Journal of Digital Contents Society Vol. 13 No. 3 Sep. 2012(pp. 317-324)
http://dx.doi.org/10.9728/dcs.2012.13.3.317

# R&D과제의 기술분류를 이용한 사업간 유사도 분석 기법에 관한 연구

김주호*, 김영자**, 김종배***

**요 약**

최근 R&D 투자효율성 제고를 목표로 사업 간의 유사중복 조정에 대한 중요성이 강조되고 있으나, 과제 혹은 예산요구서 내용 등을 텍스트 기반으로 비교하는 기존 유사검색 방식은 내용의 품질 편차 등으로 인해 유의미한 유사성 도출에 제한점이 있다. 이러한 텍스트 기반의 키워드 추출을 통한 유사검색 한계성을 극복하기 위한 방안으로 본 연구에서는 사업 간 유사도 분석 시 과제의 기술분류를 활용한다. 국가R&D사업 조사·분석 시 수집된 과제들의 과학기술표준분류를 추출하여 사업별 고유벡터 모형을 생성 후 이를 이용하여 코사인 기반, 유클리디안 거리기반 알고리즘을 통해 각 사업 간 유사도를 측정하였으며 기존 키워드 추출방식으로 유사도를 측정한 결과와의 비교를 통해 연구 효율성을 검증하였다.

# A study on Similarity analysis of National R&D Programs using R&D Project's technical classification

Ju-Ho Kim*, Young-Ja Kim**, Jong-Bae Kim***

**Abstract**

Recently, coordination task of similarity between national R&D programs is emphasized on view from the R&D investment efficiency. But the previous similarity search method like text-based similarity search which using keyword of R&D projects has reached the limit due to deviation of document's quality. For the solve the limitations of text-based similarity search using the keyword extraction, in this study, utilization of R&D project's technical classification will be discussed as a new similarity search method when analyzed of similarity between national R&D programs. To this end, extracts the Science and Technology Standard Classification of R & D projects which are collected when national R&D Survey & analysis, and creates peculiar vector model of each R&D programs. Verify a reliability of this study by calculate the cosine-based and Euclidean distance-based similarity and compare with calculated the text-based similarity.

**Keywords : Similarity, R&D Program, Vector Model, Euclidean, Technical classification**

## 1. 서론

정부는 과학기술 진흥을 통한 국가 경쟁력 확보를 위해 신성장동력, 녹색기술, 기초·원천 분야의 전략적 투자 확대 등을 통해 정부의 R&D(Research and Development, 이하 R&D) 투자규모를 지속적으로 확대하고 있다. 정부 R&D예산은 2008년 11조원에서 2012년 16조원으로 연평균 9.6%라는 큰 폭의 증가추세를 보이고 있으며, 2010년 투자규모 기준 세계 7위 수준에 도달하였다[11].

최근 R&D 투자규모의 증가와 함께 다수 부처의 경쟁적인 사업 추진 등으로 인한 유사·중복사업 추진에 대한 문제가 지적되어, 정부는 사업·장비 등의 유사 중복성 해소를 통해 불필요한 재정지출을 최소화하여 R&D 투자효율성을 제고하려는 노력을 지속적으로 하고 있다. 그러

※ 제일저자(First Author): 김주호
접수일:2012년 08월 07일, 수정일:2012년 09월 03일
완료일:2012년 09월 11일
* 한국과학기술기획평가원 ahwi@kistep.re.kr
** 한국과학기술기획평가원 yj0219@kistep.re.kr
*** 숭실대학교대학원(교신저자) kjb123@empas.com

나, 사업의 유사·중복 판단 기준이 불명확하고 예산배분·조정과정을 통해 사후적인 유사·중복을 조정하게 되는 등의 적시성 문제와 R&D 사업 예산 요구서나 과제내용의 주요 키워드를 추출·비교하는 텍스트기반의 유사성 검토방법의 한계 등으로 인하여 유사·중복 조정 업무 실효성에 문제가 제기되고 있는 실정이다.

이에 본 논문에서는 기존의 과제단위의 텍스트 기반 유사검색 방식의 한계를 극복하고자 각 사업 내 다수 과제들의 과학기술표준분류를 이용하여 사업의 고유벡터(vector)모델을 생성한 후 이를 사업간 유사도 측정에 활용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 텍스트 마이닝 및 유사도 측정연구에 관한 연구를 살펴보고, 3장에서는 R&D과제의 과학기술표준분류를 이용한 사업 시그니처 모델 생성 방안에 대해 설명한다. 4장에서는 생성된 시그니처 모델을 활용하여 기존 텍스트 방식, 코사인 기반(Cosine-based), 유클리디안 거리기반(Euclidean distance-based)방식으로 R&D사업간 유사도를 계산한다. 또한, 융합기술 과제의 기술분야 비중에 대한 새로운 산정 방안을 제시하고, 이를 활용한 유사도 측정결과를 비교하여 시사점을 도출한다. 끝으로 5장에서는 본 논문의 내용을 요약한다.

## 2. 관련연구

### 2.1 유사도 측정 알고리즘

문서들 간의 유사도를 측정하기 위한 여러 방법 중 본 연구에서는 텍스트 기반의 키워드 추출, 비교 방식과 과학기술표준분류의 출현여부를 희소행렬(sparse matrix)로 표현하여 계산량을 줄일 수 있는 장점이 있는 코사인 측정법 및 이와 반대로 각 사업별 표준분류의 출현빈도를 고려하여 유사도를 측정할 수 있는 유클리디안 거리 측정법을 채택하여 사업간 유사도를 측정하였다[1,2,3,5,12].

텍스트 기반의 유사도 측정은 문서 내에 표현된 단어로부터 문서벡터를 추출·생성하고 형태소 분석, 불용어 제거, 어간화 작업등을 통하여 주요 키워드를 추출하는 과정을 거치게 된다[13,14]. 문서에 표현되는 단어로부터 문서벡터를 생성하고, 벡터화된 훈련 문서들을 예제로 사용하여 학습함으로써 관련된 문서에 범주를 할당하는 자동분류 알고리즘 또한 추출과정과 분류과정으로 나뉠 수 있다. 추출과정은 문서에서 태그와 불용어를 제거하고 형태소분석을 거쳐 특정한 키워드를 추출하는 전처리 과정과 가중치와 같이 문서를 벡터형식으로 표현하는 자원축소과정을 거친다. 이러한 여러 자동문서벡터 생성 기법들 중 SVM(Support Vector Machines)과 TFIDF(Term Frequency Inverse Document Frequency)가 널리 사용되고 있다[7]. 이 중 TFIDF는 각 문서 내 존재하는 단어에 대해 가중치를 부여하여 문서의 벡터로 표현 후, 두 문서간의 유사도 비교를 가능하게 하는 방법이다[2]. 본 논문에서는 TFIDF 방식을 이용하여 각 사업별 예산요구서내의 가중치가 부여된 주요 키워드를 추출하여 문서 벡터를 생성 후 사업간 유사도를 측정하였다.

$$score(d,q) = \sum_{t \in q} (tf(t,d) \cdot idf(t) \cdot Boost(t,d) \cdot norm(d))$$

(수식 1) 텍스트 기반의 유사도 측정 알고리즘

본 논문에서 사용한 텍스트 기반의 유사도 알고리즘은 공개 소스인 루씬스코어(lucene score)로서 (수식 1)과 같다. d는 문서 q는 쿼리, t는 텀(term)이며 tf(t,d)는 해당 텀이 문서에서 발생한 빈도수로서 sqrt(횟수)로 구현된다. idf(t)는 해당 텀이 전체 문서 셋에서 발생한 빈도수로서 log(문서 셋 수 / (해당 텀 빈도수+1))+1 로 산출되며, Boost(t,d)는 해당 텀에 대한 검색 가중치로서 1/sqrt(필드의 텀 갯수) * 필드가중치 * 문서가중치로 구현된다. 마지막으로 norm(d)는 문서간의 스코어를 구하는데는 직접적인 영향이 없으나 쿼리간의 비교를 위한 정규값이다.

코사인 기반 유사도는 두 벡터의 유사도가 대칭 구조인 경우에 주로 사용되는 방법으로, 두 벡터가 이루는 각의 크기에 따라 유사도가 결정된다[3,4,8].

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{n} w_{ik} \cdot w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \cdot \sum_{k=1}^{n} w_{jk}^2}}$$

(수식 2) 코사인기반 유사도 알고리즘

문서 a가 $w_{i1}, w_{i2}, \ldots, w_{iN}$, 문서 b가 $w_{j1}, w_{j2}, \ldots, w_{jN}$의 키워드를 가지고 있을 때, 두 문서의 벡터는 각각 $w_i$, $w_j$이고, 이들은 $w_i=(w_{i1}, w_{i2}, \ldots, w_{iN})$, $w_j=(w_{j1}, w_{j2}, \ldots, w_{jN})$로 표현될 때 문서 a,b 간의 코사인 유사도는 (수식 2)로 계산될 수 있다. 계산된 유사도는 항상 0부터 1까지의 값을 갖게 되며, 각 과제별 기술분류 코드가 상이하여 벡터의 많은 부분이 0값을 가지는 경우에도 큰 오차없이 유사도측정이 가능하여 본 논문과 같은 실험데이터에서 활용이 용이하다.

유클리디안 거리는 두 점 사이의 거리를 계산할 때 흔히 쓰는 방법으로서 두 문서 벡터의 상대적인 거리차를 측정하는 방법이다[6,15].

$$dist\,(d_i, d_j) = \sqrt{\sum_{k=1}^{n}(w_{ik} - w_{jk})^2}$$

(수식 3) 유클리디안 거리기반 유사도 알고리즘

두 문서의 벡터가 각각 $w_i$, $w_j$이고, 이들이 $w_i=(w_{i1}, w_{i2}, w_{i3}, w_{i4}\ldots..)$, $w_j=(w_{j1}, w_{j2}, w_{j3}, w_{j4}\ldots..)$로 표현될 때 유클리디안 거리를 이용한 유사도는 (수식 3)으로 정의된다. 유클리디안 유사도는 거리 값이 작을수록 두 문서가 유사함을 나타내며 2차원 혹은 3차원 데이터에 적용 시 직관적이고, 해석이 쉽다는 장점이 있다[9,10].

## 2.2 사업간 유사 분석 검증 사례 문제점

현재 국가R&D사업에 대한 유사 검증은 예산요구서나 사업 중기계획서상의 내용을 토대로 각 분야의 일부 전문가에 국한되어 수행되며, R&D과제는 과제 기획 시 국가연구개발사업의 관리 등에 관한 규정에 의거 국가과학기술위원회에서 매년 시행하는 조사·분석 시 수집된 R&D선행과제와의 유사성을 검토하도록 의무화하고 있다.

그러나, 국가R&D 효율성 제고를 위한 노력의 일환으로 국가R&D사업 예산요구 단계에서부터 사전 조정 및 통합 등의 노력을 경주하고 있으나 다수의 사업, 짧은 조정기간, 부처 간 이기주의 등의 이유로 인해 가시적인 실효성을 거두기 어려운 실정이다.

이에 과학기술위원회에서는 사업 통합 및 조정 대상에 대한 객관적인 기준마련을 위하여 각 부처에서 제출한 예산요구서 내용 중 연구내용, 목적, 기대효과 등에서 주요 키워드를 추출 후 이를 비교하는 방식으로 사업간 유사도 측정을 시도하였으나, 각 사업별 예산요구서의 내용이 상이하고 작성된 기술수준 편차가 심하여 계속사업임에도 불구하고 <표 1>과 같이 그 유사성이 매우 낮게 측정되는 등의 한계점을 노출하였다.

<표 1> 계속사업에 대한 유사도 측정 사례(예시)

| 계속사업 A | 2011년 예산요구서 | 2010년 예산요구서 |
|---|---|---|
| 2011년 예산요구서 | N/A | 0.03 |
| 2010년 예산요구서 | 0.04 | N/A |

## 2.3 과학기술 표준분류 체계

국가과학기술표준분류 체계는 과학관련 기술정보·인력·연구개발사업등을 효율적으로 관리할 수 있도록 분리한 체계로 2002년 처음 수립된 후 선진국의 과학기술분류 동향을 조사·분석하고 새로운 기술의 출현등을 고려하여 3년 주기로 수정·보완되고 있다. 과학기술표준분류는 자연, 생명, 인공물, 인간, 사회, 인간과학과 기술의 6대 분야, 33개 대분류, 371개 중분류, 2902개 소분류로 구성되어 있으며, 국가 R&D과제들은 모두 이 코드에 따라 분류된다.

<표 2> 융합기술 과제의 과학기술표준분류 비중 예시

| 과학기술표준분류 | | | 비중 | | |
|---|---|---|---|---|---|
| 수학 (A) | 물리학 (B) | 재료 (I) | 50 | 30 | 20 |

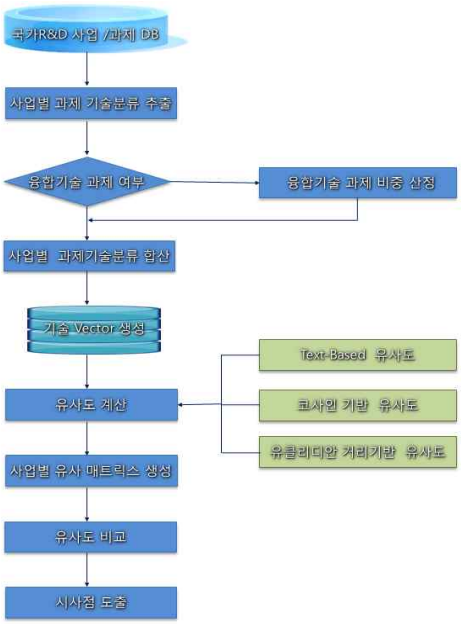
또한, 2009년부터 조사·분석 시 융합기술 과제에 대해서는 연구책임자로부터 각 과제에 해당되는 기술분야와 각 기술분야에 대한 비중을 최대 3개까지 입력받고 있다. <표 2>는 수학, 물리학, 재료에 대한 연구분야에 해당되고, 각 기술분야 비중이 5:3:2인 융합기술 과제의 과학기술표준분류 입력 예시이다.

이렇게 조사된 과학기술표준분류에 대한 적정성을 검증하기 위하여 해당 기술분야의 전문가 집단이 델파이 기법을 통해 2차적으로 검증하고 있으나, 융합기술 비중에 대한 유효성과 객관성을 검증하기는 다소 어려운 설정이다.

# 3. 기술분류를 활용한 사업 시그니처(Signature) 모델

## 3.1 사업 시그니처 모델 프레임워크

2장에서는 사업 예산요구서 내의 주요 키워드를 추출한 후 이를 통해 사업 간의 유사도를 측정하는 방식은 사업 담당자의 작성의도, 기술수준, 상이한 강조사항, 표준양식 미보유등의 사유로 인하여 계속 사업임에도 불구하고 유사도가 매우 낮게 측정됨을 확인하였다. 이는 비교 대상 문서들의 품질 차이로 인해 일부 불완전한 문서 내의 키워드를 추출하여 유사도를 측정 시 문서 간 유사도의 정확도와 객관성이 저하됨을 의미한다. 이에 본 논문에서는 국가R&D사업을 구성하고 있는 과제들의 과학기술표준분류항목을 추출하여 각 사업별 기술분류 벡터를 생성한다. 이는 각 사업을 구성하고 있는 과제들의 고유한 기술 패턴이자 각 사업별 고유 시그니처 모델(Signature Model)이며 사업간 유사도 측정 알고리즘의 적용 기반이 된다.




(그림 1) 사업 시그니처 모델 프레임워크

(그림 1)은 위에서 기술한 사업 시그니처 모델의 프레임워크이다. 국가R&D 사업·과제 DB에서 각 과제의 과학기술표준분류를 추출 후 단일 기술분야 과제는 기술분야의 출현빈도를, 융합기술 과제인 경우는 각 기술분야별 비중을 산정 후 이를 고려하여 출현빈도를 산정한다. 이렇게 산정된 각 과제들의 과학기술표준분류 출현빈도를 상위 사업별로 합산하여 각 사업의 고유 기술벡터를 생성, 이를 이용하여 각각의 알고리즘 적용을 통해 사업간 유사도 행렬을 작성한다.

## 3.2 과학기술표준분류에 대한 기술벡터 생성방안

사업을 구성하고 있는 과제들의 과학기술표준분류를 활용하여 각 사업의 고유한 기술벡터를 생성하기 위해서는 융합기술과제 여부에 따라 각 기술분류의 출현빈도 산정방식이 달라진다. 단일기술과제의 경우 각 기술분야의 출현여부에 따라 1 또는 0으로 산정하면 되지만, 하나 이상의 과학기술표준분류로 구성된 융합기술과제의 경우 각 기술분류가 0 이상 1미만의 값을 갖기 때문에 연구자가 입력한 기술분류에 대한 비중을 재 산정하여 정확도를 제고한다.

<표 3>은 융합기술과제에 대한 기술분야비중 산정 알고리즘이다. Step1은 분석 과제 중 단일 기술분야인 과제들만 과학기술표준분류별로 재 분류하는 과정이다. Step2에서는 각 기술분류별로 재 분류된 과제들의 연구내용에서 키워드를 추출하여 이를 각 과학기술표준분야별 키워드 집합으로 병합하는 과정이다. 마지막으로 Step3에서는 융합기술과제에 대하여 동일한 방법으로 키워드를 추출 후 Step2에서 생성된 과학기술표준분류별 키워드 집합에 포함되는지 비교하여 각 기술분야에 대한 출현빈도를 산정하는 과정이다.

<표 3> 기술분야 비중산정 pseudo code

***Step 1: etxtact technique keyword sets from non-fusion category project***

```
S as stack of whole Project Sets
1. procedure1(ProjectStack S)
2. while S is empty
3.    target ← S.pop
4.    if target.technique.count is 1
5.          Procedure2(target, target.technique)
6.    endif
7. end
```

***Step 2 : Make keyword sets per Technique***

```
Kw(T) as Queue of keywords of technique T
(each Kw(T) has already created before execution)
1. procedure2(Project S , Technique T)
2.    Create a Stack KwTemp
3.    KwTemp ← ExtractKeywordStackFrom(S.context)
4.    Kw(T).inserts(KwTemp)
5. end
```

***Step 3 : Calculate each technique weight ratio of fusion category project Weights(Tn) as array of each nth technique of this project***

```
1. procedure3(S as stack of whole Project Sets)
2. while S is not empty
3.    target ← S.pop
4.    if target.tehique.count is not 1
5.       Create a Stack Kwtemp
6.       KwTemp ← ExtractKeywordStackFrom(S.context)
7.       Create a Array Weights
8.        while KwTemp is not empty
9.              Temp ← KwTemp.Pop
10.             create a number i
11.             i ← 1
12.             while i < S.technique.counts+1
13.                   if Kw(Ti) is contain Temp
14.                         Weights(Ti) ← Weights(Ti)+1
15.                         i ← i+1
16                    endif
17.             end do
18.       end do
19.    end if
20. end do
21. end
```

A부터 P까지의 고유 식별 코드를 갖는 과학기술표준분류를 T, 단일 기술과제의 키워드 추출을 통해 생성된 과학기술표준분류별 키워드 집합을 $Kw_{(T)}$라 할 때, 과제별로 추출된 키워드가 $Kw_1$, $Kw_2$, ....$Kw_n$ 일 때 키워드의 집합을 S라 하면 각각의 집합은 아래와 같이 표현된다.

T = { A, B, C, D ~ P }

$Kw_{(T)}$ = { $Kw_{T1}$, $Kw_{T2}$, ,... $Kw_{Tn}$ }

S = { $Kw_1$, $Kw_2$, ... $Kw_n$ } 단, $Kw \in Kw_{(T)}$

융합기술 과제의 기술분야별 비중은 과제에서 추출된 키워드를 A부터 P까지 모든 과학기술표준분류별 키워드 집합과 비교 후 전체 일치건수 중 각 기술분류별로 일치되는 건수를 이용하여 비중을 산정하며 이는 아래 (수식 4)로 표현될 수 있다. 위 수식을 통해 산정된 융합기술 과제의 기술분야별 비중과 단일기술 과제의 비중을 사업단위로 합산 시 각 사업의 고유한 기술벡터가 생성된다.

$$Weight(S,T) = \frac{n(S \bigcap Kw_T)}{\sum_{T=A}^{P} n(S \bigcap Kw_T)}$$

(수식 4) 융합기술과제의 기술비중 산정 알고리즘

<표 4> 코사인 분석 시 기술분류 벡터 구조

| 구분 | 환경 | 건설/교통 | …… | 에너지/자원 | 기계 |
|---|---|---|---|---|---|
| 사업1 | 1 | 1 | …… | 1 | 0 |
| 사업2 | 0 | 1 | …… | 1 | 1 |
| 사업3 | 1 | 0 | …… | 0 | 1 |
| 사업4 | 0 | 0 | …… | 1 | 1 |

<표 4>는 코사인 유사도 알고리즘으로 유사도 산정을 위한 사업별 기술벡터 구조에 대한 예시로서 각 사업의 고유한 기술분류 벡터는 사업 내 과제들의 과학기술표준분류 존재여부만을 식별하여 해당 표준분류가 존재하면 1, 존재하지 않으면 0으로 표시하므로 각 사업별 기술벡터가 0 또는 1의 정수로 구성된다.

<표 5> 유클리디안 거리 분석 시 기술분류 벡터 구조

| 구분 | 환경 | 건설/교통 | …… | 에너지/자원 | 기계 |
|---|---|---|---|---|---|
| 사업1 | 2.0 | 3.5 | …… | 2.2 | 0.0 |
| 사업2 | 8.6 | 3.0 | …… | 2.4 | 0.8 |
| 사업3 | 3.2 | 6.4 | …… | 5.0 | 6.6 |
| 사업4 | 12.8 | 5.8 | …… | 4.4 | 7.2 |

<표 5>는 유클리디안 유사도 알고리즘으로 유사도 산정을 위한 사업별 기술벡터 구조에 대한 예시로서 유클리디안 거리로 분석 시에는 출현빈도와 각 기술분야의 비중을 고려하기 위하여 해당 표준분류가 존재 시 1에 비중을 곱하여 기재 후 합산하므로 기술벡터가 소수점 형태로 구성된다.

## 4. 실험 및 결과분석

### 4.1 유사검색 모델별 사업간 유사도 측정 결과

본 연구의 분석자료는 국가과학기술위원회에서 2002년부터 매년 국가R&D조사·분석을 통해 수집된 3,421개 사업 및 287,902개의 과제정보 DB이다. 본 논문에서는 제시된 모델의 타당성을 입증하기 위하여 과년도 해양관련 8개 사업의 108개 과제와 바이오관련 9개 사업의 488개 과제의 과학기술표준분류로 한정하였다.

키워드기반의 유사도 측정방식에서는 해양 및 바이오관련 주제의 유사사업을 선정 후 각 사업 목적, 내용 등이 포함된 각 사업의 예산요구서를 DB화 후 주요 키워드를 추출하여 유사도를 측정하였다. 또한, 동일한 사업을 대상으로 제시된 사업 시그니처 모델을 기반으로 코사인, 유클리디안 방식으로 유사도를 측정하였으며, 융합기술 과제의 기술분야 비중은 제시된 개선방안을 적용하여 재 산정 후 유클리디안 방식을 통해 유사도를 재 측정하였다. 각 모델 별 사업간 유사도 행렬결과는 <표 6><표 7>과 같다.

### 4.2 유사검색 모델별 실험결과 비교 분석

위 실험에서는 동일한 사업 및 과제를 모수로 하여 키워드, 코사인, 유클리디안 거리기반으로 사업간 유사도 행렬을 측정하였다.

키워드 비교방식으로는 동일한 주제의 사업임에도 불구하고 사업간 유사성을 거의 유추할 수 없을 만큼 미비한 유사도가 측정된 반면, 본 논문에서 제시된 기술벡터를 활용한 사업 시그니처 모델을 통해서는 기존의 텍스트 기반의 키워드 추출방식에 비해 각 사업간 유사도가 높게 측정됨을 확인하였다. 코사인 방식에서는 기술코드 출현여부만을 식별하여 기술벡터를 생성했기 때문에 사업 내 과제들의 기술분류에 대한 분포나 비중들을 고려하지 않았고, 유클리디안 거리 방식으로 사업간 유사도 측정 시 각 과제의 기술분류 출현빈도와 비중을 기술벡터에 고려하였다. 또한, 유클리디안 거리 기반에 활용될 기술벡터 생성 시 본 논문에서 제시된 키워드 출현빈도를 통한 기술분류 비중 산정 알고리즘을 활용하여 사용자가 입력한 기술분류 비중 활용 대비 더 높은 사업간 유사도를 측정할 수 있었다.

본 실험을 통해 사업간 유사도 측정 시 과제의 기술분류를 활용하는 것이 기존의 텍스트 방식의 유사도 측정에 대한 한계를 극복하기 위한 대안이 될 수 있음은 확인하였으나, 사업의 유사성 판단 기준을 기술분류 구성으로 국한할 것인지 혹은 사업 내 기술분류 비중까지 고려할 것인지는 예산심의 과정에서 각 사업별 특성을 고려하여 판단하여야 한다.

<표 6> 해양관련 사업 간 각 유사도 측정 모델별 유사도 행렬

| 구분 | 사업 1 | | | 사업 2 | | | 사업 3 | | | 사업 4 | | | 사업 5 | | | 사업 6 | | | 사업 7 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T | C | E | T | C | E | T | C | E | T | C | E | T | C | E | T | C | E | T | C | E |
| 사업1 | N/A | | | N/A | | | N/A | | | N/A | | | N/A | | | N/A | | | N/A | | |
| 사업2 | 0.00 | 0.04 | 7.71 | | | | | | | | | | | | | | | | | | |
| 사업3 | 0.20 | 0.00 | 4.49 | 0.00 | 0.17 | 8.17 | | | | | | | | | | | | | | | |
| 사업4 | 0.00 | 0.00 | 5.83 | 0.00 | 0.04 | 8.63 | 0.00 | 0.08 | 4.93 | | | | | | | | | | | | |
| 사업5 | 0.00 | 0.00 | 10.15 | 0.00 | 0.17 | 11.79 | 0.00 | 0.00 | 10.73 | 0.00 | 0.00 | 11.36 | | | | | | | | | |
| 사업6 | 0.40 | 0.00 | 8.75 | 0.10 | 0.14 | 9.54 | 0.30 | 0.31 | 8.61 | 0.10 | 0.11 | 9.08 | 0.00 | 0.00 | 13.10 | | | | | | |
| 사업7 | 0.10 | 0.00 | 8.42 | 0.00 | 0.04 | 10.70 | 0.10 | 0.00 | 9.11 | 0.00 | 0.00 | 9.84 | 0.60 | 0.20 | 12.42 | 0.10 | 0.00 | 11.80 | | | |
| 사업8 | 0.50 | 0.09 | 5.26 | 0.10 | 0.09 | 8.59 | 0.40 | 0.11 | 3.47 | 0.00 | 0.08 | 6.31 | 0.10 | 0.11 | 11.14 | 1.10 | 0.06 | 9.65 | 0.20 | 0.27 | 9.02 |

* T : Text based (using lucene score)  C : cosine / E : Euclidean distance

<표 7> 바이오 관련 사업 간 각 유사도 측정 모델별 유사도 행렬

| 구분 | 사업 1 | | | 사업 2 | | | 사업 3 | | | 사업 4 | | | 사업 5 | | | 사업 6 | | | 사업 7 | | | 사업 8 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T | C | E | T | C | E | T | C | E | T | C | E | T | C | E | T | C | E | T | C | E | T | C | E |
| 사업 1 | N/A | | | N/A | | | N/A | | | N/A | | | N/A | | | N/A | | | N/A | | | N/A | | |
| 사업 2 | 0.00 | 0.00 | 8.40 | | | | | | | | | | | | | | | | | | | | | |
| 사업 3 | 0.00 | 0.05 | 10.76 | 0.00 | 0.00 | 10.49 | | | | | | | | | | | | | | | | | | |
| 사업 4 | 0.00 | 0.26 | 64.10 | 0.10 | 0.17 | 63.49 | 0.10 | 0.17 | 61.02 | | | | | | | | | | | | | | | |
| 사업 5 | 0.00 | 0.00 | 6.29 | 0.00 | 0.00 | 5.74 | 0.00 | 0.00 | 8.89 | 0.00 | 0.03 | 64.85 | | | | | | | | | | | | |
| 사업 6 | 0.00 | 0.07 | 7.18 | 0.10 | 0.06 | 6.61 | 0.00 | 0.18 | 9.10 | 0.00 | 0.30 | 63.87 | 0.00 | 0.07 | 3.79 | | | | | | | | | |
| 사업 7 | 0.00 | 0.15 | 58.31 | 0.00 | 0.02 | 60.65 | 0.00 | 0.08 | 60.77 | 0.00 | 0.29 | 84.86 | 0.00 | 0.02 | 60.36 | 0.00 | 0.15 | 59.92 | | | | | | |
| 사업 8 | 0.00 | 0.06 | 6.23 | 0.00 | 0.00 | 5.69 | 0.00 | 0.13 | 8.78 | 0.00 | 0.07 | 64.86 | 0.00 | 0.00 | 1.16 | 0.00 | 0.00 | 3.93 | 0.00 | 0.04 | 60.42 | | | |
| 사업 9 | 0.00 | 0.33 | 15.00 | 0.10 | 0.08 | 18.52 | 0.00 | 0.00 | 19.76 | 0.00 | 0.27 | 64.67 | 0.00 | 0.00 | 17.70 | 0.00 | 0.10 | 17.80 | 0.00 | 0.20 | 50.11 | 0.00 | 0.00 | 17.68 |

* T : Text based (using lucene score)  C : cosine / E : Euclidean distance

## 5. 결론

국가R&D사업에 대한 투자 효율성 제고가 요구가 급증하고 있는 현 시점에서 사업간 유사성 분석은 소수 일부 전문가집단의 판단에 의존해 왔으며, 과제단위의 유사성 검증은 텍스트 기반의 키워드 매칭에 국한되어 시행되어 그 실효성 및 객관성에 있어 한계를 노출해 왔다.

이에 본 연구에서는 단순 텍스트 기반의 키워드 매칭방식에서 탈피하여 보다 정형화된 데이터인 과학기술표준분류 항목을 기반으로 사업 고유 특성을 나타낼 수 있는 기술벡터를 생성·활용하여 사업간 유사도를 산정하였으며, 실험결과 이러한 과제의 기술분류를 이용 시 기존 키워드 비교를 통한 유사 검색방식보다 실제 다수의 국가R&D사업에 대한 1차적인 유사검증 시 유용함을 입증하였다. 또한, 최근 융합기술 과제 급증에 따라 이의 검증에 소요되는 시간적·비용적 효율성과 기술비중에 대한 정확성 문제가 대두되고 있어 기존 전문가 검증방식에 대안으로 키워드 추출 및 비교를 통한 융합기술비중 산정 방식을 제시하였다.

## 참 고 문 헌

[1] Kittiphattanabawon, Theeramunkong, Nantajeewarawat, "News Relation Discovery Based on Association Rule Mining with Combining Factors", IEICE Transactions on Information and Systems, Vol.E94D, pp.404-415, MAR, 2011

[2] Wen Zhang, Taketoshi Yoshida, Xijin Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification", Expert Systems with Applications, Vol.38, pp.2758-2765, MAR, 2011

[3] Amine. A, Elberrichi. Z, Simonet. M, "Evaluation of Text Clustering Methods Using WordNet", International ARAB Journal of Information Technology, Vol.7, pp.349-357, OCT, 2010

[4] S. Yoon, S. Kim, and S. Park. "A link-based similarity measure for scientific literature. In Proc. of Int'l. Conf. on World Wide Web, pp.1213-1214, April, 2010.

[5] Egghe.L,"New Relations Between Similarity Measures for Vectors Based on Vector Norms", Journal of the American society for Information science and technology, Vol.60, pp.232-239, FEB, 2009

[6] V.Baladi and B. Vallee, "A note on "Euclidean algorithms are Gaussian", Journal of Number Theory, Vol.

129, No.10, OCT, 2009
[7] Tata. S, Patel, JM, "Estimating the selectivity of tf-idf based cosine similarity predicates", Sigmod Record, Vol.36, pp.75-80, DEC, 2007
[8] Deshpande, M. and Karypis,, G."Item-based top-N recommendation Algorithms", ACM Transactions on Information Systems, Vol.22, No.1, pp.143-177. Jan, 2004
[9] Jain, A.K., Murty, M. N. and Flynn, P. J., "Data clustering: A review", ACM Computing Surveys, Vol.31, No.3, pp.264-321, SEP, 1999
[10] Mao,J. and Jain, A. K., "A self-organizing network for hyperellipsoidal clustering(HEC)", IEEE Transactions on Neural Network, Vol.7, No.1, pp.16-29, MAR, 1996
[11] 고용수, 김치용, 김성수, "정부의 연구개발(R&D) 투자 규모 및 향후 투자방향에 대한 제언", KISTEP, 2012
[12] 윤석호, 김상욱, "논문 데이터베이스를 위한 텍스트 기반 유사도 계산 방안", 정보처리학회논문지, Vol 18 D, No.5, pp.317-322, 2011
[13] 이광희, "지식지도 작성을 위한 연구", 한국학술진흥재단, Dec, 2007
[14] 이경일, 서형국, 안태성, "텍스트마이닝 기반 고정밀 검색 시스템", 한국정보처리학회, Vol 11,No.2, pp.88-97, 2004
[15] 송미란, 김교정, "사용자 그룹을 이용한 효과적인 정보 여과 및 학습방법에 관한 연구", 숙명여자대학교, 2000

**김 주 호**

2007년 :고려대학교 정보공학석사
2012년~현재 :숭실대학교 IT정책경영학 박사과정
1997년~2005년 : 데이콤/윈즈 시스템 엔지니어
2006년~현 재 :한국과학기술기획평가원 부연구위원
관심분야 : 정보화 평가, 기술수용모델, 지식지도, AHP 등

**김 영 자**

2011년 :동국대학교 컴퓨터공학석사
2011년~현 재 :한국과학기술기획평가원 연구원
관심분야 : 컴파일러, 객체지향 프로그래밍, 유사측정 알고리즘 등

**김 종 배**

2002년 :숭실대학교 대학원 석사
2006년 :숭실대학교 대학원 박사
2004년~2006년 :남서울대학교 컴퓨터학과 겸임교수
2006년~현 재 :서울여자대학교 컴퓨터학부 겸임교수
2011년~현 재 :숭실대학교 대학원 겸임교수
관심분야 : 소프트웨어 개발 방법론, 에이전트 시스템, 오픈소스 SW