R&D 과제의 유사도 및 중복도 측정 시스템에 관한 연구

최국현* · 강용석** · 김종희*** · 신용태**** · 김종배***** ◦
*,**,***,****,***** ◦ 숭실대학교

A System for Measuring the Similarity and Redundancy of R&D Project

Kook-Hyun Choi* · Yong-Suk Kang** · Jong-Hee Kim*** · Yong-Tae Shin**** · Jong-Bae Kim***** ◦
*,**,***,****,***** ◦ Soong-sil University

E-mail : *khchoi@tsline.co.kr, **postwin@gmail.com, ***skyimo@hanmail.net, ****shin@ssu.ac.kr,
***** ◦ kjb123@ssu.ac.kr

요 약

R&D 과제간의 유사성과 중복성을 분석하는 것은 정부 예산의 효율적 투자를 위해 중요하다. 정부 R&D 과제의 기획 시, 예산의 중복 지원을 방지하기 위해 연구 관리 전담기관, 관련 부처 및 정부 차원에서 연구 과제의 중복성을 검토하고 있다. 그러나, 기존의 유사도 분석은 신규 과제 제안서와 기존의 R&D 과제 제안서를 키워드 중심으로 비교, 검색하는 방식에 의존하고 있어, 과제명의 일부 수정, 기술상의 단순 대치 등의 경우, 유사도를 정확히 측정하지 못하는 취약점이 존재한다. 본 연구에서는, R&D 과제 문서의 경우에, 이 문서들을 구별할 수 있는 특징으로써 특허 정보를 활용하고자 한다. 특허 정보는 정부 R&D 특허동향조사사업(http://ipas.rndip.re.kr)을 통해 공표된 자료를 기반으로 한다. 본 연구에서는 신규과제가 입력되었을 때, 특허 정보를 이용하여 R&D 과제간의 유사성 및 중복성을 분석할 수 있는 방안을 제시하고자 한다. 이를 위해, 집합 이론 및 확률 이론을 기반으로 한 유사도 측정 모델을 제시한다. 또한 제시한 측정 모델을 실제 시스템으로 구현하여 중복문서를 식별하고 이들의 유사도를 계산하여 보여준다.

ABSTRACT

The analysis of the similarities and redundancies among R&D projects is important for the efficient investment of government budgets. When government R&D projects are planned, the redundancies of research tasks are examined by institutions specializing in research management, relevant offices and departments, and the government to prevent redundant funding. However, as existing similarity analyses depend on methods wherein new task proposals and existing R&D project proposals are compared and looked up based on keywords. This results in vulnerability wherein similarity cannot be accurately measured in the event of partial modifications of the task name or technical substitutions. This study aims to use patent information as characteristics by which R&D project documents can be identified. The patent data used is based on materials officially published by the government's R&D patent trend survey project (http://ipas.rndip.re.kr). The study aims to propose a method by which patent information can be used to analyze the similarity and redundancy among R&D projects when new projects are entered. For this purpose, a similarity measurement model based on set theory and probability theory is presented. The presented measurement model is implemented into an actual system to identify redundant documents, and calculate and show their similarity.

키워드

R&D, 유사성, 중복성, 특허, 집합이론, 확률이론

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. 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][3][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.

Ⅱ. 관련 연구

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. 기존의 유사도 분석 모델에 관한 연구로는 핑거프린트(Fingerprint)로 차원 감소(Dimensionality Reduction)하여 이를 문서의 특징으로 사용하는 방식[6], 다중 레벨 인덱싱 구조[7]를 이용하는 방식, 문서에 등장하는 단어를 나열하고, 이 중 일부를 문서의 특징으로 추출하는 방식[8][9], 초기분석에 의해 이웃(Nearest Neighbors)을 찾은 뒤 상세한 분석을 수행하는 방식[10], R&D 과제 측면에서 유사도를 측정하기 위한 알고리즘인 포괄성형망 모델[11] 등이 있다.

Ⅲ. 유사도 측정 모델

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.

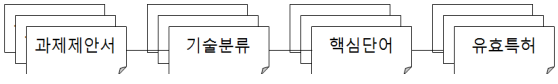

그림 1. 특허기술동향조사 정보의 구성

본 연구에서는 이러한 자료들을 중심으로 신규 과제가 입력되었을 때, 과제간의 유사도를 계산한다. 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, 집합이론에 따른 유사도는 2개의 과제 간 유효특허 중복 비율이 높을수록 유사도가 높다는 측면을 반영하는 척도이다. 집합기반 유사도의 측정을 위한 공식은 수식(1)과 같다.

집합기반 유사도 = 2개 유효특허의 교집합 / 2개 유효특허의 합집합 (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. 확률이론(Probability Theory)[13]은 확인되지 않은 결과에 대한 가능성을 표현할 수 있다. 확률기반 유사도의 측정을 위한 공식은 수식(2)와 같다.

- *A의 조건에서 A와 B가 상이할 확률 => p(A≠B | A) = P(A - B) / P(A) =>* ***"A가 B와 상이할 확률"***
- *B의 조건에서 A와 B가 상이할 확률 => p(A≠B | A) = P(B-A) / P(B) =>* ***"B가 A와 상이할 확률"***
- *1.0 - "A가 B와 상이할 확률" =>* ***"A가 B와 유사할 확률"***
- *1.0 - "B가 A와 상이할 확률" =>* ***"B가 A와 유사할 확률"***
- *"A가 B와 유사할 확률" x "B가 A와 유사할 확률" => "A와 B가 유사할 확률" =>* ***"확률이론 기반 유사도"*** (2)

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.

- ㅇ *A의 조건에서 A와 B가 상이할 확률 => p(A≠B | 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" =>* ***"확률이론 기반 유사도 (전체)"***
- ㅇ *Max("A가 B와 유사할 확률", "A와 B가 유사할 확률") =>* ***"확률이론기반 유사도 (부분)"*** (3)

Ⅳ. 적용 결과 및 시사점

제안한 유사도 측정방법을 기반으로, 본 연구는 156개 과제, 160,218개의 유효특허를 기반으로 유효특허기반 과제유사도 측정 하였다. 측정 결과, 집합의 관점에서 유효특허의 합집합 중 약 13%가 일치하는 것으로 해석되었고, 2개 과제의 유사할 확률은 약 42%로 나타났다.

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. 또한, 제안한 모델은 다양한 척도를 제시하고 있고, 이를 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 도출해야 할 필요성이 있다.

참고문헌

[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 (2011): pp.457-465.

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

[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 (2001): pp.275-290.

[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 Sep. 2012(pp. 317-324).

[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.