InSciTe 서비스 명세서

Specifications on InSciTe Service

[ISRL-SD001-004]

한국과학기술정보연구원

Korea Institute of Science and Technology Information

문서 정보

사업명	정보유통 핵심기술 연구개발 및 활용체제 강화
과제명	Technology Intelligence Service 개발
형상번호	ISRL-SD001-004
파일명	ISRL-SD001-004(이미경-101101-InSciTe 서비스명세서).doc
작성자	이미경(jerryis@kisti.re.kr), KISTI 정보유통본부 정보기술연구실
최종수정일	2010-11-01
상태	Release

개정 이력

개정일자	버전	개정내역	작성자	확인자
20100908	001	베타버전 TIS 데이터셋정리	이미경	정한민
20100913	002	베타버전 TIS 데이터셋정리_내용보완	이미경	김평
20101029	003	릴리즈버전 InSciTe 서비스 명세저 작성	이미경	정한민
20101101	004	InSciTe 서비스 명세서 업데이트	이미경	정한민

목 차

1. 데이터 크기 및 입수 방법

(1) 논문: 323,153 건

톰슨사에서 관리하는 2010 년 저널리스트(SCI, SCIE 및 기타) 6 천 여종을 대상으로,

- Green Technology 분야 관련 저널 검색 선정
  - 가) 저널명과 컨텐츠워드(저널의 분야키워드)를 검토하여 수작업으로 키워드 선택
  - 나) 특정키워드(Green*, environment*, ecosystem*, energy*, climatic*)를 타이틀과 컨텐츠워드로 가지는 저널 선택
  - 다) 선택된 저널의 저널명과 컨텐츠워드를 확인해서 수작업으로 부적격 저널 삭제 (예-greenland..)
  - 라) 총 296 개의 저널 중 Web of Sciecne 에서 결과가 나오는 저널 109 종을 최종 선택
  - 마) 선택된 저널의 모든 논문을 데이터 셋으로 활용

(2) 특허: 351,690 건

InSciTe 베타버전용 1 차 데이터 적재(70,829 건)후, 2 차 데이터 적재

- 1 차 데이터 적재 절차
  - 가) 특정키워드(Green*, environment*, ecosystem*, energy*, climatic*)로 특허의 메타 정보 검색 결과 데이터 셋 활용
    - ✓ 한국, 미국, 유럽, 국제(PCT), 일본의 특허 4000 건씩 내려받음(NDSL 검색으로 내려받을수 있는 최대 건수)
    - ✓ 논문 검색 키워드와 동일한 키워드로 검색
    - ✓ 국제는 분류가 WIPO(World Intellectual Property Organization) 되어 있음
    - ✓ 특허 상태 구분 없음
- 2 차 데이터 적재 절차

가) Green, environment, ecosystem, energy, climate, smart, grid, fuel cell, hybrid, carbon, recycling, solar, renewable, hydrogen, 녹색, 환경, 생태계, 에너지, 기후, 스마트, 그리드, 연료전지, 하이브리드, 카본, 탄소, 재활용, 태양, 재생, 수소를 키워드로 가지는 한국, 미국, 유럽, 국제(PCT), 일본의 특허를 검색하여 적재

나) 검색된 40 만건의 특허 논문 중, 1 차 데이터와 중복되는 데이터를 제외하고 적재.

(3) 기술명: 66,638 건

(1)에서 확보된 논문 데이터의 메타데이터 중에서 사용자키워드(author keyword, 사용자가 논문 등록시 직접 입력한 키워드)와 인덱스 키워드(keyword plus, 톰슨 자동 인덱서 결과)를 추출

가) 사용자 키워드와 인덱스 키워드 중에 유일한 용어를 기술명으로 선택

나) 단일어(15,062 건), 복합어(57,481 건)을 기술명으로 채택(72,543 건)

다) 수작업으로 불용어 처리 (예-국가명, 일반명사 등)하여 66,638 건 기술명으로 선정

(4) Triple 데이터: 21,659,732 건

성과물, 기술명, 연구자, 기관, 국가 등의 데이터를 클래스 인스턴스를 생성하고 아래 확장규칙에 따라 트리플 데이터로 확장

- 확장규칙

[tis1: (?x :hasCreatorInfo ?y)(?y :hasCreator ?z)->(?x :hasAuthor ?z)]

[tis2: (?x :hasCreatorInfo ?y)(?y :hasInventor ?z)->(?x :hasInventorE ?z)]

[tis3: (?x :hasCreatorInfo ?y)(?y :hasApplicant ?z)->(?x :hasApplicantE ?z)]

[tis4_1: (?a :hasCreatorInfo ?b)(?b :hasInstitution ?c)->(?a :hasInstitutionInfo ?c)]

[tis4: (?a :hasInstitutionInfo ?b)(?b :hasLocation ?c)->(?a :hasLocationE ?c)]

[tis5: (?x :hasTopicArea ?y)(?y :hasTopic ?z)->(?x :hasTopicE ?z)]

[tis6: (?x :hasPublicationInfo ?y)(?y :hasPublication ?z)->(?x :hasPublicationE ?z)]

[tis7: (?x :hasCreator ?y)(?x :hasInstitution ?z)->(?y :hasInstitutionE ?z)]

[tis8: (?x :hasAuthor ?y)(?x :hasAuthor ?z)->(?y :hasCoauthor ?z)]

[tis9_1: (?x :hasLocation ?y)(?y rdf:type :Nation)->(?x :hasNationInfo ?y)]

[tis9: (?x :hasInstitutionE ?y) (?y :hasNationInfo ?z)->(?x :hasNationE ?z)]

[tis10: (?x :hasTopicRelation ?y)(?y :hasTopic ?z)->(?x :isRelatedTo ?z)]

2. 기술 관계 추출

(1) 기술 관계 추출 대상 데이터

- SINDI 의 텍스트마이닝 기술을 이용

  가) PubMed 의 논문데이터에서 저널명에 에너지, 환경과 관련된 키워드로 검색하여 저널 후보를 선택한후, 최근 10 년이내 (2000~2009 년) 데이터의 초록을 대상 데이터로 선택

  나) 기술명 66,638 건을 대상 데이터에서 검색하여 태깅

  다) 태킹한 정보를 대상으로 기술관계를 추출

(2) 기술 관계 추출 방법

- SINDI 팀의 텍스트마이닝 기술을 이용하여 기술과 기술간의 의미적 구조를 추출한 것을 관계정보로 활용

  가) 기술관계 추출 데이터에서 2 개의 기술명이 나온 문장을 추출

  나) 문장의 관계를 Predicate-argument 방법을 통해 최소 연결 관계를 찾음

다) 문장내에서 두개체를 유의미한 관계로 맺을수 있는 어휘적 자질(의미적 구조)를 추출함 (“과학기술분야 전자문헌의 의미적 심층 분석을 위한 지식발견 통합 플랫폼 개발연구” 보고서, 29 페이지 참조)

(예제) The purpose of the study was to assess the relation between the simultaneous exposure to alcohol and consumption of micronutrients that have protective properties agaist colorectal cancer.

✓ 위의 문장에서 기술명 태깅

The purpose of the study was to assess the relation between the simultaneous exposure to alcohol and consumption of micronutrients that have protective properties agaist colorectal cancer.

✓ 두 기술명간의 최단 의미적 구조 추출

The purpose of the study was to assess the relation between the simultaneous exposure to alcohol and consumption of micronutrients that have protective properties agaist colorectal cancer.

✓ 기술명과 의미적 구조만 기술 관계로 추출함

micronutrients have properties agaist colorectal cancer

- 예정작업: 학습을 통한 관계정보의 정제

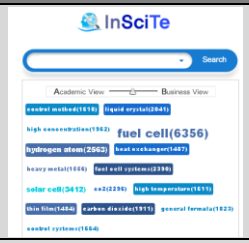
(3) 기술 관계 추출 데이터

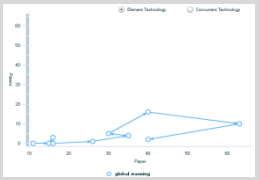
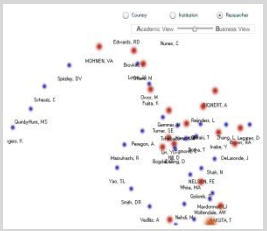
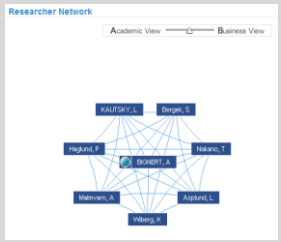
- SINDI 에서 넘겨준 기술 관계 데이터 16 억건의 데이터 중, 중복되는 데이터는 기술관계 가중치로 계산하여 유일한 관계 1,570,404 건 등록 (중복건수 133,329 건)
- 추출된 관계는 기술의 색인 데이터로 적재되고, Technology Network 서비스에서 호출하여 사용

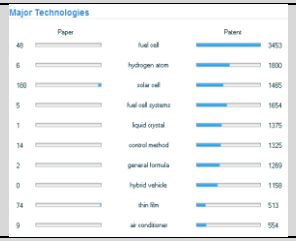
3. InSciTe API 구성

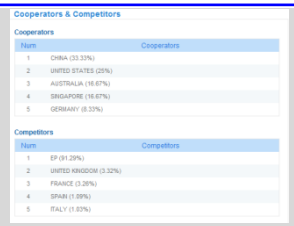
전체 API: 총 38 개

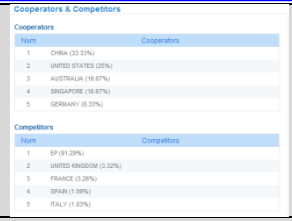
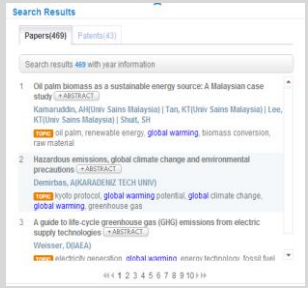
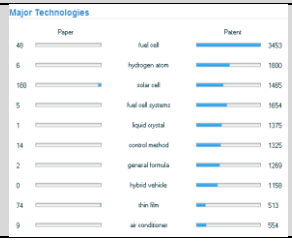
- 검색 API: 18 개 (Search&Inference API 1 건 포함)
- 추론 API: 18 개 (Search&Inference API 1 건 포함)
- 기타 API: 3 개

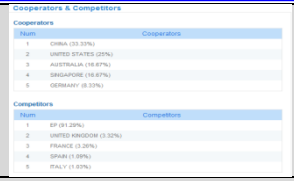
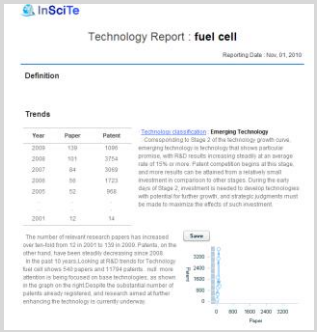
Service	Screen	Method	Input	Description	Service Type
Intro		**getURIInfo**	String keyword	입력어 키워드에 대한 개체판별	Search
		getTechCloud	int year int rate int topN	기술명 태그 클라우드	Search
		getAutoKeyword	String keyword	자동완성	Search
Agent-Technology Map		**getAgentTechMap**	String techURI String xType String yType int startYear int endYear	기술-주체 맵	Inference & search

Technology Network		**getTechNetwork**	String techURI String techName String vType int startYear int endYear int rate int count	기술관계도 네트워크(요소,기능,유사,프로젝트 등을 보여줌)	Search
Technology Trends		**getTimeline**	String techURI String techName String Type int count	기술의 주체별 시계열 그래프	Search
Agent Network		**getAgentGroup**	String techURI String scope int startYear int endYear int rate int count	연구주체(국가, 기관, 연구자)의 협력 연구 그룹-공저자그룹	Inference
		**getAgentNetwork**	String techURI String personURI String scope int startYear int endYear int rate	연구주체(국가, 기관, 연구자)의 연구자네트워크	Inference

Technology Basic Service		**getTechTrend**	String techURI String scope	기술의 성향 그래프	Search
		getIrruptedTech	String techURI String kind String scope	신규 기술 그래프	Search
		**getSearchList**	String uri String keyword String cate int displayNum int pageNum	검색결과리스트 (uri, keyword 는 다중값일 경우 '|' 사용)	Search
		getSearchListCount	String uri String keyword	검색결과 카운트 (uri, keyword 는 다중값일 경우 '|' 사용)	Search
Nation Basic Service		**getAgentStatus**	String techURI String agentURI String scope	성과물의 현황	Search
		**getTechbyAgent**	String agentURI String scope	선택된 국가의 기술의 요소기술별 건수	Inference
		getTechbyAgentR	String agentURI String scope int count	선택된 국가의 기술의 요소기술별 건수 (논문, 특허 분리)	Inference
		getPersonbyAgent	String techURI String agentURI String scope	주체의 소속연구자	Inference

		**getInstbyNation**	String techURI String agentURI	국가의 소속기관	Inference
		getRelatedAgent	String techURI String agentURI String scope	연구주체 경쟁협력관계	Inference
		**getSearchList**	String uri String keyword String cate int displayNum int pageNum int startYear int endYear	검색결과리스트 (uri, keyword 는 다중값일 경우 '|' 사용)	Search
Institution Basic Service		**getAgentStatus**	String techURI String agentURI String scope	성과물의 현황	Search
	베타버전 API	**getInstitutionMetadata**	String institutionURI	연구기관의 정보	Search
	베타버전 API	**getTechbyAgent**	String agentURI String scope	연구기관의 기술	Inference
	베타버전 API	**getPersonbyAgent**	String techURI String agentURI String scope	연구기관의 연구자	Inference

		**getRelatedAgent**	String techURI String agentURI String scope	경쟁/협력관계	Inference
		**getSearchList**	String uri String keyword String cate int displayNum int pageNum int startYear int endYear	검색결과리스트 (uri, keyword 는 다중값일 경우 '|' 사용)	Search
Person Basic Service		**getAgentStatus**	String techURI String agentURI String scope	성과물의 현황	Search
	베타버전 API	**getPersonMetadata**	String personURI	연구자의 정보	Search
		**getTechbyAgent**	String agentURI String scope	연구자의 기술	Inference

		**getRelatedAgent**	String techURI String agentURI String scope	경쟁/협력자	Inference
		**getSearchList**	String uri String keyword String cate int displayNum int pageNum int startYear int endYear	검색결과리스트 (uri, keyword 는 다중값일 경우 '|' 사용)	Search
Report		**getRNDTrend**	String techURI	getTechTrend(techURI, "R")로 대체	Inference
		getPersonbyTech	String techURI	기술관련 연구자	Inference
		getWiki	String keyword	기술정의	DB
		getRelatedTech	String techURI int count	관련기술	Inference
		getPersonPerYear	String techURI	연도별 연구자수	Inference
		getTopPersonPerYear	String techURI	연도별 최고연구자	Inference
	기술 보고서	**getReportTech**	String techURI String techName	기술보고서	기타
	기관 보고서	**getReportInst**	String instURI	기관보고서	기타

			String instName		