연구제안서 스크리닝 시스템의 설계에 관한 연구

On the Design of R&D Proposal Screening System

최창우*, 김선우*, 김혜리*, 박용태**

< 목 차 >

Ⅰ. 서 론
Ⅱ. 관련 방법론
2.1 텍스트 마이닝
2.2 사례기반 추론
Ⅲ. 스크리닝 시스템.
3.1 전체적 프로세스
3.2 변수정의
3.3 변수의 가중치 결정
3.4 사례구축 베이스
3.5 유사 연구의 도출
Ⅳ. 결론 및 추후 연구과제

Abstract

As the size and scope of R&D investment explodes, the strategic and managerial importance of R&D proposal screening becomes highlighted. This point is particularly true for a large-scale research center that deals with multi-product and multi-technology R&D projects. Despite the importance, however, previous research has focused on project evaluation and selection stage. In this research, we propose a R&D proposal screening system. The main objective of the system is to filter R&D proposals that are identified to be duplications of past or existing projects. To this end, the algorithm of the system employs text mining, multivariate statistical method, and case-based reasoning.

키워드 : 연구개발 관리, 연구개발 제안서, 스크리닝, 유사성, 텍스트 마이닝, 사례기반추론

* : 서울대학교 산업공학과 대학원
** : 서울대학교 산업공학과 교수

I. 서론

기술경쟁의 심화에 따른 연구개발 활동의 확대로 연구개발 프로젝트의 수와 범위도 증가하고 있다. 특히, multi-technology, multi-product로 대변되는 대규모 연구소에서 관리되는 연구개발 프로젝트의 수는 기하급수적으로 증가하고 있다. 이 과정에서 새로운 연구제안서와 과거에 수행했던 프로젝트의 유사도를 검토하여 동일 내지 유사 과제를 걸러내는 작업은 중복투자에 의한 인력과 비용의 낭비를 방지하는 첫걸음이 된다.

이러한 중요성에도 불구하고 기존 연구는 주로 프로젝트 평가 및 선정 방법론에 집중되었다. 이러한 방법론은 전문가에 의한 직관적 평가와 같은 정성적 방법에서 수리적 기법을 활용한 정량적 방법까지 다양하게 제시되어 왔다. 예를 들면, 효용을 극대화하는 자원의 효율적 배분을 목표로 하는 수리적 방법론, 투자에 비례하는 이익의 극대화를 목표로 하는 경제성 분석 방법론, 개인이나 조직의 의사 결정 과정의 합리성을 목적으로 하는 의사결정 방법론, 프로젝트를 선정하는 의사 결정자의 알고리즘을 구현하여 자동적으로 프로젝트를 선별하고 자원을 할당하는 것을 목적으로 하는 인공지능 방법론, 프로젝트의 포트폴리오를 최적화하기 위해서 제시된 다양한 기법을 활용하는 포트폴리오 기법 등이 프로젝트의 평가 및 선정을 위한 방법론으로 제시되어 왔다.[7][6][4][3][8][16]

프로젝트 평가 및 선정은 일차적인 스크리닝이 이루어진 다음에 수행되는 이차적 단계로 일차 단계를 통과한 제안서를 대상으로 한다. 그러나 설립연도가 오래된 대규모 연구소의 경우 선정 대상 과제의 수가 너무 많아 개별적인 평가가 어려울 뿐만 아니라 제안된 연구가 과거에 수행하였던 연구와 비슷하거나 중복이 되는지에 대한 여부조차 파악하지 못하는 경우가 많다.

본 연구에서는 연구개발 프로젝트 시작 단계인 제안서 검토과정에서 과거의 연구개발 프로젝트와의 유사도를 비교하여 일차적인 스크리닝을 수행하는 시스템을 제안한다. 연구 내용은 다음과 같다. 첫째, 본 연구에서 사용된 방법론인 텍스트 마이닝, 사례기반 추론들에 대한 설명과 본 연구에서의 활용 방향을 제시한다. 둘째, 연구제안서 스크리닝 시스템의 알고리즘을 제안한다. 마지막으로 제안된 알고리즘의 유용성과 한계에 대한 시사점을 제시한다.

II. 관련 방법론

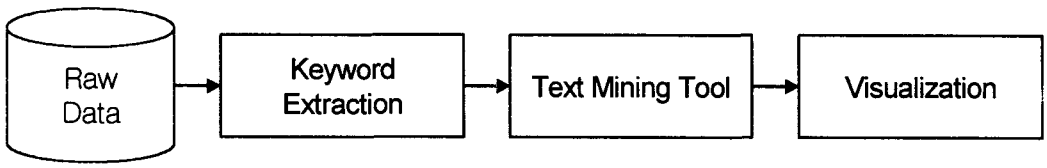
2.1 텍스트 마이닝

연구 개발 제안서는 일반적으로 일정한 형식에 따라 기술한다. 본 연구에서도 연구제안서 및 기존 연구 관련자료 모두 이 형식에 의해 작성 및 보관되어 있다고 가정한다. 그러나 자연언어로 서술된 text document 형식의 제안서와 기존 연구자료를 분석하기 위해서는 의미 있는 정보를 추출하는 작업이 선행되어야 한다. 텍스트 마이닝은 비구조적인 자료에서 의미 있는 패턴과 정보를 추출하는 유용한 도구이다. 사전적으로는 "많은 양의 문서 자료에서 의미 있는 정보를 추출하거나 분석하여 지식을 발견하는 프로세스"로 정의할 수 있다.[5] 추출된 정보는 핵심

단어가(key word) 기본이 되며 핵심 단어에 기초하여 어구를 정의하며 라벨(label)을 표기한다. 핵심단어는 사전에 통제된 핵심 어휘 사전을 정의 할 수도 있고 자동적으로 문서의 학습과정에서 정의 할 수도 있다.[1] 그러나, 사전을 정의하는 방법은 문서의 특성에 따라 유연성이 떨어지는 문제를 안고 있다. 특히 기술과제의 경우 연구 개발이 진행되면서 기술 도메인도 변하게 되며 대규모 연구소는 매우 다양한 분야의 연구가 병행되므로 핵심 단어는 지속적으로 변경되어야 한다. 따라서 본 연구에서는 자료 문서들을 학습하여 자동적으로 핵심 단어를 추출하고 이것을 연구제안서의 유사성 평가 변수로 사용하였다.

유사한 문서를 찾기 위해 문서를 군집하는 방법은 일반 군집(Regular Association), 개념 계층(Concept Hierarchy), 전체 문서 마이닝(Full Text Mining)의 세 가지 방법이 있다. 일반 군집은 자연언어처리 기법을 활용하여 문서의 구조화된 목록을 작성하고, 일정한 알고리즘에 따라 제약 조건을 만족시키는 문서들의 집합을 추출한다. 개념 계층은 문서에 등장한 개념들의 계층을 작성하고, 문서간의 개념 계층을 비교하여 문서들의 관계를 파악한다. 전체 문서 마이닝은 핵심단어 목록이 아닌 문서의 전체 내용을 대상으로 한다. 반복적인 문서의 구조는 유의미한 정보를 제공한다는 전제하에 term extraction이나 part-of-speech tagging과 같은 자연언어 처리 기법을 활용한다. 또한 전체 문서로부터 지식을 추출하며 그림이나 그래프로 표현하는 과정에 대한 자동화된 프로세스가 연구되었다.[12] 본 연구에서는 기존 문서의 구조화된 목록을 작성하거나 문서들의 계측을 작성하는 단계를 거치는 것이 아니라 전체 문서의 내용을 대상으로 하여 유사한 기존의 연구를 찾기 위해 전체 문서 마이닝의 방법을 적용한다.

텍스트 마이닝 시스템은 원하는 정보와 문서의 종류에 따라 다양하지만, 일반적인 구조는 [그림 1]과 같다. 자연언어로 서술된 문서에 우선 추출된 핵심단어로 라벨을 부여한다. 그 다음에 텍스트 마이닝 도구를 활용하여 정보를 추출한다. 마지막으로 정보를 통합하여 시각화한다.



[그림 1] 텍스트 마이닝 시스템의 기본 구조

TextAnlayst, Fact, Document Exlpoer, NeruDoc 등이 범용으로 사용되는 텍스트 마이닝 시스템이다. 본 연구에서는 TextAnlayst를 사용하였다. TextAnlayst는 대량의 문서를 검색, 요약, 분석할 수 있으며, 의미 기반의 추론과 특정 주제에 대한 문서 분석도 가능하다. 신경망 네트워크와 자연언어 처리를 모두 활용하기 때문에 비구조적인 문서의 분석에서 정확도가 높다.[11]

2.2 사례기반 추론

사례기반 추론은 명확한 이론적 근거가 없는 상황에서 문제를 해결하기 위해 널리 사용되는 기법이다. 이 기법은 인간이 새로운 문제를 직면할 경우 과거의 경험을 통하여 비슷한 사례를

찾아 재사용하는 방식을 응용한 것이다. 사례기반추론의 접근 방법에 따라 Textual CBR, Conversational CBR, Structural CBR로 구분할 수 있다.[2] Textual CBR은 수 백가지 이하의 문서형태의 사례에 적용되며, Conversational CBR은 의사결정에 필요한 질문항목과 이에 대한 답변 항목을 통하여 추론을 한다. Structural CBR은 구조화된 도메인 모델을 통하여 사례기반을 구축하고 추론하는 기법이다.

<표 1>은 각각의 사례기반 추론의 특성을 보여준다. Textual CBR은 텍스트 형식의 문서를 통하여 사례베이스를 구축하며 헤더(header)에 따라 구조화된다. 헤더 정보를 이용하여 유사한 문서를 바탕으로 새로운 문서를 작성할 수 있도록 한다. 초기 투자가 간소한 반면에 결과물에 대한 품질을 관리하기 위한 유지관리가 어렵다는 특징이 있다. Conversational CBR은 질문과 답변 리스트를 통하여 사례 베이스를 구축하며 일반적인 데이터 구조를 필요로 하지 않는다. 대화항목을 통하여 질문과 답변리스트를 찾아나가는 방식으로 사례를 추론하며 해당 사례에 대한 가능한 범위의 대안을 제시한다. Structural CBR은 DB에 기록된 레코드를 바탕으로 사례를 추론한다. 레코드에 기록된 각 변수의 속성값을 바탕으로 유사한 사례를 찾아내며, 레코드 값을 통하여 각 사례를 설명한다.

본 연구에서는 이러한 접근 방법 중 전체 문서에 대한 마이닝 기법과 Structural CBR을 활용하여 새로운 연구개발 프로젝트 제안서와 유사한 과거의 제안서를 찾는 프로세스를 제안한다.

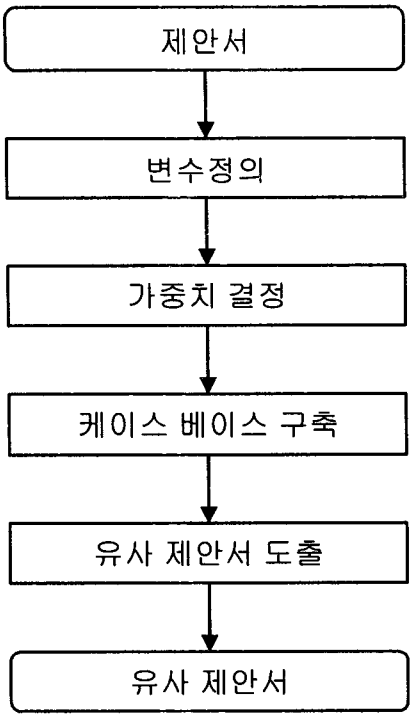
<표 1> 사례기반추론 접근방법의 비교

	Textual CBR	Conversational CBR	Structural CBR
Reuse existing materia.	Very Low	High	Medium
Initial modeling	High	Low	High
Case creation	Low	High	Medium
Tuning CBR	Very High	High/impossible	Medium-Low
Quality control	Very High	High	Low
Maintenance	Low: Case base High: Retrieval	High	Very Low

Ⅲ. 스크리닝 시스템

3.1 전체적 프로세스

유사한 기존의 연구개발 과제를 추출하기 위한 알고리즘의 전체적 프로세스는 [그림 2]와 같다. 첫째, 텍스트 마이닝(Text Mining)을 통하여 연구소내 전체 연구개발 결과 보고서 등과 같은 연구개발 관련자료에서 키워드(Keyword)를 추출한다. 추출된 키워드는 유사 연구의 비교를 위한 변수로 사용한다. 둘째, 선정된 변수의 가중치(Weight)를 결정한다. 변수의 가중치는 개별 키워드 중에서 해당 분양에서 유사 연구를 검토할 때 사용되는 키워드의 중요도에 따라 결정된다. 셋째, 제안서로부터 추출된 키워드로 구성되어 있는 사례 베이스(Case Base)를 구축한다. 넷째, 추출된 과제를 확인하여 새로운 연구개발 제안서와 중복되는 연구인가를 판단한다.



[그림 2] 스크리닝 알고리듬 과정도

3.2 변수 정의

먼저 기존 연구 결과 관련자료를 대상으로 텍스트 마이닝을 수행한다. 이를 통해 주요 키워드를 추출하여 키워드의 리스트를 작성하며 이를 새로운 제안서와 기존 연구의 유사도를 비교하기 위한 변수로 활용한다 (<표 3 참조>).

문서 내에서 텍스트 마이닝을 통하여 키워드를 추출하는 작업은 많은 수의 키워드수를 도출해 낸다. 그러나 모든 키워드를 변수로 활용하는 경우 너무 많은 키워드의 수로 인하여 변수의 수가 불필요하게 증가하므로 유사도가 높은 키워드만을 변수로 사용한다. 본 연구에서는 텍스트 마이닝을 통하여 추출된 키워드의 가중치 값을 기준으로 cut-off 값을 선정한다. cut-off 값은 전체 제안서 텍스트와의 관련성이 0.9(0 이상 1 이하)이상인 키워드를 대상으로 변수를 선정한다. 선정된 변수는 해당 기업의 연구소에서 연구되고 있는 분야를 대표할 수 있는 키워드로 판단할 수 있으며 이들을 통하여 새로운 제안서와 기존 연구의 유사도를 비교한다.

<표 3> 전제 제안서에 대한 텍스트 마이닝을 통해 추출된 키워드리스트와 속성

속성	키워드 1	키워드 2	키워드 3	키워드 4	키워드 5	키워드 6	…
빈도	n_1	n_2	n_3	n_4	n_5	n_6	…
가중치	w_1	w_2	w_3	w_4	w_5	w_6	…

3.3 변수의 가중치 결정

개별 변수로 사용되는 키워드는 연구주제와의 연관성이 모두 다르다. 따라서 해당 연구소에서 이미 수행된 연구결과 관련자료를 바탕으로 각 변수의 키워드에 대한 가중치를 결정한다. 일반적으로 가중치 결정을 위해서는 크게 두 가지 방법이 사용된다.[9] 첫째는 해당 분야의 전문가의 직관에 의한 방법이며, 둘째는 개별 변수의 중요도의 순위를 매겨 해당 순위에 따른 가중치를 부여하는 방법이다. 본 연구에서는 두 방법의 장단점을 비교하여 두 번째 방법인 텍스트 마이닝을 통한 키워드 가중치를 변수의 가중치로 활용하는 방안을 사용한다. <표 4>에 나타난 바와 같이 직관적인 방법에 의한 가중치의 부여는 유연성이 높고 상대적으로 해석이 용이하다는 장점이 있으나 변수의 개수가 너무 많을 경우 직관적으로 가중치를 정하는 것은 많은 시간이 소요되며, 전문가에 따라 다른 값을 활용할 수 있는 문제가 발생하기 때문에 텍스트 마이닝을 통해 구해진 키워드의 가중치 값을 변수의 가중치로 활용하는 방안을 선택한다.

<표 4> 가중치 정의 방법의 비교

	직관적 방법	자동화 방법(텍스트 마이닝)
장점	· 각 키워드에 대한 가중치 적용의 유연성 높음 · 직관에 의해 결정되므로 해석의 상대적인 용이 · 각 가중치의 의미 파악 용이	· 자동화되어 일괄적으로 적용 가능 · 미리 규정된 알고리즘 활용으로 인한 결과의 안정성
단점	· 많은 시간의 소모 · 해당 분야의 전문가에 따라 다른 값의 활용	· 각 키워드의 해석의 어려움 · 유사 제안서의 유사도 판단의 근거를 제시하기 어려움

3.4 사례 베이스 구축

사례 베이스를 구축하기 위해 해당 연구소에서 수행되었던 연구결과 관련 자료에 대하여 각 변수의 값을 입력한다. 각 변수 값의 입력을 위해 텍스트 마이닝을 통하여 키워드를 추출해 낸다. 텍스트 마이닝에서는 키워드의 리스트, 전체 문서에 대한 가중치, 키워드 사이의 가중치, 키워드의 빈도수 정보를 얻는다. 이 중 앞에서 정의된 변수 값으로 사용할 수 있는 것은 키워드의 빈도수와 전체 문서에 대한 가중치 값을 사용할 수 있다. 그러나 절대 빈도값은 문서에 따라 분량의 차이가 발생하므로 전체 스케일의 조정이 이루어진 가중치 값을 사용한다.
<표 5>는 키워드를 추출하여 각 키워드의 가중치 값을 정의된 변수에 입력한 예를 보여준다. 여기에서 초기에 추출된 변수는 모든 문서를 대상으로 키워드를 추출한 것이므로 전체 변수의 개수가 해당 문서에서 추출된 키워드보다 많게 된다. 따라서 정의된 변수들 중에 해당 문서에서는 나타나지 않는 키워드가 있어 변수의 값을 지정하지 못하는 경우가 발생할 수 있다.

<표 5> 사례 베이스의 구축 예

Keywords	변수 1	변수 2	변수 3	변수 4	변수 5	변수 6	변수 7	변수 8	…
Weight	99	99	98	99	96	94	99	95	…
문서 1			99	98		17	99		…
문서 2		97	4	98		9			…
문서 3			95			49	99		…
문서 4	97		38		48	4		98	…
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

3.5 유사 연구의 도출

새로운 연구와 유사한 기존 연구를 찾기 위해 새로운 연구에 대한 제안서와 기존 연구에 대한 관련 문서의 유사도를 사례 베이스를 통하여 비교한다. 사례 베이스를 통해 새로운 제안서와 기존 연구와의 유사성을 비교하기 위해서는 유사성에 대한 정의(operational definition)와 판단 기준이 필요하다. 본 연구에서는 유사도를 측정하기 위해 (식 1)과 같이 Numerical Evaluation Function을 활용한다.[9] 제안서의 유사도는 각 변수의 가중치와 각 변수의 거리의 차이의 합을 전체 변수의 가중치로 나누어 계산하는 방식을 사용한다.

$$S_i = \frac{\sum_{i=1}^{n} w_i \times dis(I_t, I_r)}{\sum_{i=1}^{n} w_i} \qquad \text{(식 1)}$$

여기서, S_i : 제안서와 기존 연구간의 유사도

w_i : i번째 변수의 가중치

I_t : 기존 연구의 i번째 변수 값

I_r: 새로운 제안서의 i번째 변수 값

$dis(I_t, I_r)$: 기존 연구와 새로운 제안서간의 i번째 변수 거리

<표 6>은 (식 1)을 통하여 계산되는 유사 연구의 도출 결과를 예시적으로 보여준다. 새로운 제안서를 기존의 연구와 비교했을 경우 '기존 연구 6'이 가장 유사한 결과를 보여주며, '기존 연구 8'이 두 번째로 유사한 결과를 나타내고 있다. 이중 변수 7이 가장 유사도가 높으며, 변수 8 또한 높은 유사성을 보이는데 큰 영향을 미친 것으로 파악할 수 있다.

이와 같이 도출되는 기존의 유사연구는 새로운 연구개발을 프로젝트를 시작하기 전에 기존 연구소 내에서 수행되었던 프로젝트들과 새로 시작하고자 하는 프로젝트 사이의 중복을 검토하여 일차적인 스크리닝을 위한 유용한 방법으로 활용할 수 있다. 즉, 키워드를 기반으로 새로운 제안서와 유사한 기존 연구 프로젝트를 찾아서 해당 제안서가 기존 연구와 얼마나 유사한지를 판단하여 일차적인 의사결정을 할 수 있는 것이다.

<표 6> 유사한 기존 연구의 도출

변수	가중치	새 제안서	기존연구 6	기존연구 8
변수 1	99	99	96	31
변수 2	96	98		
변수 3	99		99	99
변수 4	99	92	15	93
변수 5	99	24		
변수 6	94	4		
변수 7	99	99	95	98
변수 8	99	16	99	36
변수 9	99	98	20	
⋮	⋮	⋮	⋮	⋮

Ⅳ. 결론 및 추후 연구과제

본 연구는 많은 수의 연구를 진행하는 multi-technology, multi-product 방식의 대규모 연구소에서 새로운 연구 제안서의 중복성을 검토하기 위한 알고리즘을 제안하였다. 이 알고리듬을 기반으로 개발되는 스트리닝 시스템은 첫째, 프로젝트 선정 여부를 결정하는 일차적 기준으로 활용될 수 있고, 둘째, 향후 진행될 프로젝트의 방향 설정 및 결과 예측에 도움을 줄 수 있다.
그러나 본 연구는 몇 가지 점에서 한계를 가지고 있다. 첫째, 제안서 및 연구개발 관련자료에서 추출된 키워드 정보가 제안서 및 연구개발 관련자료의 실질적인 내용을 얼마나 설명할 수 있는가 하는 문제이다. 이는 텍스트 마이닝을 통하여 추출되는 정보가 키워드의 빈도 및 관계에 대한 정보이므로 문서의 문맥적인 내용은 정확하게 반영하기 어렵기 때문이다. 둘째, 유사성이 존재할 경우 어떠한 측면에서 유사한지에 대한 정확한 설명이 어렵다. 두 문서의 유사성 관계를 정확하게 파악하기 위해서는 전문가들에 의한 내용 검토 내지 통계기법의 적용을 통한 추가적인 정보 가공이 필요하다. 셋째, 본 연구는 스트리닝을 위한 아이디어 및 알고리즘을 제시하는 수준에 그치고 있으므로 향후 실제 시스템을 구현하는 작업에 대한 연구가 진행될 것이다.

< 참고문헌>

[1] Apte, C., Damerau, F., and Weiss, S., "Automated learning of decision rules for text categorization", *ACM Transaction on Information Systems*, Vol. 12, No. 3, pp. 233-251, 1994.

[2] Bergmann, R. et al, *Developing Industrial Case-Based Reasoning Applications; The INRECA-Methodology*, Springer-Verlag Berlin Heidelberg, 1999.

[3] Bohanec, M., Rajkovie V., Semolic B. and Pogacnik A., "Knowledge-based portfolio analysis for project evaluation", *Information & Management*, vol. 28, no. 5, pp. 293-302, 1995.

[4] Brenner, M., "Practical R&D project prioritization", *Research Technology Management*, vol. 37, no. 5, pp. 38-42, 1994.

[5] Cardie, C., "Empirical Methods in information Extraction", *AI Magazine*, Vol.18, No. 4, pp. 65-79, 1997.

[6] Chun, Y., "Sequential decisions under uncertainty in the R&D project selection problem", *IEEE Transactions on Engineering Management*, vol. 40, pp. 404-413, Nov. 1994.

[7] Czajkowski, A. and Jones, S., "Selecting interrelated R&D projects in space planning technology", *IEEE Transactions on Engineering Management*, vol. 33, pp. 17-24, Feb. 1986.

[8] Kocaoglu, D. and Iyigun, M., "Strategic R&D project selection and resource allocation with a decision support system application", *Proceeding 1994 IEEE International Engineering Management Conference*, pp. 225-232, 1994.

[9] Kolodner, J., *Case-Based Reasoning*, Morgan Kaufmann Pulishers, San Mateo, 1993.

[10] Mandakovic, T. and Souder, W., "Experiments with microcomputers to facilitate the use of project selection models", *Journal of Engineering and Technology Management*, vol. 7 no. 1, pp. 1-16, 1990.

[11] Semio Corporation, "Text mining and the Knowledge Management Space", *DM Review*, Dec. 1999.

[12] Zhu, D. and Porter, A., "Automated extraction and visualization of information for technological intelligence and forecasting", *Technological Forecasting and Social Change*, Vol. 69, pp. 495-506, 2002.