거래에서 반드시 확인이 필요한 문서이다. 중요하게 고려되고 있다. 이러한 부동산 관련 문서는 사용자가 일일이 인터넷 사이트에 접속하여 수집해야 하므로, 문서를 수집하는데 불편함이 따랐다.

[0003] 부동산 관련 문서와 관련된 다른 문제점은 부동산 관련 문서에서 사용되는 용어는 전문 용어가 다수 포함되어 있어 기반 지식이 없는 일반인이 이해하기 어려운 경우가 많다는 점이다. 이에 따라 부동산 관련 문서에 포함된 어려운 용어를 이해하기 위해 사용자가 별도로 검색을 하거나 전문가 등에게 질의를 해야 하는 등의 불편함이 존재한다.

[0004] 또한, 부동산 관련 문서는 변동 사항에 따른 영향이 큰 정보임에도 불구하고, 사용자가 부동산 매물과 관련된 변동 사항에 대해 즉각적인 알람을 받기 어려우며, 부동산 관련 문서의 변동 사항을 쉽게 파악하기에도 어려움이 있다.

[0005] 부동산 관련 문서는 대부분 특정 양식에 따라 작성되는 문서로, 추적이 필요한 항목 외에 문서에 기본적으로 포함된 정보량이 많으며, 일반 사용자가 필요로 하는 내용을 파악하는데 많은 시간이 소요되고, 가독성이 낮은 문제가 있다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 부동산 문서의 메타 데이터 수집 및 문서 이미지 히스토그램을 활용하여 주기적, 지속적으로 부동산 문서의 변동을 추적하고 변동 항목을 감지할 수 있는 부동산 문서 추적 항목 변동 감지 시스템 및 방법을 제공하기 위한 것이다.

[0007] 또한, 본 발명은 부동산 문서 파일을 정기적으로 업데이트하여 변동 여부를 자동으로 판단하여 알람을 전송할 수 있는 부동산 문서 추적 항목 변동 감지 시스템 및 방법을 제공하기 위한 것이다.

[0008] 또한, 본 발명은 부동산 문서 파일의 변동된 부분만을 추출하고, 부동산 용어 사전을 구축하여 기반 지식이 없는 일반인들이 이해하기 쉬운 용어와 문장으로 변환하여 간결하고 가독성 있게 부동산 문서 변동 항목을 제공하고, 부동산 전문지식이 부족한 일반인들의 전문 분야 문서 이해를 도움으로써 부동산 문서에 대한 접근성을 증대시키기 위한 것이다.

[0009] 또한, 본 발명은 수집된 부동산 문서 파일을 텍스트화하기 전에 변동 여부를 감지하고 변동 항목에 대해서만 텍

스트 변환을 수행하여 부동산 문서 항목 변동 감지 시에 불필요한 시스템 리소스 할당을 줄이기 위한 것이다.

과제의 해결 수단

- [0010] 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 시스템은 부동산 매물 관련 웹사이트에서 부동산 매물에 관련된 등기부등본, 토지대장, 및 건축물 관리 대장 중의 적어도 하나를 포함하는 PDF 파일 또는 이미지 형식 파일의 부동산 관련 문서를 정기적으로 수집하도록 구성되는 데이터 수집부; 부동산 데이터베이스에서 상기 부동산 매물과 관련된 부동산 데이터를 독출하고, 상기 부동산 데이터와 상기 부동산 관련 문서를 각각 히스토그램 변환하여 대조함으로써 변동 부분을 추출하도록 구성되는 변동 감지부; 및 상기 부동산 관련 문서 중 상기 변동 부분에서 이미지 영역을 광학 문자 인식을 기반으로 텍스트화하여 부동산 메타 데이터를 텍스트로 변환하여 추출하도록 구성되는 텍스트 변환부를 포함할 수 있다.
- [0011] 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 시스템은 부동산에 관련된 다양한 부동산 전문 용어와, 각 부동산 전문 용어의 유사어로서 상기 각 부동산 전문 용어 보다 낮은 난이도를 가지는 일상 부동산 용어 간의 대응 관계를 나타내도록 구축되는 부동산 용어 사전; 및 상기 부동산 용어 사전을 기반으로, 상기 부동산 메타 데이터의 텍스트 중에서 추출된 부동산 전문 용어를 부동산 일상 용어로 변환하여 출력하도록 구성되는 텍스트 재생성부를 더 포함할 수 있다.
- [0012] 상기 텍스트 재생성부는 상기 부동산 용어 사전을 기반으로 변환된 부동산 일상 용어들의 리스트와 기 정의된 문장들 간의 유사도를 분석하고; 상기 유사도를 기준으로 상기 기 정의된 문장들 중에서 상기 부동산 일상 용어들과 관련된 복수개의 후보 문장을 추출하고; 추출된 상기 복수개의 후보 문장에 대해 각각 문장 간결도 및 문장 가독성을 분석하고; 그리고 상기 문장 간결도 및 상기 문장 가독성을 기반으로 단일 문장을 결정하여 출력하도록 구성될 수 있다.
- [0013] 상기 변동 감지부는 상기 부동산 데이터 및 상기 부동산 관련 문서의 이미지를 그레이스케일로 변환하여 임계치를 기준으로 노이즈를 제거하고; 노이즈가 제거된 상기 부동산 데이터 및 상기 부동산 관련 문서에 대해 각각 형태학 연산을 통해 표의 프레임을 추출하고; 부동산 데이터 및 부동산 관련 문서에 대해 각각 설정된 커널 조건을 충족하는 데이터 값을 남기고 각 표의 셀에 대해 레이블링을 수행하고; 레이블링된 부동산 데이터 및 부동산 관련 문서의 영역을 매칭하여 각 영역 별로 히스토그램으로 변환하여 일치 여부를 비교함으로써 변동 여부를 판단하고; 상기 부동산 데이터 및 상기 부동산 관련 문서 중 일치하지 않는 영역인 경우, 변동 사항이 있는 영역만 추출하여 상기 텍스트 변환부로 전달하고; 그리고 상기 부동산 데이터 및 상기 부동산 관련 문서의 일치 영역은 텍스트 변환 모듈로 전달하는 프로세스를 중단하도록 구성될 수 있다.
- [0014] 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 방법은 데이터 수집부에 의해, 부동산 매물 관련 웹사이트에서 부동산 매물에 관련된 등기부등본, 토지대장, 및 건축물 관리 대장 중의 적어도 하나를 포함하는 PDF 파일 또는 이미지 형식 파일의 부동산 관련 문서를 정기적으로 수집하는 단계; 변동 감지부에 의해, 부동산 데이터베이스에서 상기 부동산 매물과 관련된 부동산 데이터를 독출하고, 상기 부동산 데이터와 상기 부동산 관련 문서를 각각 히스토그램 변환하여 대조함으로써 변동 부분을 추출하는 단계; 및 텍스트 변환부에 의해, 상기 부동산 관련 문서 중 상기 변동 부분에서 이미지 영역을 광학 문자 인식을 기반으로 텍스트화하여 부동산 메타 데이터를 텍스트로 변환하여 추출하는 단계를 포함한다.
- [0015] 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 방법은 부동산에 관련된 다양한 부동산 전문 용어와, 각 부동산 전문 용어의 유사어로서 상기 각 부동산 전문 용어 보다 낮은 난이도를 가지는 일상 부동산 용어 간의 대응 관계를 나타내는 부동산 용어 사전을 구축하는 단계; 및 텍스트 재생성부에 의해, 상기 부동산 용어 사전을 기반으로, 상기 부동산 메타 데이터의 텍스트 중에서 추출된 부동산 전문 용어를 부동산 일상 용어로 변환하여 출력하는 단계를 더 포함할 수 있다.
- [0016] 상기 부동산 일상 용어로 변환하여 출력하는 단계는 상기 부동산 용어 사전을 기반으로 변환된 부동산 일상 용어들의 리스트와 기 정의된 문장들 간의 유사도를 분석하는 단계; 상기 유사도를 기준으로 상기 기 정의된 문장들 중에서 상기 부동산 일상 용어들과 관련된 복수개의 후보 문장을 추출하는 단계; 추출된 상기 복수개의 후보 문장에 대해 각각 문장 간결도 및 문장 가독성을 분석하는 단계; 및 상기 문장 간결도 및 상기 문장 가독성을 기반으로 단일 문장을 결정하여 출력하는 단계를 포함할 수 있다.
- [0017] 상기 변동 부분을 추출하는 단계는 상기 부동산 데이터 및 상기 부동산 관련 문서의 이미지를 그레이스케일로 변환하여 임계치를 기준으로 노이즈를 제거하는 단계; 노이즈가 제거된 상기 부동산 데이터 및 상기 부동산 관련 문서에 대해 각각 형태학 연산을 통해 표의 프레임을 추출하는 단계; 부동산 데이터 및 부동산 관련 문서에

대해 각각 설정된 커널 조건을 충족하는 데이터 값을 남기고 각 표의 셀에 대해 레이블링을 수행하는 단계; 레이블링된 부동산 데이터 및 부동산 관련 문서의 영역을 매칭하여 각 영역 별로 히스토그램으로 변환하여 일치 여부를 판단하여 변동 여부를 판단하는 단계; 상기 부동산 데이터 및 상기 부동산 관련 문서 중 일치하지 않는 영역인 경우, 변동 사항이 있는 영역만 추출하여 상기 텍스트 변환부로 전달하는 단계; 및 상기 부동산 데이터 및 상기 부동산 관련 문서의 일치 영역은 텍스트 변환 모듈로 전달하는 프로세스를 중단하는 단계를 포함할 수 있다.

[0018] 본 발명의 실시예에 따르면, 상기 부동산 문서 추적 항목 변동 감지 방법을 실행시키도록 컴퓨터로 판독 가능한 기록 매체에 기록된 컴퓨터 프로그램이 제공된다.

발명의 효과

[0019] 본 발명의 실시예에 의하면, 부동산 문서의 메타 데이터 수집 및 문서 이미지 히스토그램을 활용하여 주기적, 지속적으로 부동산 문서의 변동을 추적하고 변동 항목을 감지할 수 있으며, 부동산 문서 파일을 정기적으로 업데이트하여 변동 여부를 자동으로 판단하여 알림을 전송할 수 있다.

[0020] 또한, 본 발명의 실시예에 의하면, 부동산 문서 파일의 변동된 부분만을 추출하고, 부동산 용어 사전을 구축하여 기반 지식이 없는 일반인들이 이해하기 쉬운 용어로 변환하여 간결하고 가독성 있게 부동산 문서 변동 항목을 제공할 수 있으며, 일반인들의 전문 분야 문서 이해를 도움으로써 부동산 문서에 대한 접근성을 증대시킬 수 있다.

[0021] 또한, 본 발명의 실시예에 의하면, 수집된 부동산 문서 파일을 텍스트화하기전에 변동 여부를 감지하고 변동 항목에 대해서만 텍스트 변환을 수행한 후 문장으로 변환하여 사용자에게 변동 항목과 관련된 정보를 제공함으로써, 부동산 문서 항목 변동 감지 시에 불필요한 시스템 리소스가 할당되는 것을 방지할 수 있다.

도면의 간단한 설명

[0022] 도 1은 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 시스템의 구성도이다.

도 2는 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 방법의 순서도이다.

도 3은 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 방법을 설명하기 위한 개념도이다.

도 4는 도 2의 단계 S20을 나타낸 순서도이다.

도 5는 도 2의 단계 S40을 나타낸 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0023] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다.

[0024] 본 명세서에서 사용되는 '~모듈, ~부'는 적어도 하나의 기능이나 동작을 처리하는 단위로서, 예를 들어 소프트웨어, FPGA 또는 하드웨어 구성요소를 의미할 수 있다. '~모듈, ~부'에서 제공하는 기능은 복수의 구성요소에 의해 분리되어 수행되거나, 다른 추가적인 구성요소와 통합될 수도 있다. 본 명세서의 '~모듈, ~부'는 반드시 소프트웨어 또는 하드웨어에 한정되지 않으며, 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고, 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 이하에서는 도면을 참조하여 본 발명의 실시예에 대해서 구체적으로 설명하기로 한다.

[0025] 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 시스템 및 방법은 PDF 파일이나 이미지 형식 파일로 된 부동산 데이터의 항목 변동을 추적하고, 변동 부분을 텍스트화하여 부동산 메타 데이터를 추출한 후 이를 전문적인 지식이 없는 일반인이 이해하기 쉬운 형태의 문장으로 변환하여 출력하는 기능을 제공할 수 있다.

[0026] 도 1은 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 시스템의 구성도이다. 도 2는 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 방법의 순서도이다. 도 3은 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 방법을 설명하기 위한 개념도이다.

- [0027] 도 1 내지 도 3을 참조하면, 본 발명의 실시예에 따른 부동산 문서 추적 항목 변동 감지 시스템(100)은 데이터 수집부(110), 변동 감지부(120), 부동산 데이터베이스(130), 텍스트 변환부(140), 부동산 용어 사전(150), 텍스트 재생성부(160), 및 알림부(170)를 포함할 수 있다.
- [0028] 데이터 수집부(110)는 부동산 매물 관련 웹사이트에서 부동산 매물에 관련된 등기부등본, 토지대장, 및/또는 건축물 관리 대장을 포함하는 PDF 파일 또는 이미지 형식(jpg, jpeg, png, tiff 등) 파일의 부동산 관련 문서(10)를 정기적으로 수집할 수 있다(S10).
- [0029] 데이터 수집부(110)는 크롤러(crawler)에 의해 부동산 매물 정보를 제공하는 대법원 사이트와 같은 부동산 관련 웹사이트에서 웹 크롤링(web crawling)에 의해 PDF 파일, 이미지 형식 파일 등의 부동산 관련 문서(10)를 주기적으로 수집할 수 있다.
- [0030] 변동 감지부(120)는 문서 이미지 히스토그램을 기반으로, 기존에 보유하고 있는 부동산 데이터와 새롭게 업로드된 부동산 데이터를 비교하여 일치 여부를 판단하고, 부동산 데이터에 변동 부분이 있는 경우 변동 영역만을 추출하여 텍스트 변환부(140)로 전달할 수 있다(S20).
- [0031] 변동 감지부(120)의 히스토그램 변환부(122)는 부동산 데이터베이스(130)에서 부동산 매물과 관련된 부동산 데이터를 독출하고, 부동산 데이터베이스(130)에서 독출된 부동산 데이터와 데이터 수집부(110)에 의해 수집된 부동산 관련 문서를 각각 히스토그램(histogram) 데이터(20)로 변환할 수 있다.
- [0032] 히스토그램 변환부(122)는 부동산 데이터와 부동산 관련 문서의 대응되는 영역 별로, 영역 내 이미지의 픽셀값들에 대해 X축(가로축) 및/또는 Y축(세로축) 방향을 기준으로 하는 히스토그램 분포를 측정함으로써 히스토그램 변환 처리를 수행할 수 있다.
- [0033] 변동 감지부(120)의 변동 영역 추출부(124)는 부동산 데이터(기존 부동산 관련 문서)로부터 변환된 제1 히스토그램 데이터(30)와, 데이터 수집부(110)에 의해 수집된 부동산 관련 문서(10)로부터 변환된 제2 히스토그램 데이터(40)를 비교하여 영역 별로 대조함으로써 변동 부분(50)을 추출할 수 있다.
- [0034] 데이터 수집부(110)에 의해 수집된 부동산 관련 문서와, 변동 감지부(120)에 의해 부동산 관련 문서로부터 변환된 히스토그램 데이터는 부동산 데이터베이스(130)에 갱신되어 저장될 수 있다.
- [0035] 텍스트 변환부(140)는 데이터 수집부(110)에 의해 수집된 부동산 관련 문서 중 변동 부분(50)에서 이미지 형식으로 된 영역을 광학 문자 인식(OCR ; Optical Character Recognition)을 기반으로 텍스트화하여 부동산 메타데이터를 텍스트로 변환하여 추출할 수 있다(S30).
- [0036] 광학 문자 인식 기술은 알려진 바와 같이, 전처리, 텍스트 영역 추출, 텍스트 인식 등의 처리를 통해 수행될 수 있다. 전처리 과정에서는 히스토그램 평활화(명암 대비 개선), 이진화(이미지의 픽셀값을 0 ~ 255로 나타내고, 이를 0과 1로 변환하여 의미 없는 영역을 제거) 등을 수행하여 이미지 분석에 용이한 상태로 이미지 품질을 개선할 수 있다.
- [0037] 텍스트 영역 인식은 기계학습(machine learning) 기반으로 이루어지며, 전체 이미지 영역에서 텍스트에 해당하는 영역(ROI ; region of interest)을 잘라내는 처리 작업이 수행될 수 있다. 텍스트 인식은 기계학습 기반으로 이루어질 수 있으며, 기계학습을 통해 각 문자들의 특징점을 이용하여 각 문자들이 실제 어떤 문자를 의미하는지 예측할 수 있다. 또한, 텍스트 인식 정확도 향상을 위하여 템플릿 매칭을 통하여 잘못된 문자를 수정하는 처리가 수행될 수 있다.
- [0038] 부동산 용어 사전(150)은 부동산에 관련된 다양한 부동산 전문 용어와, 각 부동산 전문 용어의 유사어에 해당하며 부동산 전문 용어 보다 낮은 난이도를 가지는 일상 부동산 용어 간의 대응 관계를 나타내도록 구축될 수 있다. 부동산 용어 사전(150)의 몇몇 예시는 아래의 표 1과 같다.

표 1

기존 용어 (부동산 전문 용어)	변환 용어 (일상 부동산 용어)
전용면적	개인 사용 면적
분양면적	공용 + 개인 사용 면적
건폐율	토지 내 건축 가능 토지 비율
연면적	토지 내 건축물의 모든 바닥면적의 합
지목	토지의 용도
공시지가	국토교통부가 감정한 토지의 가격

- [0040] 텍스트 재생성부(160)는 부동산 용어 사전(150)을 기반으로, 부동산 메타 데이터의 텍스트 중에서 추출된 어려운 부동산 전문 용어를 보다 간단한 부동산 일상 용어로 변환하고, 자연스러운 문장으로 변환하여 출력할 수 있다(S40).
- [0041] 알람부(170)는 부동산 관련 문서에서 부동산 일상 용어로 변환된 변동 부분(변동 내역)에 관한 정보를 사용자에게 문자메시지, 메일, SNS, 음성 등의 다양한 방식으로 알릴 수 있다. 이때 사용자는 부동산 관련 전문지식이 없더라도, 일반인들이 쉽게 이해할 수 있는 형태로 제공되는 부동산 변동 항목을 알람부(170)에서 제공하는 정보를 통해 손쉽게 파악할 수 있다.
- [0042] 상술한 바와 같은 본 발명의 실시예에 의하면, 부동산 문서의 메타 데이터 수집 및 문서 이미지 히스토그램을 활용하여 주기적, 지속적으로 부동산 문서의 변동을 추적하고 변동 항목을 감지할 수 있으며, 부동산 문서 파일을 정기적으로 업데이트하여 변동 여부를 자동으로 판단하여 알람을 전송할 수 있다.
- [0043] 또한, 본 발명의 실시예에 의하면, 부동산 문서 파일의 변동된 부분만을 추출하고, 부동산 용어 사전을 구축하여 기반 지식이 없는 일반인들이 이해하기 쉬운 용어로 변환하여 간결하고 가독성 있게 부동산 문서 변동 항목을 제공할 수 있으며, 일반인들의 전문 분야 문서 이해를 도움으로써 부동산 문서에 대한 접근성을 증대시킬 수 있다.
- [0044] 또한, 본 발명의 실시예에 의하면, 수집된 부동산 문서 파일을 텍스트화하기전에 변동 여부를 감지하고 변동 항목에 대해서만 텍스트 변환을 수행하여 부동산 문서 항목 변동 감지 시에 불필요한 시스템 리소스 할당을 줄일 수 있다.
- [0045] 도 4는 도 2의 단계 S20을 나타낸 순서도이다. 도 1 내지 도 4를 참조하면, 변동 감지부(120)는 먼저 부동산 데이터베이스(130)에서 독출된 부동산 데이터 및 데이터 수집부(110)에 의해 수집된 부동산 관련 문서의 이미지를 그레이스케일(grayscale)로 변환하여 설정된 임계치(threshold)를 기준으로 노이즈를 제거할 수 있다(S21).
- [0046] 변동 감지부(120)는 노이즈가 제거된 부동산 데이터 및 부동산 관련 문서에 대해 각각 형태학(morphology) 연산을 통해 부동산 관련 문서에 포함된 표의 프레임을 추출한 다음(S22), 설정된 커널 조건을 충족하는 데이터 값을 남기고 각 표의 셀에 대해 레이블링(labeling)을 수행할 수 있다(S23).
- [0047] 변동 감지부(120)는 세로 및 가로 방향으로 설정된 최소 길이를 조건으로 가지는 커널(filter)을 생성하고, 해당 커널의 조건을 충족하는 픽셀의 데이터 값을 남긴 후, 커널에 의해 추출되는 각 영역(표의 셀)에 대하여 레이블링 처리를 수행할 수 있다.
- [0048] 다음으로, 변동 감지부(120)는 레이블링된 부동산 데이터 및 부동산 관련 문서의 각 영역을 매칭하고, 매칭된 각 영역 별로 히스토그램으로 변환하여 일치 여부를 비교함으로써 두 히스토그램 데이터를 비교, 대조하여 부동산 문서의 변동 여부를 판단하고, 변동 부분을 감지할 수 있다(S24).
- [0049] 이에 따라, 부동산 관련 문서에서 변동 항목으로 인해 부동산 관련 문서 내의 데이터 배열이 변경되더라도, 설정된 조건에 따라 부동산 데이터와 부동산 고관련 문서의 대응되는 영역을 매칭시켜 각 영역 별로 히스토그램 변환을 통해 변동 여부를 파악하여 부동산 문서의 변동 항목을 정확하게 감지할 수 있다.
- [0050] 변동 감지부(120)는 부동산 데이터 및 부동산 관련 문서 중 일치하지 않는 영역인 경우, 변동 사항이 있는 영역만 추출하여 텍스트 변환부(140)로 전달할 수 있다(S25). 이와 달리, 변동 감지부(120)는 부동산 데이터 및 부동산 관련 문서의 일치 영역은 텍스트 변환부(140)로 전달하는 프로세스를 중단할 수 있다(S26).
- [0051] 한편, 특정 부동산 매물에 대해 최초로 생성되는 부동산 관련 문서인 경우, 기존 부동산 데이터가 존재하지 않으므로, 최초 생성되는 부동산 관련 문서에 대해서는 변동 감지부(120)를 거치지 않고 텍스트 변환부(140)로 바로 전달되도록 할 수 있다.
- [0052] 도 5는 도 2의 단계 S40을 나타낸 순서도이다. 도 1 내지 도 3 및 도 5를 참조하면, 텍스트 재생성부(160)는 부동산 용어 사전(150)을 기반으로 변환된 부동산 일상 용어들의 리스트와 기 정의된 문장들 간의 유사도를 분석할 수 있다(S42). 이때 부동산 일상 용어들의 리스트와 기 정의된 문장들 간의 유사도는 예를 들어 코사인 유사도(Cosine similarity) 등에 의해 산출될 수 있다.
- [0053] 기 정의된 문장들은 부동산 일상 용어들을 이용하여 부동산과 관련된 정보를 설명하는 문장들로 정의될 수 있다. 이와 같은 기 정의된 문장들은 부동산 정보를 제공하는 서버의 관리자에 의해 입력될 수도 있고, 인터넷

에서 크롤링에 의해 수집될 수도 있다.

- [0054] 텍스트 재생성부(160)는 분석한 유사도를 기준으로 기 정의된 문장들 중에서 부동산 일상 용어들과 관련된 복수개의 후보 문장을 추출할 수 있다(S44). 텍스트 재생성부(160)는 추출된 복수개의 후보 문장에 대해 각각 문장 간결도 및 문장 가독성을 분석한 다음(S46), 분석한 문장 간결도 및 문장 가독성을 기반으로, 복수개의 후보 문장 중에서 단일 문장을 결정하여 출력할 수 있다(S48).
- [0055] 문장 간결도와 문장 가독성은 문장의 길이(문장에 포함된 단어의 개수), 문장의 형태소 분석 결과, 문장 내 쉼표 개수, 문장에 포함된 용어들에 대해 각각 설정된 난이도 값의 합산 값, 합성어 개수, 부정어 개수 등을 기반으로 산출될 수 있다.
- [0056] 이와 같은 본 발명의 실시예에 의하면, 부동산 일상 용어들과 관련되어 정의된 다수의 문장들 중에서 문장 간결도와 문장 가독성이 가장 높은 하나의 문장을 결정하여 출력함으로써, 부동산 관련 전문 지식이 없는 일반인이 부동산 관련 문서의 변동 사항을 쉽게 이해하도록 도울 수 있다.
- [0057] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPGA(Field Programmable Gate Array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다.
- [0058] 처리 장치는 운영 체제 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리요소를 포함할 수 있음을 이해할 것이다.
- [0059] 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing configuration)도 가능하다. 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다.
- [0060] 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.
- [0061] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다.
- [0062] 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CDROM, DVD와 같은 광기록 매체(optical media) 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0063] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 그러므로, 다른 구현들, 다른 실시예들 및 청구범위와 균등한 것들도 후술하는 청구범위의 범위에 속

한다.

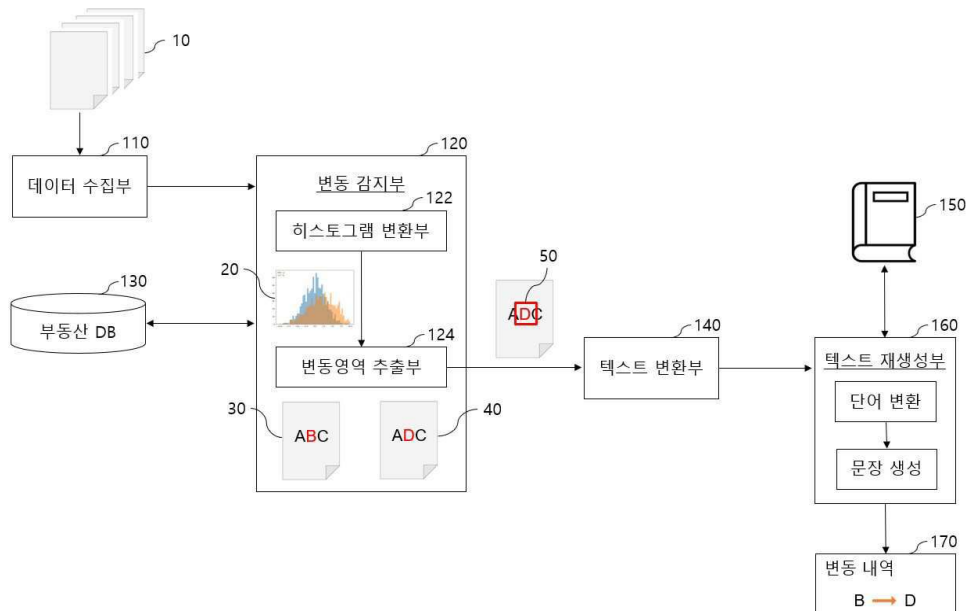
부호의 설명

[0064]

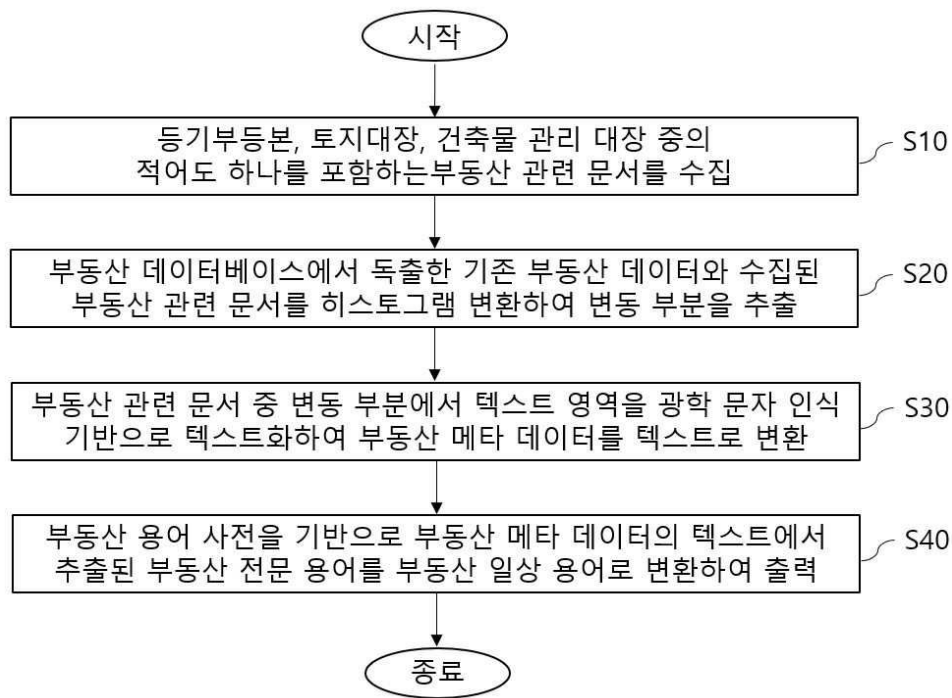
- 10 : 부동산 관련 문서
- 20 : 히스토그램
- 30 : 제1 히스토그램 데이터
- 40 : 제2 히스토그램 데이터
- 50 : 변동 부분
- 100 : 부동산 문서 추적 항목 변동 감지 시스템
- 110 : 데이터 수집부
- 120 : 변동 감지부
- 122 : 히스토그램 변환부
- 124 : 변동 영역 추출부
- 130 : 부동산 데이터베이스
- 140 : 텍스트 변환부
- 150 : 부동산 용어 사전
- 160 : 텍스트 재생성부
- 170 : 알림부

도면

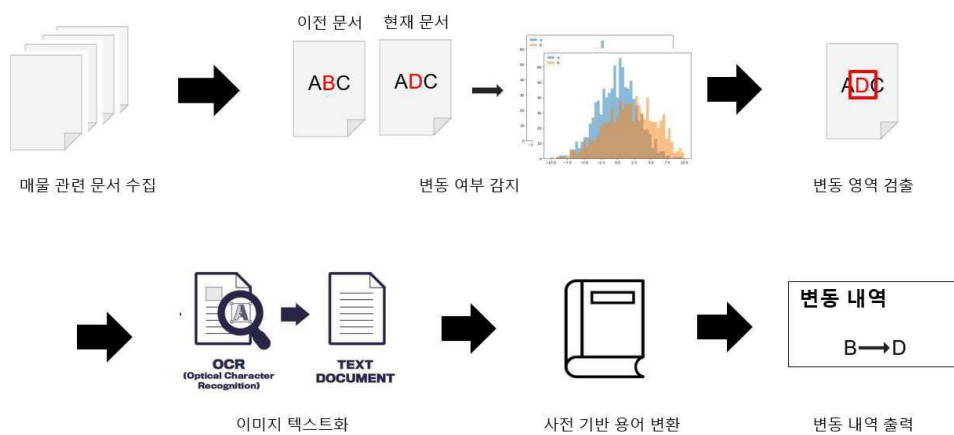
도면1



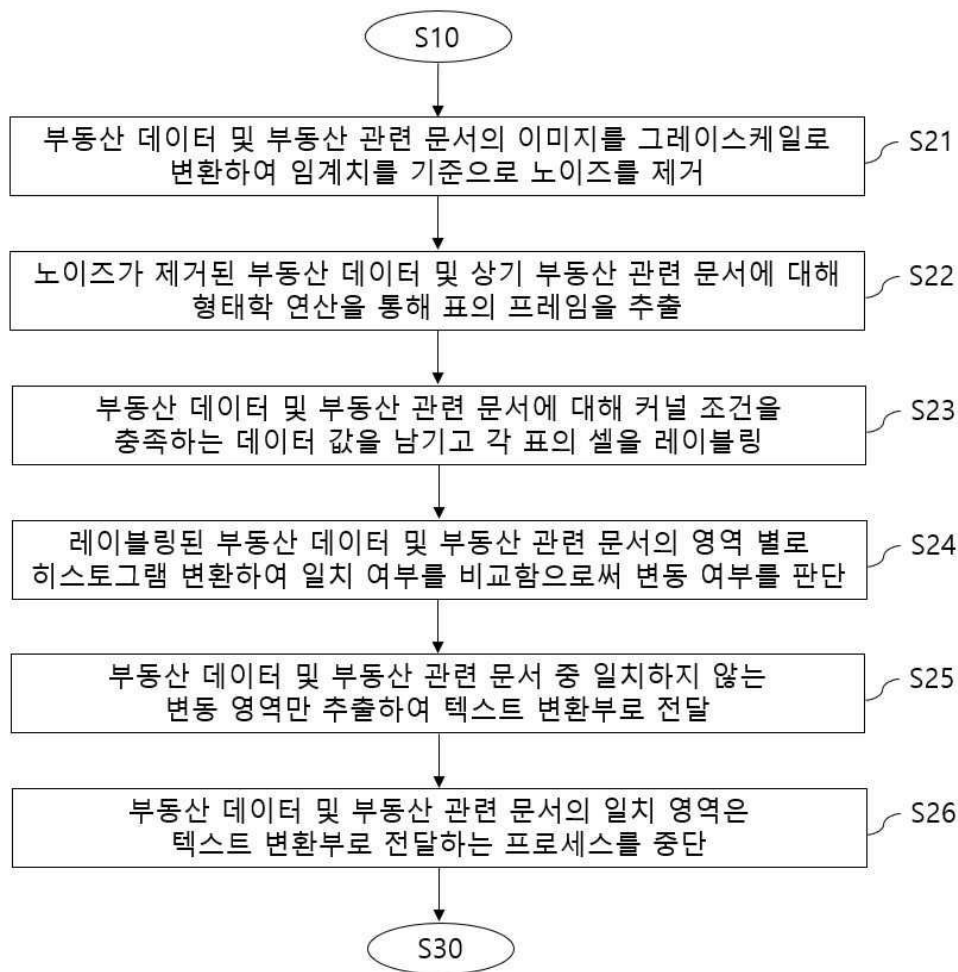
도면2



도면3



도면4



도면5

