(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0053334
(43) 공개일자 2020년05월18일

(51) 국제특허분류(Int. Cl.)
G06Q 10/06 (2012.01) *G06F 40/00* (2020.01)
G06Q 10/10 (2012.01) *G06Q 50/26* (2012.01)
(52) CPC특허분류
G06Q 10/063112 (2013.01)
G06F 40/00 (2020.01)
(21) 출원번호 10-2018-0136832
(22) 출원일자 2018년11월08일
심사청구일자 2018년11월08일

(71) 출원인
연세대학교 원주산학협력단
강원도 원주시 흥업면 연세대길 1
(72) 발명자
남영광
강원도 원주시 흥업면 세동길 51, 102동 206호(원주매지청솔아파트)
황상원
서울특별시 성북구 동소문로134길 172-4
(뒷면에 계속)
(74) 대리인
오위환, 나성곤, 정기택

전체 청구항 수 : 총 20 항

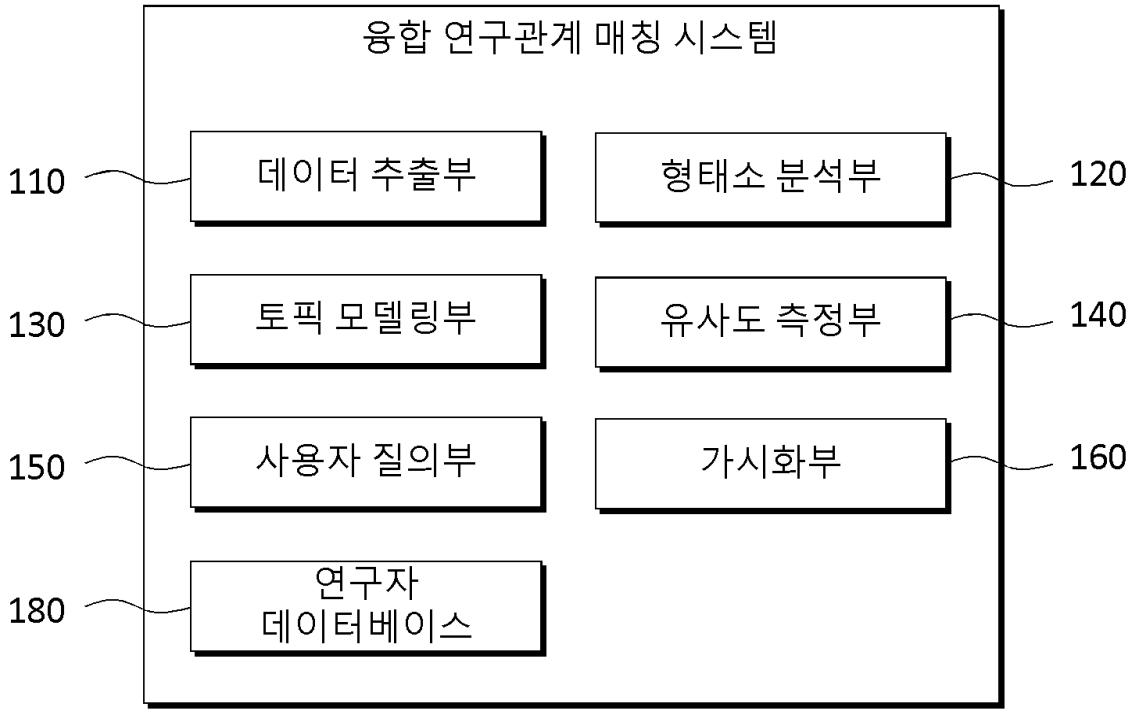
(54) 발명의 명칭 **융합 연구 촉진을 위한 연구원 맵 구축 방법 및 시스템**

(57) 요 약

본 발명은 논문, 지식재산권과 같은 연구자료를 비교 분석하여 연구분야가 서로 상이한 연구자 간의 융합 연구 가능성을 분석하고, 융복합 연구 관계를 가지는 연구자에 대한 가시화 맵을 구축하는 시스템에 관한 기술로서, 연구자료에서 연구자와 연구개요를 추출하여 데이터베이스를 구축하는 데이터 추출부와, 연구개요에서 명사 및

(뒷면에 계속)

대 표 도 - 도1



형용사를 추출하여 단어세트를 생성하는 형태소 분석부와, 말뭉치(corpus)를 이용하여 단어세트 중 연구자료의 주제가 되는 토픽 단어들을 추출하고, 관련되는 토픽 단어들을 그룹지어 토픽그룹을 형성하는 토픽 모델링부와, 사용자로부터 검색문을 입력받는 사용자 질의부와, 검색문과 토픽그룹을 이용하여 중심연구자를 검색하고, 중심연구자의 연구자료에서 추출된 토픽그룹과 다른 연구자료에서 추출된 토픽그룹을 비교하는 것으로 중심연구자와 유사한 연구를 수행하는 것으로 판단되는 관련연구자를 탐색하는 유사도 측정부와, 유사도 측정부에서 검색된 중심연구자와 관련연구자가 서로 연결된 연구자 네트워크를 구성하여 표시하는 가시화부를 포함한다.

(52) CPC특허분류
G06Q 10/105 (2013.01)
G06Q 50/26 (2013.01)

(72) 발명자
서강원
경기도 하남시 대청로 79 대명강변타운아파트 118동 504호

류원철
경기도 용인시 처인구 한터로152번길 45 피렌체아파트 109동 603호

이 발명을 지원한 국가연구개발사업
과제고유번호 NRF-2014M3C4A7030505
부처명 미래창조과학부
연구관리전문기관 한국연구재단
연구사업명 차세대정보·컴퓨팅기술개발사업
연구과제명 문맥인지기반 SW 재사용 기술 개발
기 여 율 1/1
주관기관 성균관대학교
연구기간 2014.07.01 ~ 2019.06.30

명 세 서

청구범위

청구항 1

연구자료에서 연구자와 연구개요를 추출하여 데이터베이스를 구축하는 데이터 추출부와,

상기 연구개요에서 명사 및 형용사를 추출하여 단어세트를 생성하는 형태소 분석부와,

말뭉치(corpus)를 이용하여 상기 단어세트 중 상기 연구자료의 주제가 되는 토픽 단어들을 추출하고, 관련되는 토픽 단어들을 그룹지어 토픽그룹을 형성하는 토픽 모델링부와,

사용자로부터 검색문을 입력받는 사용자 질의부와,

상기 검색문과 상기 토픽그룹을 이용하여 중심연구자를 검색하고, 상기 중심연구자의 연구자료에서 추출된 토픽그룹과 다른 연구자료에서 추출된 토픽그룹을 비교하는 것으로 상기 중심연구자와 유사한 연구를 수행하는 것으로 판단되는 관련연구자를 탐색하는 유사도 측정부와,

상기 유사도 측정부에서 검색된 중심연구자와 관련연구자가 서로 연결된 연구자 네트워크를 구성하여 표시하는 가시화부를 포함하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 2

제1항에 있어서,

상기 데이터 추출부는 연구자료가 이미지로 구성된 문서인 경우 이미지 프로세싱을 수행하여 이미지에서 문자를 추출하고,

연구자료가 이미지화된 논문인 경우, 초록이 시작되는 지점에 시작좌표와, 초록이 종료되는 지점에 종료좌표를 설정하고, 상기 종료좌표의 x좌표 값은 초록 문단의 최우측의 x좌표로 치환한 후 상기 시작좌표와 상기 종료좌표를 양 끝점으로 하는 사각 범위 내에 포함된 텍스트를 추출하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 3

제2항에 있어서,

상기 데이터 추출부는 단어 간의 공백을 판단하기 위해 공백 임계 너비가 기 설정되고, 단어의 좌측 또는 우측에 위치한 공백의 너비가 상기 공백 임계 너비를 초과하면 해당 텍스트 라인이 종료된 것으로 판단하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 4

제3항에 있어서,

상기 데이터 추출부는 단어(W), 추출된 단어의 수(n), 전체 단어(WA), i번째 추출된 단어(W_i)일 때,

상기 W_i의 영역은 WC_i($x1$, $y1$, $x2$, $y2$), 상기 W_i의 시작 좌표 WCS_i는 ($x1_{W_i}$, $y1_{W_i}$), 상기 W_i의 종료 좌표 WCE_i는 ($x2_{W_i}$, $y2_{W_i}$)가 되며,

$W_i \in WA$의 조건을 만족할 때, i-1번째와 i번째 단어 간의 공백 너비 SX는 $x2_{W_i} - x1_{W_{i-1}}$로 산출되는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 5

제1항에 있어서,

상기 형태소 분석부는 연구개요에 포함된 문장을 문자(char) 단위로 읽고, 해당 문자가 가진 정수(int) 값을 추출한 후, 상기 정수 값이 0x3131보다 크고 0xD7A3보다 작으면 연구개요가 한글인 것으로 판단하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 6

제5항에 있어서,

상기 형태소 분석부는 연구개요가 한글이면 코모란(KOMORAN)을 이용하여 형태소를 분석하여 일반명사, 고유명사, 한자인 단어를 추출하여 단어세트로 구성하고,

연구개요가 영어이면 CoreNLP를 이용하여 형태소를 분석하여 명사, 형용사인 단어를 추출하여 단어세트로 구성하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 7

제1항에 있어서,

상기 토픽 모델링부는 단어세트를 이용하여 말뭉치(corpus)를 구성하고, 상기 말뭉치를 잠재적 디리클레 할당(Latent Dirichlet Allocation) 알고리즘에 적용하여 토픽그룹을 생성하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 8

제7항에 있어서,

상기 토픽 모델링부는 각 연구자료에 K개의 토픽 단어 중 하나를 임의로 할당하고,

각 연구자료(d), 각 연구자료(d)에 포함된 전체 단어(w), 전체 단어(w)에 존재하는 토픽 단어(t)에 대해,

각 연구자료(d)의 단어세트(w) 중 토픽 단어(t)의 비율 p(t|d)를 연산하고,

모든 연구자료 중 토픽 단어(t)가 할당된 비율 p(w|t)를 연산하며,

p(t|d)와 p(w|t)의 곱에 따라 토픽 단어(t)를 신규하게 선택하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 9

제1항에 있어서,

상기 토픽 모델링부는 검색문과 관련된 연구자가 검색되면 해당 연구자가 포함된 연구자료에서 토픽그룹을 추출하고, 검색문과 매칭되는 토픽 단어를 연결하여 검색문-토픽 매핑을 실시하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 10

제9항에 있어서,

상기 토픽 모델링부는 검색문의 의미가 모호할 수 있는 문제를 해소하기 위해, 토픽그룹의 토픽 단어와 논문에 개시된 키워드를 조합하여 유사도 비교를 수행하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 11

제1항에 있어서,

상기 유사도 측정부는 중심연구자의 토픽그룹과, 다른 연구자의 토픽그룹을 자카드(jaccard) 알고리즘, SL(Scaled Levenshtein) 알고리즘 및 Soft TF/IDF 알고리즘을 이용하여 유사도를 연산하는 것으로 관련연구자를 탐색하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 12

제11항에 있어서,

상기 유사도 측정부는 연구자의 수(N)에 따라 연구자의 유사도 연산을 $\frac{N(N-1)}{2}$ 회 수행하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 13

제11항에 있어서,

상기 유사도 측정부는 자카드 알고리즘, SL 알고리즘 및 Soft TF/IDF 알고리즘의 유사도 값이 모두 1이면 중심연구자와 비교된 연구자를 유사한 연구를 수행하는 관련연구자인 것으로 결정하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 14

제11항에 있어서,

상기 유사도 측정부는 중심연구자와 다른 연구자를 비교하여 자카드 알고리즘, SL 알고리즘 및 Soft TF/IDF 알고리즘의 유사도 값 중 하나라도 0이면 상기 다른 연구자를 관련연구자에서 제외하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 15

제11항에 있어서,

상기 유사도 측정부는 중심연구자와 다른 연구자를 비교하여 자카드 알고리즘, SL 알고리즘 및 Soft TF/IDF 알고리즘 중 두 가지 이상의 유사도 값이 0.5 미만이면 의미적 유사성이 없는 것으로 판단하고 상기 다른 연구자를 관련연구자에서 제외하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 16

제1항에 있어서,

상기 연구자 네트워크는 연구자의 이름이 기재된 도형과, 어느 한 도형과 다른 도형을 연결하고, 상기 어느 한 도형 및 상기 다른 도형과 대응되는 연구자 간의 관련된 토픽 단어의 수에 대응하여 굵기가 결정되는 연결선을 포함하는 가시화부를 포함하고,

상기 가시화부는 상기 도형이 선택되면 상기 도형의 외주면에 해당 연구자의 토픽 단어를 표시하고, 상기 연결선이 선택되면 연구자 간에 관련된 토픽 단어를 표시하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 17

연구자료에서 연구자와 연구개요를 추출하여 데이터베이스를 구축하는 데이터 추출부와,

상기 연구개요에서 명사 및 형용사를 추출하여 단어세트를 생성하는 형태소 분석부와,

말뭉치(corpus)를 이용하여 상기 단어세트 중 상기 연구자료의 주제가 되는 토픽 단어들을 추출하고, 관련되는 토픽 단어들을 그룹지어 토픽그룹을 형성하는 토픽 모델링부와,

사용자로부터 검색문을 입력받는 사용자 질의부와,

상기 검색문과 상기 토픽그룹을 이용하여 중심연구자를 검색하고, 상기 중심연구자의 연구자료에서 추출된 토픽그룹과 다른 연구자료에서 추출된 토픽그룹을 비교하는 것으로 상기 중심연구자와 유사한 연구를 수행하는 것으로 판단되는 관련연구자를 탐색하는 유사도 측정부와,

상기 유사도 측정부에서 검색된 중심연구자와 관련연구자가 서로 연결된 연구자 네트워크를 구성하여 표시하는 가시화부를 포함하고,

상기 데이터 추출부는 연구자료가 이미지로 구성된 문서인 경우 이미지 프로세싱을 수행하여 이미지에서 문자를 추출하고, 연구자료가 이미지화된 논문인 경우, 초록이 시작되는 지점에 시작좌표와, 초록이 종료되는 지점에

종료좌표를 설정하고, 상기 종료좌표의 x좌표 값은 초록 문단의 최우측의 x좌표로 치환한 후 상기 시작좌표와 상기 종료좌표를 양 끝점으로 하는 사각 범위 내에 포함된 텍스트를 추출하며,

상기 데이터 추출부는 단어 간의 공백을 판단하기 위해 공백 임계 너비가 기 설정되고, 단어의 좌측 또는 우측에 위치한 공백의 너비가 상기 공백 임계 너비를 초과하면 해당 텍스트 라인이 종료된 것으로 판단하고,

상기 데이터 추출부는 단어(W), 추출된 단어의 수(n), 전체 단어(WA), i번째 추출된 단어(W_i)일 때, 상기 W_i의 영역은 WC_i(x1, y1, x2, y2), 상기 W_i의 시작 좌표 WCS_i는 ($x1_{W_i}$, $y1_{W_i}$), 상기 W_i의 종료 좌표 WCE는 ($x2_{W_i}$, $y2_{W_i}$)가 되며, $W_i \in WA$의 조건을 만족할 때, i-1번째와 i번째 단어 간의 공백 너비 SX는 $x2_{W_i} - x1_{W_{i-1}}$ 로 산출되고,

상기 형태소 분석부는 연구개요에 포함된 문장을 문자(char) 단위로 읽고, 해당 문자가 가진 정수(int) 값을 추출한 후, 상기 정수 값이 0x3131보다 크고 0xD7A3보다 작으면 연구개요가 한글인 것으로 판단하며,

상기 형태소 분석부는 연구개요가 한글이면 코모란(KOMORAN)을 이용하여 형태소를 분석하여 일반명사, 고유명사, 한자인 단어를 추출하여 단어세트로 구성하고, 연구개요가 영어이면 CoreNLP를 이용하여 형태소를 분석하여 명사, 형용사인 단어를 추출하여 단어세트로 구성하며,

상기 토픽 모델링부는 단어세트를 이용하여 말뭉치(corpus)를 구성하고, 상기 말뭉치를 잠재적 디리클레 할당(Latent Dirichlet Allocation) 알고리즘에 적용하여 토픽그룹을 생성하며,

상기 토픽 모델링부는 각 연구자료에 K개의 토픽 단어 중 하나를 임의로 할당하고, 각 연구자료(d), 각 연구자료(d)에 포함된 전체 단어(w), 전체 단어(w)에 존재하는 토픽 단어(t)에 대해, 각 연구자료(d)의 단어세트(w) 중 토픽 단어(t)의 비율 p(t|d)를 연산하고, 모든 연구자료 중 토픽 단어(t)가 할당된 비율 p(w|t)를 연산하며, p(t|d)와 p(w|t)의 곱에 따라 토픽 단어(t)를 신규하게 선택하며,

상기 토픽 모델링부는 검색문과 관련된 연구자가 검색되면 해당 연구자가 포함된 연구자료에서 토픽그룹을 추출하고, 검색문과 매칭되는 토픽 단어를 연결하여 검색문-토픽 매핑을 실시하며,

상기 토픽 모델링부는 검색문의 의미가 모호할 수 있는 문제를 해소하기 위해, 토픽그룹의 토픽 단어와 논문에 개시된 키워드를 조합하여 유사도 비교를 수행하고,

상기 유사도 측정부는 중심연구자의 토픽그룹과 다른 연구자의 토픽그룹을 자카드(jaccard) 알고리즘, SL(Scaled Levenshtein) 알고리즘 및 Soft TF/IDF 알고리즘을 이용하여 유사도를 연산하는 것으로 관련연구자를 탐색하며,

상기 유사도 측정부는 연구자의 수(N)에 따라 연구자의 유사도 연산을 $\frac{N(N-1)}{2}$ 회 수행하고,

상기 유사도 측정부는 자카드 알고리즘, SL 알고리즘 및 Soft TF/IDF 알고리즘의 유사도 값이 모두 1이면 중심연구자와 비교된 연구자를 유사한 연구를 수행하는 관련연구자인 것으로 결정하며,

상기 유사도 측정부는 중심연구자와 다른 연구자를 비교하여 자카드 알고리즘, SL 알고리즘 및 Soft TF/IDF 알고리즘의 유사도 값 중 하나라도 0이면 상기 다른 연구자를 관련연구자에서 제외하고,

상기 유사도 측정부는 중심연구자와 다른 연구자를 비교하여 자카드 알고리즘, SL 알고리즘 및 Soft TF/IDF 알고리즘 중 두 가지 이상의 유사도 값이 0.5 미만이면 의미적 유사성이 없는 것으로 판단하고 상기 다른 연구자를 관련연구자에서 제외하며,

상기 연구자 네트워크는 연구자의 이름이 기재된 도형과, 어느 한 도형과 다른 도형을 연결하고, 상기 어느 한 도형 및 상기 다른 도형과 대응되는 연구자 간의 관련된 토픽 단어의 수에 대응하여 굵기가 결정되는 연결선을 포함하는 가시화부를 포함하고,

상기 가시화부는 상기 도형이 선택되면 상기 도형의 외주면에 해당 연구자의 토픽 단어를 표시하고, 상기 연결선이 선택되면 연구자 간에 관련된 토픽 단어를 표시하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 18

연구자의 이름이 기재된 도형과, 어느 한 도형과 다른 도형을 연결하고, 상기 어느 한 도형 및 상기 다른 도형과 대응되는 연구자 간의 관련된 토픽 단어의 수에 대응하여 굵기가 결정되는 연결선을 포함하는 가시화부를 포함하고,

상기 가시화부는 상기 도형이 선택되면 상기 도형의 외주면에 해당 연구자의 토픽 단어를 표시하고, 상기 연결선이 선택되면 연구자 간에 관련된 토픽 단어를 표시하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 시스템.

청구항 19

데이터 추출부가 연구자료에서 연구자와 연구개요를 추출하여 데이터베이스를 구축하는 단계와,

형태소 분석부가 상기 연구개요에서 명사 및 형용사를 추출하여 단어세트를 생성하는 단계와,

토픽 모델링부가 말뭉치(corpus)를 이용하여 상기 단어세트 중 상기 연구자료의 주제가 되는 토픽 단어들을 추출하고, 관련되는 토픽 단어들을 그룹지어 토픽그룹을 형성하는 단계와,

사용자 질의부가 사용자로부터 검색문을 입력받는 단계와,

유사도 측정부가 상기 검색문과 상기 토픽그룹을 이용하여 중심연구자를 검색하는 단계와,

상기 유사도 측정부가 상기 중심연구자의 연구자료에서 추출된 토픽그룹과 다른 연구자료에서 추출된 토픽그룹을 비교하는 것으로 상기 중심연구자와 유사한 연구를 수행하는 것으로 판단되는 관련연구자를 탐색하는 단계와,

가시화부가 상기 중심연구자와 상기 관련연구자가 서로 연결되는 연구자 네트워크를 구성하여 표시하는 단계를 포함하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 방법.

청구항 20

사용자 질의부가 사용자로부터 검색문을 입력받는 단계와,

가시화부가 검색된 연구자의 이름이 기재된 도형과, 어느 한 도형과 다른 도형을 연결하고, 상기 어느 한 도형 및 상기 다른 도형과 대응되는 연구자 간의 관련된 토픽 단어의 수에 대응하여 굵기가 결정되는 연결선을 모니터장치에 표시하는 단계를 포함하는 것을 특징으로 하는 융합 연구 촉진을 위한 연구원 맵 구축 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 융합 연구 촉진을 위한 연구원 맵 구축 방법 및 시스템에 관한 것으로서, 보다 상세하게는 논문, 지식재산권과 같은 연구자료를 비교 분석하여 연구분야가 서로 상이한 연구자 간의 융합 연구 가능성을 분석하고, 융복합 연구 관계를 가지는 연구자에 대한 가시화 맵을 구축하는 시스템에 관한 기술이다.

배 경 기 술

[0002] 현대 사회는 융합, 복합, 혁신, 창조 등의 획기적인 변화를 강조하는 정책과 제도가 확산되고 있다. 과학 기술적인 측면에서 융복합 기술은 현대 과학기술 혁신의 보편적인 현상으로 자리 잡았다. 최근 개발 및 상업화에 성공하고 있는 대부분의 제품과 서비스가 융복합 기술의 산물로 인식되고 있으며, 기업, 연구기관, 대학, 정부 등 모든 국가 과학기술 혁신 주체가 융합기술의 혁신에 몰두하고 있다. 세계적으로도 각 나라의 정부 차원에서 기술 융합 연구의 중요성을 깊이 인식하고 융복합 연구 진흥을 위한 부서의 설치와 융복합 연구개발사업 등의 관련 정책을 추진하고 있다. 특히 기업과 대학은 융복합 기술의 혁신 주체로서 주목을 받고 있으며, 이에 대한 많은 정책적인 지원도 제공되고 있다.

[0003] 이와 같은 연구자 간의 관계도 구축을 위해 논문 'Analytical Study on the Relationship between Centralities of Research Networks and Research Performances'는 co-author, 저자 동시 인용, 저자 서지 결합 네트워크에 나타난 중심성과 연구성과의 연관성을 분석하였는데, 이것은 연구자의 연구 성과 분석에 중점을 둔 연구로 융복합 연구 수행 가능 여부에 대한 확인이 어렵다. 그리고 스탠포드 자연언어 처리 그룹(The

Stanford Natural Language Processing Group)은 HRD(Human Resource Development) 연구동향 분석을 위해 기업교육 및 산업교육 연구 관련 분야의 핵심어 기반 네트워크 분석을 수행하였는데, 해당 연구에서는 연구자가 직업 노드와 링크를 생성하여 한정된 분야에 대한 분석을 수행하였기 때문에, 실시간으로 다양한 분야의 융복합 연구자 네트워크 분석 및 확인에는 적합하지 않다.

선행기술문헌

특허문헌

[0004] (특허문헌 0001) 등록특허공보 제10-1426765호

발명의 내용

해결하려는 과제

[0005] 이에 본 발명은 상기와 같은 종래의 제반 문제점을 해소하기 위해 제안된 것으로, 본 발명의 목적은 논문, 지식재산권과 같은 연구자료를 비교 분석하여 연구분야가 서로 상이한 연구자 간의 융합 연구 가능성을 분석하고, 융복합 연구 관계를 가지는 연구자에 대한 가시화 맵을 구축하는 시스템을 제공하기 위한 것이다.

[0006] 또한, 본 발명의 다른 목적은 융복합 기술의 발전에 영향을 줄 수 있는 대학의 역할에 주목하여, 기업으로 하여금 대학과의 융복합 산학협력을 수행할 수 있도록 정보를 제공해주는 연구자 사이의 관계도를 구축하는 것이다.

과제의 해결 수단

[0007] 상기와 같은 목적을 달성하기 위하여 본 발명의 기술적 사상에 의한 융합 연구 촉진을 위한 연구원 맵 구축 시스템은 연구자료에서 연구자와 연구개요를 추출하여 데이터베이스를 구축하는 데이터 추출부와, 상기 연구개요에서 명사 및 형용사를 추출하여 단어세트를 생성하는 형태소 분석부와, 말뭉치(corpus)를 이용하여 상기 단어세트 중 상기 연구자료의 주제가 되는 토픽 단어들을 추출하고, 관련되는 토픽 단어들을 그룹지어 토픽그룹을 형성하는 토픽 모델링부와, 사용자로부터 검색문을 입력받는 사용자 질의부와, 상기 검색문과 상기 토픽그룹을 이용하여 중심연구자를 검색하고, 상기 중심연구자의 연구자료에서 추출된 토픽그룹과 다른 연구자료에서 추출된 토픽그룹을 비교하는 것으로 상기 중심연구자와 유사한 연구를 수행하는 것으로 판단되는 관련연구자를 탐색하는 유사도 측정부와, 상기 유사도 측정부에서 검색된 중심연구자와 관련연구자가 서로 연결된 연구자 네트워크를 구성하여 표시하는 가시화부를 포함하는 것을 특징으로 한다.

[0008] 또한, 상기 데이터 추출부는 연구자료가 이미지로 구성된 문서인 경우 이미지 프로세싱을 수행하여 이미지에서 문자를 추출하고, 연구자료가 이미지화된 논문인 경우, 초록이 시작되는 지점에 시작좌표와, 초록이 종료되는 지점에 종료좌표를 설정하고, 상기 종료좌표의 x좌표 값은 초록 문단의 최우측의 x좌표로 치환한 후 상기 시작좌표와 상기 종료좌표를 양 끝점으로 하는 사각 범위 내에 포함된 텍스트를 추출하는 것을 특징으로 할 수 있다.

[0009] 또한, 상기 데이터 추출부는 단어 간의 공백을 판단하기 위해 공백 임계 너비가 기 설정되고, 단어의 좌측 또는 우측에 위치한 공백의 너비가 상기 공백 임계 너비를 초과하면 해당 텍스트 라인이 종료된 것으로 판단하는 것을 특징으로 할 수 있다.

[0010] 또한, 상기 데이터 추출부는 단어(W), 추출된 단어의 수(n), 전체 단어(WA), i번째 추출된 단어(W_i)일 때, 상기 W_i의 영역은 $WC_i(x1, y1, x2, y2)$, 상기 W_i의 시작 좌표 WCS_i는 ($x1_{W_i}$, $y1_{W_i}$), 상기 W_i의 종료 좌표 WCE_i는 ($x2_{W_i}$, $y2_{W_i}$)가 되며, $W_i \in WA$의 조건을 만족할 때, i-1번째와 i번째 단어 간의 공백 너비 SX_i는 $x2_{W_i} - x1_{W_{i-1}}$ 로 산출되는 것을 특징으로 할 수 있다.

[0011] 또한, 상기 형태소 분석부는 연구개요에 포함된 문장을 문자(char) 단위로 읽고, 해당 문자가 가진 정수(int) 값을 추출한 후, 상기 정수 값이 0x3131보다 크고 0xD7A3보다 작으면 연구개요가 한글인 것으로 판단하는 것을 특징으로 할 수 있다.

[0012] 또한, 상기 형태소 분석부는 연구개요가 한글이면 코모란(KOMORAN)을 이용하여 형태소를 분석하여 일반명사, 고

유명사, 한자인 단어를 추출하여 단어세트로 구성하고, 연구개요가 영어이면 CoreNLP를 이용하여 형태소를 분석하여 명사, 형용사인 단어를 추출하여 단어세트로 구성하는 것을 특징으로 할 수 있다.

[0013] 또한, 상기 토픽 모델링부는 단어세트를 이용하여 말뭉치(corpus)를 구성하고, 상기 말뭉치를 잠재적 디리클레 할당(Latent Dirichlet Allocation) 알고리즘에 적용하여 토픽그룹을 생성하는 것을 특징으로 할 수 있다.

[0014] 또한, 상기 토픽 모델링부는 각 연구자료에 K개의 토픽 단어 중 하나를 임의로 할당하고, 각 연구자료(d), 각 연구자료(d)에 포함된 전체 단어(w), 전체 단어(w)에 존재하는 토픽 단어(t)에 대해, 각 연구자료(d)의 단어세트(w) 중 토픽 단어(t)의 비율 p(t|d)를 연산하고, 모든 연구자료 중 토픽 단어(t)가 할당된 비율 p(w|t)를 연산하며, p(t|d)와 p(w|t)의 곱에 따라 토픽 단어(t)를 신규하게 선택하는 것을 특징으로 할 수 있다.

[0015] 또한, 상기 토픽 모델링부는 검색문과 관련된 연구자가 검색되면 해당 연구자가 포함된 연구자료에서 토픽그룹을 추출하고, 검색문과 매칭되는 토픽 단어를 연결하여 검색문-토픽 매핑을 실시하는 것을 특징으로 할 수 있다.

[0016] 또한, 상기 토픽 모델링부는 검색문의 의미가 모호할 수 있는 문제를 해소하기 위해, 토픽그룹의 토픽 단어와 논문에 개시된 키워드를 조합하여 유사도 비교를 수행하는 것을 특징으로 할 수 있다.

[0017] 또한, 상기 유사도 측정부는 중심연구자의 토픽그룹과, 다른 연구자의 토픽그룹을 자카드(jaccard) 알고리즘, SL(Scaled Levenshtein) 알고리즘 및 Soft TF/IDF 알고리즘을 이용하여 유사도를 연산하는 것으로 관련연구자를 탐색하는 것을 특징으로 할 수 있다.

[0018] 또한, 상기 유사도 측정부는 연구자의 수(N)에 따라 연구자의 유사도 연산을 $\frac{N(N-1)}{2}$ 회 수행하는 것을 특징으로 할 수 있다.

[0019] 또한, 상기 유사도 측정부는 자카드 알고리즘, SL 알고리즘 및 Soft TF/IDF 알고리즘의 유사도 값이 모두 1이면 중심연구자와 비교된 연구자를 유사한 연구를 수행하는 관련연구자인 것으로 결정하는 것을 특징으로 할 수 있다.

[0020] 또한, 상기 유사도 측정부는 중심연구자와 다른 연구자를 비교하여 자카드 알고리즘, SL 알고리즘 및 Soft TF/IDF 알고리즘의 유사도 값 중 하나라도 0이면 상기 다른 연구자를 관련연구자에서 제외하는 것을 특징으로 할 수 있다.

[0021] 또한, 상기 유사도 측정부는 중심연구자와 다른 연구자를 비교하여 자카드 알고리즘, SL 알고리즘 및 Soft TF/IDF 알고리즘 중 두 가지 이상의 유사도 값이 0.5 미만이면 의미적 유사성이 없는 것으로 판단하고 상기 다른 연구자를 관련연구자에서 제외하는 것을 특징으로 할 수 있다.

[0022] 또한, 상기 연구자 네트워크는 연구자의 이름이 기재된 도형과, 어느 한 도형과 다른 도형을 연결하고, 상기 어느 한 도형 및 상기 다른 도형과 대응되는 연구자 간의 관련된 토픽 단어의 수에 대응하여 굵기가 결정되는 연결선을 포함하는 가시화부를 포함하고, 상기 가시화부는 상기 도형이 선택되면 상기 도형의 외주면에 해당 연구자의 토픽 단어를 표시하고, 상기 연결선이 선택되면 연구자 간에 관련된 토픽 단어를 표시하는 것을 특징으로 할 수 있다.

[0023] 한편, 상기와 같은 목적을 달성하기 위하여 본 발명의 기술적 사상에 의한 융합 연구 촉진을 위한 연구원 맵 구축 방법은 데이터 추출부가 연구자료에서 연구자와 연구개요를 추출하여 데이터베이스를 구축하는 단계와, 형태소 분석부가 상기 연구개요에서 명사 및 형용사를 추출하여 단어세트를 생성하는 단계와, 토픽 모델링부가 말뭉치(corpus)를 이용하여 상기 단어세트 중 상기 연구자료의 주제가 되는 토픽 단어들을 추출하고, 관련되는 토픽 단어들을 그룹지어 토픽그룹을 형성하는 단계와, 사용자 질의부가 사용자로부터 검색문을 입력받는 단계와, 유사도 측정부가 상기 검색문과 상기 토픽그룹을 이용하여 중심연구자를 검색하는 단계와, 상기 유사도 측정부가 상기 중심연구자의 연구자료에서 추출된 토픽그룹과 다른 연구자료에서 추출된 토픽그룹을 비교하는 것으로 상기 중심연구자와 유사한 연구를 수행하는 것으로 판단되는 관련연구자를 탐색하는 단계와, 가시화부가 상기 중심연구자와 상기 관련연구자가 서로 연결되는 연구자 네트워크를 구성하여 표시하는 단계를 포함하는 것을 특징으로 한다.

발명의 효과

[0024] 본 발명에 의한 융합 연구 촉진을 위한 연구원 맵 구축 방법 및 시스템에 따르면,

[0025] 첫째, 본 발명은 서로 유사한 연구를 수행하거나, 유사 기술을 이용하는 연구원들을 서로 매칭하여 연구원 네트워크를 구성하므로 연구자들 간의 협력을 강화할 수 있게 됨으로써 융복합 기술 발전을 지원할 수 있게 된다.

[0026] 둘째, 연구자료가 이미지화된 문서라 하더라도 자동으로 이미지 프로세싱을 수행하여 주요한 연구개요를 추출하게 되므로 연구자료의 형식에 한정되지 않고 분석이 가능하게 된다.

[0027] 셋째, 이미지 프로세싱 중 공백 임계 너비를 이용함으로써 단어 간의 띄어쓰기, 줄바꿈, 문단의 종료를 식별할 수 있게 되어, 정확하게 필요한 연구개요를 추출할 수 있게 된다.

[0028] 넷째, 연구자료의 언어를 식별함으로써 언어별로 최적의 형태소 분석 시스템을 적용할 수 있게 되고, 이로써 언어별로 최적의 단어세트를 획득할 수 있게 된다.

[0029] 다섯째, 단어세트를 LDA 알고리즘을 이용하여 분석함으로써 연구자료별로 토픽 단어 및 각 토픽 단어의 빈도수를 알 수 있게 되고, 이 정보를 이용하여 유사 연구를 수행하는 연구자와 관련된 연구자료들을 용이하게 탐색할 수 있게 된다.

[0030] 여섯째, 관련연구자 탐색 시 서로 상이한 성능 특징을 가진 자카드 알고리즘, SL 알고리즘, Soft TF/IDF 알고리즘을 복합적으로 이용하므로 유사도 판단에 있어서 다양한 접근이 시도되고, 이로써 다양한 관점의 유사성을 가지는 연구자와 연구자료를 탐색할 수 있게 된다.

[0031] 일곱째, 중심연구자와 관련연구자 간의 관계가 연구자 네트워크에 표시되어 직관적으로 유사 연구를 수행하는 연구자를 발견할 수 있고, 연구자를 클릭(선택)하면 해당 연구자의 연구자료에서 추출된 토픽 단어들이 표시됨에 따라, 해당 연구자의 연구분야와 활용 기술 등을 용이하게 파악할 수 있게 된다.

도면의 간단한 설명

[0032] 도 1은 본 발명의 일 실시예에 따른 융합 연구 촉진을 위한 연구원 맵 구축 시스템의 구성도.

도 2는 논문의 초록 영역과 키워드 영역에 사각 범위 설정된 것을 나타낸 예시 도면.

도 3은 연구개요에서 추출된 단어세트, 단어세트를 이용하여 생성된 토픽그룹, 검색문과 토픽그룹을 매칭하는 과정을 나타낸 예시 도면.

도 4는 검색문에 의해 검색된 연구자와, 각 연구자의 토픽 단어, 토픽 단어를 선택할 때 표시되는 연구자료의 정보의 표시 예를 나타낸 도면.

도 5는 도 4에서 특정 연구자의 토픽을 선택한 후 재검색할 때 표시되는 연구자 네트워크를 나타낸 예시 도면.

도 6은 연구자 중 하나를 클릭(선택)할 때 주요 토픽 단어가 도형 주변에 표시되고, 도형 주변의 회색 영역을 클릭할 때 표시되는 전체 토픽 단어 리스트를 나타낸 예시 도면.

도 7은 본 발명의 일 실시예에 따른 융합 연구 촉진을 위한 연구원 맵 구축 방법에서 연구자료로부터 토픽그룹이 추출되는 과정을 나타낸 순서도.

도 8은 본 발명의 일 실시예에 따른 융합 연구 촉진을 위한 연구원 맵 구축 방법에서 검색문이 입력된 후 연구자 네트워크가 구성되기까지의 과정을 나타낸 순서도.

도 9는 S120 단계의 세부 과정을 나타낸 순서도.

도 10은 S140 단계의 세부 과정을 나타낸 순서도.

도 11은 S260 단계의 세부 과정을 나타낸 순서도.

발명을 실시하기 위한 구체적인 내용

[0033] 첨부한 도면을 참조하여 본 발명의 실시예들에 의한 융합 연구 촉진을 위한 연구원 맵 구축 방법 및 시스템에 대하여 상세히 설명한다. 본 발명은 다양한 변경을 가할 수 있고 여러 가지 형태를 가질 수 있는바, 특정 실시예들을 도면에 예시하고 본문에 상세하게 설명하고자 한다. 그러나 이는 본 발명을 특정한 개시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.

[0034] 또한, 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0035] 본 발명의 일 실시예에 따른 융합 연구 촉진을 위한 연구원 맵 구축 시스템은 어느 한 컴퓨터 장치에 모든 구성이 포함되거나, 다수의 컴퓨터 장치에 구성들이 분산되고 컴퓨터 장치들이 유기적으로 연결되는 것으로 실시될 수 있다.

[0036] 도 1을 참조하면, 본 발명의 일 실시예에 따른 융합 연구 촉진을 위한 연구원 맵 구축 시스템은 연구자료에서 연구자와 연구개요를 추출하여 데이터베이스를 구축하는 데이터 추출부와, 상기 연구개요에서 명사 및 형용사를 추출하여 단어세트를 생성하는 형태소 분석부와, 말뭉치(corpus)를 이용하여 상기 단어세트 중 상기 연구자료의 주제가 되는 토픽 단어들을 추출하고, 관련되는 토픽 단어들을 그룹지어 토픽그룹을 형성하는 토픽 모델링부와, 사용자로부터 검색문을 입력받는 사용자 질의부와, 상기 검색문과 상기 토픽그룹을 이용하여 중심연구자를 검색하고, 상기 중심연구자의 연구자료에서 추출된 토픽그룹과 다른 연구자료에서 추출된 토픽그룹을 비교하는 것으로 상기 중심연구자와 유사한 연구를 수행하는 것으로 판단되는 관련연구자를 탐색하는 유사도 측정부와, 상기 유사도 측정부에서 검색된 중심연구자와 관련연구자가 서로 연결된 연구자 네트워크를 구성하여 표시하는 가시화부를 포함하는 것을 특징으로 한다.

[0037] 또한, 본 발명의 일 실시예는 연구자료, 상기 연구자료에서 추출된 연구자 이름, 연구개요, 토픽그룹 등이 하나의 세트로 구성되어 저장되는 연구자 데이터베이스를 더 포함한다.

[0038] 연구자료에는 논문, 연구과제, 특허, 실용신안, 저작권 등이 포함된다. 연구자료는 PDF, Ms word, 아래아 한글 등 공지된 형식의 파일이 이 실시예에 입력되는 것으로 실시될 수 있다.

[0039] 연구자는 논문에 개시된 저자, 특허나 실용신안에 개시된 발명자가 될 수 있다.

[0040] 연구개요에는 논문의 제목, 지식재산권의 명칭, 연구분야의 명칭, 이용되는 원리, 법칙, 알고리즘, 현상 등의 고유 명사 등이 포함된다.

[0041] 데이터 추출부는 연구자료에서 분석에 필요한 정보를 추출한다. 만약, 연구자료가 논문인 경우, 논문에서 초록(abstract) 및 키워드(keywords)에 대응하는 영역을 추출한다.

[0042] 데이터 추출부는 연구자료가 이미지 파일 또는 이미지화된 PDF파일과 같이 이미지로 구성된 문서인 경우 이미지 프로세싱을 수행하여 이미지에서 문자를 추출한다. 논문이 이미지화된 문서인 경우, 초록이 시작되는 지점에 시작좌표와, 초록이 종료되는 지점에 종료좌표를 설정하고, 상기 종료좌표의 x좌표 값은 초록 문단의 최우측의 x좌표로 치환한 후 상기 시작좌표와 상기 종료좌표를 양 끝점으로 하는 사각 범위 내에 포함된 텍스트를 추출한다(도 2 참조).

[0043] 이 실시예의 데이터 추출부는 Apache(Apache Software Foundation)에서 제작한 PDFBox를 이용하여 이미지화된 논문 PDF 파일을 분석함으로써 텍스트를 추출하였다. 추출된 텍스트에서 초록 또는 abstract라는 단어가 발견되면, 해당 페이지에 초록이 포함된 것으로 판단하며, 초록 내용을 추출하기 위한 사각형영역(region of the abstract)을 설정한다. 마찬가지로, 추출된 텍스트에 키워드 또는 keywords가 존재하면 키워드가 포함된 페이지로 판단하고, 키워드 단어를 추출하기 위한 사각형영역을 설정하였다.

[0044] 초록 내용에 사각형영역을 설정하기 위해 PDFBox 라이브러리를 이용하여 초록의 시작좌표(x1, y1)와 종료좌표(x2, y2)를 추출한다. 초록의 시작좌표는 '초록' 또는 'Abstract'가 등장하는 페이지에서 이 단어의 좌측 상단 x, y 좌표가 된다. 초록의 종료좌표는 초록 내용의 최우측 x 좌표와 하단의 y 좌표로 지정된다. 일반적으로, 이미지형식의 논문에서 초록 내용은 줄바꿈을 통해 복수의 텍스트 라인(text line) 형식으로 기재되므로, 초록의 최우측에 대한 x 좌표는 텍스트 라인 중 마지막 단어의 최우측 x 좌표가 가장 큰 값이 종료좌표의 x 좌표가 된다.

[0045] 또한, 데이터 추출부는 단어 간의 공백을 판단하기 위해 공백 임계 너비를 기 설정하고, 단어의 좌측 또는 우측에 위치한 공백의 너비가 상기 공백 임계 너비를 초과하면 해당 텍스트 라인이 종료된 것으로 판단한다. 단어를 W, 추출된 단어의 수를 n, 전체 단어를 WA, i번째 추출된 단어를 W_i라 정의할 때, W_i의 영역은 $WC_i(x1,\ y1,$

$x2$, $y2$), W_i의 시작 좌표 WCS_i는 ($x1_{W_i}$, $y1_{W_i}$), W_i의 종료 좌표 WCE는 ($x2_{W_i}$, $y2_{W_i}$)가 된다. $W_i \in WA$의 조건을 만족할 때, i-1번째와 i번째 단어 간의 공백 너비 SX는 $x2_{W_i} - x1_{W_{i-1}}$로 산출된다. 이 값들 중 최대값 SX_{max}는 단어 사이의 공백으로 인식되는 공백 임계 너비로서, 1번째부터 n번째까지의 공백 SX 중 가장 큰 값으로 설정된다. 공백의 정의는 특정 픽셀의 RGB 값이 검은색(0x000000)보다 흰색(0xFFFFFF)에 가까운 경우를 의미한다. 이 공백 임계 너비를 이용하여 각 단어가 종료되는 x 좌표 이후의 공백이 다음 단어와 분리하기 위한 띄어쓰기인지, 줄바꿈을 위한 텍스트 라인의 끝 인지, 혹은 한 페이지의 우측 여백을 나타내는 것인지 판단할 수 있다. 어떤 단어의 우측 공백 너비가 SX_{max} 보다 크면, 그 단어의 종료 좌표 WCE_i($x2_{W_i}$, $y2_{W_i}$)의 $x2_{W_i}$를 초록 내용의 최우측 x 좌표로 설정한다.

[0046] 초록의 하단 끝 y 좌표를 구하는 방법도 우측 끝 x 좌표를 구하는 방법과 유사하다. 각 단어의 다음 줄에 있는 단어와의 높이 SY_i는 $y2_{W_i} - y1_{W_{i-1}}$으로 구할 수 있고, 이들 중 임계 높이 공백 SY_{max}를 구하여 하단의 공백들을 순차적으로 분석한다. 하단의 공백이 SY_{max} 보다 크면, 초록이 종료되거나 문서 하단의 여백으로 판단하여 해당 단어의 종료 좌표 WCE_i($x2_{W_i}$, $y2_{W_i}$)의 $y2_{W_i}$를 초록의 최하단 y 좌표로 설정한다. 추출 된 x, y 좌표가 종료좌표로 설정되면, 시작좌표와 종료좌표를 각각 좌상측 꼭짓점과 우하측 꼭짓점으로 하는 사각형영역를 형성한 후 사각형 영역 내에 위치하는 텍스트에 대해 이미지 프로세싱을 실시한다. 추출된 텍스트는 초록으로 최종 결정된다.

[0047] 논문의 키워드도 초록의 사각형영역을 설정하는 방법과 동일한 방법으로 시작좌표와 종료좌표를 추출하고, 사각형영역을 설정하여 이미지 프로세싱을 수행하면 키워드에 개시된 단어들을 추출할 수 있게 된다.

[0048] 줄바꿈 후 소개(Introduction, Intro), 연구 범위(Research scope) 등의 다른 문단의 시작을 의미하는 단어가 나타나면, 초록의 내용이 종료된 것으로 판단한다.

[0049] 또한, 데이터 추출부는 추출된 초록 및 키워드 내용 중 불필요한 단어 및 문자를 제거한다. 불필요한 단어나 문자에는 초록, Abstract, 키워드, Keywords, Key words 등이 포함되고, 콤마(',')와 같은 문자도 제거한다.

[0050] 형태소 분석부는 먼저 연구개요의 문자가 한글인지 또는 영어인지 분류한다. 형태소 분석부는 연구개요에 포함된 문장을 문자(char) 단위로 읽고, 해당 문자가 가진 정수(int, 문자 코드) 값을 추출한다. 일반적으로 한글로 기재된 연구자료에는 영어가 혼용되나, 영어로 기재된 연구자료에는 한글이 혼용되는 경우가 적으므로 각 문자에 대응하는 문자 코드(정수)가 0x3131보다 크고 0xD7A3보다 작으면 연구개요 및 연구자료가 한글인 것으로 판단한다. 0x3131보다 크고 0xD7A3보다 작은 값의 문자코드가 없는 연구자료는 영어인 것으로 판단한다.

[0051] 연구개요의 언어가 판단되면 형태소 분석부는 연구개요를 대응되는 언어 분석 방법으로 분석을 실시한다. 이 실시예는 기 공지된 형태소 분석 라이브러리를 이용하였으며, 형태소 분석이 완료되어 출력되는 품사 태깅 정보를 바탕으로 분석에 필요한 단어들을 수집하였다.

[0052] 도 3을 참조하면, 연구개요가 한글인 경우, 이 실시예는 Shineware의 코모란(KOMORAN)을 이용하여 형태소를 분석하였다. 코모란은 사용자 사전파일에 포함된 일반명사 63개, wiki 타이틀을 분석한 고유명서 307,435개를 기반으로 형태소 분석을 수행한다. 코모란은 형태소 분석을 실시하여 태그 유형이 일반명사(NNG), 고유명사(NNP), 한자(SH)인 단어를 추출한다. 또한, 어떠한 태그 마지막에 하이픈(SS)이 있는 경우는 줄바꿈에 의해 단어가 분리된 것으로 판단하고 이전 태그의 단어와 이후 태그의 단어를 결합하여 하나의 태그를 만든다.

[0053] 연구개요가 영어인 경우, 이 실시예는 Stanford NLP Group의 CoreNLP(ver. 3.8)를 이용하여 형태소를 분석하였다. CoreNLP는 형태소 분석을 실시하여 태그 유형이 명사(NN), 형용사(JJ)인 단어를 추출한다.

[0054] 형태소 분석부에서 추출된 일반명사, 고유명사, 한자, 명사, 형용사들이 해당 연구자료의 단어세트가 된다. 만약, 연구자료가 논문이면 단어세트에는 데이터 추출부에 추출된 키워드의 단어도 포함된다.

[0055] 토픽 모델링부는 단어세트에 포함된 단어를 개별적으로 클러스터링(clustering)하여 대표되는 주제의 토픽그룹을 구성한다. 토픽 모델링부는 말뭉치(corpus)를 이용하는데, 이때 말뭉치는 단어세트를 이용하여 구성된다. 상기 말뭉치에 잠재적 디리클레 할당(Latent Dirichlet Allocation, 이하 LDA) 알고리즘을 적용한다.

[0056] LDA 알고리즘은 자연어 혹은 단어들의 집합으로 구성된 텍스트 문서 집합에서 각 문서에 존재하는 주제들을 추

출한다. 즉, LDA 알고리즘은 자연어로 구성된 텍스트 문서 집합으로부터 생성확률모델(Generative Probabilistic Model)을 통해 확률 토픽 모델을 유도하는 알고리즘으로, 각 문서에 어떤 토픽들이 존재하는지에 대한 확률 모델이다.

[0057] 모든 문서는 토픽(주제)을 가지고 있고, 문서들은 다수개의 토픽들과 관련되어 있으며, 문서에 등장하는 단어들은 그 토픽들을 이루기 위한 요소로 간주된다. LDA 알고리즘은 문서에 사용된 단어들이 토픽을 구성하고, 토픽이 결합하여 문서를 구성하는 형태로 모델링한다. 그리고 문서 내에서 단어들 간의 동시등장(co-occurrence) 빈도를 확률화하는 방법을 이용하여 숨겨진 토픽들을 도출한다. LDA 알고리즘은 문서 내에 등장하는 단어의 순서에 상관하지 않고 단어의 출현 횟수만을 고려한다. 토픽별 단어 수의 분포를 기반으로 각 문서에서 출현하는 단어 수의 분포를 분석하고, 해당 문서가 어떤 토픽들을 다루고 있을지 예측한다.

[0058] 이 실시예는 생성된 말뭉치를 LDA 알고리즘 분석을 위한 문서로 설정하고, 상기 말뭉치에서 토픽 모델링 과정을 거쳐 토픽을 추출한다. 예를 들어, JHotDraw라는 프로젝트를 초기 분석하여 'Mycobacterium, tuberculosis-induced, granulocyte-macrophage, colony, stimulating, factor, ...' 등의 단어로 구성된 말뭉치가 생성된다고 가정할 때, 이 말뭉치를 LDA 알고리즘을 이용하여 토픽 모델링을 수행하면, declining, macrophage, cells 등의 단어들이 각각의 분산값 0.154, 0.74, 0.065 과 함께 추출된다. 분산값은 각 단어가 전체 말뭉치에서 차지하는 중요도로 볼 수 있으며, 일정 수치 이상의 값을 가진 단어들을 활용하여 해당 프로젝트의 주요 기능 혹은 특징(feature) 목록으로 설정한다.

[0059] 또한, LDA 알고리즘은 깁스 샘플링 기반으로 구성된다. 깁스 샘플링은 각 연구자료에 K개의 토픽 단어 중 하나를 임의로 할당한다. 이로써 각 문서는 토픽 단어와 해당 토픽 단어의 분포를 갖게 된다. 토픽 단어의 분포 값은 오류가 있는 값이므로 개선을 위해 추가 프로세스를 진행한다. 각 연구자료(d), 각 연구자료(d)에 포함된 단어세트(w), 단어세트(w)에 존재하는 토픽 단어(t)에 대해 두 가지 계산을 수행한다. 첫째, 연구자료(d)의 단어세트(w) 중 토픽 단어(t)의 비율 p(t|d)를 연산한다. 둘째, 모든 연구자료 중에서 토픽 단어(t)가 할당된 비율 p(w|t)를 연산한다. 이후, p(t|d)와 p(w|t)의 곱에 따라 토픽 단어(t)를 신규하게 선택한다. 이 생성모델(generative model)에 따르면, 이것은 토픽 단어(t)가 단어세트(w)를 생성할 확률이라 볼 수 있으므로 현재 각 연구자료의 토픽 단어를 해당 확률에 따라 다시 설정한다. 즉, 이 단계에서는, 현재 측정되고 있는 단어 외에 토픽 단어가 전부 알맞게 할당되었다고 가정하고, 확률을 계산하여 현재 단어를 갱신한다. 이와 같은 일련의 과정들을 충분히 반복하여 안정적인 상태가 되면 문서에 존재하는 토픽 단어와 그 분포를 확인할 수 있다.

[0060] LDA 알고리즘 기반 토픽 모델링은 각 연구자료가 k 개의 토픽 단어 중 하나 이상을 포함하는 것을 가정한다. 모델링의 결과물인 토픽그룹은 임의의 토픽 단어의 집합이다. 예를 들어, 신문에서 토픽 모델링을 수행하여 제1그룹은 김경호, 바이올린, 드럼, 음악회라는 단어가 추출되고, 제2그룹은 독도, 북한, 정상회담, 핵이라는 단어가 추출되었다고 가정할 때, 제1그룹은 '음악'과 관련된 토픽그룹이 되고, 제2그룹은 '정치'와 관련된 토픽그룹이 된다. 즉, 토픽 모델링부는 분포 값을 기준으로 단어들을 클러스터링 하여 특정 주제를 대표하는 토픽 단어의 집합을 구성하여 토픽그룹을 형성한다.

[0061] 토픽 모델링부는 토픽 모델링을 수행하기 위해 MALLET Topic Modeling Toolkit 라이브러리를 이용한다.

[0062] 토픽 모델링부는 검색문과 관련된 연구자가 검색되면 해당 연구자가 포함된 연구자료에서 토픽그룹을 추출하고, 검색문과 매칭되는 토픽 단어를 연결하여 검색문-토픽 매핑을 실시한다.

[0063] 단일 혹은 복합 단어로 구성된 검색문의 의미가 모호할 수 있는 문제를 해소하기 위해, 토픽그룹의 토픽 단어와 논문에 개시된 키워드를 조합하여 유사도 비교를 수행한다.

[0064] 토픽그룹의 단어 중 검색문에 포함된 단어가 있으면, 해당 토픽그룹을 검색문의 단어와 연결한다. 예를 들어, 검색문이 'GM-CSF, MEK1, Mycobacterium tuberculosis, P38 MAPK, PI3-K'인 경우, 제1토픽그룹은 protein, k/mek, suggest, kinase, induction과 매핑 될 수 있다. 제2토픽그룹은 kinase, mapk, treated, inmma, increase과 매핑 될 수 있다. 제3토픽그룹은 gm-csf, infection, factor, mediated, up-refulation과 매핑 될 수 있다. 제4토픽그룹은 bymtb, mma, mapk-associatedsignaling, mek, thp과 매핑 될 수 있다. 제2토픽그룹은 mapk가 검색문의 P38 MAPK와 매핑되고, 제3토픽그룹은 gm-csf가 검색문의 GM-CSF와 매핑된다. 이와 같은 매핑 과정으로 검색문-토픽 맵이 완성된다(도 3 참조).

[0065] 사용자 질의부는 사용자가 작성한 검색문을 입력받는다. 검색문은 단일 단어 또는 복합 단어로 구성이 가능하다. 입력된 검색문은 형태소 분석부에서 품사 태깅을 수행하여 검색에 활용 가능한 단어로 최적화 된다.

[0066] 사용자 질의부는 모니터 장치에 연구자의 이름이나 단어가 포함된 검색문을 입력할 수 있는 구성을 제공한다.

이 실시예는 연구자의 이름을 입력하는 연구자 입력상자와 단어를 입력하는 키워드 입력상자를 별개로 제공하였다(도 4 참조). 사용자 질의부에 검색문이 입력되면 가시화부가 모니터에 연구자 네트워크를 구성하여 표시한다.

[0067] 연구자 입력상자에 이름의 일부를 입력하면, 해당 문자가 포함된 연구자 이름이 자동완성 목록에 표시된다. 자동완성 목록은 입력상자 바로 아래에 레이어 형태로 출력된다. 자동완성 목록에 표시된 연구자 이름 중 어느 하나를 선택하면 해당 연구자의 학교, 학과, e-mail 등이 추가로 표시될 수 있다.

[0068] 키워드 입력상자에는 검색을 원하는 단어를 입력할 수 있다.

[0069] 또한, 사용자 질의부는 검색 대상이 되는 연구자료를 논문, 과제, 지식재산권 중에서 선택할 수도 있다.

[0070] 유사도 측정부의 관련연구자 탐색은 중심연구자의 토픽그룹과, 다른 연구자의 토픽그룹을 자카드(jaccard) 알고리즘, SL(Scaled Levenshtein) 알고리즘 및 Soft TF/IDF 알고리즘을 이용하여 유사도를 연산하는 것으로 실시된다.

[0071] 자카드 알고리즘, SL 알고리즘 및 Soft TF/IDF 알고리즘의 유사도 값이 모두 1이면 중심연구자와 비교된 연구자를 유사한 연구를 수행하는 관련연구자인 것으로 결정된다.

[0072] 유사도 측정부는 크게 두 가지 기능을 수행한다. 첫째, 검색문을 기반으로 중심연구자 및 단어 검색을 수행한다. 둘째, 각 연구자의 연구자료에서 추출된 토픽그룹을 기반으로 연구자 간의 연구 유사도를 측정한다.

[0073] 중심연구자 및 단어의 검색은 형태소 분석이 수행된 검색문을 이용하여 데이터베이스 중에서 like 검색(문자열 검색)을 수행하는 것으로 실시될 수 있다. 검색 결과는 연구자 목록 및 단어 목록으로 생성되며, 이 목록은 가시화부로 전달된다.

[0074] 연구자 간의 유사도는 토픽 모델링부에서 토픽 모델링 수행 후 생성되는 토픽그룹을 기반으로 측정된다. 유사도 측정을 위해 자카드 알고리즘, SL 알고리즘, Soft TF/IDF 알고리즘이 이용된다.

[0075] 자카드는 집합 간의 유사도를 검사하는 방법이다. 자카드 유사도 J(A,B)는 두 집합의 교집합 크기를 두 집합의 합집합 크기로 나눈 값으로 정의되며, 그 관계는 수학식1과 같이 나타낼 수 있다.

[0076] [수학식1] $$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|}$$

[0077] SL 알고리즘은 편집 거리 알고리즘으로 알려진 레벤슈타인(Levenshtein) 알고리즘의 결과 값을 보정한 것이다. SL 알고리즘은 문자열 a, b에 대해 a와 b가 같아지기 위해 몇 번의 연산을 수행해야 하는지 계산한다. 여기서 연산은 삽입, 삭제 및 대체를 나타낸다. 두 문자열 a, b에 대해 |a|와 |b|가 각각 문자열 a, b의 길이를 나타내는 경우, $lev_{a,b}(|a|,|b|)$는 수학식2와 같다.

$$lev_{a,b}(i,j)=\begin{cases}\max(i,j) & \\ \min\begin{cases}lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i\neq b_j)}\end{cases} & \end{cases}$$

$$if\ \min(i,j)=0,$$

[0078] [수학식2] $$otherwise.$$

[0079] 수학식2의 결과 값은 정수로 출력된다. 단어의 길이 및 연산량에 따라 값의 변동폭이 크기 때문에 SL 알고리즘은 결과 값을 보정하여 0 내지 1 사이의 값을 생성한다. 보정 식은 $lev_{a,b}(i,j)$의 값이 d이고, 문자열 a, b의 길이 중 큰 값을 n으로 둘 경우, 수학식3과 같다.

[0080] [수학식3] $$SL=1+(d/n)$$

[0081] Soft TF/IDF 알고리즘은 TF/IDF에 부분 매치를 고려한 가중치 측정방법이다. TF/IDF는 다수의 문서로 구성된 문서 집합이 있을 때, 통계적으로 특정 단어가 특정 문서에서 차지하는 중요도를 나타낸다. TF는 문서 내에 등장

하는 단어의 빈도(term frequency)를 나타내며, 이 값이 높을수록 중요 단어로 고려된다. 해당 단어가 다른 문서에서도 자주 사용된다면 흔히 사용되는 것을 의미하는데, 이것을 문서 빈도(document frequency, DF)라 한다. 이 값의 역수를 역 문서 빈도(inverse document frequency, IDF)라고 하며, TF/IDF 알고리즘의 결과는 TF와 IDF를 곱한 값이 된다. 다만, TF/IDF 알고리즘은 오탈자를 고려하지 않아 단어의 토큰이 조금만 상이해도 다른 단어로 인식한다. 이것을 보정하기 위해 Soft TF/IDF 알고리즘은 문자 및 토큰 기반의 유사도 측정을 모두 수행한다. Soft TF/IDF 알고리즘의 토큰 유사도 측정을 위해 Jaro, JaroWinkler 등의 문자열 비교 알고리즘과 Threshold 값을 입력으로 지정할 수 있다.

[0082] 유사도 연산을 위해 토픽 모델링부에서 전달 받은 N명의 연구자의 토픽그룹을 순차적으로 비교한다. 연구자 간의 비교는 총 $\frac{N(N-1)}{2}$ 회 수행된다. 이때, 자카드 알고리즘 유사도는 S_{JAC}, SL 알고리즘의 유사도는 S_{SL}, Soft TF/IDF 알고리즘의 유사도는 S_{SOFT}이다.

[0083] 제1연구자(중심연구자)와 제2연구자를 비교하여 세 알고리즘의 유사도 결과 중 하나라도 0이 나오면 제2연구자를 관련연구자에서 제외한다. 만약, 모든 비교 값이 0이 나오는 경우에는 토픽그룹의 단어를 비교한다. 이 값에서도 0이 나오면 제1연구자와 제2연구자는 유사도는 없는 것으로 판단하고 제2연구자는 관련연구자에서 제외된다.

[0084] 유사도 비교 시 $S_{JAC} + S_{SL} + S_{SOFT} = 3$, 즉 세 알고리즘의 유사도 값이 모두 1이 나오면 제1연구자와 제2연구자의 유사도가 매우 높은 것으로 판단한다. 또한, 토픽그룹의 비교 후 일정 값 이상의 수치가 나오면 제2연구자를 관련연구자로 결정한다. 제2연구자가 관련연구자로 결정된 후 토픽그룹의 단어를 비교하여 유사 강도를 연산한다.

[0085] 이 외의 값에 대해서는 세 알고리즘의 유사도 값을 비교하여 판단한다.

[0086] 세 알고리즘 중 두 가지 이상의 유사도 값이 0.5 미만이면 제2연구자는 관련연구자 후보에서 제외된다. 유사도 값이 0.5 미만인 것은 의미적 유사성을 가진다고 보기 어려운 단어들이다. 실험결과, 유사도 값 0.5 미만은 단지 동일한 문자가 존재할 경우에 출력되는 값으로 확인되었다.

[0087] 반면, 세 알고리즘 중 두 가지 이상의 유사도 값이 0.5 이상이면 제2연구자는 관련연구자가 될 확률이 높은 것으로 판단하고, 토픽그룹의 단어들의 비교를 통하여 최종적으로 관련연구자 여부를 결정한다. 관련연구자로 결정되면 토픽그룹의 단어를 비교하여 유사 강도를 연산한다.

[0088] 관련연구자의 정보는 연구자 데이터베이스에 저장된다.

[0089] 도 4 내지 도 6을 참조하면, 이 실시예의 사용자 질의부 및 가시화부는 JSP 및 JQuery 기반의 웹 화면으로 구현되었다.

[0090] 모니터 장치에 표시되는 사용자 질의부에 대응되는 기능에 사용자가 연구자의 이름이나 단어가 포함된 검색문을 입력하면, 가시화부는 연구자 네트워크를 구성하여 모니터에 표시되게 한다.

[0091] 예를 들어, 사용자 질의부에 연구자의 이름이 입력된 후 검색이 실시되면, 가시화부는 해당 연구자의 논문, 과제, 지식재산권에서 추출된 주요 토픽이 빈도수 또는 초성 순서로 출력되게 한다. 단어 검색이 실시되면, 가시화부는 해당 단어를 포함하는 연구자료의 연구자와, 해당 연구원의 관련연구자료의 실적이 출력되게 한다. 결과 목록에는 연구자 네트워크와 관련 토픽 및 연구자의 정보를 확인할 수 있는 아이콘이 함께 출력된다.

[0092] 검색 결과 목록에서 특정 토픽과 함께 표시되는 네트워크 아이콘을 클릭하면, 가시화부는 해당 토픽을 기반으로 특정 연구자와 관련 있는 연구자를 실선으로 연결한 네트워크 화면을 출력한다. 네트워크는 연구자들의 토픽을 기반으로 연산된 유사 강도 값으로 구성된다.

[0093] 화면의 상단에는 연구자 이름과 네트워크를 구성하는 대표 토픽을 출력한다. 화면의 우측 상단에는 키워드가 추출되는 영역을 논문, 과제, 지식재산권으로 필터링할 수 있는 버튼이 있으며, 선택 및 해제에 따라 키워드가 출력되는 산출물의 영역을 제한 및 설정할 수 있다.

[0094] 연구자 네트워크 중 원, 다각형과 같은 도형은 각각 연구자를 나타낸다. 도형에는 연구자의 이름이 기재된다. 도형을 클릭(선택)하면 해당 연구자의 토픽 단어가 도형 외주면에 표시된다. 표시되는 공간의 제약으로 해당 연구자의 토픽 단어를 상위 빈도수에 따라 일부만 표시하고, 도형 주위의 회색 영역을 클릭하면 나머지 토픽 단어

가 추가로 표시된다. 토픽 단어의 우측에는 연구자료에서 등장한 빈도수가 표시되며, 토픽 단어를 클릭하면 해당 토픽 단어를 포함하는 연구자료의 제목, 키워드, 저자, 관련 토픽 목록 등이 표시된다.

[0095] 연구자를 연결하여 네트워크가 형성되게 하는 연결선은 어느 한 도형 및 다른 도형과 대응되는 연구자 간의 관련된 토픽 단어의 수(유사 강도)에 대응하여 굵기가 결정된다. 연결선에는 연구자 간에 관련된 토픽 단어의 수가 표시되며, 연결선이 클릭(선택)되면 관련된 토픽 단어의 목록이 표시된다.

[0096] 화면 좌측의 윈도우에는 사용자가 선택(클릭)한 연구자의 전체 토픽이 빈도순서로 출력된다. 윈도우의 하단에 있는 텍스트박스에 단어를 입력하고, 텍스트박스 하단에 있는 검색 버튼을 선택하거나, 윈도우에 표시된 토픽 중 일부의 체크박스에 체크하고 검색 버튼을 선택하면, 입력된 단어나 선택된 토픽을 기반으로 새로운 연구자 네트워크가 구성된다.

[0097] 화면 하단에는 연구자의 학과를 그룹지어 출력할 수 있는 학과 필터가 위치한다. 각 학과를 클릭하면 동일한 학과는 동일한 색으로 배경이 출력된다.

[0098] 이어서, 본 발명의 일 실시예에 따른 융합 연구 촉진을 위한 연구원 맵 구축 방법을 설명한다.

[0099] 도 7 및 도 8을 참조하면, 본 발명의 일 실시예에 따른 융합 연구 촉진을 위한 연구원 맵 구축 방법은 데이터 추출부가 연구자료에서 연구자와 연구개요를 추출하여 데이터베이스를 구축하는 단계(S120)와, 형태소 분석부가 상기 연구개요에서 명사 및 형용사를 추출하여 단어세트를 생성하는 단계(S140)와, 토픽 모델링부가 말뭉치(corpus)를 이용하여 상기 단어세트 중 상기 연구자료의 주제가 되는 토픽 단어들을 추출하고, 관련되는 토픽 단어들을 그룹지어 토픽그룹을 형성하는 단계(S160)와, 사용자 질의부가 사용자로부터 검색문을 입력받는 단계(S220)와, 유사도 측정부가 상기 검색문과 상기 토픽그룹을 이용하여 중심연구자를 검색하는 단계(S240)와, 상기 유사도 측정부가 상기 중심연구자의 연구자료에서 추출된 토픽그룹과 다른 연구자료에서 추출된 토픽그룹을 비교하는 것으로 상기 중심연구자와 유사한 연구를 수행하는 것으로 판단되는 관련연구자를 탐색하는 단계(S260)와, 가시화부가 상기 중심연구자와 상기 관련연구자가 서로 연결되는 연구자 네트워크를 구성하여 표시하는 단계(S280)를 포함한다.

[0100] 도 9를 참조하면, S120 단계는 구체적으로, 연구자료가 논문, 연구과제, 지식재산권 등 중에서 어떠한 종류의 문서인지 분류하는 단계(S121)와, 분류된 문서가 이미지화된 문서인지 식별하는 단계(S122)를 포함한다.

[0101] 만약, 논문이 이미지화된 경우, 문서를 대상으로 이미지 프로세싱을 실시하여 '초록' 및 '키워드'에 대응되는 단어를 탐색하는 단계(S126)와, '초록' 및 '키워드'가 개시된 영역에 사각 범위를 설정한 후 사각 범위 내의 텍스트를 추출하는 단계(S127)를 포함한다.

[0102] 만약, 논문이 텍스트를 추출할 수 있는 상태인 경우, '초록' 및 '키워드'의 텍스트를 추출한다(S128).

[0103] S127 단계 또는 S128 단계에서 추출된 텍스트에서 불필요한 단어나 문자, 즉 초록, Abstract, 키워드, Keywords, Key words, 콤마(',')와 같은 문자를 제거하는 단계(S129)를 더 포함한다.

[0104] S121 단계에서 연구자료가 논문이 아닌 경우, 대응되는 방법으로 연구자료에서 연구개요가 추출된다(S123).

[0105] 도 10을 참조하면, S140 단계는 구체적으로, 연구개요가 한글인지 또는 영어인지 분류하는 단계(S145)와, 연구개요가 한글인 경우, 코모란을 이용하여 연구개요에서 일반명사, 고유명사, 한자를 추출하고, 추출된 단어를 단어세트로 구성하는 단계(S146)와, 연구개요가 영어인 경우, CoreNLP를 이용하여 연구개요에서 명사, 형용사를 추출하고, 추출된 단어를 단어세트로 구성하는 단계(S147)를 포함한다.

[0106] 도 11을 참조하면, S260 단계는 구체적으로, 중심연구자의 토픽그룹과 다른 연구자의 토픽그룹을 자카드 알고리즘, SL 알고리즘, Soft TF/IDF 알고리즘을 이용하여 유사도 연산하는 단계(S264)와, 세 알고리즘의 유사도 값을 기 설정된 분류에 따라 분류하는 단계(S265)와, 기 설정된 분류에 따라 다른 연구자를 관련연구자로 선택하거나 관련연구자에서 제외하는 단계(S266)를 포함한다.

[0107] 기 설정된 분류에 따라 다른 연구자를 관련연구자로 선택하거나 관련연구자에서 제외하는 단계(S266)는, 자카드 알고리즘, SL 알고리즘, Soft TF/IDF 알고리즘 모두 유사도 값이 1이면 다른 연구자를 관련연구자 후보로 선택하는 단계(S266a), 자카드 알고리즘, SL 알고리즘, Soft TF/IDF 알고리즘 중 두 개 이상의 유사도 값이 0.5 이상이면 다른 연구자를 관련연구자 후보로 선택하는 단계(S266b), 자카드 알고리즘, SL 알고리즘, Soft TF/IDF 알고리즘 모두 유사도 값이 0이면 다른 연구자를 관련연구자에서 제외하는 단계(S266c), 자카드 알고리즘, SL 알고리즘, Soft TF/IDF 알고리즘 중 두 개 이상의 유사도 값이 0.5 미만이면 다른 연구자를 관련연구자에서 제

외하는 단계(S266d)를 포함한다.

[0108] S266a 및 S266b 단계에서 관련연구자 후보로 선택된 다른 연구자는 중심연구자와 토픽그룹의 토픽 단어들을 서로 비교하여 유사 강도가 기 설정된 기준 이상이 될 때 관련연구자로 선정된다(S268).

[0109] 이상에서 본 발명의 바람직한 실시예를 설명하였으나, 본 발명은 다양한 변화와 변경 및 균등물을 사용할 수 있다. 본 발명은 상기 실시예를 적절히 변형하여 동일하게 응용할 수 있음이 명확하다. 따라서 상기 기재 내용은 하기 특허청구범위의 한계에 의해 정해지는 본 발명의 범위를 한정하는 것이 아니다.

부호의 설명

[0110] 110 : 데이터 추출부

120 : 형태소 분석부

130 : 토픽 모델링부

140 : 유사도 측정부

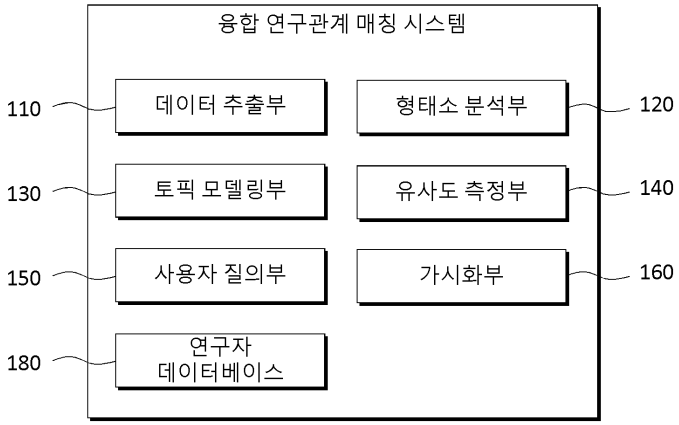
150 : 사용자 질의부

160 : 가시화부

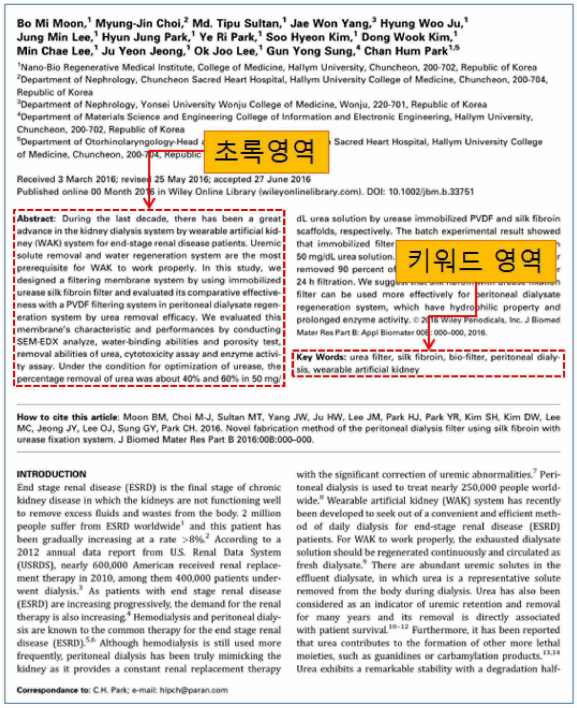
180 : 연구자 데이터베이스

도면

도면1



도면2



Bo Mi Moon,[1] Myung-Jin Choi,[2] Md. Tipu Sultan,[1] Jae Won Yang,[3] Hyung Woo Ju,[1] Jung Min Lee,[1] Hyun Jung Park,[1] Ye Ri Park,[1] Soo Hyeon Kim,[1] Dong Wook Kim,[1] Min Chae Lee,[1] Ju Yeon Jeong,[1] Ok Joo Lee,[1] Gun Yong Sung,[4] Chan Hum Park[1,5]

[1]Nano-Bio Regenerative Medical Institute, College of Medicine, Hallym University, Chuncheon, 200-702, Republic of Korea
[2]Department of Nephrology, Chuncheon Sacred Heart Hospital, Hallym University College of Medicine, Chuncheon, 200-704, Republic of Korea
[3]Department of Nephrology, Yonsei University Wonju College of Medicine, Wonju, 220-701, Republic of Korea
[4]Department of Materials Science and Engineering College of Information and Electronic Engineering, Hallym University, Chuncheon, 200-702, Republic of Korea
[5]Department of Otorhinolaryngology-Head & Sacred Heart Hospital, Hallym University College of Medicine, Chuncheon, 200-704, Republic

Received 3 March 2016; revised 25 May 2016; accepted 27 June 2016
Published online 00 Month 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/jbm.b.33751

Abstract: During the last decade, there has been a great advance in the kidney dialysis system by wearable artificial kidney (WAK) system for end-stage renal disease patients. Uremic solute removal and water regeneration system are the most prerequisite for WAK to work properly. In this study, we designed a filtering membrane system by using immobilized urease silk fibroin filter and evaluated its comparative effectiveness with a PVDF filtering system in peritoneal dialysate regeneration system by urea removal efficacy. We evaluated this membrane's characteristic and performances by conducting SEM-EDX analyze, water-binding abilities and porosity test, removal abilities of urea, cytotoxicity assay and enzyme activity assay. Under the condition for optimization of urease, the percentage removal of urea was about 40% and 60% in 50 mg/dL urea solution by urease immobilized PVDF and silk fibroin scaffolds, respectively. The batch experimental result showed that immobilized filter [illegible] 50 mg/dL urea solution [illegible] removed 90 percent of [illegible] 24 h filtration. We suggest [illegible] filter can be used more effectively for peritoneal dialysate regeneration system, which have hydrophilic property and prolonged enzyme activity. © 2016 Wiley Periodicals, Inc. J Biomed Mater Res Part B: Appl Biomater 00B: 000–000, 2016.

Key Words: urea filter, silk fibroin, bio-filter, peritoneal dialysis, wearable artificial kidney

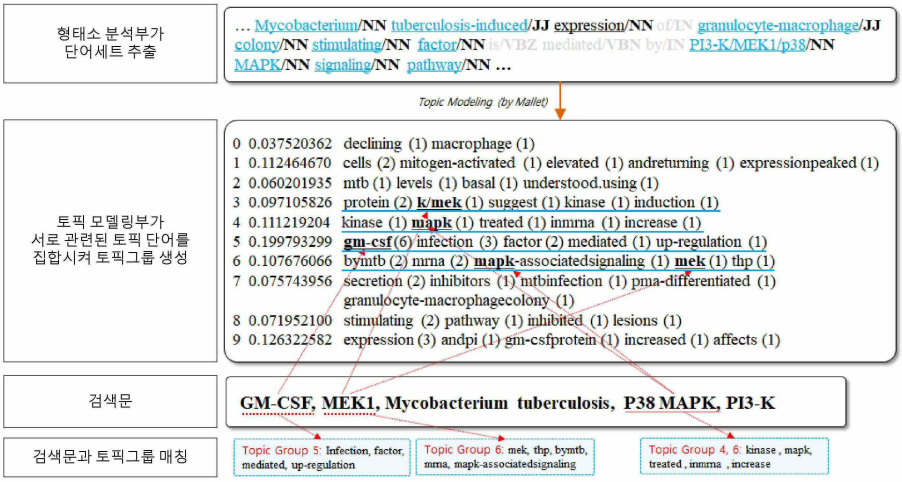
How to cite this article: Moon BM, Choi M-J, Sultan MT, Yang JW, Ju HW, Lee JM, Park HJ, Park YR, Kim SH, Kim DW, Lee MC, Jeong JY, Lee OJ, Sung GY, Park CH. 2016. Novel fabrication method of the peritoneal dialysis filter using silk fibroin with urease fixation system. J Biomed Mater Res Part B 2016:00B:000–000.

INTRODUCTION

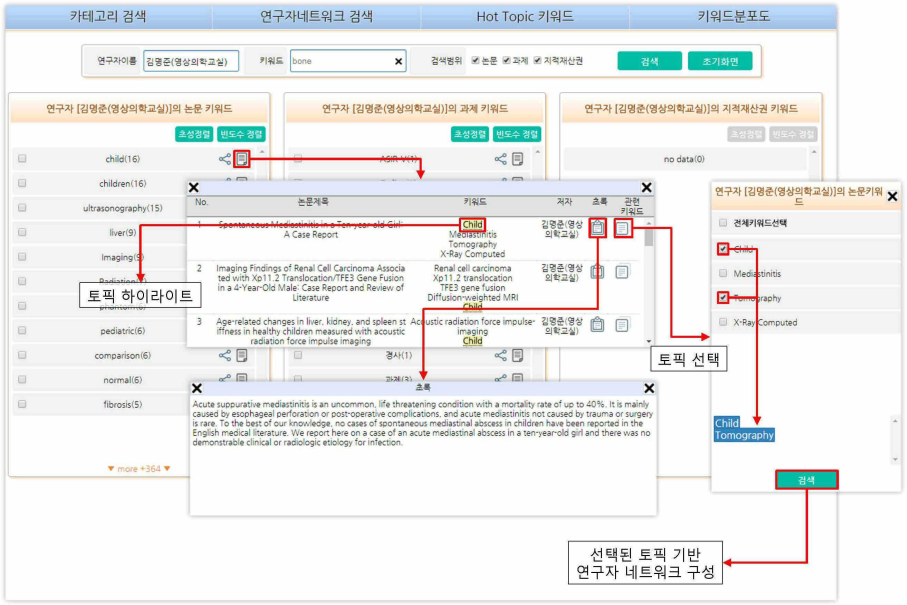
End stage renal disease (ESRD) is the final stage of chronic kidney disease in which the kidneys are not functioning well to remove excess fluids and wastes from the body. 2 million people suffer from ESRD worldwide[1] and this patient has been gradually increasing at a rate >8%.[2] According to a 2012 annual data report from U.S. Renal Data System (USRDS), nearly 600,000 American received renal replacement therapy in 2010, among them 400,000 patients underwent dialysis.[3] As patients with end stage renal disease (ESRD) are increasing progressively, the demand for the renal therapy is also increasing.[4] Hemodialysis and peritoneal dialysis are known to the common therapy for the end stage renal disease (ESRD).[5,6] Although hemodialysis is still used more frequently, peritoneal dialysis has been truly mimicking the kidney as it provides a constant renal replacement therapy with the significant correction of uremic abnormalities.[7] Peritoneal dialysis is used to treat nearly 250,000 people worldwide.[8] Wearable artificial kidney (WAK) system has recently been developed to seek out of a convenient and efficient method of daily dialysis for end-stage renal disease (ESRD) patients. For WAK to work properly, the exhausted dialysate solution should be regenerated continuously and circulated as fresh dialysate.[9] There are abundant uremic solutes in the effluent dialysate, in which urea is a representative solute removed from the body during dialysis. Urea has also been considered as an indicator of uremic retention and removal for many years and its removal is directly associated with patient survival.[10–12] Furthermore, it has been reported that urea contributes to the formation of other more lethal moieties, such as guanidines or carbamylation products.[13,14] Urea exhibits a remarkable stability with a degradation half-

Correspondance to: C.H. Park; e-mail: hlpch@paran.com

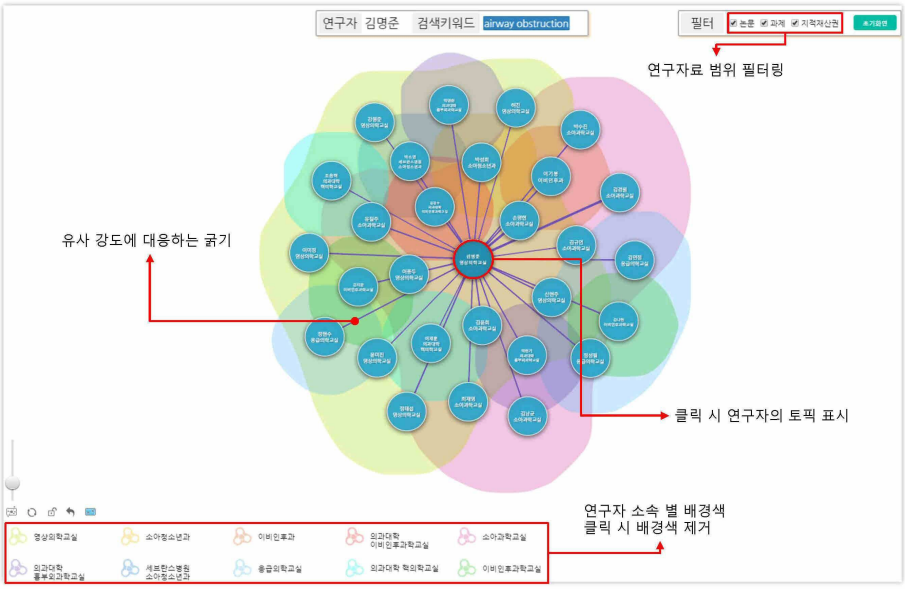
도면3



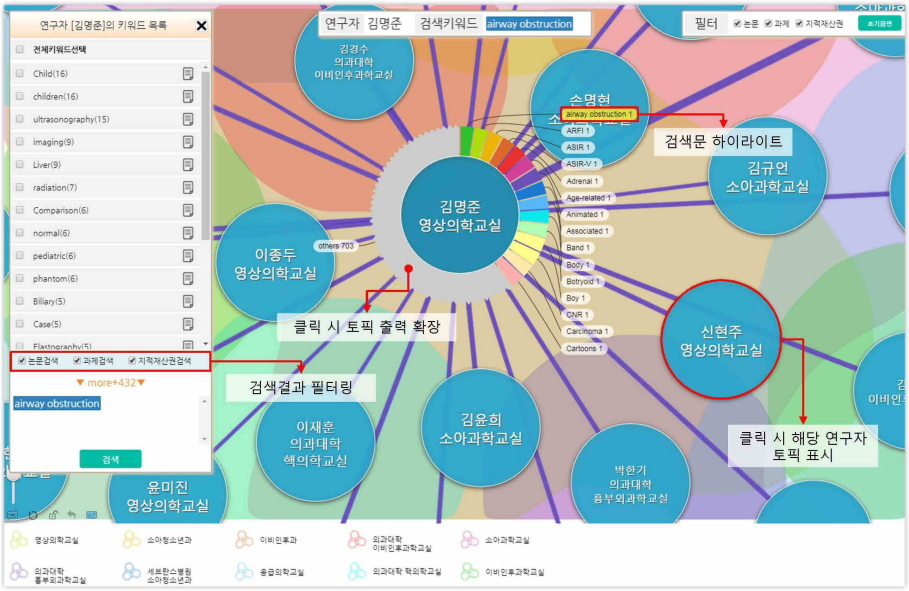
도면4


카테고리 검색
연구자네트워크 검색
Hot Topic 키워드
키워드분포도
토픽 하이라이트
토픽 선택
선택된 토픽 기반
연구자 네트워크 구성

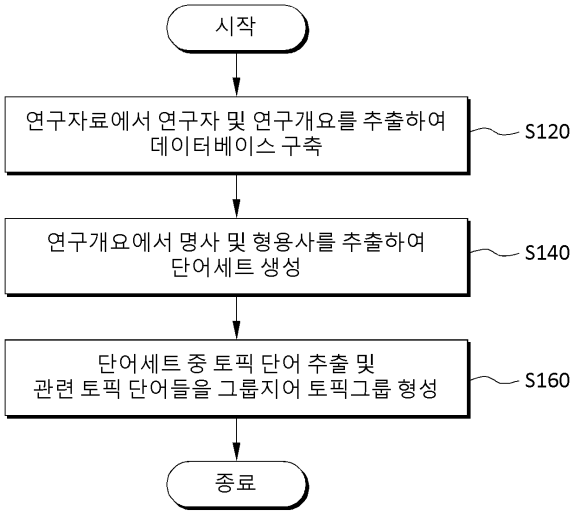
도면5


연구자 김명준
검색키워드 airway obstruction
필터
연구자료 범위 필터링
유사 강도에 대응하는 굵기
클릭 시 연구자의 토픽 표시
연구자 소속 별 배경색
클릭 시 배경색 제거

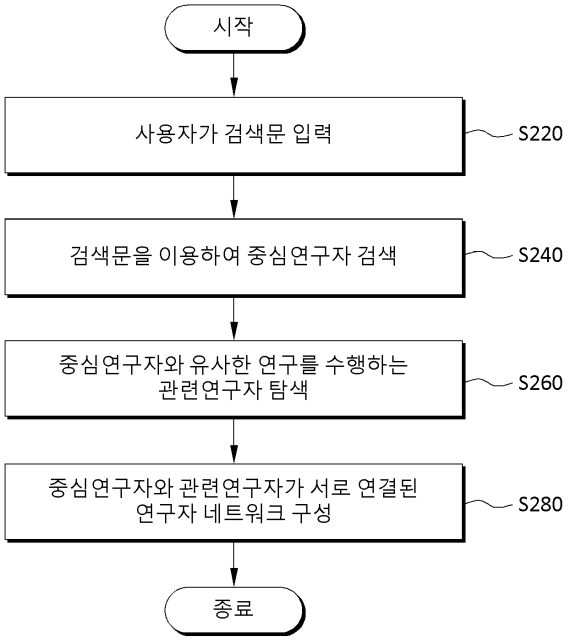
도면6


연구자 [김명준]의 키워드 목록
연구자 김명준
검색키워드 airway obstruction
필터
검색문 하이라이트
김명준
영상의학교실
손명현
김규언
소아과학교실
이종두
영상의학교실
신현주
영상의학교실
김윤희
소아과학교실
이재훈
의과대학
핵의학교실
윤미진
영상의학교실
박한기
의과대학
흉부외과학교실
클릭 시 토픽 출력 확장
검색결과 필터링
클릭 시 해당 연구자
토픽 표시
airway obstruction
검색

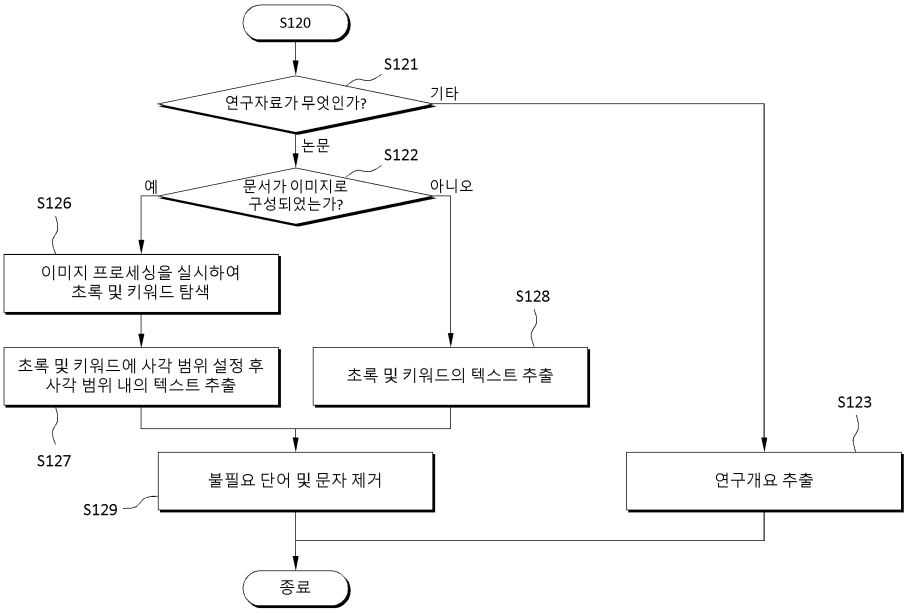
도면7


시작
연구자료에서 연구자 및 연구개요를 추출하여
데이터베이스 구축
S120
연구개요에서 명사 및 형용사를 추출하여
단어세트 생성
S140
단어세트 중 토픽 단어 추출 및
관련 토픽 단어들을 그룹지어 토픽그룹 형성
S160
종료

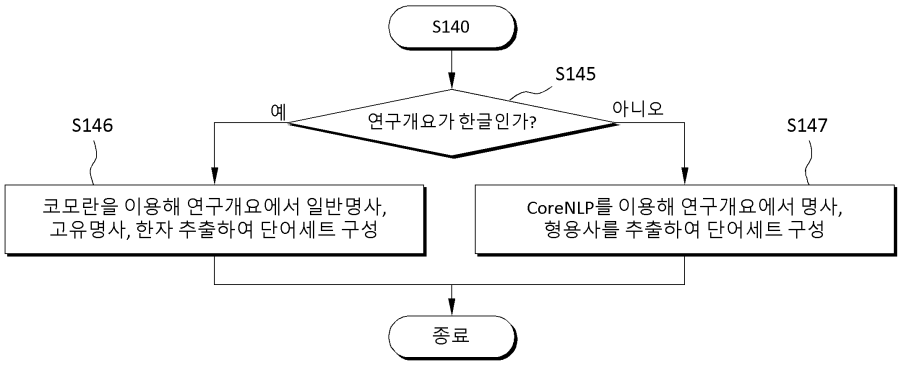
도면8


시작
사용자가 검색문 입력
S220
검색문을 이용하여 중심연구자 검색
S240
중심연구자와 유사한 연구를 수행하는
관련연구자 탐색
S260
중심연구자와 관련연구자가 서로 연결된
연구자 네트워크 구성
S280
종료

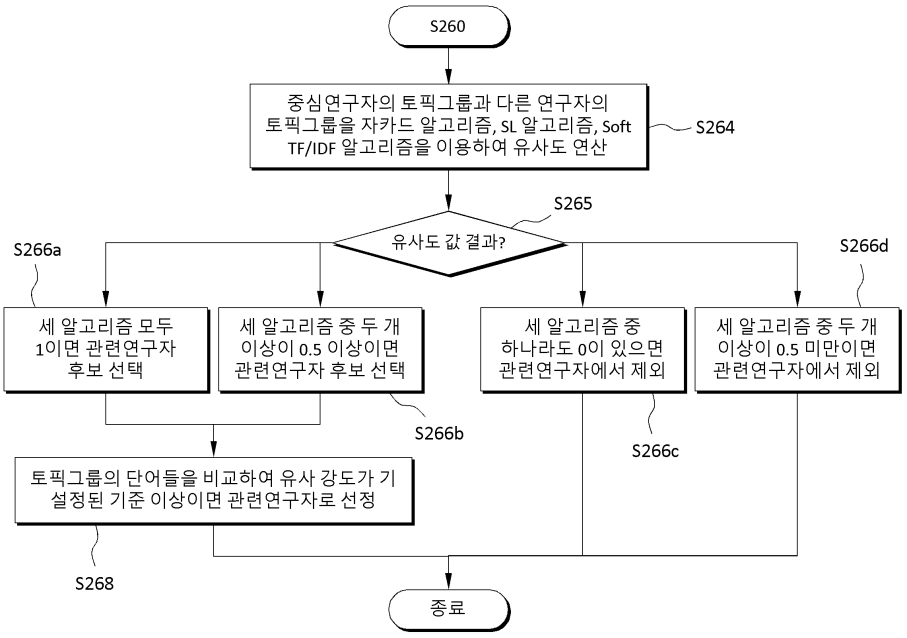
도면9


S120
S121
연구자료가 무엇인가?
기타
논문
S122
문서가 이미지로
구성되었는가?
예
아니오
S126
이미지 프로세싱을 실시하여
초록 및 키워드 탐색
S128
초록 및 키워드에 사각 범위 설정 후
사각 범위 내의 텍스트 추출
초록 및 키워드의 텍스트 추출
S123
S127
불필요 단어 및 문자 제거
연구개요 추출
S129
종료

도면10


S140
S145
예
연구개요가 한글인가?
아니오
S146
S147
코모란을 이용해 연구개요에서 일반명사, 고유명사, 한자 추출하여 단어세트 구성
CoreNLP를 이용해 연구개요에서 명사, 형용사를 추출하여 단어세트 구성
종료

도면11


S260
중심연구자의 토픽그룹과 다른 연구자의 토픽그룹을 자카드 알고리즘, SL 알고리즘, Soft TF/IDF 알고리즘을 이용하여 유사도 연산
S264
S265
유사도 값 결과?
S266a
S266d
세 알고리즘 모두 1이면 관련연구자 후보 선택
세 알고리즘 중 두 개 이상이 0.5 이상이면 관련연구자 후보 선택
세 알고리즘 중 하나라도 0이 있으면 관련연구자에서 제외
세 알고리즘 중 두 개 이상이 0.5 미만이면 관련연구자에서 제외
S266b
S266c
토픽그룹의 단어들을 비교하여 유사 강도가 기 설정된 기준 이상이면 관련연구자로 선정
S268
종료