텍스트 문서 기반
연관 법령 검색 방법에 관한 연구

연세대학교 대학원
정보산업공학과
이 유 나

텍스트 문서 기반
연관 법령 검색 방법에 관한 연구

지도 김 우 주 교수

이 논문을 석사 학위논문으로 제출함

2016년 12월 25일

연세대학교 대학원
정보산업공학과
이 유 나

연세대학교
YONSEI UNIVERSITY

이유나의 석사 학위논문으로 인준함

심사위원 김 우 주 인

심사위원 정 병 도 인

심사위원 박 상 언 인

연세대학교 대학원

2016년 12월 25일

감사의 글

연구실 첫 출근 때 어색해하며 자기소개 하던 날이 엊그제 같은데 석사 생활을 마무리하려니 2년이라는 시간이 너무 빨리 흘러간 것 같습니다. 우선 스마트 시스템 연구실의 기둥이신 김우주 교수님! 교수님의 아낌없는 가르침 덕분에 무사히 석사 생활을 마무리할 수 있었고, 알찬 2년을 보낼 수 있었던 것 같아 너무 감사드립니다. 저의 졸업 연구에 날카로운 지적과 많은 조언을 해주신 박상언 교수님과 정병도 교수님께도 감사의 예를 표합니다.

2년 동안 가족보다도 더 많은 시간을 보냈고, 함께한 시간만큼 잊지 못할 추억도 정도 많이 쌓인 연구실 식구들에게도 감사의 말을 전하고 싶습니다. 춤추는 모습이 인상적이고 똑소리나게 일도 잘하는 라이언 닮은 상헌오빠, 말은 모질게 하지만 어려움이 있을 때마다 자신의 일처럼 도와주는 츤데레 희원오빠, 막히는 일이 있을 때면 먼저 도움의 손길을 내밀어 주는 착한 해민오빠, 프로젝트를 함께 진행하면서 저의 열정을 일깨워 주신 광일오빠, 연구실의 맏언니로써 동생들의 말에 귀기울여주고 굳은 일도 도맡아하는 오키나와 추억을 함께한 지현언니, 연구실의 첫 발을 같이 들여서 동기처럼 의지했던 샐러드 파트너 도경이, 연구실의 어느 누구보다도 코딩을 잘하고 애용하는 쇼핑몰이 같아서 놀랬던 승희, 1학기 라는게 믿기지 않을 정도로 프로젝트에서 맡은 일을 능숙하게 처리하는 동그리 안경이 잘 어울리는 혜수, 저의 경상도 사투리 본능을 일깨워주는 ZARA 츄리닝 바지가 매력적인 영택오빠, 곧 입학 동기로 만나 동고동락 할 민태오빠. 덕분에 즐거운 연구실 생활을 할 수 있었고, 많은 것을 깨닫고 배울 수 있는 시간이었습니다.

그리고 지금은 졸업했지만 저의 석사 1·2학기를 함께 보낸 석재오빠, 민재

오빠, 홍매언니, 향단언니, 이연언니에게도 고마운 마음을 전하고 싶습니다. 또 다른 연구실이지만 같은 학기라 졸업 논문과 관련한 희로애락을 나눌 수 있어서 많이 의지가 되었던 성범오빠에게도 고맙다는 말을 전하고 싶습니다.

마지막으로 제가 제일 존경하고 사랑하는 부모님! 항상 저를 믿고 묵묵히 응원해주셔서 지금의 제가 있는 것 같습니다. 너무 고맙고 사랑합니다. 또 이제 곧 대학생이 되는 동생 원석이와 사랑하는 할머니께도 고마움과 사랑을 전합니다. 그밖에 세상에 둘도 없는 친구들에게도 고맙다는 말을 전하며 졸업의 기쁨을 나누고 싶습니다.

저를 지켜봐주시는 분들의 마음에 보답할 수 있도록 하루하루 성장하도록 노력하겠습니다. 진심으로 감사드립니다.

2016년 12월

이유나 올림

차 례

< 그림 차례 >

< 표 차례 >

국문요약

텍스트 문서 기반 연관 법령 검색 방법에 관한 연구

연세대학교 일반대학원
정보산업공학 전공
이유나

일반적으로 R&D 과제는 기획부터 상업화까지 수년이 소요되는데, 그 과정에서 연관된 법령이 개정되거나 새로 제정될 수 있다. 이러한 법령의 변경은 R&D 과제에 영향을 미쳐 최악의 경우 투자 실패로 이어지는 결과를 가져오게 된다. 또한 R&D 과제와 관련된 정책 및 법령을 검색할 수 있는 국내 연구 개발 실적이 전무한 형편이기 때문에, 법령 분야의 전문가가 아닌 연구자들은 어떤 법령이 자신의 프로젝트와 연관되어 있는지 알기가 어려운 실정이다.

따라서 본 연구는 R&D 과제와 관련된 정책 및 법령을 실시간으로 검색할 수 있는 방법론을 제안하였다. R&D 과제를 위해 작성된 연구계획서를 입력 받아 중심성 점수(Centrality Score)를 사용하여 연구계획서 벡터를 생성하고, 4000여개의 법령 TF-IDF 벡터와 코사인 유사도(Cosine Similarity)를 비교하여 높은 유사도 순으로 법령을 보여주게 된다. 이 때, 중심성 점수(Centrality Score)는 현행 법령들의 본문으로 학습된 Word2Vec 모델의 단어간 similarity를 사용한 Weighted-PageRank 값이다. 연관 법령 검색 방법으로 총 5가지 방법론을 제시하였다. 한국철도기술연구원에서 제공한 3개의 연구계

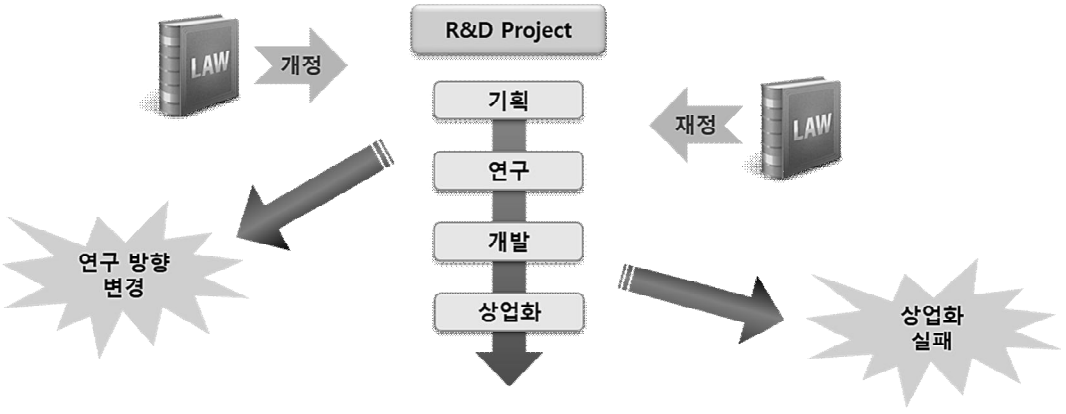
획서로 실험을 진행하였으며, 한국철도기술연구원의 담당자들이 제공한 관련 법령 목록을 평가지표로 삼았다.

최종적으로 본 연구는 R&D 과제 연구자에게 R&D 사업과 관련된 정책 및 법령을 실시간으로 제공함으로써 기획단계에서 연구 방향을 결정하는데 참고 자료로 사용될 수 있을 것이다. 또한 연구 기간 동안 정책 및 법령이 변경되는 것을 실시간으로 알 수 있기 때문에 변경된 정책 및 법령에 의해 기술이 실용화·사업화되지 못하는 경우를 빠르게 대비함으로써 큰 경제적 손실을 막을 수 있을 것이다. 궁극적으로 R&D 과제의 질을 향상시킬 수 있을 것이다.

Keyword: 단어 벡터 모델, 중심성 점수, TF-IDF, 코사인 유사도, Word2Vec, Weighted-PageRank, 온톨로지

1. 서론

R&D 기술은 그 특성상 많은 인력과 비용이 투입되어야 하며 짧은 시간 안에 개발될 수 없기 때문에 다년간에 걸쳐 연구를 진행하게 된다. 따라서 R&D 과제의 긴 연구 기간 동안 법령이나 법정계획이 변경되어 기술이 실용화·사업화되지 못하게 되면, 큰 경제적 손실이 발생할 수밖에 없다. 기술을 사용할 수 있도록 법령을 다시 변경하거나, 기술을 새로운 정책제도에 맞추어 수정하기 위해서 불필요한 추가 비용이 발생할 수 있으며, 심지어는 추가되는 비용에 상관없이 실용화될 수 없는 경우가 발생할 수도 있다. 실제로 2013년 국정감사에서 곡선 선로 전용 열차인 틸팅 열차를 개발하기 위해 투자한 860억이 공중분해 됐다는 지적을 받았다. 틸팅 열차는 중앙선 등 곡선선로로 인해 고속운행이 불가능한 지역에 투입하는 곡선 선로용 R&D 신기술 열차로 10년간 860억을 들여 개발했다. 하지만 2011년 '제2차 국가철도망 구축계획'에서 곡선 선로가 직선화 작업으로 변경되면서 기술이 무용지물이 되었다.



[그림 1-1] R&D 과제와 법령 간의 관계

그러므로 R&D 과제에 참여하는 연구자들은 과제를 수행하는데 필요한 법령들에 대해서 잘 이해하고 있어야 한다. 하지만 현재는 R&D 과제와 관련된 정책 및 제도를 검색할 수 있는 국내 연구 개발 실적이 전무한 형편이기 때문에, 법령 분야의 전문가가 아닌 연구자들은 어떤 법령이 자신의 프로젝트와 연관되어 있는지 알기가 어려운 실정이다. 따라서 R&D 과제와 관련된 정책 및 법령을 실시간으로 적용할 수 있는 검색 방법론의 필요성이 대두되었고, 본 연구는 R&D 연구계획서에 기반한 연관 법령 검색 방법론을 제안한다.

본 논문은 다음과 같이 구성한다. 2장에서는 관련 연구들에 대해 소개한다. 3장에서는 본 연구에서 제시하는 연관 법령 검색 프로세스에 대한 개요와 연관 법령 검색 방법론을 제시한다. 4장에서는 한국철도기술연구원에서 제공한 3개의 연구계획서를 기반으로 분석한 결과와 평가이며, 마지막으로 5장은 본 연구의 결론과 향후 연구 방향에 대해서 기술한다.

2. 관련 연구

2.1 Term Vector Model

Term Vector Model은 텍스트 문서를 식별자(index term)들의 벡터로 나타내는 대수적인 모델이다. 이 때, 단어 가중치(term weight)값을 계산하는 방법은 여러 가지가 있으며, 본 연구에서 사용한 방법은 아래와 같다.

2.1.1 TF-IDF

TF-IDF(Term Frequency - Inverse Document Frequency)는 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다.

TF(term frequency)는 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값으로, 이 값이 높을수록 문서에서 중요하다고 생각할 수 있다. 하지만 단어 자체가 문서군 내에서 자주 사용되는 경우, 이것은 그 단어가 흔하게 등장한다는 것을 의미한다. 이것을 DF(document frequency)라고 하며, 이 값의 역수를 IDF(inverse document frequency)라고 한다.

TF-IDF는 TF와 IDF를 곱한 값이다. 특정 문서 내에서 단어 빈도가 높을수록, 그리고 전체 문서들 중 그 단어를 포함한 문서가 적을수록 TF-IDF값이 높아지며, 수식은 아래와 같다.

$$TF-IDF(t,d,D) = tf(t,d) * \log\frac{|D|}{1+|d\in D: t\in d|}$$

- $tf(t,d)$: 문서d에서 단어t의 빈도
- $|D|$: 문서집합D의 총 문서 수
- $|d\in D: t\in d|$: 단어t가 포함된 문서의 수

2.1.2 Centrality Score

Centrality Score는 네트워크에서 각 개체 혹은 노드의 상대적 중요도를 나타내는 척도이다. 즉, 문서 내에서 단어들이 갖는 중요도를 표현할 때 사용한다. 이 Centrality Score는 지수로 계산되는데, 본 연구에서는 단어 간의 Word2Vec similarity 지수를 단어 네트워크로 설정하여 Weighted-PageRank 알고리즘을 적용한 결과 값을 사용한다.

2.1.2.1 Weighted-PageRank

기존 Weighted-PageRank 알고리즘은 웹 문서들의 참조관계에 기반하여 문서들의 중요도를 산출하는 방식이며, 수식은 아래와 같다.

$$WP(V_i) = (1-d) + d * \sum_{V_j\in in(V_i)} \frac{W_{ji}}{\sum_{V_k\in out(V_j)} W_{jk}} WP(V_j)$$

- $WP(V_i)$: V_i의 $Weighted-Pagerank$값
- d : 이탈 확률($damping\,factor$)로 0과 1 사이의 실수 값이며, 보통 0.85로 설정

- $in(V_i)$: V_i로 들어가는 $edge$를 가진 노드 집합
- W_{ji} : V_j에서 V_i로 가는 $edge$의 가중치
- $out(V_j)$: V_j에서 나가는 $edge$를 가진 노드 집합
- W_{jk} : V_j에서 V_k로 가는 $edge$의 가중치

본 연구에서는 Weighted-PageRank 값을 구할 때 웹 문서들의 참조관계에 기반하지 않고, 현행 법령들로 학습된 Word2Vec 모델에서 단어들 간의 Word2Vec similarity 지수를 단어 네트워크로 사용한다. 변형된 수식은 아래와 같다.

$$WP(V_i) = (1-d) + d * \sum_{V_j \in in(V_i)} \frac{Word2Vec(j,i)}{\sum_{V_k \in out(V_j)} Word2Vec(j,k)} WP(V_j)$$

- $Word2Vec(i,j)$: 단어 i와 단어 j의 $similarity$ 값

현행 법령들로 학습된 Word2Vec 모델을 사용하여 단어 네트워크를 생성하면 단어의 의미가 포함된 semantic한 단어 네트워크 생성이 가능해진다. 또한 법령 내에서 단어들 간의 similarity를 구할 수 있기 때문에 보다 법령을 잘 설명하는 네트워크 생성이 가능해진다.

2.1.2.2 Word2Vec similarity

Weighted-PageRank 값을 구하기 위한 입력 값인 Word2Vec similarity 지수란 4000여개 현행 법령 본문으로 학습된 Word2Vec 모델에서 단어 벡터들 간의 Cosine Similarity를 계산한 값이다. 수식은 아래와 같다.

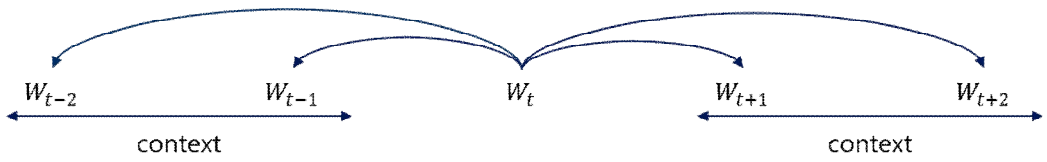
$$similarity(u,v) = \frac{W_u * W_v}{|W_u| * |W_v|}$$

· W_i : 단어 i의 n차원 벡터

본 연구에서 Word2Vec similarity 지수는 Centrality Score를 구해야 하는 모든 단어에서 2개를 뽑는 조합의 수만큼 구해지게 된다.

2.1.2.3 Word2Vec

Word2Vec이란 신경망(Neural Network)을 기반으로 단어를 N차원의 벡터로 학습하는 방법이며, 비슷한 문맥에 자주 출현하게 되는 단어들의 벡터 값이 유사하도록 학습한다. 따라서 단어의 벡터가 단순히 수치적인 의미가 아닌 N차원 공간에서 단어가 가지고 있는 의미적인 위치를 나타낸다. 모델 학습 방법은 [그림 2-1]과 같으며, 주어진 문서에서 단어 W_t가 $context(c)$에 결합할 확률을 최대화해서 학습한다.



$$\arg\max_{\theta} \prod_{(w,c)\in D} p(c|w;\theta)$$

[그림 2-1] Word2Vec 모델 학습 방법

2.2 Cosine Similarity

Cosine Similarity는 내적공간의 두 벡터 간 각도의 cosine값을 이용하여 벡터간의 유사도를 측정할 때 이용된다. 이 값은 벡터의 크기가 아닌 방향의 유사도를 판단하는 목적으로 사용되며, 두 벡터의 방향이 완전히 같을 경우 1, 서로 독립적인 경우 0, 완전히 반대 방향인 경우 -1의 값을 갖게 되며, 수식은 아래와 같다.

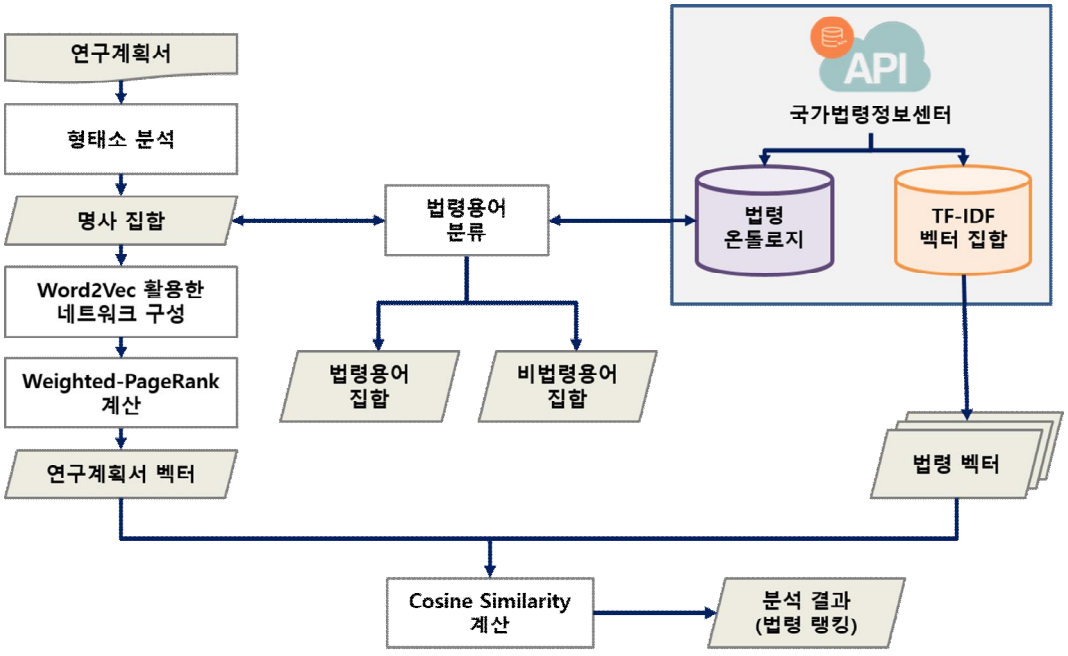
$$\cos(\theta) = \frac{A * B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

· A_i, B_i : 벡터 A와 B의 $components$

본 논문에서는 Term Vector Model들 간의 유사도를 비교하여 우선순위 지표를 계산할 때 사용한다.

3. 연구 방법론

본 연구의 전체적인 프로세스는 [그림 3-1]과 같다. 연구계획서가 입력되면 형태소 분석을 거쳐 명사들만을 획득한다. 명사 집합은 법령용어 또는 법령용어가 아닌 일반용어로 분류되는데, 그 기준은 국가법령정보센터의 정보를 기준으로 구축된 법령 온톨로지에 들어있는지 아닌지로 판단한다. 한편 연구계획서 내의 명사 집합은 미리 4000여개 현행 법령 본문으로 학습해둔 Word2Vec 모델에서 단어 간 similarity를 활용해 단어 네트워크를 구성하게 되고, Weighted-PageRank 알고리즘을 통해 계산된 값을 이용해 벡터로 표현한다. 최종적으로 이 벡터는 법령들의 TF-IDF 벡터와 Cosine Similarity를 계산하여 최종적인 분석 결과를 나타내게 된다.



[그림 3-1] 연관 법령 검색 프로세스

3.1 문서의 벡터 표현 방법

본 절에서는 연구계획서 및 법령을 벡터로 표현하는 방법을 설명한다.

3.1.1 연구계획서 벡터 표현

Centrality Score는 Weighted-PageRank 알고리즘을 사용하며, 어떤 노드가 네트워크에서 얼마나 중요한 위치를 나타내고 있는가를 의미한다. 즉, 문서 내에서 상대적으로 중요한 단어를 찾기 위해 사용한다. 본 연구는 '문서에서 추출된 모든 단어들은 문서의 의미를 함축하고 있는 정도가 서로 다르다'는 판단을 전제하고, 연구계획서 벡터를 연구계획서에 등장하는 단어들의 Centrality Score로 표현한다.

연구계획서는 많은 데이터를 담고 있지만 본 연구에서는 연구계획서의 핵심을 담고 있는 7개의 섹션의 데이터만 사용한다. 7개의 섹션은 연구 개발 개요, 최종목표, 연구내용 및 범위, 연도별 주요내용, 연구 성과, 활용방안, 핵심어이다. 섹션들은 문장으로 구성되어 있으며, 이 문장들을 KLT 형태소 분석기를 사용하여 명사만 추출하고, 불용어 사전을 적용하여 너무 의미 없는 명사를 필터링한다.

불용어 사전은 법령 본문에 등장하는 모든 단어의 TF(Term Frequency)를 계산해보고, TF의 결과가 7500개 이상이거나 상대적으로 무의미하다고 판단되는 단어로 구성된다. 총 226개 용어가 불용어로 등록되었으며, 대표적 예시로는 '대통령령', '장관', '부령' 등과 같이 법 원문의 내용과는 관계가 없는 단어들이 있다. 아래 [표 3-1]은 불용어 예시이다.

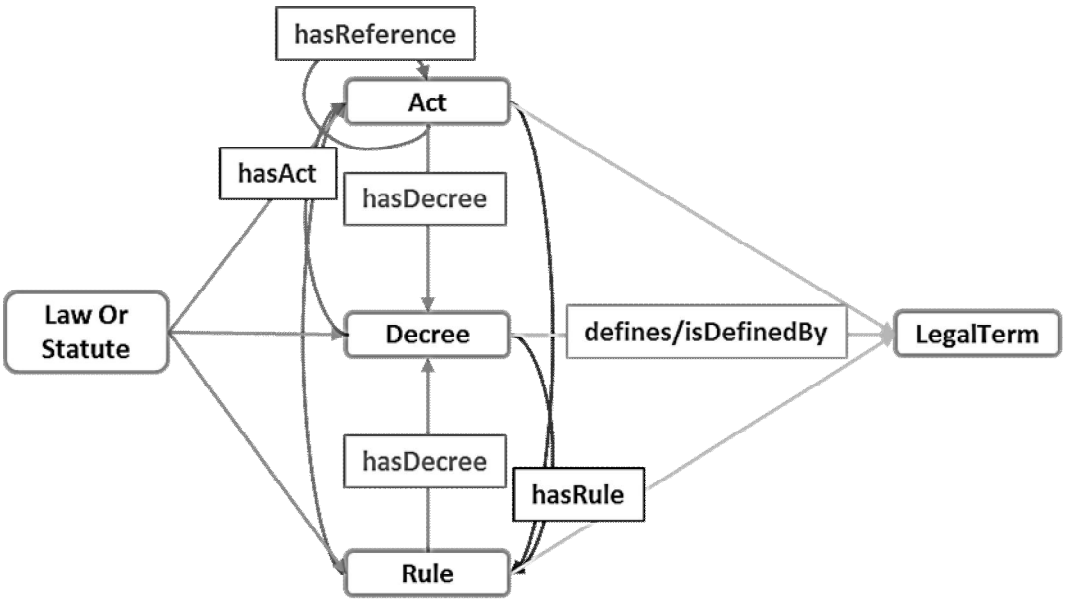
[표 3-1] 불용어 예시

단어	TF
경우	205,665
개정	112,813
사항	103,351
해당	93,958
다음	85,536
필요	71,315
호의	63,999
규정	57,269
대통령령	45,217
제출	35,426
사람	32,870
환경부장관	5,492
농림축산식품부장관	4,505
국토교통부령	3,133
기획재정부령	2,531

다음으로 필터링된 명사 집합을 법령용어 또는 법령용어가 아닌 일반용어로 분류하는데, 그 기준은 국가법령정보센터의 정보를 기준으로 구축된 법령 온톨로지에 들어있는지 아닌지로 판단한다.

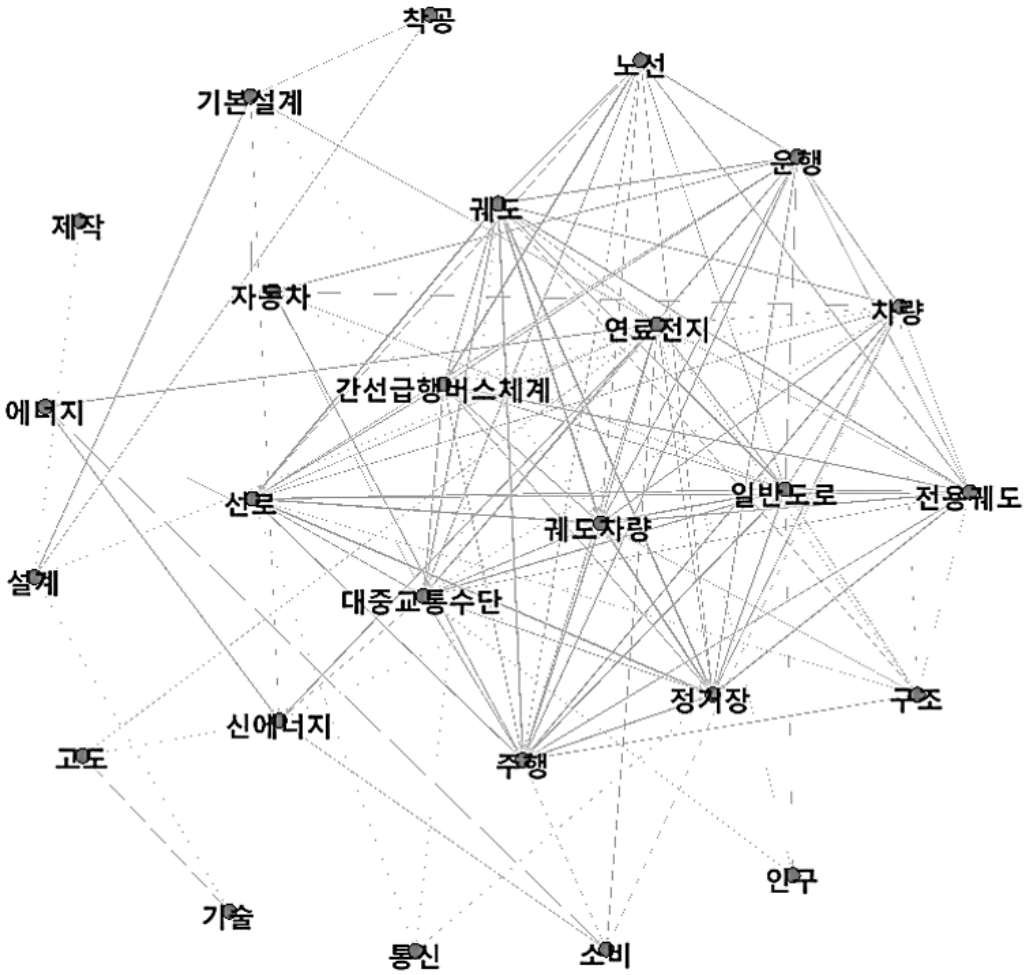
법령 온톨로지는 국가법령정보 공동활용 사이트(http://open.law.go.kr/)에서 제공하는 법령용어 API와 법령 API를 활용하여 법령용어 사전 전체와 법령 사전 전체를 실시간으로 수집하여 만든 온톨로지 형태의 지식DB이다. 구축된 온톨로지 스키마는 [그림3-2]와 같다. 둥근 사각형은 클래스(Class)고, 각 사

각형을 잇는 화살표는 오브젝트 프로퍼티(Object Property)이다. 「Act」는 법률, 「Decree」는 대통령령, 「Rule」은 부처훈령, 「LegalTerm」은 법령용어 데이터를 담고 있는 클래스이다. 온톨로지에서 필요한 데이터를 검색하기 위해서는 SQL 형태의 SPARQL을 사용한다. SPARQL은 W3C의 표준으로서 RDF로 표현된 데이터를 찾기 위해 사용되는 쿼리(Query)이다. 본 연구에서도 온톨로지에 저장된 데이터를 조회하기 위해 SPARQL 쿼리를 사용한다.



[그림 3-2] 온톨로지 스키마

다음으로 추출된 모든 명사로 단어 네트워크를 생성한다. 노드는 단어이며, 에지는 현행 법령들의 본문으로 학습된 Word2Vec 모델에서 단어들의 similarity이다. 이 때, ‘similarity가 0.6 이상이면 두 단어는 관계가 있다.’는 이론에 따라서 threshold를 0.6으로 설정한다. 생성된 단어 네트워크 예시는 [그림 3-3]과 같다.



[그림 3-3] 단어 네트워크

Word2Vec 모델의 학습 데이터는 법령 온톨로지에 들어 있는 현행 법령 4000여개 본문이다. 학습 방법은 Skip-gram을 사용한다. Skip-gram이란 한 단어를 입력 값으로, 설정한 Window size에 의해 결정되는 Context와 결합 확률이 최대가 되도록 단어의 벡터를 N차원으로 학습시키는 방법이다. 서로 관계가 있

는 단어는 비슷한 문맥에서 출현하게 되고, 학습 후 벡터의 위치가 비슷해진다는 판단을 전제한다. 학습시킨 단어의 벡터를 사용하면 [그림 3-4]와 같이 단어들의 similarity를 구할 수 있다.

```
wv_model_general.similarity('노선', '운행')
```

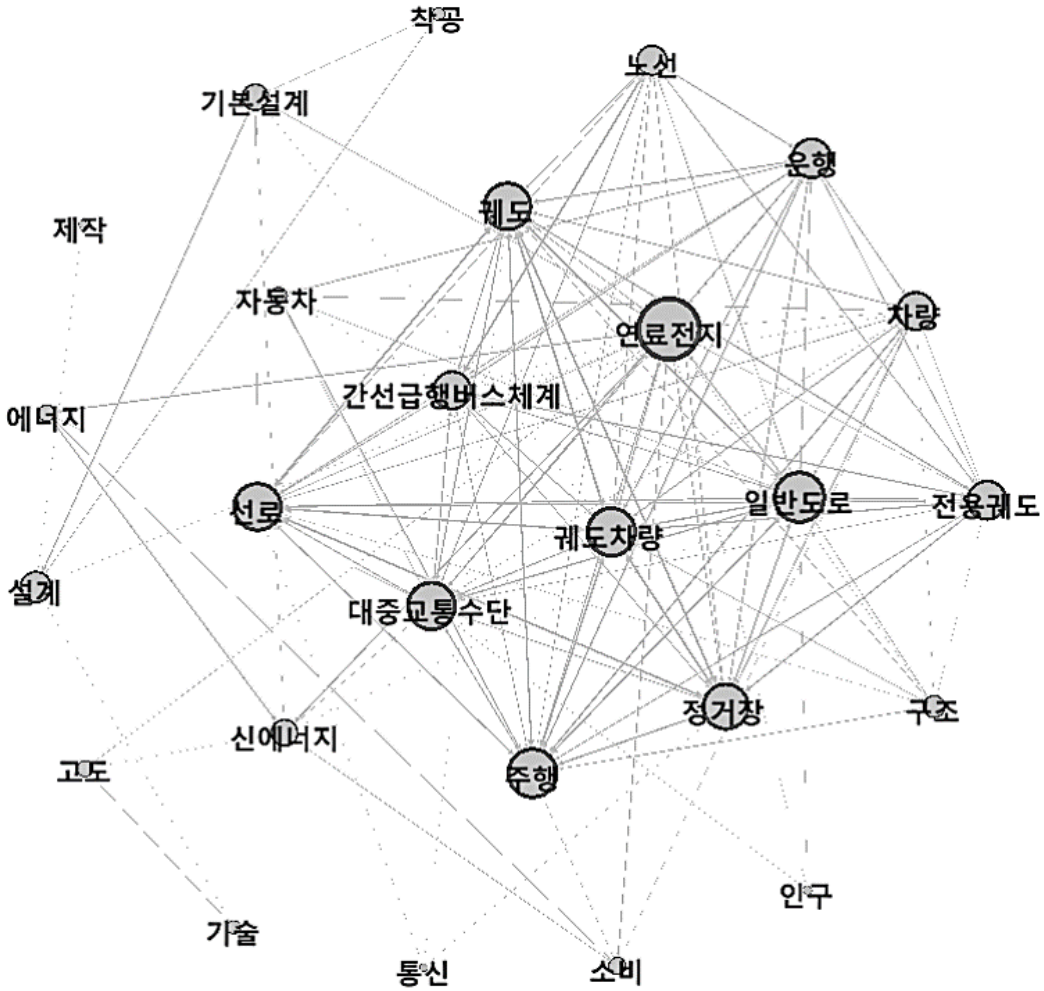
```
0.7385030359961533
```

```
wv_model_general.most_similar('노선')
```

```
[('운행계통', 0.8900529146194458),
 ('운행구간', 0.8336923718452454),
 ('종점', 0.8079904317855835),
 ('여객자동차운송사업자', 0.8062624335289001),
 ('구간', 0.8059366941452026),
 ('노선버스', 0.8038692474365234),
 ('운행횟수', 0.8012652397155762),
 ('기점', 0.7933613061904907),
 ('운행경로', 0.7926549315452576),
 ('운행시간', 0.7784532308578491)]
```

[그림 3-4] 단어들의 similarity 예시

다음으로 앞서 계산된 similarity를 입력 값으로 Python의 네트워크 분석 툴인 NetworkX를 이용하여, Weighted-PageRank 값을 계산한다. 이 때, 임의 탐색자와 관련된 이탈 확률인 damping 값을 0.85, 끊임없이 반복되는 재귀 호출을 정지하는 기준인 epsilon 값은 0.001, 그래프 방향성은 undirected(역방향)로 설정한다. 이러한 과정을 통해 계산된 Weighted-PageRank 단어 네트워크 예시는 [그림 3-5]와 같다. 해당 그림에서는 '연료전지'라는 단어가 가장 중요한 단어로 볼 수 있다.



[그림 3-5] Weighted-PageRank 단어 네트워크

마지막으로 각 단어의 Weighted-PageRank 값으로 연구계획서 벡터를 생성한다. 생성된 연구계획서 벡터 예시는 [표 3-2]와 같다.

[표 3-2] 연구계획서 벡터

	자동차	궤도	노선	에너지	연료전지	...	
<	0.004	0.132	0.104	0.009	0.382	...	>

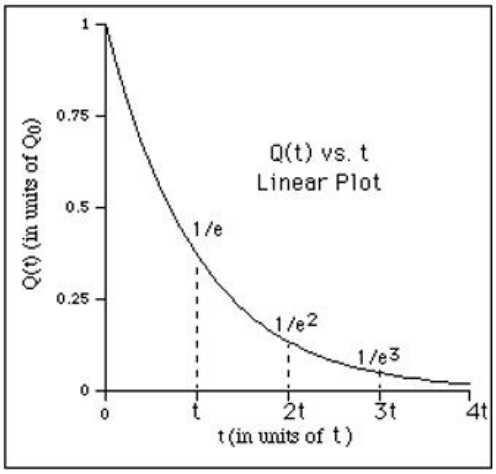
3.1.2 법령 벡터 표현

일반적으로 TF-IDF는 문서에서 특정 단어의 가중치를 구하는 공식으로 TF 값은 해당 단어가 해당 문서 내에 얼마나 많이 등장하는지를 나타내고, IDF값은 해당 단어가 다른 문서들에 비해 해당 문서 내에서 얼마나 희귀한지를 의미한다. 본 연구는 4000여개 법령 벡터를 각 법령 본문에 등장하는 단어들의 TF-IDF 값으로 표현한다. 즉, 해당 단어가 해당 법령에서 갖는 가중치를 의미한다. 예를 들어 '전용궤도'란 단어를 「궤도운송법 시행규칙」, 「궤도운송법 시행령」에 대한 가중치를 구하면 13.26과 8.57이다. 이를 통해 '전용궤도'란 단어는 「궤도운송법 시행령」보다 「궤도운송법 시행규칙」에서 보다 중요한 의미를 갖는 것을 알 수 있다.

3.2 벡터 변형

연구계획서에 등장하는 단어가 너무 일반적인 단어일 경우, 연구계획서의 특성을 잘 반영하지 못하기 때문에 영향력을 약화시켜주고자 연구계획서 벡터를 변형한다. 영향력을 약화시켜주는 방법으로는 Exponential Decay function의 아이디어를 활용한다.

Exponential Decay function이란 시간에 따라 지수적으로 중요도를 감소시키는 함수이며, 아래 [그림 3-6]의 미분 방정식으로 표현될 수 있다.



- $\frac{dN}{dt} = -\lambda N$
- $N(t) = N_0 e^{-\lambda t}$
  - $N_0 = N(0)$ = 초기값
  - λ = decay 상수
  - t = 시간

[그림 3-6] Exponential Decay function

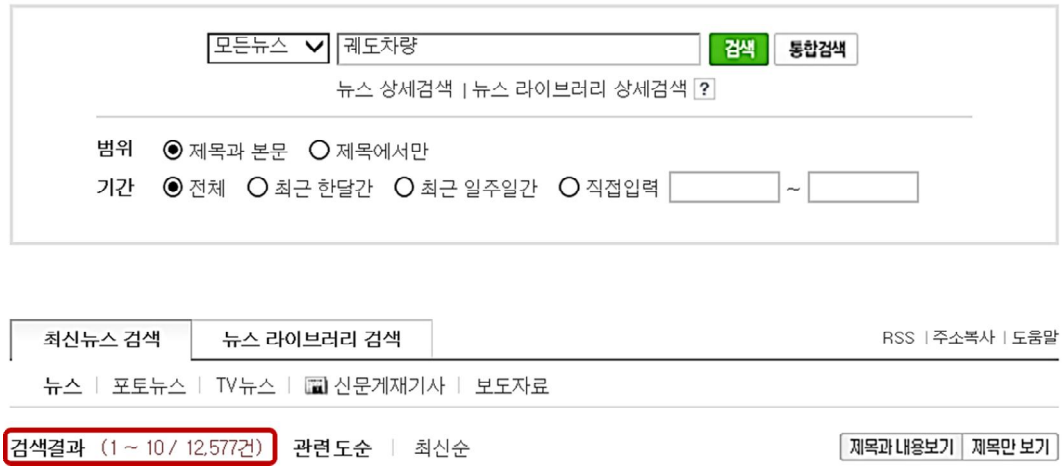
본 논문에서는 Exponential Decay function의 지수적으로 중요도를 감소시키는 아이디어를 활용하여 단어와 단어 간의 similarity를 조정한다. 상대적으로 너무 일반적인 단어일 경우, 뉴스 검색 결과 수가 높다는 것을 전제하고, 기존 Exponential Decay function의 시간을 나타내는 t를 네이버 뉴스에서 각 단어의 검색 결과 수로 대체한다. 이렇게 되면 높은 검색 결과일수록 지수에 큰 값이 곱해지기 때문에 일반적인 단어의 weight를 낮출 수 있게 된다. 변형된 수식은 아래와 같다.

$$W_{i,j} = e^{-\lambda(f_i + f_j)}$$

· f_i : 네이버 뉴스에서 단어 i의 검색 결과 수

· f_j : 네이버 뉴스에서 단어 j의 검색 결과 수

네이버 뉴스 검색 결과란 네이버 뉴스 창(http://news.naver.com/)에서의 검색 결과를 의미한다. 예를 들어, [그림 3-7]은 네이버 뉴스에서 '궤도차량' 이라는 단어의 검색 결과이다. 검색 결과 수는 12,577이며 이 값이 f_j 즉, $f_{\text{궤도차량}}$에 해당하는 값이다.



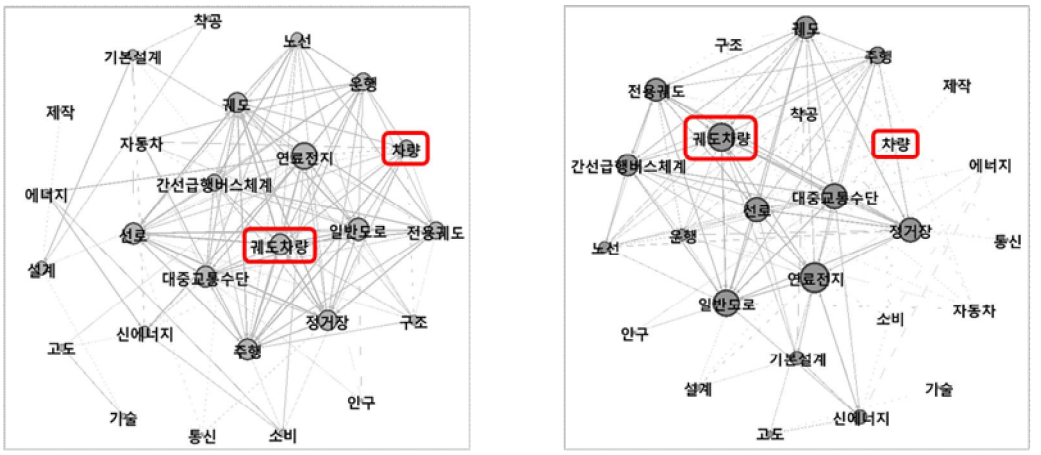
[그림 3-7] 네이버 뉴스 검색 결과 예시

위의 변형된 Exponential Decay function을 Weighted-PageRank 알고리즘에 적용한 수식은 아래와 같다. 기존 Word2Vec 모델에서 두 단어의 similarity 값에 $e^{-\lambda(f_{\text{단어1}}+f_{\text{단어2}})}$를 곱하면 된다.

$$WP(V_i) = (1-d) + d * \sum_{V_j \in in(V_i)} \frac{weight(i,j)}{\sum_{V_k \in out(V_j)} weight(j,k)} WP(V_j)$$

- $weight(i,j) = Word2Vec(i,j) * e^{-\lambda(f_i+f_j)}$
- $weight(j,k) = Word2Vec(j,k) * e^{-\lambda(f_j+f_k)}$

[그림 3-8]은 기존 Weighted-PageRank 알고리즘을 적용한 단어 네트워크와 Exponential Decay function을 적용한 Weighted-PageRank 알고리즘을 적용한 단어 네트워크이다. '궤도차량'과 같은 특수한 단어는 영향력이 강화되었으며, '차량'과 같은 너무 일반적인 단어는 영향력이 약화된 것을 알 수 있다.



기존 Weighted-Pagerank 단어 네트워크 *Decay를 적용한 Weighted-Pagerank 단어 네트워크*

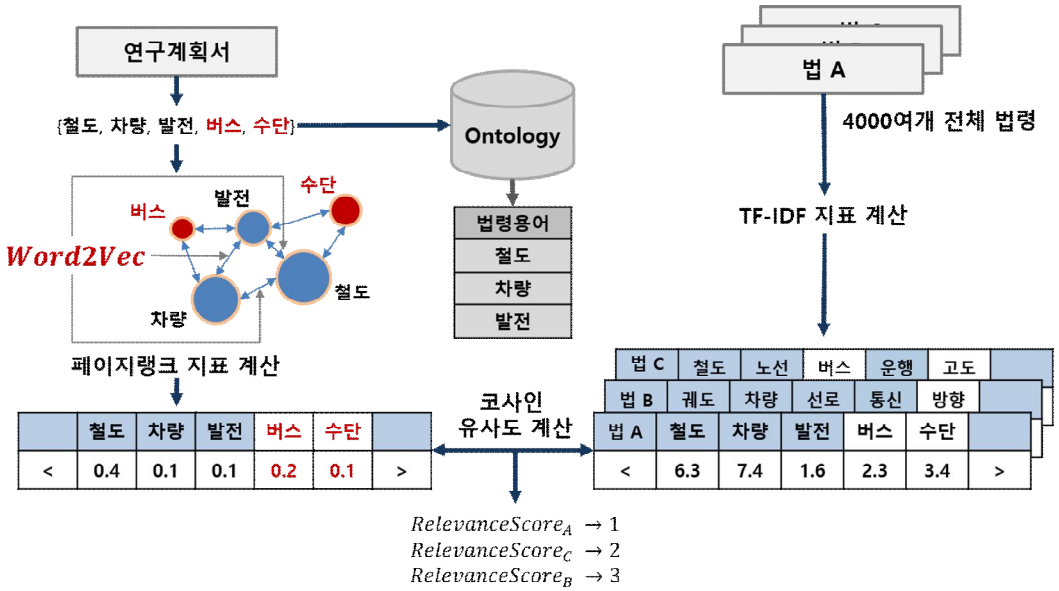
[그림 3-8] 단어의 영향력 변화 예시

한편, 빈도 지표의 영향력에 변화를 주고자 법령 벡터도 변형한다. 법령 본문에서 동일한 단어가 너무 많이 등장하여 $tf(t,d)$의 값이 민감하게 반응하는 상황에 대응하고자 $\log tf(t,d)$를 사용한다. 변형된 수식은 아래와 같다.

$$\log tf(t,d) - idf(t,d,D) = \log tf(t,d) * idf(t,d,D)$$

3.3 연관 법령 검색 방법론

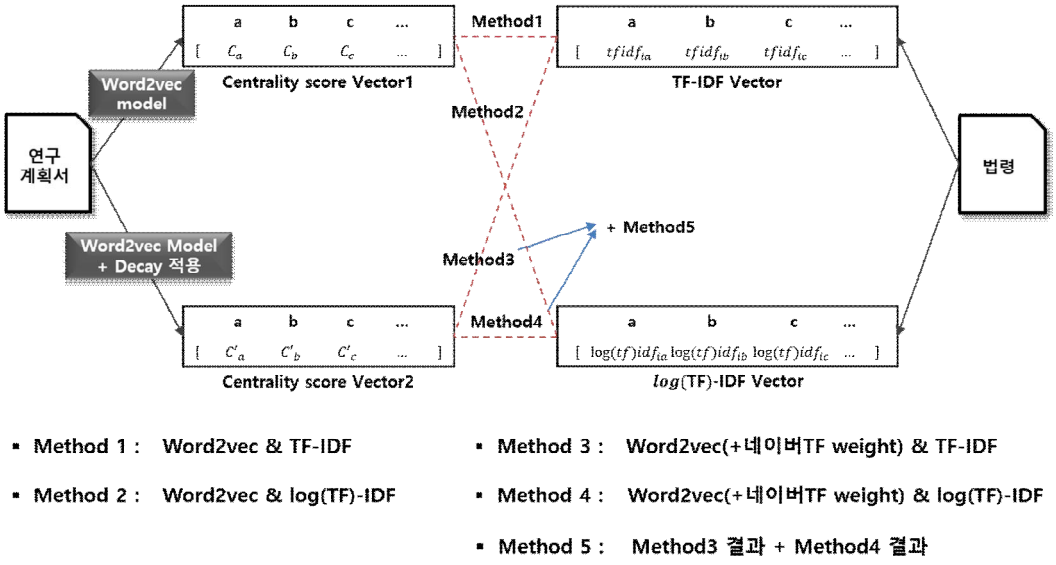
연구계획서와 관련 있는 법령들을 결정하기 위해서 연구계획서 벡터와 4000여개의 법령 TF-IDF 벡터 간 Cosine Similarity를 계산한다. [그림 3-9]는 연구계획서 입력 단계부터 Consin Similarity 계산까지의 과정을 나타낸다.



[그림 3-9] 연관 법령 검색 과정 예시

본 연구에서는 연구계획서와 관련 있는 법령들을 결정하기 위한 5가지 방법론을 제안한다. 제안된 방법론은 [그림 3-10]과 같다. Method 1은 기존 Word2Vec 모델의 similarity를 사용한 연구계획서 벡터와 법령들의 TF-IDF 벡터의 Cosine Similarity를 비교한다. Method 2는 기존 Word2Vec 모델의 similarity를 사용한 연구계획서 벡터와 법령들의 log(TF)-IDF 벡터의 Cosine Similarity를 비교한다. Method 3은 기존 Word2Vec 모델의 similarity에

Exponential Decay function을 적용한 연구계획서 벡터와 법령들의 TF-IDF 벡터의 Cosine Similarity를 비교한다. Method 4는 기존 Word2Vec 모델의 similarity에 Exponential Decay function을 적용한 연구계획서 벡터와 법령들의 log(TF)-IDF 벡터의 Cosine Similarity를 비교한다. Method 5는 Method 3의 결과와 Method 4의 결과를 앙상블(Ensemble) 시킨 방법론이다. Exponential Decay function을 적용한 연구계획서 벡터가 결과에 미치는 영향력이 log(TF)-IDF 법령 벡터가 결과에 미치는 영향력보다 컸기 때문에 Exponential Decay function을 적용한 연구계획서 벡터를 사용하는 Method 3과 Method 4의 결과를 앙상블(Ensemble) 시킨다. 영향력 평가 결과는 본 논문의 4.3절에서 다루도록 한다.



[그림 3-10] 제안 방법론

Method 5의 앙상블(Ensemble) 방법 예시는 [그림 3-11]과 같다. 각 법령들의 Method 3에서의 순위와 Method 4에서의 순위의 평균을 구한 뒤, 그 값으로 다시 순위를 구한다. 예를 들어, '에너지법'은 Method 3에서의 순위는 6위, Method 4에서의 순위는 3위이므로 Method 5에서 순위를 정하기 위한 값으로 (6+3)/2=4.5를 사용하게 된다.

Method 3	
순위	법령
1	철도건설규칙
2	궤도운송법 시행규칙
3	도시철도건설규칙
4	철도산업발전기본법
5	궤도운송법
6	에너지법

Method 4	
순위	법령
1	철도건설규칙
2	도시철도건설규칙
3	에너지법
4	철도산업발전기본법
5	궤도운송법 시행규칙
6	궤도운송법

Method 5		
순위	법령	(M3순위 + M4순위) / 2
1	철도건설규칙	(1+1) / 2 = 1
2	도시철도건설규칙	(3+2) / 2 = 2.5
3	궤도운송법 시행규칙	(2+5) / 2 = 3.5
4	철도산업발전기본법	(4+4) / 2 = 4
5	에너지법	(6+3) / 2 = 4.5
6	궤도운송법	(5+6) / 2 = 5.5

[그림 3-11] 앙상블(Ensemble) 방법 예시

4. 실험

한국철도기술연구원에서 제공한 3개의 연구계획서로 실험을 진행하였다. 3개의 연구계획서는 '신에너지 Bimodal 저상굴절 차량 개발', '고속화물 열차 및 여객/화물 복합기술 개발', 'BIM 기반의 철도인프라 관리 표준기술 개발'이다. 실험은 각 연구계획서를 입력 받아 5개의 방법론을 적용하여 4000여개의 법령을 연구계획서와의 유사도 순으로 순위를 매기고, 실제로 관련 있는 법령이 몇 위에 랭킹이 되어있는지 평가하였다.

실험환경은 Python 언어를 사용하여 구성하였다. Word2Vec 모델의 학습을 위해 Python 기반 라이브러리인 gensim을 사용하였고, Weighted-PageRank 알고리즘을 계산하기 위해 Python의 네트워크 분석 툴인 NetworkX를 사용하였다.

4.1 평가지표

평가지표는 한국철도기술연구원의 관계자들이 우선순위가 매겨진 법령 목록을 보고 직접 평가한 것을 기준으로 삼았다. 법령을 '적합', '보통', '부적합' 3단계로 구분하였다. 적합은 해당 연구계획서와 정확히 관련된 법령이며, 보통은 연구계획서와의 관련성이 적합의 절반을 의미한다. 또한 부적합은 해당 연구계획서와의 관련성이 거의 없음을 의미한다.

4.2 성능검증지표

성능 검증을 위해 총 3가지 지표를 사용하였다.

첫 번째는 '순위 합(Rank Sum)'이다. '적합'에 해당하는 법령들의 순위를 합한 값이다. 수식은 아래와 같다.

$$Rank\,Sum = \sum_{i=1} rank_i * R_i$$

- $rank_i$: 연구계획서와 법령i의 유사도 순위
- $R_i = \begin{cases} 1 : \text{법령}\,i\text{가 '적합'일 때} \\ 0 : \text{법령}\,i\text{가 '적합'이 아닐 때} \end{cases}$

두 번째는 '가중 순위 합(Weighted Rank Sum)'이다. '적합'에 해당하는 법령들의 순위 합과 '보통'에 해당하는 순위에 0.5 가중치를 곱한 값들의 합이다. 수식은 아래와 같다.

$$Weighted\,Rank\,Sum = \sum_{i=1} rank_i * R_i^{'}$$

- $rank_i$: 연구계획서와 법령i의 유사도 순위
- $R_i^{'} = \begin{cases} 1 & : \text{법령}\,i\text{가 '적합'일 때} \\ 0.5 & : \text{법령}\,i\text{가 '보통'일 때} \\ 0 & : \text{법령}\,i\text{가 '적합','보통'이 아닐 때} \end{cases}$

세 번째는 '11-point 평균 정확률(11-point Average Precision)'이다. 11개의 재현율(Recall) 구간(0.0, 0.1, 0.2, … , 1.0)에서 각 구간별로 보간법

(interpolation)이 적용된 정확률(Precision) 값을 구하고 이를 평균한 값이다. 정확률(Precision)은 계산하려는 해당 법령을 기준으로 앞 순위 법령 중 '적합'에 해당한 법령의 개수를 계산하려는 해당 법령의 순위로 나누어 계산한다.

'순위 합(Rank Sum)'과 '가중 순위 합(Weighted Rank Sum)'은 낮을수록 좋은 성능을 나타내고, '평균 정확률-재현율(Average Precision-Recall)'은 높을수록 좋은 성능을 나타낸다.

4.3 Exponential Decay function과 log(TF)-IDF의 영향력 평가

본 논문의 3.3절에서 Exponential Decay function을 적용한 연구계획서 벡터가 결과에 미치는 영향력이 log(TF)-IDF 법령 벡터가 결과에 미치는 영향력보다 컸기 때문에 Exponential Decay function을 적용한 연구계획서 벡터를 사용하는 Method 3과 Method 4의 결과를 앙상블(Ensemble) 시킨 Method 5를 제안한다고 언급했다. 영향력 평가 결과는 [그림 4-1]과 [그림 4-2]와 같다. [그림 4-1]은 Exponential Decay function의 영향력 평가 결과이고, [그림 4-2]는 log(TF)-IDF의 영향력 평가이다. '순위 합(Rank Sum)'과 '가중 순위 합(Weighted Rank Sum)'의 향상도를 평가해 본 결과, 대체적으로 Exponential Decay function을 적용한 연구계획서 벡터를 사용하는 것이 더욱 좋은 결과를 도출했다.

법령용어	bimodal		ctx		bim	
	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum
word2vec & TF-IDF	13609	15874	1694	5930.5	20532	27977.5
word2vec(+Decay 적용) & TF-IDF	12831	15040.5	1670	5816	20836	28397.5
향상도	-778	-833.5	-24	-114.5	304	420

법령용어	bimodal		ctx		bim	
	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum
word2vec & log(TF)-IDF	9798	11331.5	718	3322	16630	23048.5
word2vec(+Decay 적용) & log(TF)-IDF	9708	11215.5	706	3454	17391	24019
향상도	-90	-116	-12	132	761	970.5

법령용어 + 비법령용어	bimodal		ctx		bim	
	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum
word2vec & TF-IDF	11289	12269	1967	5445	19675	32236.5
word2vec(+Decay 적용) & TF-IDF	9930	10758	1647	4820.5	17912	29736.5
향상도	-1359	-1511	-320	-624.5	-1763	-2500

법령용어 + 비법령용어	bimodal		ctx		bim	
	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum
word2vec & log(TF)-IDF	16379	17919.5	2445	7138	24953	38404
word2vec(+Decay 적용) & log(TF)-IDF	13505	14711.5	2008	6120	22266	34880.5
향상도	-2874	-3208	-437	-1018	-2687	-3523.5

[그림 4-1] Exponential Decay function 영향력 평가 결과

법령용어	Bimodal		CTX		BIM	
	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum
word2vec & TF-IDF	13609	15874	1694	5930.5	20532	27977.5
word2vec & log(TF)-IDF	9798	11331.5	718	3322	16630	23048.5
향상도	-3811	-4542.5	-976	-2608.5	-3902	-4929

법령용어	Bimodal		CTX		BIM	
	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum
word2vec(+Decay 적용) & TF-IDF	12831	15040.5	1670	5816	20836	28397.5
word2vec(+Decay 적용) & log(TF)-IDF	9708	11215.5	706	3454	17391	24019
향상도	-3123	-3825	-964	-2362	-3445	-4378.5

법령용어 + 비법령용어	Bimodal		CTX		BIM	
	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum
word2vec & TF-IDF	11289	12269	1967	5445	19675	32236.5
word2vec & log(TF)-IDF	16379	17919.5	2445	7138	24953	38404
향상도	5090	5650.5	478	1693	5278	6167.5

법령용어 + 비법령용어	Bimodal		CTX		BIM	
	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum	rank_sum	weighted_rank_sum
word2vec(+Decay 적용) & TF-IDF	9930	10758	1647	4820.5	17912	29736.5
word2vec(+Decay 적용) & log(TF)-IDF	13505	14711.5	2008	6120	22266	34880.5
향상도	3575	3953.5	361	1299.5	4354	5144

[그림 4-2] log(TF)-IDF 영향력 평가 결과

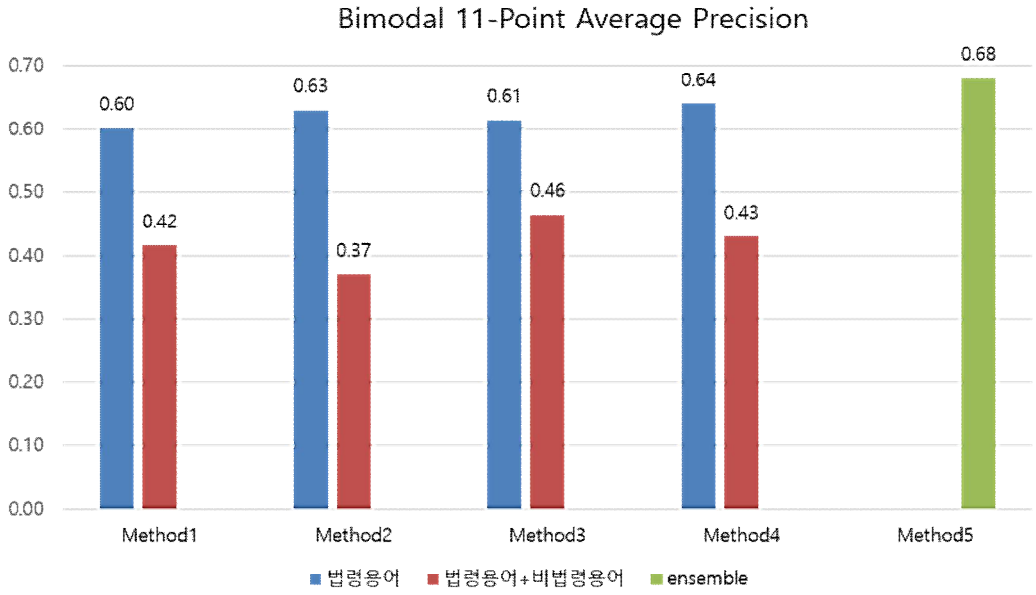
4.4 실험결과

[표 4-1]은 ‘신에너지 Bimodal 저상굴절 차량 개발(Bimodal)’ 연구계획서의 ‘순위 합(Rank Sum)’과 ‘가중 순위 합(Weighted Rank Sum)’ 결과이다. Method3과 Method4의 총 4개의 결과를 앙상블(Ensemble) 시킨 Method5의 성능이 우수함을 알 수 있다.

[표 4-1] ‘Bimodal’ 연구계획서 순위 합 결과

Bimodal	법령용어		법령용어+비법령용어		Ensemble (Method5)	
	Rank Sum	Weighted Rank Sum	Rank Sum	Weighted Rank Sum	Rank Sum	Weighted Rank Sum
Method1	13609	15874	11289	12269		
Method2	9798	11331.5	16379	17919.5		
Method3	12831	15040.5	9930	10758	5476	6111.5
Method4	9708	11215.5	13505	14711.5		

[그림 4-3]은 '신에너지 Bimodal 저상굴절 차량 개발(Bimodal)' 연구계획서의 '11-point 평균 정확률(11-point Average Precision)' 결과이다. 이 또한 Method3과 Method4의 총 4개의 결과를 앙상블(Ensemble) 시킨 Method5의 성능이 우수함을 알 수 있다.



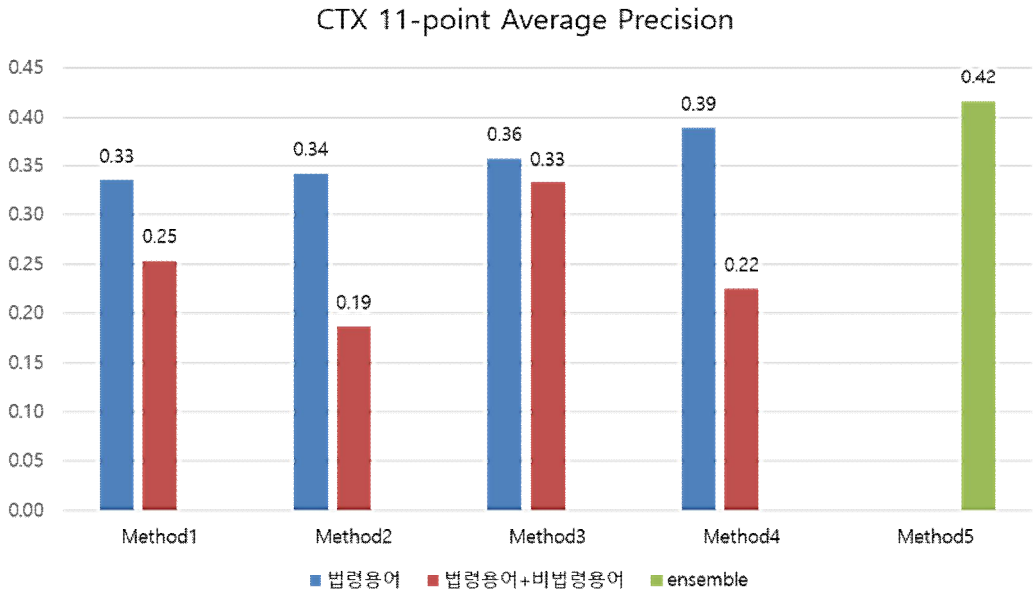
[그림 4-3] 'Bimodal' 연구계획서 11-Point 평균 정확률

[표 4-2]는 ‘고속화물 열차 및 여객/화물 복합기술 개발(CTX)’ 연구계획서의 ‘순위 합(Rank Sum)’과 ‘가중 순위 합(Weighted Rank Sum)’ 결과이다. Method3과 Method4의 총 4개의 결과를 앙상블(Ensemble) 시킨 Method5의 성능이 우수함을 알 수 있다.

[표 4-2] ‘CTX’ 연구계획서 순위 합 결과

CTX	법령용어		법령용어+비법령용어		Ensemble (Method5)	
	Rank Sum	Weighted Rank Sum	Rank Sum	Weighted Rank Sum	Rank Sum	Weighted Rank Sum
Method1	1694	5930.5	1967	5445		
Method2	718	3322	2445	7138		
Method3	1670	5816	1647	4820.5	579	1922.5
Method4	706	3454	2008	6120		

[그림 4-4]는 '고속화물 열차 및 여객/화물 복합기술 개발(CTX)' 연구계획서의 '11-point 평균 정확률(11-point Average Precision)' 결과이다. 이 또한 Method3과 Method4의 총 4개의 결과를 앙상블(Ensemble) 시킨 Method5의 성능이 우수함을 알 수 있다.



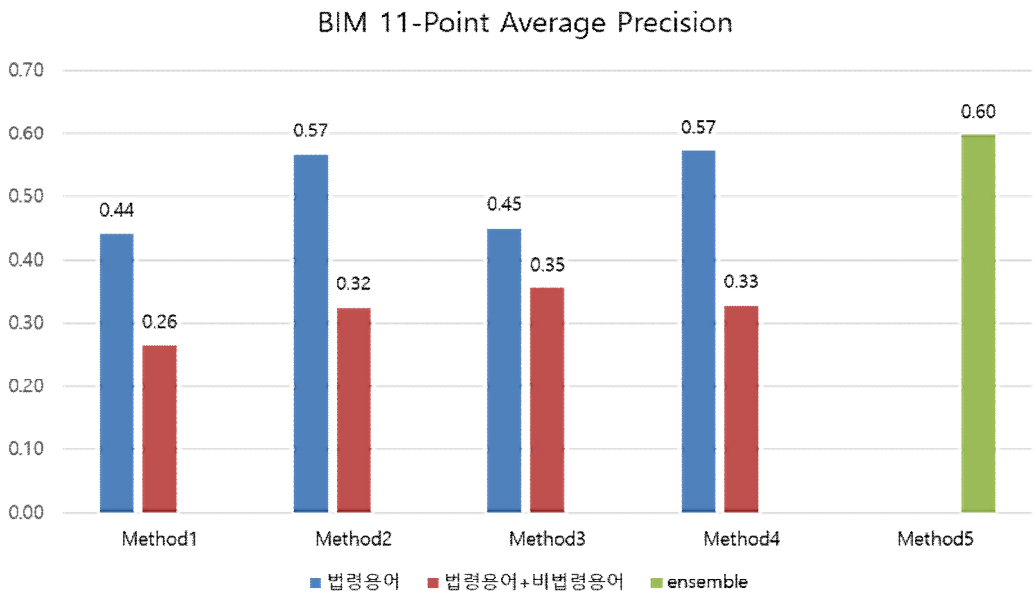
[그림 4-4] 'CTX' 연구계획서 11-Point 평균 정확률

[표 4-3]은 ‘BIM 기반의 철도인프라 관리 표준기술 개발(BIM)’ 연구계획서의 ‘순위 합(Rank Sum)’과 ‘가중 순위 합(Weighted Rank Sum)’ 결과이다. Method3과 Method4의 총 4개의 결과를 앙상블(Ensemble) 시킨 Method5의 성능이 우수함을 알 수 있다.

[표 4-3] ‘BIM’ 연구계획서 순위 합 결과

BIM	법령용어		법령용어+비법령용어		Ensemble (Method5)	
	Rank Sum	Weighted Rank Sum	Rank Sum	Weighted Rank Sum	Rank Sum	Weighted Rank Sum
Method1	20532	27977.5	19675	32236.5		
Method2	16630	23048.5	24953	38404		
Method3	20836	28397.5	17912	29736.5	9766	15030.5
Method4	17391	24019	22266	34880.5		

[그림 4-5]는 'BIM 기반의 철도인프라 관리 표준기술 개발(BIM)' 연구계획서의 '11-point 평균 정확률(11-point Average Precision)' 결과이다. 이 또한 Method3과 Method4의 총 4개의 결과를 앙상블(Ensemble)시킨 Method5의 성능이 우수함을 알 수 있다.



[그림 4-5] 'BIM' 연구계획서 11-Point 평균 정확률

5. 결론 및 향후 연구 방향

5.1 결론

본 연구는 R&D 연구계획서와 관련된 정책 및 법령을 실시간으로 검색할 수 있는 방법론을 제안하였다. 한국철도기술연구원에서 제공한 연구개발계획서로 실험 및 평가를 진행하였다. 실험 결과, Exponential Decay function을 적용한 연구계획서 벡터를 사용하는 Method 3과 Method 4의 결과를 앙상블(Ensemble) 시킨 Method 5의 성능이 가장 좋았다.

국가법령정보 공동활용 사이트(http://open.law.go.kr/)에서 제공하는 법령용어 API와 법령 API를 활용하여 법령용어 사전 전체와 법령 사전 전체를 실시간으로 수집하여 Word2Vec 모델링에 사용하기 때문에 법령의 개정과 재정을 파악할 수 있어 R&D 사업의 실용화와 질을 높을 수 있을 것이다. 또한 법령 및 법령용어와 관련된 데이터를 온톨로지 형태의 지식 DB로 구축하였는데, 온톨로지 기술은 표준화되어 있기 때문에 본 연구에서 구축한 온톨로지는 다른 유사 온톨로지를 만들 때 변형하여 쉽게 적용할 수 있을 것이다.

5.2 향후 연구 방향

본 연구에서는 현행 법령 본문으로 학습된 Word2Vec 모델을 사용한다. 따라서 법령 내에서 단어들 간의 similarity를 구할 수 있게 된다. 이 Word2Vec 모델이 학습한 단어별 유사어를 사용하여 연구계획서 벡터 변형을 고려해 볼 수 있다. 연구계획서에 등장하는 단어들 중에서 비법령 용어를 Word2Vec 모델

에서 가장 유사하다고 판단되는 법령용어로 대체시켜 새로운 연구계획서 벡터를 생성하는 것이다. 이 새로운 연구계획서 벡터는 모두 법령용어로만 이루어졌기 때문에 이 벡터를 사용한다면 보다 관련된 법령을 잘 뽑을 수 있을 것이라 기대된다.

참고문헌

[1] Annett Mitschick, "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.

[2] Carlos M. Toledo, Mariel A. Ale, Omar Chiotti and Mar ı a R. Galli, "An Ontology-driven Document Retrieval Strategy for Organizational Knowledge Management Systems", Electronic Notes in Theoretical Computer Science 281 pp.21-34, 2011.

[3] T Mikolov, K Chen, G Corrado, J Dean “Efficient estimation of word representations in vector space”, arXiv preprint arXiv:1301.3781, 2013.

[4] Brin, S. and Page, L., "The anatomy of a large-scale hypertextual Web search engine", Computer Networks and ISDN Systems 30, pp.107-117, 1998.

[5] G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing”, Communications of the ACM vol.18 nr.11, pp.613-620, 1975.

[6] Jongbae kim, Jungwon Byun, Dongju Sun, Taegyun Kim, Yung Kim, “A model for Maesuring the R&D Project Similarity using Patent Information,” J. Korea Inst. Inf. Commun. Eng. Vol.18 No.5, pp.1013-1021, 2014.

[7] D.L. Lee, Huei Chuang, K. Seamons, “Document ranking and the

vector-space model”, IEEE Software Vol.14 Issue.2, 1997.3.

[8] Goldberg Yoav, Levy Omer, "word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method", arXiv:1402.3722, 2014.

[9] 김우주, 네트워크 중심성 이론, 카오스북, 2015.

[10] 손동원, 소셜 네트워크 분석, 박영사, 2007.

[11] 김지현, 이종서, 이명진, 김우주, 홍준석, “법령정보 검색을 위한 생활용어와 법률용어 간의 대응관계 탐색 방법론”, 지능정보연구 제18권 제3호, pp.137-152, 2012.9.

[12] 장인호, “온톨로지 기반 법률 검색시스템의 구축 및 평가에 관한 연구”, 한국문헌정보학회지, 제45권 제2호, pp.345-366, 2011.5.

[13] 백종법, 이수원, “웹 마이닝을 활용한 법령정보검색 지원 시스템”, 정보과학학회논문지:소프트웨어 및 응용 제40권 제7호, pp.395-404, 2013.7.

[14] W3C, Document Object Model(DOM), htt://www.w3.org/DOM/

[15] 법제처, http://www.law.go.kr/main.html

ABSTRACT

A text document-based search algorithm of relevant statutes for accuracy improvement

Youna Lee
Dept. of Information & Industrial Engineering
The Graduate School
Yonsei University

In general, the R&D projects can take many years from planning to commercialization. In the process, relevant statutes can be revised and newly enacted. The changes like this will affect the R&D project, and in the worst case, can cause investment failures. Also, since there is currently no methodology to search policies and statutes related to a R&D Project, researchers who are not experts in the field of law are hard to know what laws are associated with their projects.

Therefore, we propose a methodology that can search policies and statutes related to R&D subjects in real time. First, it receives R&D plan, and generates the vector of the R&D plan using the centrality score. Next, the vector of the R&D plan compares the 4,000 or more legal TF-IDF vector with the cosine similarity, and it shows the

relevant statutes in order of high similarity. At this time, the centrality score is a Weighted-PageRank value using the similarity between the words of the Word2Vec model learned in the text of the current statutes. We presented five methodologies by related statutes search method. We conducted experiments on three research plans provided by the Korea Railway Technology Research Institute. The evaluation index is a related statutes list provided by the Korea Railway Technology Research Institute.

Finally, since we can provide policy and statutes related to R&D project to R&D researchers, they use them as reference materials to determine the direction of research at the planning stage. In addition, since we collect related statutes in real time, We expect that researchers can review relevant statutes during project period, preventing their R&D projects from commercialization failure caused by statutes.

Keyword: Term Vector Model, Centrality Score, TF-IDF, Cosine Similarity, Weighted-PageRank, Ontology