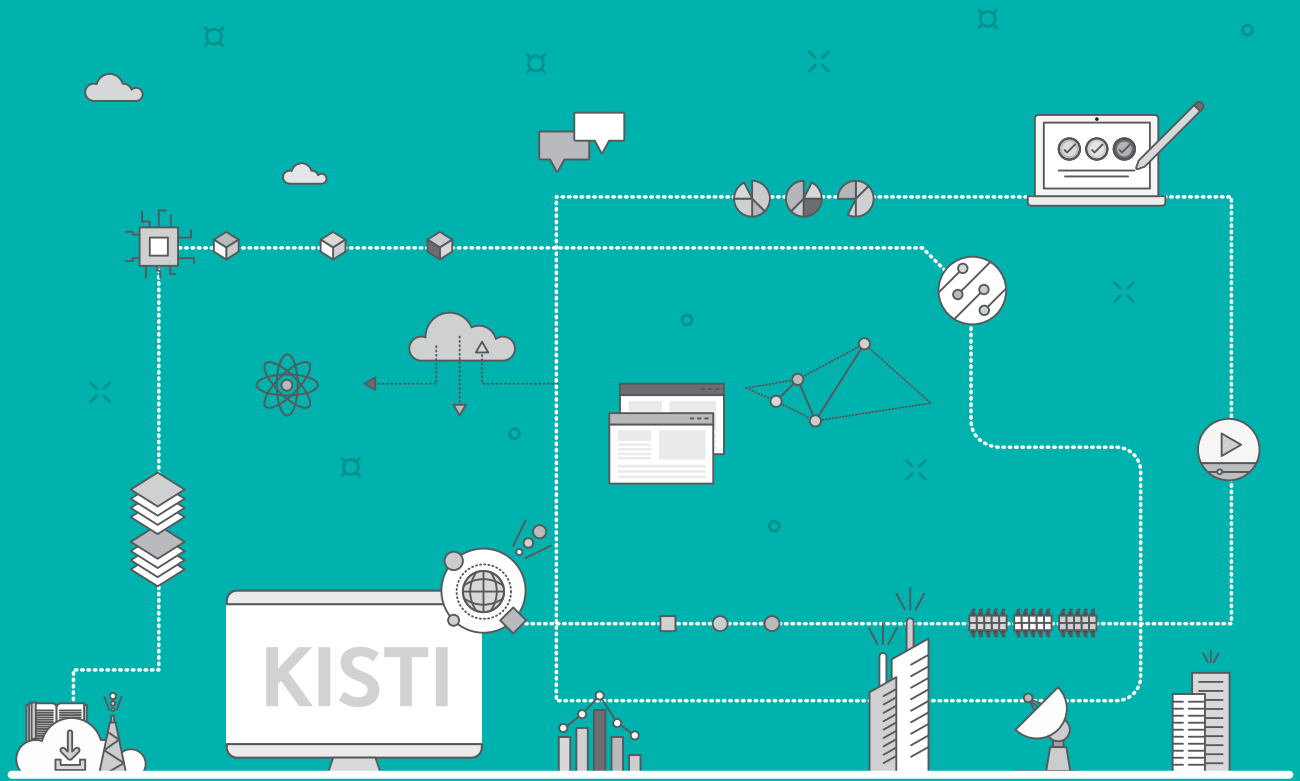
국가과학기술데이터본부

학술정보기반 개체 식별데이터 생성기술 및 데이터



Keyword	식별데이터, 검색, 문헌
연구책임자	이석형
기술 완성단계(TRL)	9단계(기술 사업화 단계)

기술개요

학술 논문, 특허 문헌, 연구 보고서 등의 문헌 데이터를 토대로 연구자 전거 DB를 구축, 연구자 전거 DB를 이용하여 사용자에게 연구자의 학술 정보 검색 서비스를 제공

기존 기술의 문제점

학술정보 검색 시스템은 서지정보와 함께 관심 연구자를 질의어로 하는 검색 서비스를 제공하고 있음, 이에 동일한 연구자가 존재, 연구자의 이름 표기가 같지 않을 수도 있어 정확도가 떨어짐

- 개별 논문 및 특허 문헌 단위까지 연구자 전거 데이터가 구축되어 있지 않은 실정임
- 연구자 전거 데이터를 구축하기 위해 많은 시간과 인력이 소모됨
- 기존 검색 시스템은 단편적인 학술 정보만 제공, 유용하고 가치 있는 통계 분석 정보를 제공하는데 한계가 있음
- 이음동의어와 동음이의어 및 띄어쓰기 오류와 오탈자로 인해 야기되는 클러스터의 정확도 저하 문제

기술 내용 및 차별성

연구자 검색 서비스 제공 방법 및 장치

기술 내용

- 연구자 전거 DB를 기초로 연구자에 대한 다양한 통계분석 정보를 제공
- 학술 논문, 특허 문헌 연구 보고서 등의 문헌 데이터를 기초로 정확하게 동일 연구자를 식별
  → 식별된 저자를 기초로 정확도 높은 연구자 전거 DB를 구축

차별성

- 관심 연구자에 관한 학술 정보를 제공하는 검색 서비스에 있어, 검색 결과의 정확도가 향상될 수 있음
- DB구축 시 소요되는 시간 비용 및 인적 비용 절감
- DB 관리자의 개입 없이도 정확도 높은 연구자 전거 DB가 구축되는 효과가 있음
- 복수의 언어로 구성된 키워드쌍이 주어진 경우에도 정확도 높은 클러스터가 구축

주요기술 구성 및 구현방법

시 작

키워드 리스트 및 문서 리스트 입력

복수의 키워드를 이용하여
초기 클러스터셋 구성

단일 키워드에서 추출된 특징에 기반한
특징 백터의 유사도를 기준으로
클러스터를 병합하는 과정을
반복하여 중간 클러스터셋을 구성

동일 문서에 포함된 복수의 키워드에서
추출된 특징에 기반한 특징 백터의
유사도를 기준으로 클러스터를 병합하는
과정을 반복하여 최종 클러스터셋을 구성

종 료

| 주요기술 구성 |

키워드 클러스터링 방법

- 키워드 리스트에 포함된 각각의 키워드는 한글 키워드, 영어키워드로 구성될 수 있음
- 키워드 식별자(id)가 부가될 수도 있으며, 문서 리스트는 각 키워드가 포함된 문서의 집합으로 구성될 수 있음
- 키워드 클러스터링 장치는 복수의 키워드쌍에 포함된 각각의 키워드쌍을 하나의 클러스터로 생성하는 과정을 반복함으로써, 초기 클러스터셋을 구성할 수 있음

| 구현방법 |

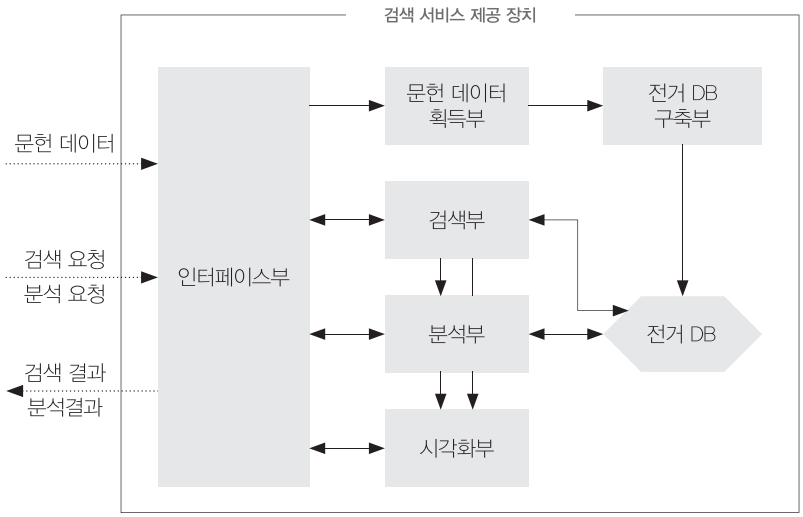
실행 시뮬레이션 소프트웨어의 통계적 특성(작업수행시간, MPI)을 기반으로 적합한 자원을 할당

① 인터페이스부는 연구자 검색 서비스 제공 장치와 다른 장치 간의 인터페이스 기능을 제공

– 하나의 검색 클라이언트 단말로부터 검색 · 분석요청을 수신하고, 검색부 도는 분석부가 제공하는 검색 · 분석결과를 검색 클라이언트 단말로 전송할 수 있음

② 문헌 데이터 획득부는 내부 도는 외부의 데이터소스로부터 다양한 유형의 문헌 데이터를 획득

③ 전거 DB 구축부는 획득된 문헌 데이터를 토대로 연구자 전거 DB를 구축



기술 테스트 현황

| 통합개체식별커널 – 인물 식별커널 정확률 1 |

- 학술정보 인물(저자), 기관(소속정보), 용어(주제어)의 의미 중의성을 해소하여 고유 식별자를 부여하는 세계적 수준의 개체식별기술을 개발하여 학술정보에 대한 개체식별 정확성 94.79% 확보
- 평가셋 현황 : 92,100건의 논문의 저자 287,352명을 대상으로 평가셋을 구축함. 동일 이름 개수는 53,526개이며 실제 인물수는 103,559명임 (1명당 2.77개의 논문)

# of papers	# of author entities	# of same name author groups	# of real authors
92,100	287,352	53,526	103,559

- 평가방법 : k–fold cross validation (k겹 교차검증)

구분	Precision(%)	Recall(%)	F1 measure(%)
교차평가셋–1	94.00	94.12	94.06
교차평가셋–2	95.40	95.43	95.41
교차평가셋–3	94.93	94.86	94.89
Average	94.78	94.80	94.79

| 통합개체식별커널 – 인물 식별커널 정확률 2 |

- 인물 관련 성과물 정보(인물 식별데이터), 소속기관 이력(기관 식별데이터), 연구주제(용어 식별데이터) 등을 해당인물이 e–mail로 직접 확인하는 이용자 참여형 방식으로 정확률 측정
- 조사 기간 : 2016년 1월 1일 ~ 2016년 3월 31일
- 응답 연구자 수 : 1,090명 (응답률 1.33%)
- 개인 저자 Precision의 평균 : 91.29% (연구자 개인 성과물의 정확률 평균치임)

구분	논문	특허	보고서	전체
정확률	95.74%	89.31%	88.83%	91.29%

활용분야 및 권리현황

| 권리현황 |

- 국내 등록특허 2건

발명의 명칭	특허번호	비고
키워드 클러스터링 방법 및 장치	10–1828995	–
연구자 검색 서비스 제공 장치 및 그 방법	10–1823463	–

추가기술정보

기술분류 슈퍼컴퓨팅 – 컴퓨팅 자원 관리

기술문의 이석형 선임연구원 (정보융합연구실) 042–869–1779 | skyi@kisti.re.kr
윤신혜 행정원 (성과확산실 기술이전 담당) 042–869–1832 | shyoon@kisti.re.kr