

온톨로지 기반 법령 검색시스템의 개발: 철도·교통 분야 연구개발사업을 중심으로

A Development of Ontology-Based Law Retrieval System: Focused on Railroad R&D Projects

원민재(Min-Jae Won)*, 김동희(Dong-He Kim)**, 정해민(Hae-Min Jung)***
이상근(Sang Keun Lee)****, 홍준석(June Seok Hong)*****, 김우주(Wooju Kim)*****

초 록

철도교통 분야의 연구개발사업은 여러 법령과 긴밀하게 관련되어 있기 때문에, 연구개발을 성공적으로 수행했더라도 법령에 의해 제약되어 연구개발 결과의 실질적인 사업화 또는 실용화를 이루어내지 못하는 사례가 발생하고 있다. 본 논문에서는 이러한 사례를 방지하기 위한 방편으로 철도교통 분야에서 진행되는 연구개발사업과 관련된 법령을 검색할 수 있는 법령검색시스템의 모델을 제시하였다. 사업 내용을 설명하는 연구개발계획서가 시스템에 입력되면 요약서의 내용을 대상으로 형태소 분석을 수행하여 명사들만을 남긴다. 국가법령정보센터에서 제공하는 법령정보공동활용서비스를 사용하여 명사들 중 법령용어를 분류하고, 법령용어와 해당 법령 용어를 정의하는 법령과의 관계를 지능형 지식 베이스인 온톨로지에 저장한다. 온톨로지에 저장된 법령들은 본 연구에서 개발한 추가적인 지표 계산과정을 거쳐 연구개발사업과 관련된 정도를 기준으로 순위가 매겨진 후, 시스템 사용자에게 제공된다. 사용자는 연구개발에 영향을 미칠 수 있는 법령을 검색할 수 있게 되어 사업 시작 전에 연구 방향을 결정하는 데 참고하거나, 사업 진행하는 과정에서 참고자료로 사용할 수 있다. 궁극적으로, 법령에 의해 철도교통 분야 연구개발사업이 실패하거나 실용화되지 못하는 경우를 사전에 방지함으로써, 사업에 투자한 예산에 의해 기대되는 충분한 기술적·경제적 효과를 얻을 수 있게 될 것이다.

ABSTRACT

Research and development projects in railroad domain are different from those in other domains in terms of their close relationship with laws. Some cases are reported that new technologies from R&D projects could not be industrialized because of relevant laws restricting

본 연구는 한국철도기술연구원 주요사업비 지원으로 수행되었습니다.

* First Author, Department of Information Industrial Engineering, Yonsei University
(wmjzzang@naver.com)

** Second Author, Korea Railroad Research Institute(kdh777@krri.re.kr)

*** Third Author, Department of Information Industrial Engineering, Yonsei University(mainartery37@gmail.com)

**** Fourth Author, Korea Railroad Research Institute(sangkeun@krri.re.kr)

***** Corresponding Author, Department of Management Information Systems, Kyonggi University
(junehong@kyonggi.ac.kr)

***** Sixth Author, Department of Information Industrial Engineering, Yonsei University
(wkim@yonsei.ac.kr)

Received: 2015-10-29, Review completed: 2015-11-19, Accepted: 2015-11-23

them. This problem comes from the fact that researchers don't know exactly what laws can affect the result of R&D projects. To deal with this problem, we suggest a model for law retrieval system that can be used by researchers of railroad R&D projects to find related legislation. Input of this system is a research plan describing the main contents of projects. After laws related to the R&D project is provided with their rankings, which are assigned by scores we developed. A ranking of a law means its order of priority to be checked. By using this system, researchers can search the laws that may affect R&D projects throughout all the stages of project cycle. So, using our system model, researchers can get a list of laws to be considered before the project they participate ends. As a result, they can adjust their project direction by checking the law list, avoiding their elaborate projects being useless.

키워드 : 철도교통, 연구개발사업, 법령 검색 시스템
Railroad, R&D, Law Retrieval System

1. 서 론

철도교통 분야의 연구개발사업 과정은 중·장기적 관점에서 보았을 때, 초기 기획부터 연구개발 이후의 실용화까지 여러 단계를 거친다. 사업이 여러 해에 걸쳐 진행되기 때문에, 연구개발 중인 기술과 관련된 법이 개정되거나 새롭게 제정되는 경우가 적지 않다. 관련 법령의 변경은 작게는 연구의 진행방향에 영향을 미칠 수 있으며, 크게는 개발한 기술을 실용화할 수 없게 되어 궁극적으로는 국가 예산의 낭비를 초래하게 된다. 많은 예산을 투입하여 연구개발사업을 수행했다고 하더라도 개발된 기술의 실질적인 실용화·사업화에 미치지 못하여 연구개발사업에 대한 투자 효과를 제대로 이끌어내지 못하고 있는 것이다. 또한 연구를 통해 새롭게 개발한 기술적 성과가 국내 산업계에 확산되지 못함으로써 국제 시장에서의 철도산업 기술경쟁력을 한 단계 끌어올릴 수 있는 기회를 잃어버리게 된다는 문제점이 있다.

이와 같은 문제를 해결하려면 연구개발사업과 관련된 법령들을 적시에 파악하여 연구개발의 내용과 연계함으로써, 연구자들이 기획단계

에서부터 사업 계획을 내실 있게 구성할 수 있도록 하는 한편 연구개발의 과정에서 적절하게 연구방향을 조절할 수 있도록 하여야 한다. 따라서 본 연구에서는 철도교통 분야 연구개발사업과 관련된 법령들을 파악하여 연구자들에게 제공할 수 있는 법령 검색 방법론을 개발하고, 실제 시스템으로 구현함으로써 그 가능성에 대해 살펴보았다. 본 논문에서 제안하는 법령 검색 시스템은, 연구개발사업의 내용을 담고 있는 연구개발계획서에 포함된 내용을 추출하고 이로부터 유관 법령의 목록을 검색한 뒤, 코사인 유사도 기반으로 계산된 점수에 따라 법령의 우선순위를 결정하여 사용자에게 제공한다. 이 때, 법령 검색 과정에서는 국가법령정보센터의 Open API를 사용하며, 법령의 정보를 지속적으로 누적시키기 위한 지식 베이스로서 온톨로지(ontology)를 구축하여 활용한다.

본 논문의 구성은 다음과 같다. 제2장에서는 본 연구와 관련된 연구 및 기술에 대해 알아본다. 제3장에서는 철도교통 분야 연구개발과 관련된 법령을 검색하는 방법론에 대해 설명하였다. 제4장에서는 개발한 방법론을 기반으로 법령 검색 시스템을 시범적으로 구축하였으며, 제5

장에서는 본 논문의 기대효과 및 향후 개선방향을 제시한다.

2. 관련 연구 및 관련 기술

본 연구에서는 철도교통 분야 법령을 저장하는 지식 데이터베이스로서 온톨로지를 활용하였다. 웹(web)으로 대표되는 정보시대에서, 정보는 그 자체로 의미를 가질 뿐만 아니라 다른 정보와의 관계에서도 의미를 가진다. 정보간의 다양한 관계를 표현하는 것은 기존의 관계형 데이터베이스를 통해서만 한계가 있었기 때문에, 새로운 지식 베이스로서 온톨로지가 등장하였다.

온톨로지는 여러 대상 간의 관계를 구조화하여 저장함으로써 사용자가 그 관계를 명확하게 파악할 수 있도록 한다. 어떤 분야(domain)에 대해 구축된 온톨로지는 해당 분야와 관련된 대상과 대상들 간의 관계를 종합적으로 표현하는 지식 베이스로서의 기능을 수행할 수 있으며, 의미적 검색 기능을 제공한다. 온톨로지를 검색 시스템에 활용한 연구로는, 도메인 온톨로지를 구축하고 유사도 기반 우선순위 검색 기법을 사용한 연구[8], 문서의 검색을 위해 온톨로지 기술을 이용한 연구[4, 7, 10], 온톨로지에 기반을 두고 구현한 검색 시스템의 개발에 관한 연구[11] 등이 있다.

온톨로지를 활용하여 법령 정보를 구조화한 연구로는, 자연어처리(Natural Language Processing, NLP) 기술을 사용하여 법률적 개념들 간의 의미 관계를 반영한 온톨로지를 구축하고 법률 검색시스템에 적용한 사례가 있다[6]. 이 연구에서는 구문분석과 통계분석을 통

해 법률 개념을 추출하고 법률 문서를 특성에 따라 분류하는 등 분야별 법률 개념들을 구조화하였다.

국내에서 법령과 관련하여 온톨로지 형태의 지식 베이스를 구축한 연구는 많지 않지만, 법률적 지식이 부족한 일반인도 원하는 바에 맞게 법령정보를 검색할 수 있도록 생활용어와 법률용어간의 대응관계를 탐색하는 데 온톨로지를 활용한 사례가 있으며[5], 법령 정보 검색 엔진을 구축하기 위해 온톨로지에 기반을 둔 모델을 제시한 사례가 있다[3].

3. 연구개발사업 관련 법령 검색 방법론

본 장에서는 철도교통 분야 연구개발사업과 관련된 법령을 검색하기 위한 방법론을 제시한다. 방법론을 적용할 대상이 되는 데이터는 ‘연구개발계획서’이다. 연구개발계획서는 사업에 착수하기 이전 시점 또는 사업 진행 중 사업 단계가 변경되는 시점에 작성되는 문서로서, 과제 전체적인 내용을 설명하고 있으며 사업명, 과제 번호, 연구(기술) 분야, 과제분류, 연구 수행형태, 주관연구기관 등의 다양한 메타데이터도 담겨있다. 본 연구에서는 연구개발계획서의 내용을 압축적으로 담고 있는 요약서에 포함된 7개 섹션(Section)의 데이터를 사용한다. 각각의 섹션으로는 연구개발 개요, 최종목표, 연구내용 및 범위, 연도별 주요 내용, 연구 성과, 활용방안, 핵심어가 있다.

본 장의 구성은 다음과 같다. 제3.1절에서는 요약서를 형태소 분석하여 조사, 동사 등 상대적으로 불필요한 단어들을 제외하고 의미적으로

중요하다고 판단되는 명사 형태의 단어들만 남기는 과정에 대해 설명한다. 제3.2절에서는 단어들 중 법에서 정의하고 있는 법령용어들을 분류해내고, 법령용어와 법령용어를 정의하는 법 간의 정보를 구하는 과정에 대해 설명한다. 이 때, 관련 법령들은 크게 국토(Territory), 물류(Distribution), 교통(Traffic), 철도(Railroad)의 4개 대분류로 한정하고, 각 분류의 대통령령, 부처령, 장관부령까지를 범위로 포함하였다. 자치법규나 행정규칙은 상위 법령에 비해 중요도가 낮다고 판단하여 제외하였다. 대분류에 속한 법령들의 개수는 아래 <Table 1>과 같다. 제3.3절에서는 온톨로지 스키마를 정의해둔 상태에서 연구개발계획서 메타데이터 및 법령용어와 법령용어를 정의하고 있는 법령 관계 등의 정보들을 채워 넣어 온톨로지를 구축한다. 마지막으로 제3.4절에서는 온톨로지를 기반으로 연구개발계획서와 관련된 법령 목록을 검색하고, 검색된 법령 간의 우선순위를 결정하기 위한 6가지 방법을 제안한다.

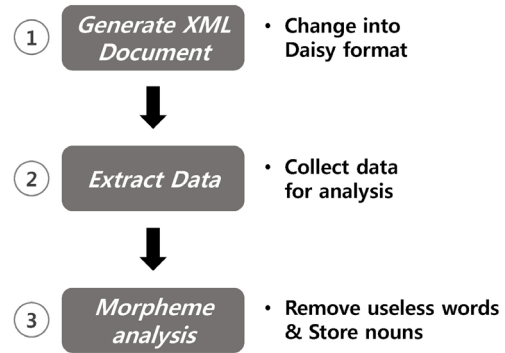
<Table 1> Number of Target Acts

Domain	Number of Acts
Territory	8
Distribution	9
Traffic	10
Railroad	8

3.1 요약서 형태소 분석

요약서 형태소 분석 단계는 연구개발계획서, 그 중 요약서로부터 정보를 추출하는 단계로, 3개의 세부단계로 구성되어 있다. 입력부터 형태소 분석까지의 진행 과정은 <Figure

1>과 같다. 연구개발계획서마다 구조가 상이하므로 본 연구에서는 입력받을 계획서를 국토교통기술연구개발계획서, 건설교통기술연구개발계획서의 두 가지 형태로만 한정하였다.



<Figure 1> Document Analysis Process

첫 번째 세부단계에서는 입력문서인 한글문서(HWP) 형태의 요약서를 XML 형태 문서로 변환한다. 한글문서는 구조가 정형화되어 있지 않아 데이터를 추출하는 것이 상대적으로 까다롭기 때문에 목표 데이터를 바로 추출하기에는 한계가 있으므로, 태그로 내용이 분류된 XML 형태의 데이터(daisy) 문서로 변환하여 입력한다.

두 번째 세부단계에서는 XML 형태의 요약서를 파싱(parsing)하여 목표 데이터를 추출한다. 본 논문에서는 파싱을 위해 자바(java) 기반의 문서객체모델(DOM, Document Object Model) 파서(parser)를 이용하였다.

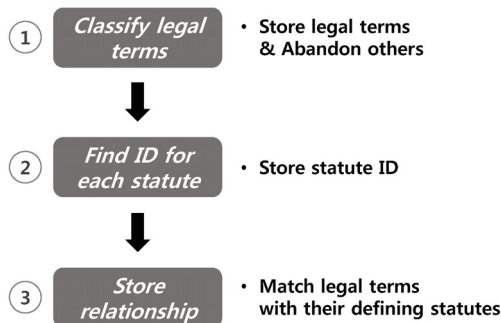
마지막 세부단계에서는 각 섹션에서 추출한 데이터를 형태소 분석하여 명사만 추출한다. 형태소 분석에는 KLT 형태소 분석기가 사용되었다. <Table 2>는 각 섹션별로 추출된 명사 집합을 나타낸 예시이다. 섹션 내에서 명사의 중복은 없지만, 각 섹션별로 명사가 중복 등장할 수는 있다.

〈Table 2〉 An Example of Word Sets for Each Section

Sections	Extracted words
1. Abstract	A, B, C, D
2. Purpose	A, C, E
3. Contents	B, E, F
.....	

3.2 법령용어 분류 및 법령 연계

법령용어 분류 및 법령 연계 단계에서는 형태소 분석 단계에서 얻어진 명사 집합을 대상으로, 법령용어 여부를 결정한다. 이 때, 국가법령정보 공동활용서비스(<http://open.law.go.kr>)에서 제공하는 Open API(Application Programming Interface) 서비스를 활용한다. 이 서비스는 국가법령정보센터(<http://law.go.kr>)의 정보를 실시간으로 제공한다. 법령용어 여부를 결정하면서 법령용어와 법령용어를 정의하는 법령 간의 관계를 파악하는 법령 연계 과정이 동시에 이루어진다. 진행 과정은 〈Figure 2〉와 같으며, 3개의 세부단계로 나누어진다.



〈Figure 2〉 Relation Extraction Process

첫 번째 세부 단계에서는 형태소 분석의 결과로 추출된 모든 명사를 법령정보센터 Open

API를 통해 검색하여 해당 명사를 정의하고 있는 법령이 존재하는지 판단한다. 법령에 의해 정의되는 명사는 법령용어로, 그렇지 않으면 비법령용어로 분류되며, 본 연구에서는 비법령용어로 판단된 명사들은 제외시킨다. 예를 들어 ‘대중교통수단’을 법령정보센터 Open API에서 검색하면, 해당 명사를 정의하고 있는 법령으로 ‘대중교통의 육성 및 이용촉진에 관한 법률’이 반환된다. 즉, ‘대중교통수단’은 법령용어로 판별된다. 하나의 법령용어가 한 개 이상의 법령에 의해 정의될 수 있으며, 하나의 법령은 여러 법령용어를 정의할 수 있다. 따라서 법령용어와 법령은 다대다(多對多)관계를 나타낸다. 〈Table 3〉는 전체 명사 중 법령용어를 분류한 결과를 나타내는 예시이다.

〈Table 3〉 An Example of Extracted Words Classified by Open API Service

Sections	Extracted words	Legal terms
1. Abstract	A, B, C, D	A, C
2. Purpose	A, C, E	A, C, E
3. Contents	B, E, F	E
.....		

두 번째 세부단계에서는 Open API 서비스를 이용하여 법령용어를 정의하고 있는 법들의 법령일련번호를 검색하여 수집한다. 예를 들어, ‘대중교통의 육성 및 이용촉진에 관한 법률’을 검색했을 경우 해당 법령의 일련번호 ‘149569’를 얻을 수 있다. 법령일련번호는 법령 본문을 조회하기 위해 필요하며, 법령 본문은 이후 법령 우선순위를 결정하기 위한 지표계산에 사용된다.

마지막 세부단계에서는 법령용어와 법령의 상관관계를 정리하여 저장한다. 예를 들어,

<Table 4>는 요약서의 법령용어들과 그 법령 용어를 정의하고 있는 법령의 일련번호가 결합되어 저장된 것을 나타내는 예시이다. 일차적으로 이 법령들이 연구개발사업과 관련되어 있다고 볼 수 있다.

<Table 4> An Example of Relations between Legal Terms and Statutes Defining Them

Legal terms	Statutes defining legal terms
A	124253, 126039
C	105231
E	114352, 149203, 150326
.....	

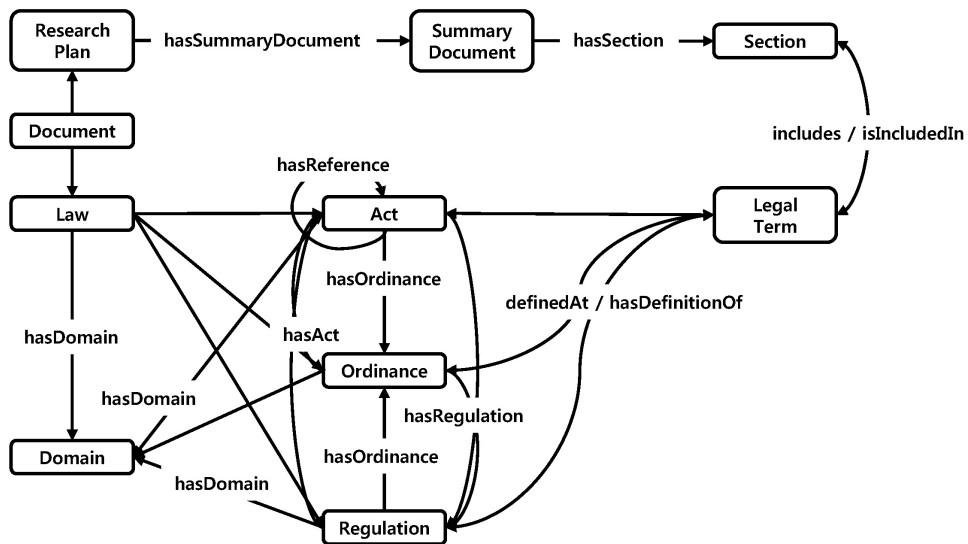
3.3 온톨로지 기반 지식 베이스 구축

이전 단계에서 얻어진 데이터는 온톨로지에 포함되어 지식 베이스가 구축된다. 이렇게 구축된 지식 베이스는 검색 과정을 거치면서 지속적

으로 확장된다. <Figure 3>은 철도기술연구원 온톨로지의 클래스간 관계를 나타낸 것이다.

법령과 관련된 주요 클래스(class)로는 법률(Act), 시행령(Ordinance), 시행규칙(Regulation)으로, 세 수준의 법령은 서로 상하위 관계가 있다. 법률은 가장 상위의 법령이며, 관계 법률의 집합과 연결되어 있다. 법령 간의 관계 정보는 국가법령정보센터의 법령 체계도에 근거하여 수집된다. 부문(Domain) 클래스는 인스턴스(instance)로 본 연구에서 한정된 4개 분야인 ‘국토’, ‘교통’, ‘물류’, ‘철도’의 4개를 고정적으로 포함하고 있다. 연구계획서(ResearchPlan) 클래스의 인스턴스는 요약서(SummaryDocument) 클래스의 인스턴스와 연결되며, 요약서 인스턴스는 7개 섹션 인스턴스들과 연결된다. 각 섹션은 그 안에서 등장한 법령용어들과 온톨로지 내에서 연결된다.

법령 검색 시스템에서 온톨로지의 역할은 요약서의 정보로부터 검색된 법령들의 상하위



<Figure 3> Ontology Schema

법 관계나 관련 법령 관계를 통해 검색 대상이 되는 법령 목록을 확장하는 것이다. 즉, 추출된 법령들의 상하위법을 조회하여 요약서 관련 법령 목록에 추가함으로써 보다 많은 법령들을 사용자에게 제공할 수 있게 된다. 예를 들어, 요약서의 관련 법령으로 도로교통법이 검색되었다면, 그 하위 법령인 도로교통법 시행령과 도로교통법 시행규칙 역시 관련 법령 목록에 추가되는 것이다. 이와 같이 법령용어, 법령용어와 그 용어를 정의하고 있는 법령 간의 관계, 법령의 상하위 관계 등이 온톨로지에 저장되어 있으면서, 새로운 연구개발계획서 입력 시 재활용되거나 추가된다. 온톨로지에서 정보를 검색하고 출력할 때는 온톨로지 질의 언어인 SPARQL을 사용한다.

3.4 법령 간 우선순위 결정

요약서와 관련된 법령 목록은 적절한 우선순위와 함께 사용자에게 제공될 때 의미를 가질 수 있으므로, 본 절에서는 우선순위 결정을

위한 6가지 방법론을 제안한다. 우선순위 결정은 각 법령에 부여된 점수를 기준으로 이루어지며, 법령 점수는 해당 법령에 의해 정의되는 법령용어를 기준으로 계산된다. <Table 5>는 법령 간 우선순위 결정에 사용될, 주요 변수에 대한 정의이다.

법령별 점수는 관련점수(RScore, Relevancy score)라고 표현하며, 아래 <Table 6>은 그 계산에 사용되는 단위 지표들의 표기와 정의를 나타낸다. 이 단위 지표들은 기본적으로 코사인 유사도에 기반을 두고 계산된다. 이 때, 법령의 경우 TF-IDF 벡터로 표현하였고, 요약서에 대해서는 <Table 6>에서 제시한 것처럼 다양한 표현방식을 적용하였다. 관련된 개념들로는 법령용어를 포함하는 섹션의 개수, 단어의 TF-IDF, 단어의 중심성 점수 등이 있다.

3.4.1 법령용어를 포함하는 섹션의 개수

법령용어를 포함하는 섹션의 개수는 요약서의 7개 섹션에서 중 특정 법령용어를 포함하고 있는 섹션의 수를 의미한다. 다수의 섹션에 등

<Table 5> Notations and Definitions for Symbols

Notation	Definition
W	An ordered multiset of words appeared in a given document.
SC	An ordered set of sections belongs to the given document.
SCW_k	An ordered multiset of words appears in k-th section in SC .
V	A set of vocabularies appeared in W .
ST	A set of statute terms from V .
ST_k	A set of statute terms from V , which appears in SCW_k .
S	A set of statutes, which defines at least one term in ST .
$D = \{d_{ij}\}$	A $ ST \times S $ matrix where d_{ij} is 1 if i-th term is ST is defined by j-th statute in S , otherwise, 0.
$D_k = \{d_{ij}^k\}$	A $ ST_k \times S $ matrix where d_{ij}^k is 1 if i-th term in ST_k is defined by j-th statute in S , otherwise, 0.

〈Table 6〉 Notations and Definitions for Unit Score Measures

Notation	Definition
$CScore$	Centrality score of I-th statute term in a given set, ST .
$c(i, j, S) = d_{i,j} \times tfidf(i, j, S)$	Relative contribution of the j-th statute in S, through the i-th statute term in a given set, ST .
$c(i, j, k, S) = d_{i,j}^k \times tfidf(i, j, S)$	Relative contribution of the j-th statute in S, through the i-th statute term in a given set, ST_k .
$cn(i, j, S) = d_{i,j} \times CScore(i, ST) \times tfidf(i, j, S)$	Relative centrality-based contribution of the j-th statute in S, through the i-th statute term in a given set, ST .
$cn(i, j, k, S) = d_{i,j}^k \times CScore(i, ST) \times tfidf(i, j, S)$	Relative centrality-based contribution of the j-th statute in S, through the i-th statute term in a given set, ST_k .

장한다면 요약서에서 비중 있게 다루는 법령 용어라고 볼 수 있다.

3.4.2 단어의 TF-IDF

문서빈도-역문서빈도(TF-IDF, Term Frequency-Inverse Document Frequency)는 문서 집합에서 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다[9]. 기본적인 개념은 등장하는 단어의 빈도가 단어의 중요도를 의미한다는 것이다. 정보 검색과 텍스트 마이닝 분야에서 흔히 가중치로 사용되어 문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정하는데 사용된다.

3.4.3 중심성 점수 계산

중심성 점수(Centrality Score)는 네트워크 이론에 기반을 두고 있는 지표로서, 네트워크에서 어떤 노드가 얼마나 중심에 위치하고 있는가를 나타낸다. 본 논문에서는 요약서의 법령용어들 중 어떤 용어가 핵심적인 것인지를 판단하고 그 정도만큼 가중치를 부여하기 위

해 중심성 점수를 계산하였다. 먼저, 법령용어 간 관계를 표현하기 위해 용어 간 밀접도를 나타내는 PMI(pointwise mutual information) 지수를 사용하였다. 이것은 하나의 단어 쌍(x, y)에 대해 계산되는 확률로서, 단어 x와 y가 각각 등장하는 확률과 동시 등장하는 확률을 사용하여 계산한다[2].

본 연구에서는 법령용어 간 PMI 값을 계산하기 위해 네이버 뉴스 검색 결과 값을 사용했다. 예를 들어, ‘궤도’와 ‘운송’ 각 단어를 네이버 뉴스에서 검색해 나온 검색결과 건수를 곱한 값을 분모로, ‘궤도’와 ‘운송’을 함께 검색해 나온 검색결과 건수를 분자로 계산한다. 이 때, ‘궤도’와 ‘운송’의 순서는 고려하지 않으므로 ‘궤도-운송’의 PMI 값과 ‘운송-궤도’의 PMI 값은 같다. 두 법령용어 간 PMI 값이 클수록 두 용어는 서로 높은 관련성이 있다고 볼 수 있다.

법령용어의 모든 쌍에 대하여 PMI를 구한 뒤에는 법령용어를 각각 노드로, PMI 지수를 두 법령용어 간의 관련된 정도로 보고 네트워크를 구성하였다. 그리고 중심성 점수 지표 중 하나인 페이지랭크(PageRank)를 계산하여 그 값을 단어별 중심성 점수로 부여한다. 페이지

랭크는 구글에서 신뢰성 있는 검색 결과를 제공하기 위해 사용하는 지표로서, 대표적인 중심성 점수이다[1]. 이 때, 에지의 가중치가 음수일 수는 없으므로 PMI 값의 범위를 양수로 만들어주기 위해 로그를 해제한 PMI 값을 적용하였다. 이렇게 계산된 중심성 점수는 요약서에서 추출된 법령용어 집합을 대상으로 계산하는 것이므로, 법령용어 중에서 중요한 순서를 의미하는 지표로 사용될 수 있다.

<Figure 4>은 입력 문서에서 데이터를 추출하는 단계에서부터 최종 법령용어별 가중치를 계산하는 단계까지의 과정 전체를 나타낸 것이다.

3.4.4 우선순위 점수

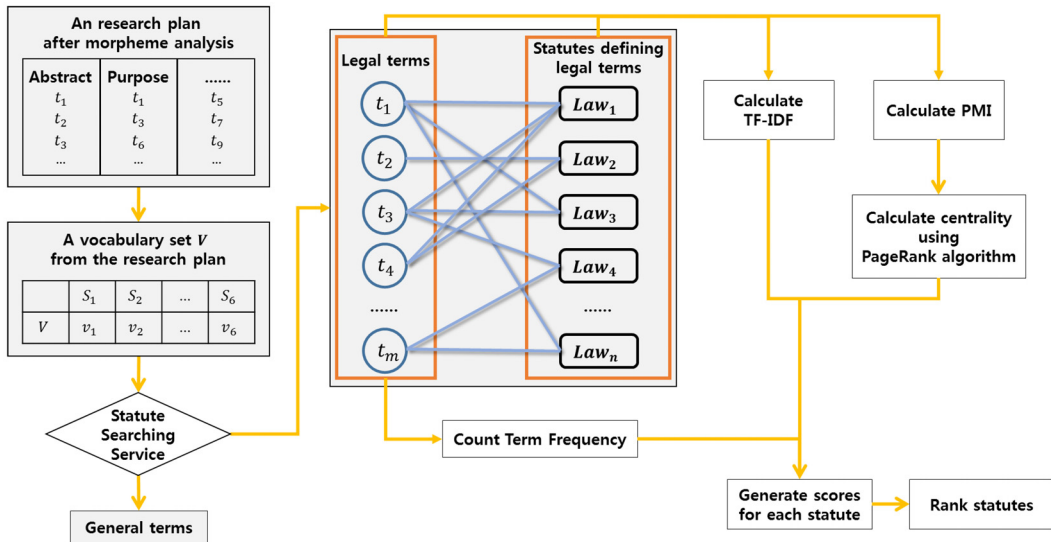
본 연구에서는 우선순위 결정을 위한 6개의 지표들을 제시하였으며, 점수 계산 방법론은 각각 3개씩 크게 2개의 계열로 나눌 수 있다. 첫 번째 계열은 법령용어의 TF-IDF 값을 중

심으로 계산한 ‘단순기여도’이다. 두 번째 계열은 TF-IDF 값에 중심성 점수를 가중해 계산하는 ‘중심성기여도’이다. 단순기여도는 ‘C’로 표기하고, 중심성기여도는 ‘CN’으로 표기한다. 각 방법에 의해 계산한 점수를 통해 법령의 우선순위를 결정한다.

첫 번째는 기여도 합(C_SUM)으로, 각 법령용어의 TF-IDF 값을 모두 더한다. 이 때 법령용어가 어느 섹션에 등장했는지는 고려하지 않고, 문서 전체에서의 등장 여부만을 기준으로 한다. 즉, 어떤 법령용어가 두 개의 섹션에서 각각 등장했어도 해당 법령용어의 빈도는 1로 간주한다.

$$RScore^{C_SUM}(j) = \sum_{i=1}^{|ST|} c(i, j, S) \quad (1)$$

두 번째는 섹션별 기여도 합(C_SUM_SC)으로 모든 섹션에 대해 법령을 정의하고 있는 법령용어의 TF-IDF 값의 합을 계산한다. 즉,



<Figure 4> Ranking Process

법령용어가 등장한 섹션의 수를 고려한다. 예를 들어 ‘정거장’을 정의하는 법으로 ‘전력기술관리법’이 있고, 정거장이 총 2개의 섹션에 포함되어 있다면, ‘전력기술관리법’의 점수는 정거장의 TF-IDF 값에 2를 곱한 값과 같다.

$$RScore^{C_SUM_SC}(j) = \sum_{k=1}^{|SC|} \sum_{i=1}^{|ST_k|} c(i, j, k, S) \quad (2)$$

세 번째는 섹션별 평균 기여도 합(C_SUM_AVG_SC)이다. 먼저 어떤 법령에 의해 정의되는 법령용어들의 TF-IDF 값을 섹션별로 더한 후, 각 섹션에 속한 법령용어의 총 개수로 나눈다. 그 다음, 섹션별로 산출된 값들을 모두 더한 값이 법령의 점수가 된다. 예를 들어, 도로교통법이 정의하는 단어인 ‘자동차’와 ‘고속도로’가 각각 ‘연구성과’와 ‘활용방안’에 포함되어 있었다면, 도로교통법의 점수는 [자동차의 TF-IDF 값을 연구성과 섹션의 법령용어 개수로 나눈 값]과 [고속도로의 TF-IDF 값을 활용방안 섹션의 법령용어 개수로 나눈 값]의 합과 같다.

$$RScore^{C_SUM_AVG_SC}(j) = \sum_{k=1}^{|SC|} \frac{\sum_{i=1}^{|ST_k|} c(i, j, k, S)}{|ST_k|} \quad (3)$$

다음에 제시할 3개 지표들은 위에서 제시된 기여도 계열 지표들에 중심성 점수를 가중하여 계산한 중심성기여도 계열의 지표들이다. 두 계열의 점수들은 계산 과정은 크게 다르지 않으며, 중심성 점수의 반영여부만 다르다. 수식 (4)~식 (6)은 순서대로 각각 중심성기여도 합(CN_SUM), 섹션별 중심성기여도 합(CN_SUM_SC), 섹션별 평균 중심성기여도 합(CN_SUM_AVG_SC)의 수식이다.

$$RScore^{CN_SUM}(j) = \sum_{i=1}^{|ST|} cn(i, j, S) \quad (4)$$

$$RScore^{CN_SUM_SC}(j) = \sum_{k=1}^{|SC|} \sum_{i=1}^{|SK_k|} cn(i, j, k, S) \quad (5)$$

$$RScore^{CN_SUM_AVG_SC}(j) = \sum_{k=1}^{|SC|} \frac{\sum_{i=1}^{|ST_k|} cn(i, j, k, S)}{|ST_k|} \quad (6)$$

이처럼 지표의 계산을 통해 요약서와 관련된 각 법령에 점수를 부여하고, 점수를 기반으로 우선순위를 제공함으로써 사용자가 어떤 법령을 더 먼저, 더 중요하게 검토하여야 하는지에 대한 정보를 제공할 수 있다.

4. 법령 검색시스템의 구현 및 평가

제3장에서 제시한 방법론을 적용하여, 연구개발계획서를 입력받아 관련 법령의 목록을 출력하는 법령검색시스템을 구현하였다. 프로그래밍 언어로는 자바를 사용하였다.

4.1 실험 개요

분석 대상으로 삼은 것은 교통체계효율화사업 하위 과제인 ‘신에너지 바이모달 저장굴절차량개발(Advanced Energy Bimodal KneelingCar Development)’과 철도기술연구사업 하위 과제인 ‘고속화물열차 및 여객/화물 복합열차 기술개발(High-Speed Passenger/Mixed Train Technical development)’에 대한 연구개발계획서이다. 최종적으로 ‘신에너지 바이모달 저장굴절차량개발’ 과제와 관련하여 78개 법령이, ‘고속화물열차 및 여객/화물 복합열차 기술개발’ 과제와 관련하여 73개 법령이 검색되었다. 검색

된 법령 목록에 6개의 지표를 각각 적용해 우선 순위를 결정하고 비교하였다.

4.2 가중치 기반 순위 모형의 성능 평가

검증 기준 지표는 한국철도기술연구원 연구원들이 검색된 법령 목록을 보고 직접 평가하였다. 해당 법령이 연구개발과 관련하여 참조하기에 적절하다고 판단될 경우 ‘적합’(relevant), 다소 적절하다고 판단될 경우 또는 무관하지는 않다고 판단할 경우 ‘보통’(normal), 해당 법령이 아예 무관하다고 판단될 경우 ‘부적합’(incongruity)의 3단계로 평가하였다. 첫 번째 입력 문서인 ‘신에너지 바이모달 저장굴절 차량개발(Advanced Energy Bimodal KneelingCar Development)’에서 검색된 총 법령 78개 중, 적합에 해당하는 법령은 22개(28%), 보통에 해당하는 법령은 8개(10%), 부적합에 해당하는 법령은 48개(61%)이며, 두 번째 입력문서인 ‘고속화물열차 및 여객/화물 복합열차 기술개발(High-Speed Passenger/Mixed Train Technical development)’에서 검색된 총 법령 73개 중, 적합에 해당하는 법령은 12개(16%), 보통에 해당하는 법령은 9개(12%), 부적합에 해당하는 법령은 52개(71%)이다.

성능을 검증하기 위해 ‘순위(Rank) 합’과 ‘평균 정확률-재현율(Average Precision-Recall)’ 총 2가지 지표들을 사용했다. 첫 번째 검증방법인 순위 합은 다시 두 가지로 세분화된다. 각각 ‘적합에 해당하는 법령의 순위들의 합’과, ‘적합에 해당하는 법령의 순위들 및 보통에 해당하는 법령의 순위에 0.5 가중치를 곱한 값들의 합’이다. 두 번째 검증방법인 평균 정확률-재현율(Average Precision-Recall)은 ‘신에너지

지 바이모달 저장굴절 차량개발’의 경우, 78개 법령에 대한 순위를 이용하여 78번의 정확률과 재현율을 계산하여 이를 평균한 값을 성능 지표로 사용하는 것이다. 정확률(Precision)은 계산하려는 해당 법령을 기준으로 앞 순위 법령 중 ‘적합’에 해당하는 법령의 개수를 계산하려는 해당 법령의 순위(Rank)로 나누어 계산하며, 재현율(Recall)은 계산하려는 해당 법령을 기준으로 앞 순위 법령 중 ‘적합’에 해당하는 법령의 개수를 ‘적합’에 해당하는 모든 법령의 개수로 나누어 계산한다.

순위 합에 대한 ‘신에너지 바이모달 저장굴절 차량개발’의 결과는 <Table 7>과 같으며, ‘고속화물열차 및 여객/화물 복합열차 기술개발’의 결과는 <Table 8>에 나타나 있다. 평가 결과는 순위들의 합이기 때문에 각 지표별 값이 작을수록 성능이 좋다는 것을 나타내며, 지표들 중에서 가장 좋은 결과는 다른 색으로 표시했다. ‘only relevant’ 지표와 ‘relevant+normal’ 지표 모두 중심점 점수를 사용한 방법들이 상위권에 있음을 알 수 있다.

<Table 7> Performance of Rank Sum
(Title: Advanced Energy Bimodal KneelingCar Development)

Score (measure)	Rank Sum	
	only relevant	relevant + normal
1. CN_SUM	398	542.5
2. CN_SUM_SC	413	545.0
3. CN_SUM_AVG_SC	415	540.5
4. C_SUM	575	765.5
5. C_SUM_SC	552	706.0
6. C_SUM_AVG_SC	546	684.5

〈Table 8〉 Performance of Rank Sum
(Title: High-Speed
Passenger/Mixed Train
Technical Development)

Score (measure)	Rank Sum	
	only relevant	relevant + normal
1. CN_SUM	216	345.0
2. CN_SUM_SC	200	331.0
3. CN_SUM_AVG_SC	212	339.5
4. C_SUM	270	439.0
5. C_SUM_SC	230	397.5
6. C_SUM_AVG_SC	253	414.5

평균 정확률-재현율은 <Table 9>와 <Table

10>에 나타나 있다. 역시 중심성 점수를 사용하는 ‘CN’ 계열 지표의 성능이 ‘C’ 계열 지표의 성능보다 상대적으로 높음을 알 수 있다.

<Figure 5>와 <Figure 6>은 각각 요약서와 관련된 법령들의 정확률과 재현율의 변화를 그린 차트이다. <Figure 5>에서 ‘CN_SUM’에 비해, ‘C_SUM_AVG_SC’와 ‘C_SUM_SC’가 상대적으로 성능이 떨어진다는 것을 알 수 있다. 또한 <Figure 6>을 보면, ‘CN_SUM_SC’ 방법이 상위권에 있으며 ‘C_SUM’ 방법이 상대적으로 하위권에 있음을 알 수 있다.

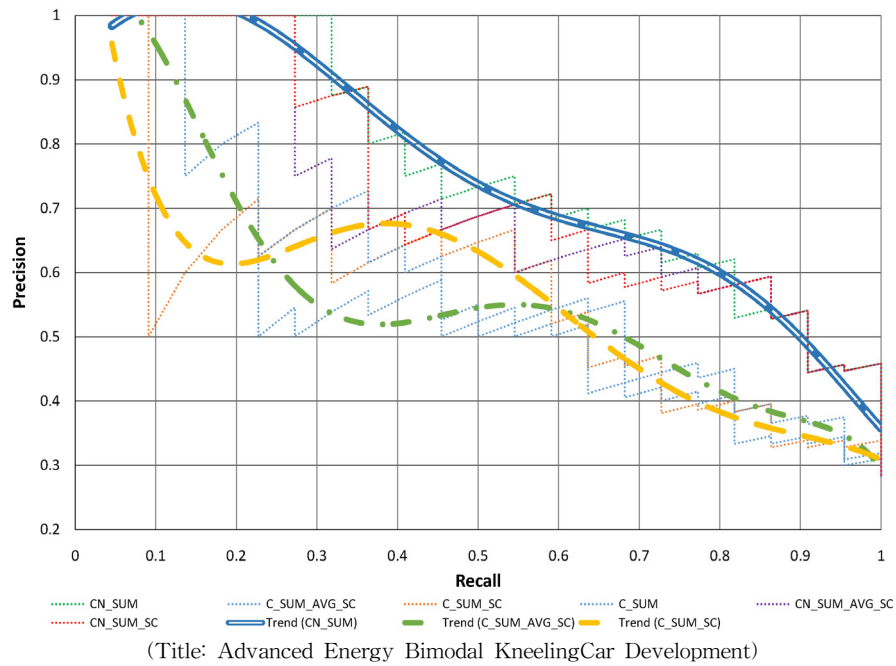
위의 결과들을 통해, 중심성 점수를 반영한 중심성기여도 계열 지표들의 성능이 TF-IDF

〈Table 9〉 Performance of Average Precision-Recall
(Title: Advanced Energy Bimodal KneelingCarDevelopment)

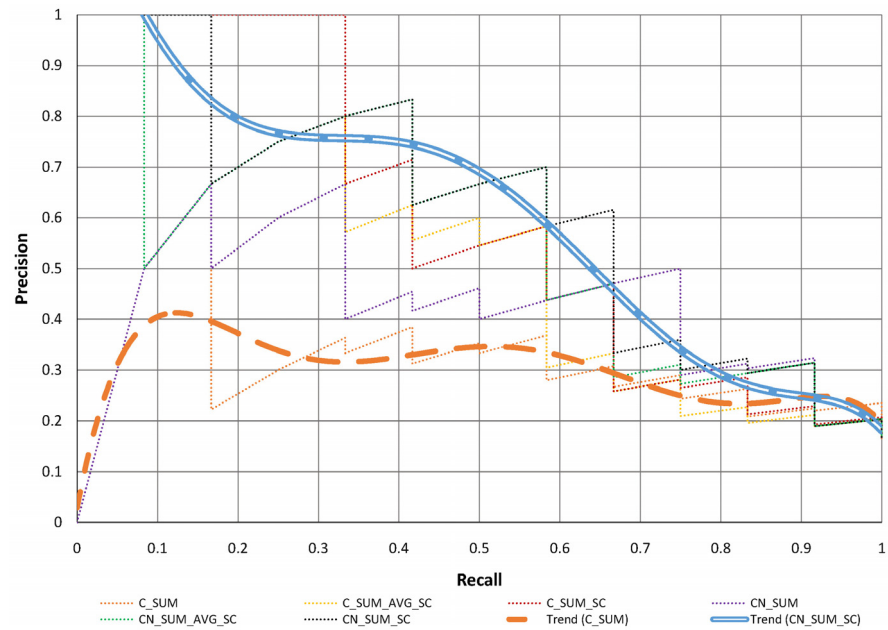
Score (measure)	Total average		Top 10 average		Top 20 average	
	precision	recall	precision	recall	precision	recall
1. CN_SUM	0.550	0.780	0.956	0.231	0.845	0.372
2. CN_SUM_SC	0.538	0.772	0.942	0.227	0.813	0.354
3. CN_SUM_AVG_SC	0.533	0.770	0.908	0.213	0.787	0.340
4. C_SUM	0.465	0.677	0.808	0.186	0.709	0.304
5. C_SUM_SC	0.464	0.691	0.713	0.168	0.672	0.306
6. C_SUM_AVG_SC	0.466	0.694	0.777	0.172	0.659	0.277

〈Table 10〉 Performance of Average Precision-Recall
(Title: High-Speed Passenger/Mixed Train Technical Development)

Score (measure)	Total average		Top 10 average		Top 20 average	
	precision	recall	precision	recall	precision	recall
1. CN_SUM	0.31	0.767	0.484	0.233	0.467	0.408
2. CN_SUM_SC	0.357	0.785	0.775	0.333	0.646	0.491
3. CN_SUM_AVG_SC	0.343	0.771	0.725	0.325	0.607	0.470
4. C_SUM	0.264	0.705	0.345	0.150	0.347	0.300
5. C_SUM_SC	0.344	0.751	0.786	0.316	0.633	0.462
6. C_SUM_AVG_SC	0.343	0.739	0.723	0.300	0.591	0.437



〈Figure 5〉 Precision-Recall Chart



〈Figure 6〉 Precision-Recall Chart

만 사용한 단순기여도 계열 지표들보다 우수하다는 것을 확인할 수 있다. 따라서 단어 중요도를 계산할 때 중심성 점수의 영향력이 크다는 것과, 중심성기여도 계열 지표들을 통해 사용자에게 더 적절한 우선순위를 제공할 수 있음을 알 수 있다.

5. 결 론

본 연구에서는 철도교통분야 연구개발사업에 영향을 미칠 수 있는 법령들을 찾아 관련성 순으로 사용자에게 제공할 수 있는 법령 검색 시스템을 제안하였다. 철도교통 분야의 연구개발계획서로부터 정보를 추출하여 연구개발사업에 영향을 미칠 수 있는 법령들의 집합을 찾고, 검색시스템의 기반이자 지식 베이스로서의 역할을 수행하는 온톨로지를 구축, 활용하도록 하였다. 또한 사용자에게 제공하는 법령들의 순위를 매기는 데 사용하는 가중치 지표를 개발하고, 그 중 가장 성능이 좋은 지표를 탐색함으로써 우선적으로 검토해야 할 법령을 상위에 제시하도록 하였다. 시스템의 결과로 도출된 법령 리스트는 철도교통분야 연구자들이 검토하는 자료로서 활용되어 철도교통 분야에서 이루어지는 연구개발사업의 성공과 사업결과의 실용화 가능성을 높이고, 투자된 예산에 대해 충분한 효과를 얻을 수 있도록 도움을 줄 것이다.

온톨로지 기술은 표준화되어있기 때문에, 본 연구에서 온톨로지 구축에 사용된 어휘 및 의미 구조는 다른 유사 온톨로지를 만들 때 변형하여 쉽게 적용할 수 있다. 즉, 법령뿐만 아니라 정책 제도 등 유사 정보에 대해서도 데이터 교환과

변환을 위한 중립적 모델로 활용되거나 확장될 수 있다는 점에서 의미가 있다고 하겠다.

다만, 본 연구에서는 철도교통 분야의 연구개발사업과 상대적으로 관련성이 높다고 판단되는 국토, 교통, 물류, 철도의 4가지 분야로 한정된 법령들만을 대상으로 검색시스템을 구현하였다. 다른 분야의 법령이나 정책·제도 역시 철도교통 분야 연구개발사업에 직간접적으로 영향을 미칠 수 있으므로, 향후 추가적인 연구를 통해 모든 법령과 정책·제도 전반에 대한 데이터를 포함하고 있는 온톨로지 및 검색 시스템을 구축할 필요성이 있다.

References

- [1] Brin, S. and Page, L., "The anatomy of a large-scale hypertextual Web search engine," Computer Networks and ISDN Systems, Vol. 30, pp. 107-117, 1998.
- [2] Church, K. W. and Hanks, P., "Word association norms, mutual information, and lexicography," Comput. Linguist, Vol. 16, No. 1, pp. 22-29, 1990.
- [3] Jo, D. W., Seo, M. J., and Kim, M. Ho., "A Study on Legal Information Retrieval Engine based on Ontology," Korea Information Science Society, pp. 1568-1570, 2015.
- [4] Kim, H. L., Kim, H. G., "Personal Electronic Document Retrieval System Using Semantic Web/Ontology Technologies," The Journal of Society for e-Business Studies,

- Vol. 12, No. 1, pp. 135-149, 2007.
- [5] Kim, J. H., Lee, J. S., Lee, M. J., Kim, W. J., and Hong, J. S., "Term Mapping Methodology between Everyday Words and Legal Terms for Law Information Search System," *Bibliographic Info: J Intell Inform Syst*, Vol. 18, No. 3, pp. 137-152, 2012.
- [6] Lame, G., "Using NLP Techniques to Identify Legal Ontology Components : Concepts and Relations," *Artificial Intelligence and Law*, Vol. 12, No. 4, pp. 169-184, 2005.
- [7] Mitschick, A., "Ontology-based Indexing and Contextualization of Multimedia Documents for Personal Information Management Applications," *International Journal on Advances in Software*, Vol. 3, No. 1, 2 pp. 31-40, 2010.
- [8] Mun, H. J., Lee, S. J., Kim, Y. J., and Woo, Y. T., "A Personalized Concept-based Retrieval Technique Using Domain Ontology," *The Journal of Society for e-Business Studies*, Vol. 12, No. 3, pp. 269-282, 2007.
- [9] Rajaraman, A. and Ullman, J. D., "Data Mining," pp. 1-17. doi:10.1017/CBO9781139058452.002. ISBN 9781139058452, 2011.
- [10] Toledo, C. M., M. A. Ale, Chiotti, O., and Galli, M. R., "An Ontology-driven Document Retrieval Strategy for Organizational Knowledge Management Systems," *Electronic Notes in Theoretical Computer Science*, Vol. 281, pp. 21-34, 2011.
- [11] Yoo, D. H. and Suh, Y. M., "An Ontology-based Hotel Search System Using Semantic Web Technologies," *Society for e-Business Studies, The Journal of Society for e-Business Studies*, Vol. 13, No. 4, pp. 71-92, 2008.

저 자 소 개



원민재
2014년
현재
관심분야

(E-mail: wmjzzang@naver.com)
삼육대학교 컴퓨터공학 (학사)
연세대학교 정보산업공학과 (석사 과정)
데이터 마이닝, 그래프 마이닝 등



김동희
1994년
2000년
2000년~현재
관심분야

(E-mail: kdh777@krri.re.kr)
인하대학교 산업공학 (석사)
인하대학교 산업공학 (박사)
한국철도기술연구원 책임연구원
철도운영시스템, 운영최적화 및 시뮬레이션,
정책기술융합플랫폼 등



정해민
2013년
현재
관심분야

(E-mail: mainartery37@gmail.com)
홍익대학교 정보산업공학 (학사)
연세대학교 정보산업공학과 (석·박사 통합과정)
데이터 마이닝, 그래프 마이닝 등



이상근
2005년
2007년
2010~2011년
2012년~현재
관심분야

(E-mail: sangkeun@krri.re.kr)
오이타대학교 건설공학 전공 (석사)
오이타대학교 환경공학 전공 (박사)
대전발전연구원 초빙연구원
한국철도기술연구원 박사후연수연구원
도시계획, 공간분석, 국토-교통정책 등



홍준석

1991년

1997년

1999~2002년

2003년~현재

관심분야

(E-mail: junehong@kyonggi.ac.kr)

KAIST 경영과학 (석사)

KAIST 테크노경영대학원 경영공학 (박사)

인제대학교 경영학부 교수

경기대학교 경영정보학과 교수

빅데이터 분석 및 플랫폼, 지능형 에이전트 등



김우주

1994년

1996~2003년

2004년~현재

관심분야

(E-mail: wkim@yonsei.ac.kr)

KAIST에서 경영과학과 (박사)

전북대학교 산업정보시스템공학과 교수

연세대학교 정보산업공학과 교수

시맨틱 웹, 지식 관리 및 인공지능 웹 서비스 등