JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 18, No. 5 : 1013~1021 May. 2014

특허 정보를 활용한 R&D 과제 유사도 측정 모델

김종배[1] · 변정원[2*] · 선동주[3] · 김태균[4] · 김융[5]

A Model for Measuring the R&D Project Similarity using Patent Information

Jong-bae Kim[1] · Jung-Won Byun[2*] · Dong-Ju Sun[3] · Tae-Gyun Kim[4] · Yung Kim[5]

[1]Graduate School of Software, Soongsil University, Seoul 156-743, Korea
[2*]Department of Computer Science and Engineering, Soongsil University, Seoul 156-743, Korea
[3]Patent Analysis Team, Korea Intellectual Property Strategy Institute, Seoul 135-980, Korea
[4]Patent Analysis Team, Korea Intellectual Property Strategy Institute, Seoul 135-980, Korea
[5]Patent Analysis Team, Korea Intellectual Property Strategy Institute, Seoul 135-980, Korea

요 약

정부의 입장에서 R&D 과제간의 유사도를 분석하는 것은 불필요한 예산의 낭비를 없애고, R&D 투자의 효과를 높이는데 있어서 매우 중요한 문제이다. 그 동안, 문서의 내용을 대표하는 키워드를 중심으로 두 문서간의 유사도를 분석하거나, 문장 단위로 유사도를 분석함으로써, R&D 과제의 중복 여부를 판단하기 위한 연구들이 시도되어 왔으나, 여러 가지 이유로 아직까지 그 정확도는 매우 낮은 실정이다. 이에, 본 연구는 기 수행된 R&D 관련 특허를 조사, 수집하는 정부 R&D 특허기술동향조사사업의 특허분석 DB를 활용하여 R&D 과제간의 유사도를 분석할 수 있는 방안을 제시하고자 한다. 이를 위해, 집합 이론 및 확률 이론을 기반으로 한 유사도 측정 모델을 제시하였다. 또한, 제시한 모델의 검증을 위해 156개 과제, 160,218개의 유효특허를 기반으로 유효특허기반 과제 유사도 측정 실험을 수행하고, 그 사례를 제시하였다.

ABSTRACT

For efficient investments of government budgets, It is important to analyze the similarities of R&D projects. So, existing studies have proposed a techniques for analyzing similarities using keywords or segments. However, the techniques have low accuracy. We propose a technique for similarities of projects using patent information. To achieve our goal, we suggest three metrics that are based some mathematic theories; set theory and probability theory. In order to validate our technique, we perform case studies that have 156 R&D projects and 160,218 patent informations.

키워드 : 연구개발과제, 과제유사도, 특허정보, 집합이론, 확률이론

Key word : R&D, Project Similarity, Patent information, Set theory, Probability theory

접수일자 : 2014. 03. 03 심사완료일자 : 2014. 03. 28 게재확정일자 : 2014. 04. 11

*** Corresponding Author** Jung-Won Byun(E-mail:jimi010327@gmail.com, Tel:+82-10-4540-0360)
Department of Computer Science and Engineering, Soongsil University, Seoul 156-743, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2014.18.5.1013 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

정부 R&D 과제에 대한 투자는 정부의 적극적인 과학기술 정책에 의해 매년 약 10%정도씩 증가하고 있는 추세이다[1]. 그러나, 각 부처의 경쟁적인 사업 추진으로 인한 예산의 낭비가 여전히 문제로 지적되고 있는 실정이다. 정부 R&D 예산의 중복 투자를 방지하고, 투자 효율을 제고하기 위해서는, 유사 과제를 제안 단계에서부터 식별해내는 것이 매우 중요하다. 이에 따라, 정부에서는 R&D 과제 기획 시, 국가연구개발 사업관리 등에 의한 규정에 따라 국가과학기술지식정보서비스를 통한 유사성 검토를 의무화하고 있다. 그러나, 이 서비스는 단순 키워드 비교에 의한 유사도 평가 방식을 사용하고 있어서, 과제명의 일부 수정, 기술상의 단순 대치 등의 경우, 그 유사도를 정확히 측정하지 못하는 한계가 있다[2-4].

본 연구는 이러한 문제를 개선하고자, 기 수행된 R&D 관련 특허를 조사, 수집하는 정부 R&D 특허기술동향조사사업의 특허분석 DB를 활용한 유사도 분석 모델을 개발한다. 실제, 이 특허기술동향조사사업(http://ipas.rndip.re.kr)은 정부가 지원하는 R&D 과제의 기획 시, 해당 기술 분야의 특허 동향을 분석함으로써, 새로운 과제의 연구 방향을 제시하고, 이미 특허가 출원된 기술에 대한 R&D 중복 지원을 방지하기 위한 목적으로 시행되고 있다. 뿐만 아니라, 기존에 기획되고 수행된 R&D 사업의 기획 정보와 성과 정보를 관리하고 있기 때문에, 이를 새로운 R&D 기획 정보와 비교, 분석함으로써 과제의 유사성을 판단할 수 있다.

이에 본 연구에서는, 기존의 문서 간 유사도 분석 기법들을 고찰하고, 이의 활용 및 개선 방법을 도출한다. 이를 통해, 특허분석 DB를 활용한 유사도 분석 모델을 개발한다. 또한, 유사도 분석을 위한 입출력 정보와, 분석 알고리즘을 제시하고, 이를 검증한다.

Ⅱ. 관련연구

R&D 과제의 유사도를 정량적으로 분석하기 위해서는 과제의 산출물, 특히 문서를 기반으로 비교하는 것이 가장 현실적인 방법이다. 그런데, 이처럼 문서를 중심으로 한 유사성의 식별은 R&D 영역에서 뿐만 아니라, 이미 다양한 영역에서 중요한 문제로 인식되어 왔다. 유사도란 “두 개체가 공통으로 지닌 정보와, 서로 다르게 지닌 정보의 양에 대한 정량적 측정”으로 정의할 수 있다. 또, 유사도는 정도에 따라, 완전중복, 부분중복, 유사중복 등으로 구분할 수 있다[5].

문서간의 유사도를 분석하기 위해서는, 하나의 문서가 다른 문서와 구별될 수 있는 특징이 있어야 하며, 이렇게 추출된 특징들이 높은 유사도를 나타낼 수 있어야만 문서간의 유사성 정도를 판단할 수 있다. 참고[6]은 높은 차원의 문서를 낮은 차원의 핑거프린트(Fingerprint)로 차원 감소(Dimensionality Reduction)하여 이를 문서의 특징으로 사용하는 대표적인 유사도 분석 모델이다. 하지만, 이는 자연어의 중의성을 반영하지 못하며, 표준 문법을 준수하지 않은 문장의 비교가 어렵다는 단점이 있다. 또한, 유사 문서의 분석은 품질뿐만 아니라 성능도 매우 중요한데, 문서 n개를 핑거프린트로 비교할 경우, 전수 검사가 필요하며, 이들의 유사도를 구하는 일은 O(n2) 정도의 복잡도가 요구되는 문제가 있다.

또 다른 방법은 다중 레벨 인덱싱 구조인데[7], 이는 문서의 라인수와 k-bit 수를 조정하여 비교 대상을 제안하는 방법이다. 이 경우 동일 핑거프린트 군집의 수가 줄어들기 때문에 빠른 비교가 가능해진다. 이와 비슷한 방법으로, 참고[8]은 문서에 등장하는 단어를 나열하고, 이 중 일부를 문서의 특징으로 추출하는 기법을 제안하였다. 현실적으로 가장 널리 쓰이는 유사도 분석 모델이지만, 키워드를 식별하기 위해 전문가의 휴리스틱한 판단이 필요하다는 단점이 있다[9].

한편, 특정 문서와 유사한 문서를 찾기 위해 모든 문서를 분석하는 것이 아니라, 초기분석에 의해 이웃(Nearest Neighbors)을 찾고, 이 문서에 대하여 더욱 상세한 분석을 수행하는 방법이 있다. 이는 분석될 문서의 범위를 줄여줌으로써 유사도 분석의 효율을 높여준다[10]. R&D 과제 측면에서 유사도를 측정하기 위한 알고리즘의 예도 있는데, 포괄성형망 모델[11]이 그것이다.

이는 키워드 기반 유사도 분석 모델의 한계점을 개선하기 위해 고안된 모델로, R&D 사업을 구성하고 있는 과제들의 과학기술표준분류항목을 추출하여, 각 사업별 기술 분류에 대한 고유 벡터를 생성하고, 각 사업을 구성하는 과제들의 고유한 기술 패턴을 지정함으로

써 이들의 유사성을 포괄성형망으로 표현한다. 그러나, 이 역시 의미를 고려하지 못한다는 한계가 있다.

Ⅲ. 특허 정보를 이용한 과제 유사도 측정 모델

3.1. 과제 유사도 측정을 위한 특허 정보의 도출

본 연구에서는 문서간의 구별을 위한 특징으로써 특허 정보를 활용한다. 특허 정보는 정부 R&D 특허기술동향조사사업을 통해 획득한 자료를 기반으로 하였는데, 그 데이터베이스에는 그림 1과 같은 정보들이 포함되어 있다.

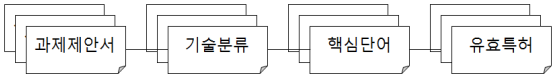


그림 1. 특허기술동향조사 정보의 구성
Fig. 1 Information of the patent technology trend survey

본 연구에서는 이러한 자료들을 중심으로 신규 과제가 입력되었을 때, 과제간의 유사도를 계산한다. 이 개념은 그림 2와 같이 표현할 수 있다. 이를 위한 가정으로서, 기존 과제 및 신규 과제는 모두 저마다의 유효특허를 가져야 한다는 것인데, 앞서 설명한 바와 같이, 이미 정부 R&D 특허기술동향조사사업을 통해 유효특허가 조사되었다는 것을 전제로 한다. 유효특허 정보 내의 "국가공보", "출원번호" 가 일치하는 경우를 동일한 특허로 인정하며, 유효특허간의 일치 정도에 따라 유사도를 측정한다.

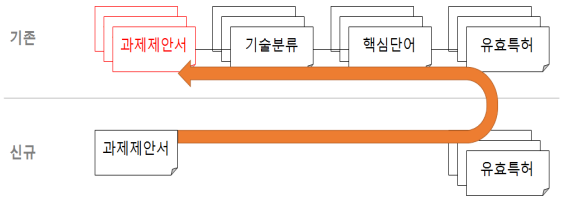


그림 2. 특허정보 기반 과제 유사도 분석 개요
Fig. 2 Patent Information Based Project Similarity Analysis Overview

3.2. 유사도 측정을 위한 척도

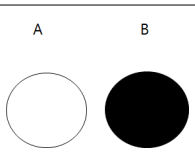
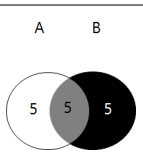
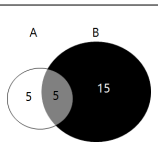
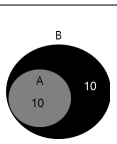
각 과제는 유효특허의 집합을 가지는데, 이러한 집합간의 일치 정도를 분석하기 위한 방법으로 집합 이론(Set Theory)[12]을 적용한다. 집합기반 유사도는 2개 집합의 합집합 특허 중, 2개 집합의 교집합 특허의 비율로 나타낸다. 즉, 집합 이론에 따른 유사도는 2개의 과제 간 유효특허 중복 비율이 높을수록 유사도가 높다는 측면을 반영하는 척도이다. 집합기반 유사도의 측정을 위한 공식은 수식(1)과 같다.

$$\textit{집합기반 유사도} = \textit{2개 유효특허의 교집합} / \textit{2개 유효특허의 합집합} \quad (1)$$

이 척도의 장점은 누구나 쉽게 이해할 수 있다는 것이며, 이를 기반으로 결과 해석의 합리성을 확보할 수 있다. 반면, 2개 집합의 합집합을 모수로 사용하기 때문에, "신규 과제가 기존 과제와 유사한 정도"와 "기존 과제가 신규 과제와 유사한 정도"가 동일한 값을 가진다. 이 때문에, "특정 과제가 다른 과제에 포함된 정도"를 표현하지 못한다는 단점이 있다. 또한, 집합의 개수가 유사도에 영향을 미치기 때문에, 유효특허의 개수에 따라 측정된 유사도 값이 달라질 수 있다는 점도 단점이다. 집합기반 유사도의 단점을 개선하고자, 두 번째 측정척도로 확률기반 유사도를 제안한다. 확률 이론(Probability Theory)[13]은 확인되지 않은 결과에 대한 가능성을 표현할 수 있다. 과제간의 유사도가 확인된 값으로서의 정답을 가질 수 없다는 측면에서, 확률 이론의 적용은 유사도의 측정과 표현에 적합할 가능성이 크다. 확률기반 유사도의 측정을 위한 공식은 수식(2)와 같다.

- *A의 조건에서 A와 B가 상이할 확률 =>* $p(A \neq B \mid A) = P(A - B) / P(A)$ *=>* ***"A가 B와 상이할 확률"***
- *B의 조건에서 A와 B가 상이할 확률 =>* $p(A \neq B \mid A) = P(B-A) / P(B)$ *=>* ***"B가 A와 상이할 확률"***
- *1.0 - "A가 B와 상이할 확률" =>* ***"A가 B와 유사할 확률"*** (2)
- *1.0 - "B가 A와 상이할 확률" =>* ***"B가 A와 유사할 확률"***
- *"A가 B와 유사할 확률" x "B가 A와 유사할 확률" => "A와 B가 유사할 확률" =>* ***"확률이론기반 유사도"***

표 1. 과제 유사도 척도의 비교
Table. 1 Comparison of the metric for similarity

구분		완전 불일치	완전 일치	부분 일치		
				동일 개수	비동일 개수	포함
						
집합 (전체)	A -> B	0	1	0.333	0.200	0.500
	B -> A	0	1	0.333	0.200	0.500
확률 (전체)	A -> B	0	1	0.250	0.125	0.500
	B -> A	0	1	0.250	0.125	0.500
확률 (부분)	A -> B	0	1	0.500	0.500	1.000
	B -> A	0	1	0.500	0.250	0.500

- ㅇ *A의 조건에서 A와 B가 상이할 확률 => $p(A \neq B \mid A) = P(A - B) / P(A)$ =>* ***"A가 B와 상이할 확률"***
- ㅇ *1.0 - "A가 B와 상이할 확률" =>* ***"A가 B와 유사할 확률"***
- ㅇ *"A가 B와 유사할 확률" x "B가 A와 유사할 확률" => "A와 B가 유사할 확률" =>* ***"확률이론기반 유사도 (전체)"*** (3)
- ㅇ *Max("A가 B와 유사할 확률", "A와 B가 유사할 확률") =>* ***"확률이론기반 유사도 (부분)"***

확률 이론 기반의 유사도 분석은, 각각의 과정이 정량적인 확률로써 계산될 수 있으며, 유효특허의 개수에 영향을 받지 않는다는 것이 장점이다. 반면, 조건부 확률 이론은 결과 해석의 합리성을 확보하기에 비교적 어렵고, 신규 과제와 기존 과제의 유사 확률이 결합되기 때문에 포함 관계를 표현하지 못한다는 단점이 있다.

한편, 확률 이론 기반 유사도 분석 및 집합 이론 기반 유사도 분석의 결과는 모두 신규 과제와 기존 과제의 포함 관계에 대해 설명하고 있지 못하다. 이 문제를 해결하기 위해서는, "신규 과제가 기존 과제와 유사할 확률" 과 "신규 과제와 기존 과제가 유사할 확률"을 종합할 수 있는 방안이 요구된다. 본 연구에서는 포함 관계의 명확한 표현을 위해, 큰 값을 채택하는 방식을 도입하였다. 단, 이때 반드시 수식(2)의 확률 이론 기반 유사도를 함께 고려하여 해석해야 한다. 이를 정리하면 수식 (3)과 같다.

3.3. 유효특허 기반의 과제 중복/유사도 척도간의 비교

앞서 제시한 유효특허 기반의 유사도 척도들을 신규 과제와 기존 과제의 유효특허 간 관계에 따라 비교하면 표 1 과 같다. 유효특허 간의 관계는 유효특허가 완전히 불일치하는 경우, 유효특허가 완전히 일치하는 경우, 유효특허가 부분적으로 일치하는 경우 등과 같이 크게 3가지의 경우로 구분된다. 이 중 부분적으로 일치하는 경우는 다시 3가지 과제간의 세부 관계로 구분할 수 있는데, 신규 과제와 기존 과제의 유효특허 개수가 동일한 경우, 유효특허 개수가 다른 경우, 포함 관계를 가지는 경우 등이 그것이다.

단, 이 3가지 유사성 척도는 선택적인 척도가 아닌 상호 보완적인 척도임에 유의해야 한다. 일반적으로 신규 과제의 측면에서 기존 과제를 바라보기 때문에(즉, A -> B), 확률을 잘못 해석할 경우 A와 B는 완전한 일치로 해석될 수 있는 소지가 있다. A와 B의 관계가 완전한 일치인지 아니면 포함 관계인지를 판단하기 위해서는 수식(2)의 확률(전체)과 함께 비교되어야 한다. 완전한 일치의 경우, 수식(2)의 확률(전체)과 수식(3)의 확률(부분)이 모두 1의 값을 가지며, 포함 관계일 경우 수식(3)의 확률(부분)만 1의 값을 가진다.

표 2. 특허 정보 기반 유사도 분석 결과
Table. 2 Similarity analysis based on patent information

분석대상	과제명	집합	확률 (전체)	확률 (부분)
작물 종자형질유전자 기능분석 및 산업화 (상위 5개)	벼 변이집단 및 생물정보를 이용한 유용 농업형질 유전자 탐색	0.137	0.427	0.485
	GMO 안전성 평가기술 및 산업화	0.036	0.180	0.186
	토양인산이용성이 증진된 형질전환벼 개발 및 실용화	0.002	0.043	0.043
	고부가 지질 생산 생물소재 개발	0.003	0.014	0.014
	특용작물 유기농 재배기술 산업화	0.003	0.011	0.011

표 3. 키워드 기반 유사도 분석 결과
Table. 3 Similarity analysis based on keyword

분석대상	과제명
작물 종자형질 유전자 기능분석 및 산업화 (상위 5개)	복숭아 신품종 이용촉진 및 재해경감 기술 개발
	국내 농식품 함유 면역증강물질 특성분석 및 생산공정개발
	고구마 생산 일관기계화 기술 개발
	유기농업기술 개발 가치평가 및 신 수요 예측
	왕겨에너지를 이용한 벼 신건조시스템 현장실증연구

Ⅳ. 사례연구

4.1. 제안한 유사도 측정 방법의 적용 가능성 및 유사도 측정 결과의 비교

제안한 유사도 측정방법을 기반으로, 본 연구는 156개 과제, 160,218개의 유효특허를 기반으로 유효특허 기반 과제유사도 측정 하였다. 측정 결과는 표 2와 같으며, 대조군으로써 키워드 중심의 과제유사결과 표 3과 비교를 하였다. 이 예시에서, “작물 종자형질유전자 기능분석 및 산업화” 과제의 경우 집합, 확률 관점에서 “벼 변이집단 및 생물 정보를 이용한 유용 농업형질 유전자 탐색”과 가장 유사한 것으로 나타났다. 집합의 관점에서 유효특허의 합집합 중 약 13%가 일치하는 것으로 해석할 수 있다. “작물 종자형질유전자 기능분석 및 산업화” 의 유효특허 중 약 48%의 유효특허가 “벼 변이집단 및 생물 정보를 이용한 유용 농업형질 유전자 탐색”과 일치하는 것으로 해석할 수 있다. 또한 2개 과제의 유사할 확률은 약 42%로 확률(전체) 값을 통해 해석할 수 있다.

4.2. 통계 분석을 활용한 방안의 제안한 척도의 검증

본 연구에서 제안한 방안은 2회의 분석을 통해 검증한다. 첫 번째 검증은, 제안한 방법의 타당성에 대한 검증이다. 이는 각 방안별 척도를 통해 얻은 결과가 공통의 이해를 가진 사람의 휴리스틱한 분석결과를 반영할 수 있는지를 의미하는 것이다[14]. 두 번째 검증은 제안한 방법의 신뢰성과 효율성에 관한 검증이다. 제안한 방안이 타당하다는 전제조건 하에 도출된 결과가 전문가의 의견과 비교하여 신뢰성을 가질 수 있는지 판단하는 것을 의미한다.

제안한 방법의 타당성에 대한 첫 번째 검증은 일반적인 지식을 가진 일반인을 대상으로 광범위한 설문조사를 수행하였다. 질문 내용은 앞서 도출된 유사성이 있을 것으로 판단된 2개의 과제 중 임의의 1개를 선정하여 이들의 유사도를 묻는 방식을 구성하였다. 과제별로 피설문자에게 주어진 정보는 과제명과 과제 내용이다. 응답은 리커트 7점 척도를 이용하였으며, 질문지 1개에 7개의 질문을 포함하였다. 설문은 웹 사이트를 통해 질의하였으며, 질의 결과는 통계 분석 소프트웨어인 SPSS 18.0 으로 분석하였다. 설문 결과의 신뢰성을 증진하기 위해 Reverse Question, Reverse Answer, Interval Request 기법을 통해 적절하지 않은 설문지를 제거하였다. 피설문자의 개인정보는 전화번호 뒷자리, 생년월일 등 2가지로 구분하였다. 설문기간은 2013년 10월부터 11월까지 30일 동안 진행하였다.

응답자 수 132명 중 총 94명의 Reverse Question, Reverse Answer, Interval Request 검증 결과 유효한 응답자를 식별하였으며, 총 4722 개의 과제간 유사성 응답을 식별하였다. 이러한 통계분석을 위한 가설은 다음과 같다.

H0 (귀무가설) : 휴리스틱한 판단과 제안한 방법 사이의 유사도 검증결과는 상관관계가 없다.

H1 (검증가설) : 휴리스틱한 판단과 제안한 방법 사이의 유사도 검증결과는 상관관계가 있다.

휴리스틱한 판단에 의한 유사성 분석 결과와 본 연구에서 제안한 방법과의 상관관계를 분석한다. 2개의 척도 모두 정량척도이므로 피어슨 상관 분석을 수행하였으며, 그림 3과 같은 분석 결과를 도출하였다.

		설문응답	집합	확률_전체	확률_부분
설문응답	Pearson 상관계수	1	.564**	.507**	.533**
	유의확률 (양쪽)		.000	.000	.000
	N	4722	1734	1734	1734
집합	Pearson 상관계수	.564**	1	.978**	.983**
	유의확률 (양쪽)	.000		.000	.000
	N	1734	1734	1734	1734
확률_전체	Pearson 상관계수	.507**	.978**	1	.999**
	유의확률 (양쪽)	.000	.000		.000
	N	1734	1734	1734	1734
확률_부분	Pearson 상관계수	.533**	.983**	.999**	1
	유의확률 (양쪽)	.000	.000	.000	
	N	1734	1734	1734	1734

그림 3. 제안한 방법의 척도와 설문응답 사이의 상관분석
Fig. 3 Correlation analysis between the proposed metric and response

상관관계 분석결과, 설문응답과 유효특허기반의 분석결과 (집합, 확률_전체, 확률_부분) 사이의 상관관계는 유의미한 것으로 나타났다. 유의확률(양쪽)의 값이 0.01 이하 (99% 신뢰수준)을 가짐을 확인할 수 있다. 이 중 설문응답과 유효특허기반의 분석결과는 상관계수가 0.5 이상을 가지므로 상관관계가 있다고 해석할 수 있다. 또한, 이 값이 양수이기 때문에 Positive Correlation)를 가진다. 이는 설문응답의 값이 높을수록(유사성이 높다고 응답할수록) 제안한 방법으로 계산된 값이 커진다는(척도의 값이 클수록) 의미를 가진다. 상관관계 분석결과 유효특허기반의 분석결과의 신뢰성과 정확성을 분석하고자 회귀분석을 수행한다. 회귀분석의 결과는 그림 4와 같다.

분석 결과, 휴리스틱한 응답결과를 종속변수로, 유효특허기반 분석결과를 독립변수로 하는 회귀모형이 유의미함을 알 수 있다. 유의확률은 0.01보다 작으며 (99% 이상 신뢰도), R 제곱은 0.671(67.1% 해석률, 정확도)을 가진다. 이는 유효특허기반 과제 유사도 분석의 결과가 휴리스틱한 판단을 99%의 신뢰구간에서 67.1%의 정확도로 해석할 수 있음을 의미한다. 32.9% (1 – 정확도) 에 대한 부분은 다른 요인에 의해 해석되어야 함을 의미한다.

모형 요약

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차
1	.819[a]	.671	.670	.024

a. 예측값: (상수), 확률_부분, 집합, 확률_전체

분산분석[b]

모형		제곱합	자유도	평균 제곱	F	유의확률
1	회귀 모형	2.009	3	.670	1174.377	.000[a]
	잔차	.986	1730	.001		
	합계	2.995	1733			

a. 예측값: (상수), 확률_부분, 집합, 확률_전체

b. 종속변수: 설문응답

그림 4. 신뢰성 및 정확성을 위한 회귀분석
Fig. 4 Regression analysis for the reliability and accuracy

분석 결과, 휴리스틱한 응답결과를 종속변수로, 유효특허기반 분석결과를 독립변수로 하는 회귀모형이 유의미함을 알 수 있다. 유의확률은 0.01보다 작으며 (99% 이상 신뢰도), R 제곱은 0.671(67.1% 해석률, 정확도)을 가진다. 이는 유효특허기반 과제 유사도 분석의 결과가 휴리스틱한 판단을 99%의 신뢰구간에서 67.1%의 정확도로 해석할 수 있음을 의미한다. 32.9% (1 – 정확도) 에 대한 부분은 다른 요인에 의해 해석되어야 함을 의미한다.

표 4. 전문가설문 양식
Table. 4 Expert questionnaire form

과제명 1	작물 종자형질유전자 기능분석 및 산업화	
과제명 2	GMO 안전성 평가기술 및 산업화	
분야	기술분야	상이함 1 2 3 4 5 6 7 유사함
	지원분야	상이함 1 2 3 4 5 6 7 유사함
	적용분야	상이함 1 2 3 4 5 6 7 유사함
내용	제목	상이함 1 2 3 4 5 6 7 유사함
	목적	상이함 1 2 3 4 5 6 7 유사함
수행	목표	상이함 1 2 3 4 5 6 7 유사함
	업적	상이함 1 2 3 4 5 6 7 유사함
	추진체계	상이함 1 2 3 4 5 6 7 유사함

주체	추진내용	상이함 1 2 3 4 5 6 7 유사함
	진행자	상이함 1 2 3 4 5 6 7 유사함
	수혜자	상이함 1 2 3 4 5 6 7 유사함
총점 (각 항목의 총점)		

제안한 방안의 두 번째 검증은 전문가에 의한 분석 결과의 검증이다. 3명의 전문가 델파이 기법을 통해 과제 유사성 분석을 수행하였다. 과제의 수는 156개 이며, 주요 영역은 농업, 식품, 원예, 식량, 연구에 해당한다. 이 중 임의의 2개 과제에 대해서 전문가에게 제공하여 이 과제간의 유사정도를 7점 척도로 표현하게 하였다. 분석을 위해 연구과제가 속한 사업제안서, 과제제안서를 제공하였다. 분석 기준은 과제의 분야, 내용, 수행, 주체 등으로 하였다. 분야의 유사성은 기술분야, 지원분야, 적용분야를 판단하였으며, 내용의 유사성은 제목, 목적, 목표, 업적을 판단하였다. 수행의 유사성은 추진체계 및 추진내용으로 판단하였으며, 주체는 진행자와 수혜자의 유사성을 기준으로 평가하였다. 평가 결과는 델파이 기법을 통해 만장일치제로 결과를 도출하였다.

		집합	확률_전체	확률_부분	전문가
집합	Pearson 상관계수	1	.943**	.989**	.994**
	유의확률 (양쪽)		.000	.000	.000
	N	1048	1048	1048	1048
확률_전체	Pearson 상관계수	.943**	1	.981**	.927**
	유의확률 (양쪽)	.000		.000	.000
	N	1048	1048	1048	1048
확률_부분	Pearson 상관계수	.989**	.981**	1	.980**
	유의확률 (양쪽)	.000	.000		.000
	N	1048	1048	1048	1048
전문가	Pearson 상관계수	.994**	.927**	.980**	1
	유의확률 (양쪽)	.000	.000	.000	
	N	1048	1048	1048	6527

그림 5. 제안한 방법의 척도와 설문응답 사이의 상관분석
Fig. 5 Correlation analysis between the proposed metric and expert survey result

전문가의 회의 결과의 신뢰도에 대해 크롬바치 알파기법을 이용하여 분석하였다. 이는 전문가가 일관된 기준으로 유사도에 대해 평가하였는지에 대해 분석할 수 있다. 분석 결과 알파값이 0.771 로써, 신뢰도 분석 결과 알파 값이 0.7 이상이며, 이는 통계적으로 95% 신뢰 수준에서 일관성을 가지고 답하였다고 할 수 있다. 전문가설문을 통해 분석된 과제 유사도 분석 결과와 본 연구에서 제안한 방안을 통해 측정된 결과 사이의 상관분석을 수행하였다. 전문가설문 결과 유효특허 기반 분석결과와 강한 상관관계를 가진다. 상관계수는 0.994, 신뢰수준은 99% 이상으로써 매우 강한 상관관계를 가진다.

그림 5의 분석결과는 다음과 같은 해석을 가질 수 있다. 첫째, 전문가 분석은 앞서 제안한 4개 분류, 11개 항목을 기준으로 유사성을 판단한다. 이는 특허 정보 기반의 분류방법과 매우 유사한 결과를 가진다고 해석할 수 있다. 그러나 키워드 기반 분석 역시 전문가 의견과 상관관계를 가지고는 있으나 매우 낮은 해석률을 가진다고 할 수 있다. 이러한 이유는 동일한 기술분류, 예를 들어 “생산성개선” 이라는 기술분류 항목이 원예에서 사용되는 방법 및 과제 내용과 농업에서 사용되는 방법 및 과제 내용이 달라지기 때문이다. 10% 정도의 해석률은 일부 원예에서 생산성 개선을 위해 사용된 핵심 내용이 농업의 생산성 개선을 위해 사용되었다고 볼 수 있다. 전문가설문 결과, 제안한 2개 방식에 대한 통계분석 결과를 정리하면 다음과 같다.

표 5. 전문가설문 통계분석 결과의 해석
Table. 5 Statistical analysis of the expert survey

○ 전문가설문 결과는 특허정보 기반 유사도 분석 결과를 지지한다.
○ 전문가설문 결과는 특허정보 기반 분석 결과의 수치가 높을수록 유사도가 높다고 판단한다.
○ 전문가설문 결과는 특허정보 기반 분석이 키워드 기반 분석에 비해 더 높은 해석률(정확도)를 가진다고 판단한다.
○ 위의 모든 결과는 통계적으로 99% 이상의 신뢰 수준을 가진다.

Ⅴ. 결 론

본 연구는 특허 정보를 이용한 과제 유사도 분석 방안을 제시하는데 목적을 둔다. 이를 위해 유사도 분석 방안에 대한 문헌들을 분석하였다. 이러한 분석 결과를 기반으로 본 연구는, 첫째, 특허 정보 기반 과제 유사도 분석 모델을 제시하였다. 특허 정보를 기반으로 과제

유사도 분석을 위한 입출력 정보를 제시하였다. 또한 이러한 정보를 활용하여 2가지 측면의 분석 방안을 제시하였다. 그리고 제시한 모델의 타당성(일반인 조사 결과)과 정확성(전문가 조사 결과)을 보였다.

본 연구를 진행하며, 몇몇 한계점을 발견하였다. 사례 검증 시 적용된 도메인은 농촌진흥청관련 2010년 ~ 2013년의 과제를 기반으로 하였다. 본 연구의 검증이 특정 도메인에 한정된 이유는, 검증 시 전문가 협의가 반드시 필요한 사항이었기 때문이다. 하지만, 향후 연구를 통해 다양한 도메인에 적용하여 그 결과를 검증한다면 제안한 과제 유사도 분석 모델의 일반성을 더욱 확고히 할 수 있을 것으로 사료된다.

제안한 모델은 다양한 척도를 제시하고 있다. 이들은 각자 고유의 의미를 가지며, 판단 기준의 다양성을 위해 반드시 필요하다. 제안한 척도들을 통해 유사한 과제를 우선순위화할 수 있지만, 척도의 해석으로 우선순위가 높거나 낮음을 표현할 수는 없다. 이러한 해석을 위해선 과거 과제들과 특허 정보가 충분히 수집되어 통계적인 해석을 도출해야할 필요성이 있다.

REFERENCES

[1] Government Research and Development Budget Analysis in the FY 2013, *Korea Institute of S&T Evaluation and Planning*, 2014-002, 2014.

[2] OkNam Jung, SungYul Rhew, JongBae Kim. "An Empirical Study on Improvement model for Measuring of Project Similarity." *Journal of Digital Contents Society*, Vol.12, No.4, pp.457-465, 2011.

[3] MyungSuk Yang, et al. "Discussion about the National Science & Technology Information Service(NTIS)." *Proceedings of the Korea Technology Innovation Society Conference,* pp.294-304, 2013.

[4] Hyung Deuk Hong. "Comparative Analysis on the Evaluation Systems of the Public R&D Programs in the Developed Countries." *Proceedings of the Korea Technology Innovation Society Conference,* pp.275-290, 2001.

[5] Bendersky, Michael, and W. Bruce Croft. "Finding text reuse on the web." *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM, 2009.

[6] Rabin, Michael O. Fingerprinting by random polynomials. Center for Research in Computing Techn., Aiken Computation Laboratory, Univ., 1981.

[7] Miihleisen, H., Tilman Walther, and Robert Tolksdorf. "Multi-level indexing in a distributed self-organized storage system." *Evolutionary Computation (CEC), 2011 IEEE Congress on. IEEE*, 2011.

[8] Chowdhury, Gobinda, and Sudatta Chowdhury. Introduction to digital libraries. Facet publishing, 2002.

[9] Ju-Ho Kim, Young-Ja Kim, Jong-Bae Kim. "A study on Similarity analysis of National R&D Programs using R&D Project's technical classification." *Journal of Digital Contents Society*. Vol. 13, No. 3, pp. 317-324, Sep. 2012

[10] Domâinguez, Josâe Ferreirâos. Labyrinth of thought: A history of set theory and its role in modern mathematics. Springer, 2007.

[11] Kang Jong Seok, Lee Hyuck Jai, Moon Yeong Ho, “Apparatus and method for configuring a comprehensive intellectual property rights star network by detecting patent similarity.”, *Korea Institute Of Science & Technology Information*, G06F 17/30, 1020070071793, 2006.

[12] Domâinguez, Josâe Ferreirâos. Labyrinth of thought: A history of set theory and its role in modern mathematics. Springer, 2007.

[13] Kolmogorov, Andreĭ Nikolaevich. "Foundations of the Theory of Probability." (1950).

[14] Freedman, David. Statistical models: theory and practice. Cambridge University Press, 2009.

김종배(Jong-Bae Kim)

2002년 8월 숭실대학교 정보과학대학원 석사
2006년 8월 숭실대학교 대학원 컴퓨터학과 박사
2001년~2012년 (주)이엔터프라이즈 대표이사
2012년~현재 숭실대학교 SW특성화대학원 교수
※관심분야 : 소프트웨어공학, 정보보호, 오픈소스소프트웨어

변정원(Jung-Won Byun)

2007년 2월 숭실대학교 미디어학부 공학사
2013년 2월 숭실대학교 컴퓨터학과 공학박사
※관심분야 : 소프트웨어공학, 사용성분석, 데이터분석, 요구공학

선동주(Dong-Ju Sun)

1998년 2월 명지대학교 대학원 기계공학과 석사
1998년~2009년 한국특허정보원 기계조사팀 그룹장
2010년~현재 한국지식재산전략원 특허동향팀 팀장
※관심분야 : 일반기계, 자동차공학, 특허정보 분석 및 활용

김태균(Tae-Gyun Kim)

2001년 2월 인하대학교 자동화공학과 학사
2001년~2004년 티맥스소프트 시스템 엔지니어
2004년~2010년 한국항공우주연구원, 건강보험심사평가원 정보화 담당
2010년~현재 한국지식재산전략원 특허동향팀 선임연구원
※관심분야 : 경영정보/경영전략, 테이터 분석 , 미들웨어

김 융(Yung Kim)

2002년 2월 경북대학교 전자공학과 학사
2013년 2월 서강대학교 기술경영학 석사
2002년~2009년 한국특허정보원 IP전략지원팀 선임연구원
2010년~현재 한국지식재산전략원 특허동향팀 선임연구원
※관심분야 : 디지털 방송, MPEG, 신호처리