



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0097669
(43) 공개일자 2019년08월21일

(51) 국제특허분류(Int. Cl.)

G06F 16/00 (2019.01)

(52) CPC특허분류

G06F 16/93 (2019.01)

G06F 2216/03 (2013.01)

(21) 출원번호 10-2018-0017372

(22) 출원일자 2018년02월13일

심사청구일자 2018년02월13일

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

박상성

서울특별시 성북구 안암로 145 (안암동5가)

김종찬

서울특별시 성북구 안암로 145 (안암동5가)

강지호

서울특별시 성북구 안암로 145 (안암동5가)

(74) 대리인

김연권

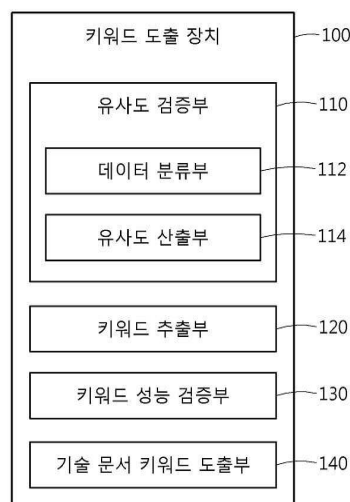
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 기술 문서 키워드를 도출하는 장치 및 방법

(57) 요약

본 발명은 기술 문서 키워드를 도출하는 장치 및 방법을 개시한다. 본 발명의 일실시예에 따르면 기술 문서 키워드를 도출하는 장치는 동일한 검색어에 기반하여 수집된 분석 데이터에 포함되는, 학술 문서와 기술 문서의 유사도를 검증하는 유사도 검증부, 상기 검증된 유사도가 기준 값보다 높을 경우, 상기 학술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 학술 공통 키워드를 추출하는 키워드 추출부 및 상기 추출된 적어도 하나 이상의 학술 공통 키워드를 기술 문서 키워드로 도출하는 기술 문서 키워드 도출부를 포함할 수 있다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 -
부처명 연구지원부
연구관리전문기관 산학협동재단
연구사업명 학술연구비 지원사업
연구과제명 산학협동 연구 활성화를 위한 R&D연계 모형 개발
기 여 율 1/3
주관기관 고려대학교 산학협력단
연구기간 2017.06.01 ~ 2018.05.30

이 발명을 지원한 국가연구개발사업

과제고유번호 2015R1D1A1A01059742
부처명 교육부
연구관리전문기관 한국연구재단
연구사업명 이공학개인지초연구지원-기본연구
연구과제명 R&D 성과확산을 위한 기술사업화 시스템 개발
기 여 율 1/3
주관기관 고려대학교 산학협력단
연구기간 2017.11.01 ~ 2018.10.31

이 발명을 지원한 국가연구개발사업

과제고유번호 F17SN26T4601
부처명 교육부
연구관리전문기관 한국연구재단
연구사업명 BK21플러스사업
연구과제명 BK21PLUS사업 5차년도 빅데이터운용사업팀 국고지원금
기 여 율 1/3
주관기관 고려대학교 산학협력단
연구기간 2017.03.01 ~ 2018.02.28

명세서

청구범위

청구항 1

동일한 검색어에 기반하여 수집된 분석 데이터에 포함되는, 학술 문서와 기술 문서의 유사도를 검증하는 유사도 검증부;

상기 검증된 유사도가 기준 값보다 높을 경우, 상기 학술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 학술 공통 키워드를 추출하는 학술 키워드 추출부;

상기 추출된 적어도 하나 이상의 학술 공통 키워드를 상기 학술 문서의 저자 키워드와 비교하여 키워드 도출 성능을 검증하는 키워드 성능 검증부;

상기 기술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 기술 공통 키워드를 추출하는 기술 키워드 추출부; 및

상기 추출된 적어도 하나 이상의 기술 공통 키워드를 기술 문서 키워드로 도출하는 기술 문서 키워드 도출부를 포함하는

기술 문서 키워드를 도출하는 장치.

청구항 2

제1항에 있어서,

학술 문서 비교 그룹, 기술 문서 비교 그룹 및 학술 및 기술 문서 비교 그룹으로 상기 수집된 분석 데이터를 분류하는 데이터 분류부; 및

상기 학술 문서 비교 그룹에서 학술 문서 간의 제1 유사도를 산출하고, 상기 기술 문서 비교 그룹에서 기술 문서 간의 제2 유사도를 산출하며, 상기 학술 및 기술 문서 비교 그룹에서 학술 문서와 기술 문서 간의 제3 유사도를 산출하는 유사도 산출부를 더 포함하는

기술 문서 키워드를 도출하는 장치.

청구항 3

제2항에 있어서,

상기 유사도 검증부는,

상기 산출된 제1 유사도, 상기 산출된 제2 유사도 및 상기 산출된 제3 유사도의 평균값을 산출하고, 상기 산출된 평균값과 상기 산출된 제1 유사도, 상기 산출된 제2 유사도 및 상기 산출된 제3 유사도 각각을 비교하며, 상기 산출된 평균값과 상기 산출된 제1 유사도, 상기 산출된 제2 유사도 및 상기 산출된 제3 유사도 각각의 차이가 상기 기준 범위에 상응할 경우, 상기 유사도를 상기 기준 값보다 높은 것으로 검증하고, 상기 차이가 상기 기준 범위를 벗어날 경우, 상기 유사도를 상기 기준 값보다 낮은 것으로 검증하는

기술 문서 키워드를 도출하는 장치.

청구항 4

제2항에 있어서,

상기 유사도 산출부는,

상기 학술 문서 비교 그룹에서 복수의 학술 문서의 제1 요약물 추출하고, 상기 추출된 제1 요약물에 대한 텍스트 마이닝을 수행하여 제1 문서 단어 행렬로 정형화하고, 상기 정형화된 제1 문서 단어 행렬을 이용하여 상기 제1 유사도를 산출하는

기술 문서 키워드를 도출하는 장치.

청구항 5

제4항에 있어서,

상기 유사도 산출부는,

상기 기술 문서 비교 그룹에서 복수의 기술 문서의 제2 요약물 추출하고, 상기 추출된 제2 요약물에 대한 텍스트 마이닝을 수행하여 제2 문서 단어 행렬로 정형화하고, 상기 정형화된 제2 문서 단어 행렬을 이용하여 상기 제2 유사도를 산출하는

기술 문서 키워드를 도출하는 장치.

청구항 6

제5항에 있어서,

상기 유사도 산출부는,

상기 학술 및 기술 문서 비교 그룹에서 복수의 학술 문서의 제3 요약물 및 복수의 기술 문서의 제4 요약물 추출하고, 상기 추출된 제3 요약물 및 상기 추출된 제4 요약물에 대한 텍스트 마이닝을 수행하여 제3 문서 단어 행렬로 정형화하고, 상기 정형화된 제3 문서 단어 행렬을 이용하여 상기 제3 유사도를 산출하는

기술 문서 키워드를 도출하는 장치.

청구항 7

제6항에 있어서,

상기 유사도 산출부는,

상기 제1 문서 단어 행렬의 행 값과 열 값을 코사인 거리(cosine distance) 수학적식에 적용하여 상기 제1 유사도를 산출하고,

상기 제2 문서 단어 행렬의 행 값과 열 값을 상기 코사인 거리(cosine distance) 수학적식에 적용하여 상기 제2 유사도를 산출하며,

상기 제3 문서 단어 행렬의 행 값과 열 값을 상기 코사인 거리(cosine distance) 수학적식에 적용하여 상기 제3 유사도를 산출하는

기술 문서 키워드를 도출하는 장치.

청구항 8

제1항에 있어서,

상기 키워드 성능 검증부는,

상기 추출된 적어도 하나 이상의 학술 공통 키워드와 상기 저자 키워드 간의 일치되는 키워드 수를 산출하고, 상기 산출된 키워드 수를 상기 추출된 적어도 하나 이상의 학술 공통 키워드의 수로 나눠서 상기 키워드 도출 성능을 검증하는

기술 문서 키워드를 도출하는 장치.

청구항 9

제8항에 있어서,

상기 기술 문서 키워드 도출부는,

상기 적어도 하나 이상의 학술 공통 키워드의 개시 빈도에 기초하여 상기 적어도 하나 이상의 학술 공통 키워드의 순위들을 결정하는

기술 문서 키워드를 도출하는 장치.

청구항 10

제9항에 있어서,

상기 기술 문서 키워드 도출부는,

상기 결정된 순위들에 기초하여 상기 적어도 하나 이상의 학술 공통 키워드 중 노이즈 분류 기준 보다 낮은 순위에 해당하는 학술 공통 키워드를 상기 적어도 하나 이상의 학술 공통 키워드에서 제외하는

기술 문서 키워드를 도출하는 장치.

청구항 11

제1항에 있어서,

상기 학술 문서를 구성하는 복수의 부분은 상기 학술 문서의 요약, 서론 및 결론 중에서 적어도 하나를 포함하고,

상기 기술 문서를 구성하는 복수의 부분은 상기 기술 문서의 요약, 서론, 결론, 청구범위 및 제목 중에서 적어도 하나를 포함하는

기술 문서 키워드를 도출하는 장치.

청구항 12

유사도 검증부에서, 동일한 검색어에 기반하여 수집된 분석 데이터에 포함되는, 학술 문서와 기술 문서의 유사도를 검증하는 단계;

학술 키워드 추출부에서, 상기 검증된 유사도가 기준 값보다 높을 경우, 상기 학술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 학술 공통 키워드를 추출하는 단계;

키워드 성능 검증부에서, 상기 추출된 적어도 하나 이상의 학술 공통 키워드를 상기 학술 문서의 저자 키워드와 비교하여 키워드 도출 성능을 검증하는 단계;

기술 키워드 추출부에서, 상기 기술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 기술 공통 키워드를 추출하는 단계; 및

기술 문서 키워드 도출부에서, 상기 추출된 적어도 하나 이상의 기술 공통 키워드를 기술 문서 키워드로 도출하는 단계를 포함하는

기술 문서 키워드를 도출하는 방법.

청구항 13

제12항에 있어서,

상기 학술 문서와 기술 문서의 유사도를 검증하는 단계는,

데이터 분류부에서, 학술 문서 비교 그룹, 기술 문서 비교 그룹 및 학술 및 기술 문서 비교 그룹으로 상기 수집된 분석 데이터를 분류하는 단계;

유사도 산출부에서, 상기 학술 문서 비교 그룹에서 학술 문서 간의 제1 유사도를 산출하는 단계;

상기 유사도 산출부에서, 상기 기술 문서 비교 그룹에서 기술 문서 간의 제2 유사도를 산출하는 단계; 및

상기 유사도 산출부에서, 상기 학술 및 기술 문서 비교 그룹에서 학술 문서와 기술 문서 간의 제3 유사도를 산출하는 단계를 포함하는

기술 문서 키워드를 도출하는 방법.

청구항 14

제13항에 있어서,

상기 학술 문서와 기술 문서의 유사도를 검증하는 단계는,

상기 산출된 제1 유사도, 상기 산출된 제2 유사도 및 상기 산출된 제3 유사도의 평균값을 산출하는 단계;

상기 산출된 평균값과 상기 산출된 제1 유사도, 상기 산출된 제2 유사도 및 상기 산출된 제3 유사도 각각을 비교하는 단계;

상기 산출된 평균값과 상기 산출된 제1 유사도, 상기 산출된 제2 유사도 및 상기 산출된 제3 유사도 각각의 차이가 상기 기준 범위에 상응할 경우, 상기 유사도를 상기 기준 값보다 높은 것으로 검증하는 단계; 및

상기 차이가 상기 기준 범위를 벗어날 경우, 상기 유사도를 상기 기준 값보다 낮은 것으로 검증하는 단계를 포함하는

기술 문서 키워드를 도출하는 방법.

청구항 15

제13항에 있어서,

상기 제1 유사도를 산출하는 단계는,

상기 학술 문서 비교 그룹에서 복수의 학술 문서의 제1 요약 추출하는 단계;

상기 추출된 제1 요약에 대한 텍스트 마이닝을 수행하여 제1 문서 단어 행렬을 정형화하는 단계; 및

상기 정형화된 제1 문서 단어 행렬을 이용하여 상기 제1 유사도를 산출하는 단계를 포함하고,

상기 제2 유사도를 산출하는 단계는,

상기 기술 문서 비교 그룹에서 복수의 기술 문서의 제2 요약 추출하는 단계;

상기 추출된 제2 요약에 대한 텍스트 마이닝을 수행하여 제2 문서 단어 행렬을 정형화하는 단계; 및

상기 정형화된 제2 문서 단어 행렬을 이용하여 상기 제2 유사도를 산출하는 단계를 포함하며,

상기 제3 유사도를 산출하는 단계는,

상기 학술 및 기술 문서 비교 그룹에서 복수의 학술 문서의 제3 요약 및 복수의 기술 문서의 제4 요약을 추출하는 단계;

상기 추출된 제3 요약 및 상기 추출된 제4 요약에 대한 텍스트 마이닝을 수행하여 제3 문서 단어 행렬을 정형화

하는 단계; 및

상기 정형화된 제3 문서 단어 행렬을 이용하여 상기 제3 유사도를 산출하는 단계를 포함하는

기술 문서 키워드를 도출하는 방법.

발명의 설명

기술 분야

- [0001] 본 발명은 특허 데이터를 효율적으로 검색하기 위한, 기술 문서 키워드를 도출하는 기술적 사상에 관한 것으로, 더욱 상세하게는 학술 문서와 기술 문서의 유사도를 검증한 후, 학술 문서의 키워드에 기초하여 기술 문서 키워드를 도출하는 장치 및 방법에 관한 것이다.

배경 기술

- [0003] 정부와 기업들은 신기술 개발을 통해 경쟁우위를 얻기 위해 노력하고 있다.
- [0004] 그러나, 신기술 개발은 성공했을 경우에 얻어지는 이익만큼이나 실패했을 경우에 입는 손실이 막대하다.
- [0005] 따라서 올바른 기술 개발 방향을 설정하고 실패 위험을 최소화하기 위한 기술 R&D 전략 수립은 신기술 개발의 필수적인 요소이다.
- [0006] 이러한 기술 R&D 전략 수립을 위한 도구로 특허 분석이 활용되고 있다. 특히 지속 가능한 기술 경영을 위한 정량적인 특허 분석의 중요성이 증가함에 따라 텍스트 마이닝을 이용한 특허 분석 방법론에 대한 연구가 활발히 수행되고 있다.
- [0007] 텍스트 마이닝을 이용한 단어 기반의 특허 분석은 노이즈 단어 제거 및 키워드 추출 과정이 선행된다.
- [0008] 따라서, 추출된 키워드는 후속 분석에 절대적인 영향을 미치며, 추출된 키워드에 대한 검증 없이는 후속 분석의 결과를 신뢰할 수 없다.
- [0009] 그러나, 지금까지 대부분의 연구는 선택된 키워드에 대한 검증 없이 키워드가 분석에 적합하다는 가정 하에 특허 분석을 수행하고 있다.
- [0010] 또한, 키워드 도출 기준은 분석 목적에 따라 달라 질 수 있다. 지금까지는 수집된 특허 데이터 셋에서 각각의 특허들을 기술적 특징에 따라 얼마나 잘 분류 및 군집할 수 있는지를 키워드 선택 기준으로 삼고 있었다.
- [0011] 집단 내의 특허들이 포함하고 있는 단어들을 비교하는 TF-IDF(Term Frequency - Inverse Document Frequency) 가중치를 이용한 방법이 대표적이다.
- [0012] 그러나, 종래 방법으로 선택된 특허 키워드들은 한 특허와 다른 특허와의 비교되는 특징을 나타낼 수는 있지만 그 특허의 고유한 특징을 나타낸다고 하기는 어렵다.
- [0013] 또한, 분석 데이터 안의 특허들을 비교하여 키워드를 선택하기 때문에 분석 범위가 정해져 있지 않은 특허 분석에서는 기존의 키워드 선택 기준은 적합하지 않다.
- [0014] 즉, 다른 특허와의 비교 없이, 단일 특허의 기술적 특징을 나타낼 수 있는 키워드 선택 기준 및 도출 방법이 필요하다.

선행기술문헌

특허문헌

- [0016] (특허문헌 0001) 한국등록특허 제10-1600870호, "통계적 방법을 이용한 핵심 키워드 추출 방법"
- (특허문헌 0002) 한국공개특허 제10-2009-0033728호, "컨텐츠 요약 정보 제공 방법 및 그 장치"

(특허문헌 0003) 한국등록특허 제10-1623860호, "문서 요소에 대한 유사도를 산출하는 방법"

(특허문헌 0004) 한국등록특허 제10-1505546호, "텍스트 마이닝을 이용한 키워드 도출 방법"

(특허문헌 0005) 미국공개특허 2012/0330955, "DOCUMENT SIMILARITY CALCULATION DEVICE"

(특허문헌 0006) 미국공개특허 2015/0234835, "KEYWORD ASSESSMENT"

발명의 내용

해결하려는 과제

- [0017] 본 발명은 기술 문서 키워드를 도출하는 장치 및 방법을 제공하는 것을 목적으로 한다.
- [0018] 본 발명은 학술 문서와 기술 문서의 유사도 검증에 기초하여 학술 문서의 저자 키워드와 학술 문서의 구성에서 추출된 학술 공통 키워드를 비교하여 학술 공통 키워드에 대한 성능을 검증하고, 이를 기초하여 기술 문서의 구성에서 추출된 기술 공통 키워드를 기술 문서 키워드로 도출하는 것을 목적으로 한다.
- [0019] 본 발명은 텍스트 마이닝 기반의 통계적 검증을 이용하여 도출된 기술 문서 키워드의 신뢰성을 향상시키는 것을 목적으로 한다.
- [0020] 본 발명은 텍스트 마이닝 기반의 통계적 검증을 이용하여 학술 문서와 기술 문서의 유사도를 검증함으로써 도출된 기술 문서 키워드의 타당성을 향상시키는 것을 목적으로 한다.

과제의 해결 수단

- [0022] 본 발명의 일실시예에 따르면 기술 문서 키워드를 도출하는 장치는 동일한 검색어에 기반하여 수집된 분석 데이터에 포함되는, 학술 문서와 기술 문서의 유사도를 검증하는 유사도 검증부, 상기 검증된 유사도가 기준 값보다 높을 경우, 상기 학술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 학술 공통 키워드를 추출하는 학술 키워드 추출부, 상기 추출된 적어도 하나 이상의 학술 공통 키워드를 상기 학술 문서의 저자 키워드와 비교하여 키워드 도출 성능을 검증하는 키워드 성능 검증부, 상기 기술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 기술 공통 키워드를 추출하는 기술 키워드 추출부 및 상기 추출된 적어도 하나 이상의 기술 공통 키워드를 기술 문서 키워드로 도출하는 기술 문서 키워드 도출부를 포함할 수 있다.
- [0023] 본 발명의 일실시예에 따르면 기술 문서 키워드를 도출하는 장치는 학술 문서 비교 그룹, 기술 문서 비교 그룹 및 학술 및 기술 문서 비교 그룹으로 상기 수집된 분석 데이터를 분류하는 데이터 분류부 및 상기 학술 문서 비교 그룹에서 학술 문서 간의 제1 유사도를 산출하고, 상기 기술 문서 비교 그룹에서 기술 문서 간의 제2 유사도를 산출하며, 상기 학술 및 기술 문서 비교 그룹에서 학술 문서와 기술 문서 간의 제3 유사도를 산출하는 유사도 산출부를 더 포함할 수 있다.
- [0024] 본 발명의 일실시예에 따르면 상기 유사도 검증부는, 상기 산출된 제1 유사도, 상기 산출된 제2 유사도 및 상기 산출된 제3 유사도의 평균값을 산출하고, 상기 산출된 평균값과 상기 산출된 제1 유사도, 상기 산출된 제2 유사도 및 상기 산출된 제3 유사도 각각을 비교하며, 상기 산출된 평균값과 상기 산출된 제1 유사도, 상기 산출된 제2 유사도 및 상기 산출된 제3 유사도 각각의 차이가 상기 기준 범위에 상응할 경우, 상기 유사도를 상기 기준 값보다 높은 것으로 검증하고, 상기 차이가 상기 기준 범위를 벗어날 경우, 상기 유사도를 상기 기준 값보다 낮은 것으로 검증할 수 있다.
- [0025] 본 발명의 일실시예에 따르면 상기 유사도 산출부는, 상기 학술 문서 비교 그룹에서 복수의 학술 문서의 제1 요약어를 추출하고, 상기 추출된 제1 요약어에 대한 텍스트 마이닝을 수행하여 제1 문서 단어 행렬로 정형화하고, 상기 정형화된 제1 문서 단어 행렬을 이용하여 상기 제1 유사도를 산출할 수 있다.
- [0026] 본 발명의 일실시예에 따르면 상기 유사도 산출부는, 상기 기술 문서 비교 그룹에서 복수의 기술 문서의 제2 요약어를 추출하고, 상기 추출된 제2 요약어에 대한 텍스트 마이닝을 수행하여 제2 문서 단어 행렬로 정형화하고, 상기 정형화된 제2 문서 단어 행렬을 이용하여 상기 제2 유사도를 산출할 수 있다.

- [0027] 본 발명의 일실시예에 따르면 상기 유사도 산출부는, 상기 학술 및 기술 문서 비교 그룹에서 복수의 학술 문서의 제3 요약 및 복수의 기술 문서의 제4 요약을 추출하고, 상기 추출된 제3 요약 및 상기 추출된 제4 요약에 대한 텍스트 마이닝을 수행하여 제3 문서 단어 행렬로 정형화하고, 상기 정형화된 제3 문서 단어 행렬을 이용하여 상기 제3 유사도를 산출할 수 있다.
- [0028] 본 발명의 일실시예에 따르면 상기 유사도 산출부는, 상기 제1 문서 단어 행렬의 행 값과 열 값을 코사인 거리(cosine distance) 수학적식에 적용하여 상기 제1 유사도를 산출하고, 상기 제2 문서 단어 행렬의 행 값과 열 값을 상기 코사인 거리(cosine distance) 수학적식에 적용하여 상기 제2 유사도를 산출하며, 상기 제3 문서 단어 행렬의 행 값과 열 값을 상기 코사인 거리(cosine distance) 수학적식에 적용하여 상기 제3 유사도를 산출할 수 있다.
- [0029] 본 발명의 일실시예에 따르면 상기 키워드 성능 검증부는, 상기 추출된 적어도 하나 이상의 학술 공통 키워드와 상기 저자 키워드 간의 일치되는 키워드 수를 산출하고, 상기 산출된 키워드 수를 상기 추출된 적어도 하나 이상의 학술 공통 키워드의 수로 나뉘서 상기 키워드 도출 성능을 검증할 수 있다.
- [0030] 본 발명의 일실시예에 따르면 상기 기술 문서 키워드 도출부는, 상기 적어도 하나 이상의 학술 공통 키워드의 개시 빈도에 기초하여 상기 적어도 하나 이상의 학술 공통 키워드의 순위들을 결정할 수 있다.
- [0031] 본 발명의 일실시예에 따르면 상기 기술 문서 키워드 도출부는, 상기 결정된 순위들에 기초하여 상기 적어도 하나 이상의 학술 공통 키워드 중 노이즈 분류 기준 보다 낮은 순위에 해당하는 학술 공통 키워드를 상기 적어도 하나 이상의 학술 공통 키워드에서 제외할 수 있다.
- [0032] 본 발명의 일실시예에 따르면 상기 학술 문서를 구성하는 복수의 부분은 상기 학술 문서의 요약, 서론 및 결론 중에서 적어도 하나를 포함하고, 상기 기술 문서를 구성하는 복수의 부분은 상기 기술 문서의 요약, 서론, 결론, 청구범위 및 제목 중에서 적어도 하나를 포함할 수 있다.
- [0033] 본 발명의 일실시예에 따르면 기술 문서 키워드를 도출하는 방법은 유사도 검증부에서, 동일한 검색어에 기반하여 수집된 분석 데이터에 포함되는, 학술 문서와 기술 문서의 유사도를 검증하는 단계, 학술 키워드 추출부에서, 상기 검증된 유사도가 기준 값보다 높을 경우, 상기 학술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 학술 공통 키워드를 추출하는 단계, 키워드 성능 검증부에서, 상기 추출된 적어도 하나 이상의 학술 공통 키워드를 상기 학술 문서의 저자 키워드와 비교하여 키워드 도출 성능을 검증하는 단계, 기술 키워드 추출부에서, 상기 기술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 기술 공통 키워드를 추출하는 단계 및 기술 문서 키워드 도출부에서, 상기 추출된 적어도 하나 이상의 기술 공통 키워드를 기술 문서 키워드로 도출하는 단계를 포함할 수 있다.
- [0034] 본 발명의 일실시예에 따르면 상기 학술 문서와 기술 문서의 유사도를 검증하는 단계는, 데이터 분류부에서, 학술 문서 비교 그룹, 기술 문서 비교 그룹 및 학술 및 기술 문서 비교 그룹으로 상기 수집된 분석 데이터를 분류하는 단계, 유사도 산출부에서, 상기 학술 문서 비교 그룹에서 학술 문서 간의 제1 유사도를 산출하는 단계, 상기 유사도 산출부에서, 상기 기술 문서 비교 그룹에서 기술 문서 간의 제2 유사도를 산출하는 단계 및 상기 유사도 산출부에서, 상기 학술 및 기술 문서 비교 그룹에서 학술 문서와 기술 문서 간의 제3 유사도를 산출하는 단계를 더 포함할 수 있다.
- [0035] 본 발명의 일실시예에 따르면 상기 학술 문서와 기술 문서의 유사도를 검증하는 단계는, 상기 산출된 제1 유사도, 상기 산출된 제2 유사도 및 상기 산출된 제3 유사도의 평균값을 산출하는 단계, 상기 산출된 평균값과 상기 산출된 제1 유사도, 상기 산출된 제2 유사도 및 상기 산출된 제3 유사도 각각을 비교하는 단계, 상기 산출된 평균값과 상기 산출된 제1 유사도, 상기 산출된 제2 유사도 및 상기 산출된 제3 유사도 각각의 차이가 상기 기준 범위에 상응할 경우, 상기 유사도를 상기 기준 값보다 높은 것으로 검증하는 단계 및 상기 차이가 상기 기준 범위를 벗어날 경우, 상기 유사도를 상기 기준 값보다 낮은 것으로 검증하는 단계를 포함할 수 있다.
- [0036] 본 발명의 일실시예에 따르면 상기 제1 유사도를 산출하는 단계는, 상기 학술 문서 비교 그룹에서 복수의 학술 문서의 제1 요약을 추출하는 단계, 상기 추출된 제1 요약에 대한 텍스트 마이닝을 수행하여 제1 문서 단어 행렬을 정형화하는 단계 및 상기 정형화된 제1 문서 단어 행렬을 이용하여 상기 제1 유사도를 산출하는 단계를 포함하고, 상기 제2 유사도를 산출하는 단계는, 상기 기술 문서 비교 그룹에서 복수의 기술 문서의 제2 요약을 추출하는 단계, 상기 추출된 제2 요약에 대한 텍스트 마이닝을 수행하여 제2 문서 단어 행렬을 정형화하는 단계 및 상기 정형화된 제2 문서 단어 행렬을 이용하여 상기 제2 유사도를 산출하는 단계를 포함하며, 상기 제3 유사도를 산출하는 단계는, 상기 학술 및 기술 문서 비교 그룹에서 복수의 학술 문서의 제3 요약 및 복수의 기술 문서

의 제4 요약을 추출하는 단계, 상기 추출된 제3 요약 및 상기 추출된 제4 요약에 대한 텍스트 마이닝을 수행하여 제3 문서 단어 행렬을 정형화하는 단계 및 상기 정형화된 제3 문서 단어 행렬을 이용하여 상기 제3 유사도를 산출하는 단계를 포함할 수 있다.

발명의 효과

- [0038] 본 발명은 학술 문서와 기술 문서의 유사도 검증에 기초하여 학술 문서의 저자 키워드와 학술 문서의 구성에서 추출된 학술 공통 키워드를 비교하여 학술 공통 키워드에 대한 성능을 검증하고, 이를 기초하여 기술 문서의 구성에서 추출된 기술 공통 키워드를 기술 문서 키워드로 도출할 수 있다.
- [0039] 본 발명은 텍스트 마이닝 기반의 통계적 검증을 이용하여 도출된 기술 문서 키워드의 신뢰성을 향상시킬 수 있다.
- [0040] 본 발명은 텍스트 마이닝 기반의 통계적 검증을 이용하여 학술 문서와 기술 문서의 유사도를 검증함으로써 도출된 기술 문서 키워드의 타당성을 향상시킬 수 있다.
- [0041] 본 발명은 학술 문서와 기술 문서의 유사도 검증에 기초하여 기술 문서 키워드를 도출함에 따라 다양한 특허 분석 방법론 및 연구 방향을 제공할 수 있다.

도면의 간단한 설명

- [0043] 도 1은 본 발명의 일실시예에 따른 기술 문서 키워드를 도출하는 장치의 구성 요소를 설명하는 도면이다.
- 도 2 내지 도 5는 본 발명의 일실시예에 따른 기술 문서 키워드를 도출하는 방법과 관련된 흐름도를 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0044] 이하, 본 문서의 다양한 실시 예들이 첨부된 도면을 참조하여 기재된다.
- [0045] 실시 예 및 이에 사용된 용어들은 본 문서에 기재된 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 해당 실시 예의 다양한 변경, 균등물, 및/또는 대체물을 포함하는 것으로 이해되어야 한다.
- [0046] 하기에서 다양한 실시 예들을 설명에 있어 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다.
- [0047] 그리고 후술되는 용어들은 다양한 실시 예들에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.
- [0048] 도면의 설명과 관련하여, 유사한 구성요소에 대해서는 유사한 참조 부호가 사용될 수 있다.
- [0049] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함할 수 있다.
- [0050] 본 문서에서, "A 또는 B" 또는 "A 및/또는 B 중 적어도 하나" 등의 표현은 함께 나열된 항목들의 모든 가능한 조합을 포함할 수 있다.
- [0051] "제1," "제2," "첫째," 또는 "둘째," 등의 표현들은 해당 구성요소들을, 순서 또는 중요도에 상관없이 수식할 수 있고, 한 구성요소를 다른 구성요소와 구분하기 위해 사용될 뿐 해당 구성요소들을 한정하지 않는다.
- [0052] 어떤(예: 제1) 구성요소가 다른(예: 제2) 구성요소에 "(기능적으로 또는 통신적으로) 연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 상기 어떤 구성요소가 상기 다른 구성요소에 직접적으로 연결되거나, 다른 구성요소(예: 제3 구성요소)를 통하여 연결될 수 있다.
- [0053] 본 명세서에서, "~하도록 구성된(또는 설정된)(configured to)"은 상황에 따라, 예를 들면, 하드웨어적 또는 소프트웨어적으로 "~에 적합한," "~하는 능력을 가지는," "~하도록 변경된," "~하도록 만들어진," "~를 할 수 있는," 또는 "~하도록 설계된"과 상호 호환적으로(interchangeably) 사용될 수 있다.

- [0054] 어떤 상황에서는, "~하도록 구성된 장치"라는 표현은, 그 장치가 다른 장치 또는 부품들과 함께 "~할 수 있는" 것을 의미할 수 있다.
- [0055] 예를 들면, 문구 "A, B, 및 C를 수행하도록 구성된(또는 설정된) 프로세서"는 해당 동작을 수행하기 위한 전용 프로세서(예: 임베디드 프로세서), 또는 메모리 장치에 저장된 하나 이상의 소프트웨어 프로그램들을 실행함으로써, 해당 동작들을 수행할 수 있는 범용 프로세서(예: CPU 또는 application processor)를 의미할 수 있다.
- [0056] 또한, '또는'이라는 용어는 배타적 논리합 'exclusive or' 이기보다는 포함적인 논리합 'inclusive or' 를 의미한다.
- [0057] 즉, 달리 언급되지 않는 한 또는 문맥으로부터 명확하지 않는 한, 'x가 a 또는 b를 이용한다'라는 표현은 포함적인 자연 순열들(natural inclusive permutations) 중 어느 하나를 의미한다.
- [0059] 도 1은 본 발명의 일실시예에 따른 기술 문서 키워드를 도출하는 장치의 구성 요소를 설명하는 도면이다.
- [0060] 이하 사용되는 '..부', '..기' 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어, 또는, 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.
- [0061] 본 발명의 일실시예에 따른 기술 문서 키워드를 도출하는 장치는 키워드 도출 장치를 포함한다.
- [0062] 도 1을 참고하면, 키워드 도출 장치(100)는 유사도 검증부(110), 키워드 추출부(120) 및 기술 문서 키워드 도출부(140)를 포함한다.
- [0063] 본 발명의 일실시예에 따르면 유사도 검증부(110)는 동일한 검색어에 기반하여 수집된 분석 데이터가 포함하는 학술 문서와 기술 문서의 유사도를 검증한다.
- [0064] 예를 들어, 학술 문서는 논문(paper)를 포함하고, 기술 문서는 특허(patent) 및 과학 도서를 포함한다.
- [0065] 본 발명의 다른 실시예에 따르면 유사도 검증부(110)는 데이터 분류부(112) 및 유사도 산출부(114)를 포함한다.
- [0066] 일례로, 데이터 분류부(112)는 학술 문서 비교 그룹, 기술 문서 비교 그룹 및 학술 및 기술 문서 비교 그룹으로 동일한 검색어에 기반하여 수집된 분석 데이터를 분류한다.
- [0067] 예를 들어, 학술 문서 비교 그룹은 동일한 검색어에 기반하여 수집된 분석 데이터에 포함되는 학술 문서들을 포함한다.
- [0068] 또한, 기술 문서 비교 그룹은 동일한 검색어에 기반하여 수집된 분석 데이터에 포함되는 기술 문서들을 포함한다.
- [0069] 더하여, 학술 및 기술 문서 비교 그룹은 동일한 검색어 기반하여 수집된 분석 데이터에 포함되는 학술 문서들과 기술 문서들을 포함한다.
- [0070] 일례로, 데이터 분류부(112)는 동일한 검색어를 이용하여 학술 문서 데이터베이스로부터 총 60건의 학술 문서를 수집하고, 동일한 검색어를 이용하여 기술 문서 데이터베이스로부터 총 60건의 기술 문서를 수집할 수 있다.
- [0071] 다음으로, 데이터 분류부(112)는 총 30건의 학술 문서를 학술 문서 비교 그룹으로 분류하고, 총 30건의 기술 문서를 기술 문서 비교 그룹으로 분류하고, 학술 문서 및 기술 문서 각각 30건씩을 학술 및 기술 문서 비교 그룹으로 분류할 수 있다.
- [0072] 일례로, 유사도 산출부(114)는 학술 문서 비교 그룹, 기술 문서 비교 그룹 및 학술 및 기술 문서 비교 그룹 각에서 그룹 내 문서 유사도를 산출할 수 있다.
- [0073] 예를 들어, 유사도 산출부(114)는 학술 문서 비교 그룹에서 학술 문서 간의 제1 유사도를 산출하고, 기술 문서 비교 그룹에서 기술 문서 간의 제2 유사도를 산출하며, 학술 및 기술 문서 비교 그룹에서 학술 문서와 기술 문서 간의 제3 유사도를 산출한다.
- [0074] 일례로, 유사도 산출부(114)는 학술 문서 비교 그룹, 기술 문서 비교 그룹 및 학술 및 기술 문서 비교 그룹 각에서 공통 구성요소에 해당하는 요약 부분을 추출하고 텍스트 마이닝(text mining)을 이용해 문서 단어 행렬 형태로 정형화한다.
- [0075] 본 발명의 일실시예에 따르면 유사도 산출부(114)는 학술 문서 비교 그룹에서 복수의 학술 문서의 제1 요약을

추출한다.

[0076] 다음으로, 유사도 산출부(114)는 추출된 제1 요약에 대한 텍스트 마이닝을 수행하여 제1 문서 단어 행렬로 정형화하고, 상기 정형화된 제1 문서 단어 행렬을 이용하여 상기 제1 유사도를 산출할 수 있다.

[0077] 일례로, 유사도 산출부(114)는 기술 문서 비교 그룹에서 복수의 기술 문서의 제2 요약을 추출하고, 추출된 제2 요약에 대한 텍스트 마이닝을 수행하여 제2 문서 단어 행렬로 정형화하고, 정형화된 제2 문서 단어 행렬을 이용하여 제2 유사도를 산출할 수 있다.

[0078] 또한, 유사도 산출부(114)는 학술 및 기술 문서 비교 그룹에서 복수의 학술 문서의 제3 요약 및 복수의 기술 문서의 제4 요약을 추출한다.

[0079] 다음으로, 유사도 산출부(114)는 추출된 제3 요약 및 추출된 제4 요약에 대한 텍스트 마이닝을 수행하여 제3 문서 단어 행렬로 정형화하고 제3 문서 단어 행렬을 이용하여 제3 유사도를 산출할 수 있다.

[0080] 예를 들어, 유사도 산출부(114)는 하기 표 1과 같이 요약 부분을 문서 단어 행렬 형태로 정형화 할 수 있다.

[0082] [표 1]

	제1단어	제2단어	제3단어	제4단어
문서 1	0	0	2	1
문서 2	1	1	0	0
문서 3	1	2	0	0
문서 4	0	1	0	1

[0083]

[0085] 다음으로, 유사도 산출부(114)는 하기 문서 단어 행렬을 하기 수학식 1에 적용하여 유사도를 산출한다.

[0087] [수학식 1]

$$\text{Cosine similarity} = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

[0088]

[0090] 수학식 1에서, Cosine similarity는 유사도를 나타낼 수 있고, A는 문서 단어 행렬에서 행의 값을 나타낼 수 있고, B는 문서 단어 행렬에서 열의 값을 나타낼 수 있다.

[0091] 예를 들어, 수학식 1은 코사인 거리(cosine distance) 수학식을 포함할 수 있다.

[0092] 일례로, 유사도 산출부(114)는 제1 문서 단어 행렬의 행 값과 열 값을 코사인 거리 수학식에 적용하여 제1 유사도를 산출할 수 있다.

[0093] 또한, 유사도 산출부(114)는 제2 문서 단어 행렬의 행 값과 열 값을 상기 코사인 거리 수학식에 적용하여 제2

유사도를 산출할 수 있다.

[0094] 또한, 유사도 산출부(114)는 제3 문서 단어 행렬의 행 값과 열 값을 코사인 거리 수학적식에 적용하여 제3 유사도를 산출할 수 있다.

[0095] 일례로, 유사도 산출부(114)는 제1 유사도, 제2 유사도 및 제3 유사도를 하기 표 2와 같이 산출할 수 있다.

[0097] [표 2]

그룹	문서 수	유사도	표준편차	표준오차
학술 문서	30	0.6473	0.694	0.0127
기술 문서	30	0.6179	0.0550	0.0100
학술 및 기술 문서	60	0.6450	0.636	0.0082
총	120	0.6388	0.638	0.0058

[0098]

[0100] 본 발명의 일실시예에 따르면 유사도 검증부(110)는 제1 유사도, 제2 유사도 및 제3 유사도의 평균값을 산출한다.

[0101] 다음으로, 유사도 검증부(110)는 산출된 평균값과 제1 유사도, 제2 유사도 및 제3 유사도 각각을 비교하며, 평균값과 제1 유사도, 제2 유사도 및 제3 유사도 각각의 차이가 기준 범위에 상응할 경우, 학술 문서와 기술 문서 간의 유사도를 기준 값보다 높은 것으로 검증한다.

[0102] 또한, 유사도 검증부(110)는 평균값과 제1 유사도, 제2 유사도 및 제3 유사도 각각의 차이가 기준 범위를 벗어날 경우, 학술 문서와 기술 문서 간의 유사도를 상기 기준 값보다 낮은 것으로 검증할 수 있다.

[0103] 즉, 유사도 검증부(110)는 각 그룹 내에서 산출되는 유사도를 비교하여 차이가 크지 않을 경우에 학술 문서와 기술 문서 간의 유사도가 높은 것으로 검증한다.

[0104] 예를 들어, 기준 범위는 유사도 간의 비교 결과에 따른 차이 값의 허용 범위에 상응할 수 있고, 기준 값은 유사도가 높고 낮음을 판단하기 위한 기준선에 상응할 수 있다.

[0105] 또한, 유사도 검증부(110)는 제1 유사도, 제2 유사도 및 제3 유사도에 대하여 분산분석(Analysis of Variance, ANOVA)을 수행하여 각 그룹별 유사도 간의 차이를 검증할 수 있다.

[0106] 즉, 본 발명은 텍스트 마이닝 기반의 통계적 검증을 이용하여 도출된 기술 문서 키워드의 신뢰성을 향상시킬 수 있다.

[0107] 즉, 본 발명은 텍스트 마이닝 기반의 통계적 검증을 이용하여 학술 문서와 기술 문서의 유사도를 검증함으로써 도출된 기술 문서 키워드의 타당성을 향상시킬 수 있다.

[0108] 본 발명의 일실시예에 따르면 키워드 추출부(120)는 학술 키워드 추출부(미도시) 및 기술 키워드 추출부(미도시)를 포함할 수 있다. 이하, 설명에서는 키워드 추출부(120)가 학술 문서에서 학술 공통 키워드를 추출하고, 기술 문서에서 기술 공통 키워드를 추출하는 구성을 설명한다.

[0109] 본 발명의 일실시예에 따르면 키워드 추출부(120)는 유사도 검증부(110)에서 검증된 유사도가 기준값보다 높을 경우, 학술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 학술 공통 키워드를 추출

할 수 있다.

- [0110] 예를 들어, 학술 문서를 구성하는 복수의 부분은 학술 문서의 요약, 서론 및 결론 중에서 적어도 하나를 포함할 수 있다.
- [0111] 본 발명의 일실시예에 따르면 키워드 추출부(120)는 기술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 기술 공통 키워드를 추출할 수 있다.
- [0112] 예를 들어, 기술 문서를 구성하는 복수의 부분은 기술 문서의 요약, 서론, 결론, 청구범위 및 제목 중에서 적어도 하나를 포함할 수 있다.
- [0113] 본 발명의 일실시예에 따르면 키워드 성능 검증부(130)는 추출된 적어도 하나 이상의 학술 공통 키워드를 학술 문서의 저자 키워드와 비교하여 키워드 도출 성능을 검증한다.
- [0114] 예를 들어, 적어도 하나 이상의 저자 키워드는 학술 문서의 저자가 학술 문서 제작 시, 학술 문서의 특징을 나타내는 단어로 지정할 수 있다.
- [0115] 일례로, 키워드 성능 검증부(130)는, 추출된 적어도 하나 이상의 학술 공통 키워드와 저자 키워드 간의 일치되는 키워드 수를 산출하고, 산출된 키워드 수를 추출된 적어도 하나 이상의 학술 공통 키워드의 수로 나눠서 키워드 도출 성능을 검증할 수 있다.
- [0116] 예를 들어, 본 발명은 학술 문서와 기술 문서의 유사도 검증에 기초하여 학술 문서의 저자 키워드를 키워드 도출 기준으로 이용해 기술 문서 키워드 도출 방법을 탐색할 수 있다.
- [0117] 예를 들어, 키워드 성능 검증부(130)는 산출된 키워드 수를 추출된 적어도 하나 이상의 학술 공통 키워드의 수로 나눠서 산출되는 정확성 비율이 키워드 성능 검증 기준값보다 높을 경우 키워드 도출 성능을 긍정 등급으로 검증할 수 있다.
- [0118] 또한, 키워드 성능 검증부(130)는 산출된 키워드 수를 추출된 적어도 하나 이상의 학술 공통 키워드의 수로 나눠서 산출되는 정확성 비율이 키워드 성능 검증 기준값보다 낮을 경우 키워드 도출 성능을 부정 등급으로 검증할 수 있다.
- [0119] 일례로 키워드 성능 검증부(130)는 적어도 하나 이상의 학술 공통 키워드의 개시 빈도에 기초하여 적어도 하나 이상의 학술 공통 키워드의 순위들을 결정할 수 있다.
- [0120] 본 발명의 일실시예에 따르면 키워드 성능 검증부(130)는 결정된 순위들에 기초하여 적어도 하나 이상의 학술 공통 키워드 중 노이즈 분류 기준 보다 낮은 순위에 해당하는 학술 공통 키워드를 적어도 하나 이상의 학술 공통 키워드에서 제외한다.
- [0121] 일례로, 키워드 성능 검증부(130)는 유효 키워드와 일치되는 키워드들의 수를 추출된 키워드들의 수로 나눠서 비율을 산출하여 키워드 도출 성능을 검증할 수 있다.
- [0122] 일례로, 키워드 성능 검증부(130)는 추출된 학술 공통 키워드들 중 유효 키워드와 일치되는 키워드들의 수를 유효 키워드의 수로 나눠서 비율을 산출하여 키워드 도출 성능을 검증할 수 있다.
- [0123] 본 발명의 일실시예에 따르면 기술 문서 키워드 도출부(140)는 키워드 도출 성능 검증 후, 추출된 적어도 하나 이상의 기술 공통 키워드를 기술 문서 키워드로 도출할 수 있다.
- [0124] 본 발명은 학술 문서와 기술 문서의 유사도 검증에 기초하여 학술 문서의 저자 키워드를 키워드 도출 기준으로 이용해 기술 문서 키워드 도출 방법을 탐색할 수 있다.
- [0125] 또한, TF-IDF를 통하여 추출되는 키워드들에 대하여 비율을 산출한 후, TF-IDF를 통하여 추출되는 키워드들의 정확도와 기술 문서 키워드 도출부(140)에 의하여 도출되는 기술 문서 키워드들의 정확도를 비교할 수 있다.
- [0126] 본 발명은 TF-IDF를 대체하여 단어의 출현 빈도와 동일하게 개시되는지의 여부를 이용하여 기술 문서 키워드를 도출할 수 있다.
- [0127] 본 발명의 다른 실시예에 따르면 키워드 도출 장치(100)는 데이터 수집부(미도시)를 더 포함할 수 있다.
- [0128] 일례로, 데이터 수집부(미도시)는 동일한 검색어에 기반하여 학술 문서 데이터베이스로부터 학술 문서들을 수집할 수 있다.

- [0129] 예를 들어, 데이터 수집부(미도시)는 동일한 검색어에 기반하여 기술 문서 데이터베이스로부터 기술 문서들을 수집할 수 있다.
- [0130] 즉, 데이터 수집부(미도시)는 학술 문서 데이터베이스(database)와 기술 문서 데이터베이스에서 동일한 검색어를 이용하여 학술 문서 및 기술 문서를 포함하는 분석 데이터를 수집할 수 있다.
- [0131] 본 발명은 학술 문서와 기술 문서의 유사도 검증에 기초하여 기술 문서 키워드를 도출함에 따라 다양한 특허 분석 방법론 및 연구 방향을 제공할 수 있다.
- [0132] 또한, 본 발명은 기술 문서의 유사도 검증에 기초하여 기술 문서 키워드를 도출함에 따라 학술 문서의 키워드를 특허 분석의 지표로 활용할 수 있다.
- [0134] 도 2는 본 발명의 일실시예에 따른 기술 문서 키워드를 도출하는 방법과 관련된 흐름도를 설명하는 도면이다.
- [0135] 구체적으로, 도 2는 본 발명의 일실시예에 따른 기술 문서 키워드를 도출하는 방법이 학술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 학술 공통 키워드를 기술 문서 키워드로 도출하는 절차를 예시한다.
- [0136] 도 2를 참고하면, 단계(201)에서 기술 문서 키워드를 도출하는 방법은 학술 문서와 기술 문서의 유사도를 검증한다.
- [0137] 즉, 기술 문서 키워드를 도출하는 방법은 동일한 검색어에 기반하여 수집된 분석 데이터에 포함되는, 학술 문서와 기술 문서의 유사도를 검증한다.
- [0138] 일례로, 기술 문서 키워드를 도출하는 방법은 동일한 검색어에 기반하여 수집된 분석 데이터에 포함되는 복수의 학술 문서와 복수의 기술 문서를 세 그룹으로 분류한다.
- [0139] 다음으로, 기술 문서 키워드를 도출하는 방법은 각 그룹에서 유사도를 산출하여 학술 문서와 기술 문서의 유사도를 검증한다.
- [0140] 단계(202)에서 기술 문서 키워드를 도출하는 방법은 학술 문서에서 학술 공통 키워드를 추출한다.
- [0141] 즉, 기술 문서 키워드를 도출하는 방법은 검증된 유사도가 기준 값보다 높을 경우, 학술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 학술 공통 키워드를 추출한다.
- [0142] 단계(203)에서 기술 문서 키워드를 도출하는 방법은 학술 공통 키워드와 저자 키워드를 비교하여 도출 성능을 검증한다.
- [0143] 즉, 기술 문서 키워드를 도출하는 방법은 학술 공통 키워드와 저자 키워드의 일치 비율을 산출하여 키워드 도출 성능을 검증한다.
- [0144] 단계(204)에서 기술 문서 키워드를 도출하는 방법은 기술 문서에서 기술 공통 키워드를 추출한다.
- [0145] 즉, 기술 문서 키워드를 도출하는 방법은 기술 문서를 구성하는 복수의 부분에서 동일하게 개시되는 적어도 하나 이상의 기술 공통 키워드를 추출한다.
- [0146] 즉, 기술 문서 키워드를 도출하는 방법은 추출된 적어도 하나 이상의 기술 공통 키워드를 기술 문서 키워드로 도출할 수 있다.
- [0147] 본 발명은 학술 문서와 기술 문서의 유사도 검증에 기초하여 학술 문서의 저자 키워드와 학술 문서의 구성에서 추출된 학술 공통 키워드를 비교하여 학술 공통 키워드에 대한 성능을 검증하고, 이를 기초하여 기술 문서의 구성에서 추출된 기술 공통 키워드를 기술 문서 키워드로 도출할 수 있다.
- [0149] 도 3은 본 발명의 일실시예에 따른 기술 문서 키워드를 도출하는 방법과 관련된 흐름도를 설명하는 도면이다.
- [0150] 도 3은 기술 문서 키워드를 도출하는 방법이 동일한 검색어에 기반하여 수집된 분석 데이터를 그룹별로 분류한 후, 그룹별 유사도를 산출하는 절차를 예시한다.
- [0151] 도 3을 참고하면, 단계(301)에서 기술 문서 키워드를 도출하는 방법은 수집된 분석 데이터를 그룹별로 분류한다.

- [0152] 즉, 기술 문서 키워드를 도출하는 방법은 학술 문서 비교 그룹, 기술 문서 비교 그룹 및 학술 및 기술 문서 비교 그룹으로 수집된 분석 데이터를 분류한다.
- [0153] 다시 말해, 기술 문서 키워드를 도출하는 방법은 분석 데이터에 포함되는 학술 문서들과 기술 문서들을 학술 문서 비교 그룹, 기술 문서 비교 그룹 및 학술 및 기술 문서 비교 그룹 중 어느 하나로 분류한다.
- [0154] 단계(302)에서 기술 문서 키워드를 도출하는 방법은 학술 문서 비교 그룹에서 제1 유사도를 산출한다.
- [0155] 즉, 기술 문서 키워드를 도출하는 방법은 학술 문서 비교 그룹에 분류된 복수의 학술 문서에서 요약 부분을 추출하고, 각 문서 별 요약 부분에 대하여 텍스트 마이닝을 수행하여 제1 문서 단어 행렬로 정형화한 후, 제1 문서 단어 행렬의 행 값과 열 값을 상기 수학식 1에 적용하여 제1 유사도를 산출한다.
- [0156] 단계(303)에서 기술 문서 키워드를 도출하는 방법은 기술 문서 비교 그룹에서 제2 유사도를 산출한다.
- [0157] 즉, 기술 문서 키워드를 도출하는 방법은 기술 문서 비교 그룹에 분류된 복수의 기술 문서에서 요약 부분을 추출한다.
- [0158] 다음으로, 기술 문서 키워드를 도출하는 방법은 각 문서 별 요약 부분에 대하여 텍스트 마이닝을 수행하여 제2 문서 단어 행렬로 정형화한 후, 제2 문서 단어 행렬의 행 값과 열 값을 상기 수학식 1에 적용하여 제2 유사도를 산출한다.
- [0159] 단계(304)에서 기술 문서 키워드를 도출하는 방법은 학술 및 기술 문서 비교 그룹에서 제3 유사도를 산출한다.
- [0160] 즉, 기술 문서 키워드를 도출하는 방법은 학술 및 기술 문서 비교 그룹에서 학술 문서의 요약 부분 및 기술 문서의 요약 부분을 추출한다.
- [0161] 다음으로, 기술 문서 키워드를 도출하는 방법은 각각 추출된 요약 부분에 대한 텍스트 마이닝을 수행하여 제3 문서 단어 행렬로 정형화하고, 정형화된 제3 문서 단어 행렬을 이용하여 제3 유사도를 산출할 수 있다.
- [0162] 즉, 기술 문서 키워드를 도출하는 방법은 제3 문서 단어 행렬의 행 값과 열 값을 상기 수학식 1에 적용하여 제3 유사도를 산출할 수 있다.
- [0163] 예를 들어, 학술 문서와 기술 문서는 동일하게 요약 부분을 포함하고 있다. 따라서, 기술 문서 키워드를 도출하는 방법은 각각의 그룹에서 요약 부분을 추출할 수 있다.
- [0165] 도 4는 본 발명의 일실시예에 따른 기술 문서 키워드를 도출하는 방법과 관련된 흐름도를 설명하는 도면이다.
- [0166] 도 4는 기술 문서 키워드를 도출하는 방법이 각 그룹에서 산출된 유사도의 평균값을 산출하고, 각 그룹에서 산출된 유사도와 그 평균값을 이용하여 분석 데이터에 포함되는 학술 문서와 기술 문서의 유사도를 판단하는 절차를 예시한다.
- [0167] 도 4를 참고하면, 단계(401)에서 기술 문서 키워드를 도출하는 방법은 제1 유사도, 제2 유사도 및 제3 유사도의 평균값을 산출한다.
- [0168] 단계(402)에서 기술 문서 키워드를 도출하는 방법은 제1 유사도와 평균값을 비교한 후, 제1 유사도와 평균값의 차이가 기준 범위에 상응할 경우, 단계(403)로 진행한다. 반면에, 기술 문서 키워드를 도출하는 방법은 제1 유사도와 평균값의 차이가 기준 범위를 벗어날 경우, 단계(406)로 진행한다.
- [0169] 즉, 기술 문서 키워드를 도출하는 방법은 제1 유사도와 평균값의 차이가 기준 범위에 포함될 경우, 단계(403)로 진행하여 제2 유사도와 비교를 진행하고, 제1 유사도와 평균값의 차이가 기준 범위를 벗어날 경우, 단계(406)에서 학술 문서와 기술 문서 간의 유사도가 기준값보다 낮은 것으로 판단한다.
- [0170] 단계(403)에서 기술 문서 키워드를 도출하는 방법은 제2 유사도와 평균값을 비교한 후, 제2 유사도와 평균값의 차이가 기준 범위에 상응할 경우, 단계(404)로 진행한다. 반면에, 기술 문서 키워드를 도출하는 방법은 제2 유사도와 평균값의 차이가 기준 범위를 벗어날 경우, 단계(406)로 진행한다.
- [0171] 즉, 기술 문서 키워드를 도출하는 방법은 제2 유사도와 평균값의 차이가 기준 범위에 포함될 경우, 단계(404)로 진행하여 제3 유사도와 비교를 진행하고, 제2 유사도와 평균값의 차이가 기준 범위를 벗어날 경우, 단계(406)에서 학술 문서와 기술 문서 간의 유사도가 기준값보다 낮은 것으로 판단한다.

- [0172] 단계(404)에서 기술 문서 키워드를 도출하는 방법은 제3 유사도와 평균값을 비교한 후, 제3 유사도와 평균값의 차이가 기준 범위에 상응할 경우, 단계(405)로 진행한다. 반면에, 제3 유사도와 평균값의 차이가 기준 범위를 벗어날 경우, 단계(406)로 진행한다.
- [0173] 즉, 기술 문서 키워드를 도출하는 방법은 제3 유사도와 평균값의 차이가 기준 범위에 포함될 경우, 단계(405)에서 학술 문서와 기술 문서 간의 유사도가 기준값보다 높은 것으로 판단하고, 해당 절차를 종료한다.
- [0174] 다시말 해, 기술 문서 키워드를 도출하는 방법은 제1 유사도, 제2 유사도 및 제3 유사도가 평균값과 큰 차이를 나타내지 않을 경우, 분석 데이터에 포함되는 학술 문서와 기술 문서의 유사도를 높다고 판단할 수 있다.
- [0175] 상술한 설명에서는 기술 문서 키워드를 도출하는 방법이 각 그룹의 유사도와 그 평균값을 비교하는 절차를 설명하였다.
- [0176] 그러나 기술 문서 키워드를 도출하는 방법은 그에 한정되지 않고, 평균값을 산출하는 단계 없이 각 그룹의 유사도를 상호 간에 대비하여 학술 문서와 기술 문서의 유사도를 판단할 수 있다.
- [0178] 도 5는 본 발명의 일실시예에 따른 기술 문서 키워드를 도출하는 방법과 관련된 흐름도를 설명하는 도면이다.
- [0179] 도 5는 기술 문서 키워드를 도출하는 방법이 학술 공통 키워드의 개시 빈도에 기초하여 노이즈 키워드를 제외하는 절차를 예시한다.
- [0180] 도 5를 참고하면, 단계(501)에서 기술 문서 키워드를 도출하는 방법은 학술 문서에서 적어도 하나 이상의 학술 공통 키워드를 추출한다.
- [0181] 즉, 기술 문서 키워드를 도출하는 방법은 학술 문서와 기술 문서 간에 유사도가 검증되는 학술 문서를 구성하는 복수의 부분에 동시에 개시되는 적어도 하나 이상의 학술 공통 키워드를 추출한다.
- [0182] 단계(502)에서 기술 문서 키워드를 도출하는 방법은 각 학술 공통 키워드의 개시 빈도를 산출한다.
- [0183] 즉, 기술 문서 키워드를 도출하는 방법은 문서 내에서 학술 공통 키워드의 개시 빈도 또는 복수의 학술 문서 내에서 학술 공통 키워드의 개시 빈도를 텍스트 마이닝에 기반하여 산출할 수 있다.
- [0184] 단계(503)에서 기술 문서 키워드를 도출하는 방법은 산출된 개시 빈도에 기초하여 적어도 하나 이상의 학술 공통 키워드의 순위를 결정한다.
- [0185] 즉, 기술 문서 키워드를 도출하는 방법은 학술 문서 내에서 학술 공통 키워드의 개시 빈도를 산출하고, 산출된 개시 빈도에 따라 각 학술 공통 키워드의 순위를 결정한다.
- [0186] 여기서, 기술 문서 키워드를 도출하는 방법은 개시 빈도가 많은 순서대로 학술 공통 키워드의 순위를 결정한다.
- [0187] 단계(504)에서 기술 문서 키워드를 도출하는 방법은 학술 공통 키워드의 순위와 노이즈 분류 기준을 비교한다.
- [0188] 즉, 기술 문서 키워드를 도출하는 방법은 학술 공통 키워드의 순위가 노이즈 분류 기준보다 상대적으로 클 경우, 단계(505)로 진행하고, 반대의 경우 단계(506)로 진행한다.
- [0189] 단계(505)에서 기술 문서 키워드를 도출하는 방법은 학술 공통 키워드를 유효 키워드로 결정하여 학술 공통 키워드를 기술 문서 키워드로 도출한다.
- [0190] 즉, 기술 문서 키워드를 도출하는 방법은 노이즈 분류 기준과 대비된 학술 공통 키워드를 기술 문서 키워드로 도출한다.
- [0191] 단계(506)에서 기술 문서 키워드를 도출하는 방법은 학술 공통 키워드를 노이즈 키워드로 결정하여 도출 대상에서 제외한다.
- [0192] 즉, 기술 문서 키워드를 도출하는 방법은 노이즈 분류 기준과 대비된 학술 공통 키워드를 추출된 학술 공통 키워드들에서 제외한다.
- [0193] 예를 들어, 노이즈 분류 기준은 결정된 학술 공통 키워드들의 순위에서 하위 1/3 위치에 상응할 수 있다.
- [0194] 즉, 기술 문서 키워드를 도출하는 방법은 추출된 학술 공통 키워드들 순위에서 상위 2/3 위치에 상응하는 학술 공통 키워드만을 유효 키워드로 결정할 수 있다.

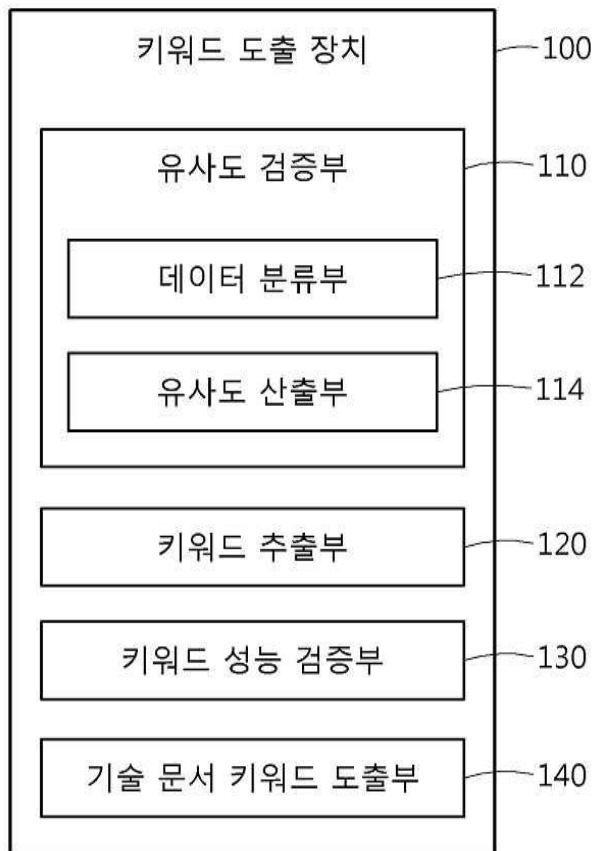
- [0196] 본 발명의 청구항 또는 명세서에 기재된 실시 예들에 따른 방법들은 하드웨어, 소프트웨어, 또는 하드웨어와 소프트웨어의 조합의 형태로 구현될(implemented) 수 있다.
- [0197] 그러한 소프트웨어는 컴퓨터 판독 가능 저장 매체에 저장될 수 있다. 컴퓨터 판독 가능 저장 매체는, 적어도 하나의 프로그램(소프트웨어 모듈), 전자 장치에서 적어도 하나의 프로세서에 의해 실행될 때 전자 장치가 본 발명의 방법을 실시하게 하는 명령어들(instructions)을 포함하는 적어도 하나의 프로그램을 저장한다.
- [0198] 이러한 소프트웨어는, 휘발성(volatile) 또는 (ROM: Read Only Memory)과 같은 불휘발성(non-volatile) 저장장치의 형태로, 또는 램(RAM: random access memory), 메모리 칩(memory chips), 장치 또는 집적 회로(integrated circuits)와 같은 메모리의 형태로, 또는 콤팩트 디스크 롬(CD-ROM: Compact Disc-ROM), 디지털 다목적 디스크(DVDs: Digital Versatile Discs), 자기 디스크(magnetic disk) 또는 자기 테이프(magnetic tape) 등과 같은 광학 또는 자기적 판독 가능 매체에, 저장될 수 있다.
- [0199] 저장 장치 및 저장 미디어는, 실행될 때 일 실시 예들을 구현하는 명령어들을 포함하는 프로그램 또는 프로그램들을 저장하기에 적절한 기계-판독 가능 저장 수단의 실시 예들이다.
- [0200] 상술한 구체적인 실시 예들에서, 발명에 포함되는 구성 요소는 제시된 구체적인 실시 예에 따라 단수 또는 복수로 표현되었다.
- [0201] 그러나, 단수 또는 복수의 표현은 설명의 편의를 위해 제시한 상황에 적합하게 선택된 것으로서, 상술한 실시 예들이 단수 또는 복수의 구성 요소에 제한되는 것은 아니며, 복수로 표현된 구성 요소라 하더라도 단수로 구성되거나, 단수로 표현된 구성 요소라 하더라도 복수로 구성될 수 있다.
- [0202] 한편 발명의 설명에서는 구체적인 실시 예에 관해 설명하였으나, 다양한 실시 예들이 내포하는 기술적 사상의 범위에서 벗어나지 않는 한도 내에서 여러 가지 변형이 가능함은 물론이다.
- [0203] 그러므로 본 발명의 범위는 설명된 실시 예에 국한되어 정해져서는 아니되며 후술하는 청구범위뿐만 아니라 이 청구범위와 균등한 것들에 의해 정해져야 한다.

부호의 설명

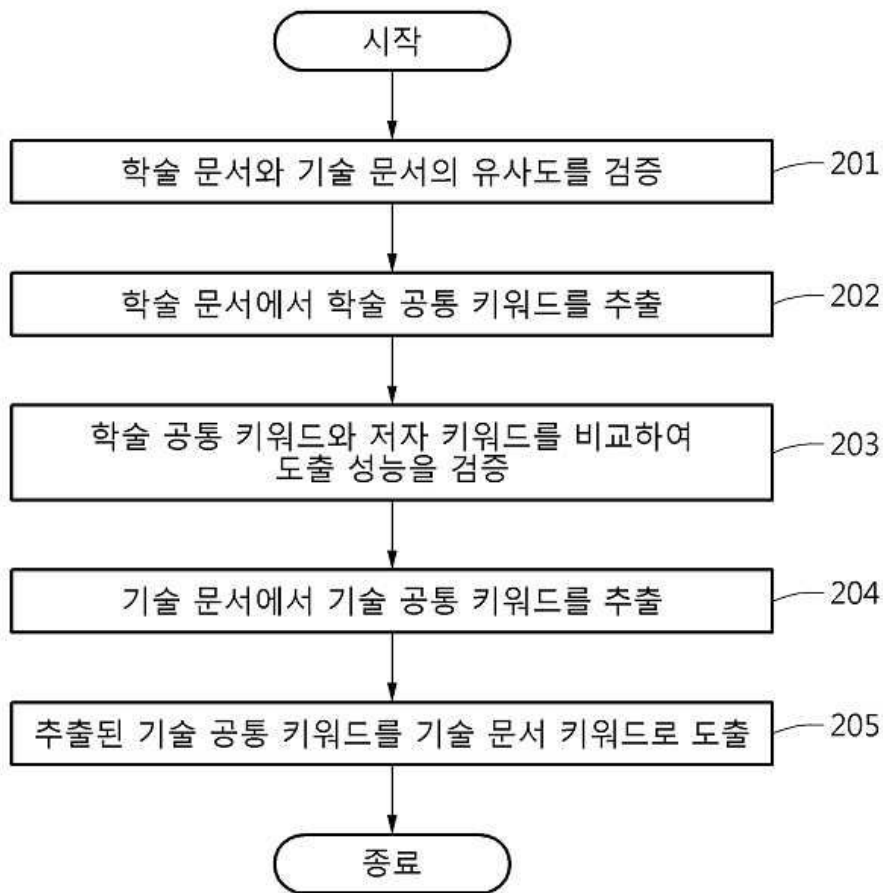
- [0205] 100: 키워드 도출 장치 110: 유사도 검증부
- 112: 데이터 분류부 114: 유사도 산출부
- 130: 키워드 추출부 130: 기술 문서 키워드 도출부

도면

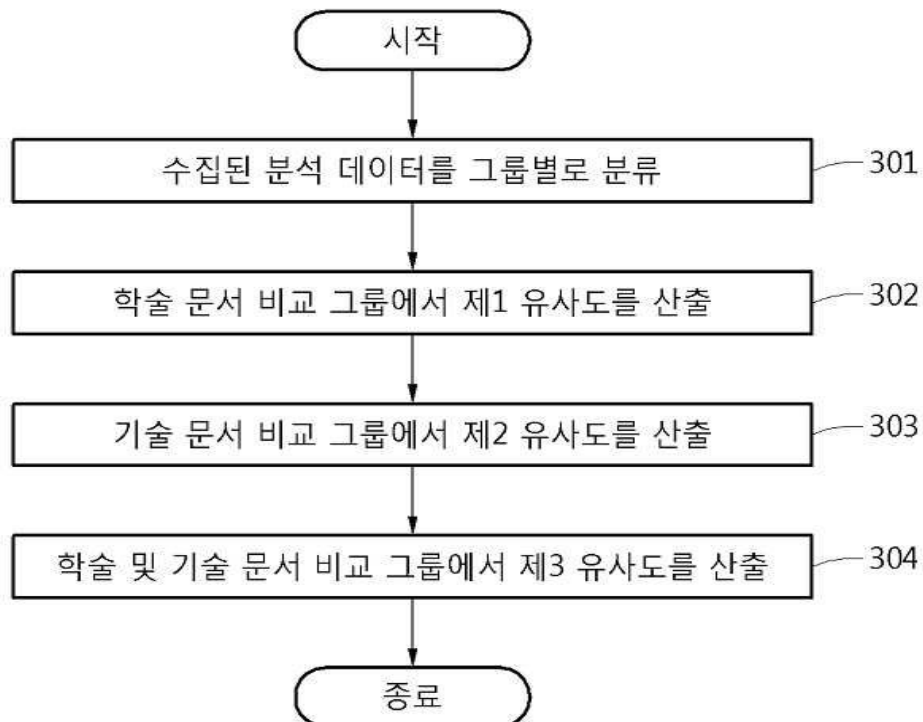
도면1



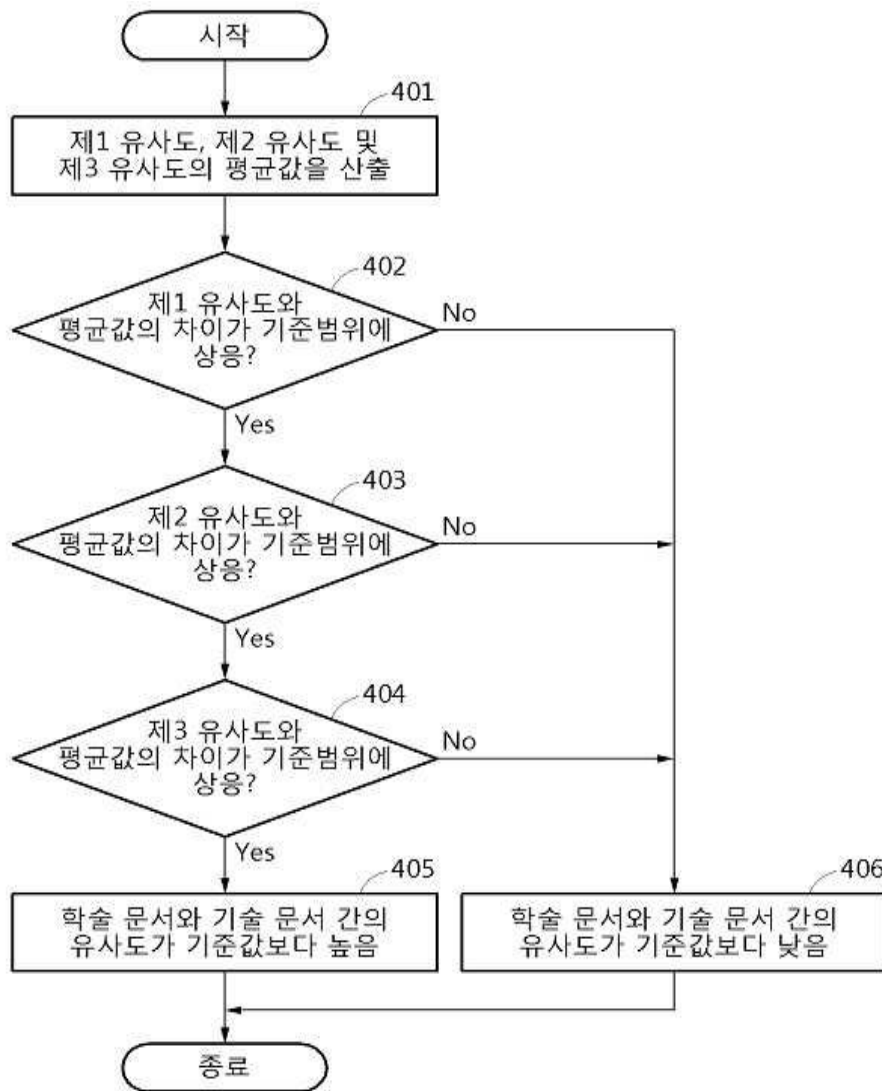
도면2



도면3



도면4



도면5

