머신러닝 기반 유효특허 분류기 개발에 관한 연구

(A Study on Development of Garbage Patent Classifier Based on Machine Learning)

장예훈, 최성철 / 서원철

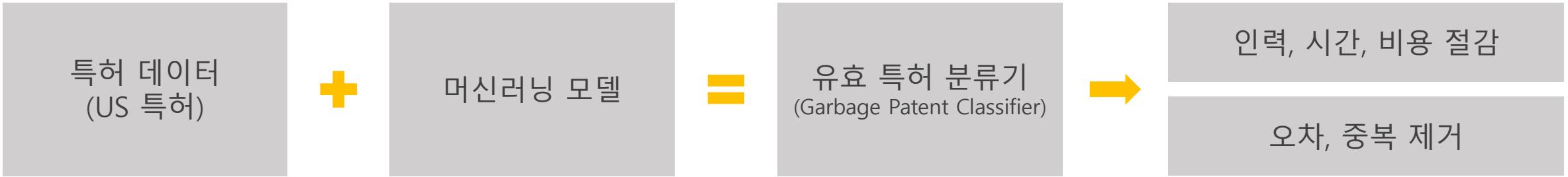
가천대학교 산업경영공학과 / 부경대학교 기술경영전문대학원

jangyh0420@gmail.com, sc82.choi@gachon.ac.kr / wcseo@pknu.ac.kr

연구 개요

연구 배경, 연구 목적 및 연구의 필요성

- **기술 연구 개발 단계에서 특허권 확보는 매우 중요**
  ➔ 해당 기술의 **유효 특허 검색 필수**
- **실제 유효 특허 수의 10~1,000배의 문서 검토 작업 후 유효 특허 여부 판별 가능**
  ➔ 막대한 인력, 시간 등의 **비용** 소모 & 수작업으로 인한 **오차**, **중복** 발생
- **특허 데이터와 머신러닝을 활용하여 유효 특허 분류기 개발**
  ➔ 비용과 오차 **감소** 효과



연구 방법

실험 데이터, 데이터 전처리, 실험, 결론

실험 데이터	- 특허나라와 WIPS DB에서 US 특허 데이터 추출
데이터 전처리	- 머신러닝 모델이 학습 가능한 형태로 데이터 형 변환 - 서지적, 네트워크, text 데이터를 가공 후 전체 데이터 set에 merge
실험	- 유효 특허 분류 예측을 위해 여러 머신러닝 모델 학습 - recall 점수 향상을 위해 다양한 기법 활용
결론	- recall 점수가 높은 유효 특허 예측 모델과 features

실험 데이터

실험 데이터 개요, 모델 성능 측정 방법

- 실험 데이터

1) 데이터 개요

	개수 (백분율)
Valid*	29 (0.0152)
Garbage**	1876 (0.9848)
Total	1905 (1)

* Valid: 특허나라에서 제공하는 '관리형 해상최종처분장 조성기술개발'의 영문 검색식을 이용하여 WIPS DB에 유효 특허 검색 조건에 맞게 검색한 모든 US 특허 Data 중 유효 US 특허와 일치하는 특허

** Garbage: 특허나라에서 제공하는 '관리형 해상최종처분장 조성기술개발'의 영문 검색식을 이용하여 WIPS DB에 유효 특허 검색 조건에 맞게 검색한 모든 US 특허 Data 중 유효 특허를 제외한 모든 특허

2) 데이터 예시

	국가코드	DB종류	특허/실용구분	문헌종류코드	발명의 명칭	요약	대표청구항	청구항수	출원번호	출원일	...	Original CPC Main	Original CPC All	Original IPC Main	Original IPC All	Original US Class Main[US]	Original US Class All[US]	Original FI[JP]	Original F-term[JP]	Original Theme Code [JP]	WIPS ON key
0	US	US	P	B2	Turboengine water wash system	A system for cleaning gas turbine engines is d...	1. An apparatus for cleaning wing mounted gas ...	64	11/644784	2006-12-22	...	NaN	NaN	B08B-009/00	B08B-009/00	134/166.R	134/166.R | 134/138	NaN	NaN	NaN	4.914000e+12
1	US	US	P	A1	TREATMENT OF TAILINGS WITH DEIONIZED SILICATE ...	A process for treating a tailings stream compr...	1. A process for treating a tailings stream co...	20	13/848244	2013-03-21	...	B09B-0003/0025	B09B-0003/0025	B09B-003/00	B09B-003/00	106/627	106/627 | 106/600 | 106/631 | 106/632 | 106/63...	NaN	NaN	NaN	5.414000e+12
2	US	US	P	A1	Method And Apparatus For Treating Tailings Usi...	There is a method described of treating tailin...	1. A method of treating tailings having a comb...	18	13/765924	2013-02-13	...	C02F-0001/48	C02F-0001/48	C02F-001/48	C02F-001/48	204/554	204/554	NaN	NaN	NaN	5.414000e+12

1905 rows x 37 columns

- 전체 데이터 중에서 극소수의 유효 특허를 찾아내고 Garbage 특허를 제거하는 모델 ➔ Recall* 점수 중요

$$* \ Recall = \frac{|TP|}{|FN| + |TP|}$$

(TP: True Positive, FN: False Negative)

* 실제 유효 특허에 대해 예측모델이 유효 특허라고 예측한 특허 개수의 비율

데이터 전처리

1차 - Text 형태의 데이터를 제외한 모든 데이터 전처리 후 실험

- 데이터 전처리

1. 전체 데이터에서 유효 특허와 Garbage 특허를 각각 **1과 0으로 Labeling** 한 Column 추가
2. IPC, US Class 등 **서지적** 데이터 → **pivot table**로 변환한 뒤 전체 데이터 set에 merge
3. 날짜정보, 국제공개번호 등의 **hyphen(-)이 포함**된 데이터 → **hyphen 제거** 및 Nan 값 채우기
4. 우선권 번호, 문헌종류 코드 등 **categorical** 데이터 → **Label Encoding**으로 간략화

ex)

	전	후
유효 특허 여부	유효 / Garbage	1 / 0
hyphen	2018-01-01	20180101
categorical	A1, B2, A1, …	1, 3, 1, …

데이터 값 변환

	A02B	B03F	D05H
patent 0	1	0	0
patent 1	0	0	0
patent 2	1	1	0

서지적 정보 pivot table

1차 실험 방법 및 결과

실험 결과 및 한계점과 해결방안

- 1차 실험 방법

1. 전체 데이터를 7:3으로 Training, Test set 분리
2. 머신러닝 모델 학습

- 1차 실험 결과

	precision	recall	f1-score
0	0.99	1.00	0.99
1	0.67	0.29	0.40
accuracy		0.99	

Decision Tree

	precision	recall	f1-score
0	0.99	1.00	1.00
1	1.00	0.57	0.73
accuracy		0.99	

Random Forest Classifier

recall 점수가
현저히 낮음

1. 유효 특허 간의 연관성 있는 데이터 X ➔ **네트워크 데이터** 추가 필요
2. 데이터의 Imbalance 문제 ➔ **SMOTE*** 기법 사용

* SMOTE(synthetic minority oversampling technique): 비율이 낮은 분류의 데이터를 만들어내는 방법, 먼저 분류 개수가 적은 쪽의 데이터의 샘플을 취한 뒤 이 샘플의 k 최근접 이웃(k neighbor)을 찾는다. 그리고 현재 샘플과 이들 k개 이웃 간의 차이(difference)를 구하고, 이 차이에 0~1 사이의 임의의 값을 곱하여 원래 샘플에 더한다. 이렇게 만든 새로운 샘플을 훈련 데이터에 추가한다. 결과적으로 SMOTE는 기존의 샘플을 주변의 이웃을 고려해 약간씩 이동시킨 점들을 추가하는 방식으로 동작한다.

2차 실험 방법 및 결과

2차 - SMOTE를 사용하여 데이터 imbalance 문제 해결

- 2차 실험 방법

1. 특허등록번호로 parsing한 **assignee, application citation, grant citation** 등의 **네트워크 데이터** 전처리 및 가공 후 전체 데이터 set에 merge
2. 전체 데이터를 7:3으로 Training, Test set 분리
3. **SMOTE 기법**을 활용하여 Training set의 유효 특허 데이터를 Garbage 특허 데이터의 개수와 동일하게 생성
4. 머신러닝 모델 학습

- 2차 실험 결과

	precision	recall	f1-score
0	0.99	0.99	0.99
1	0.36	0.57	0.44
accuracy		0.98	

Decision Tree

	precision	recall	f1-score
0	1.00	1.00	1.00
1	1.00	0.71	0.83
accuracy		1.00	

Random Forest Classifier

네트워크 데이터 추가, SMOTE 기법 사용 후 recall이 50% 이상

- Text 데이터를 추가 ➔ Recall 점수가 향상될 것이라고 예상

3차 실험 방법 및 결과

3차 - Text 데이터를 전처리 후 전체 데이터에 추가한 뒤 학습

- 3차 실험 방법

1. 특허등록번호로 parsing한 assignee, application citation, grant citation 등의 네트워크 데이터 전처리 및 가공 후 전체 데이터 set에 merge
2. **Text 데이터** (Abstract, Claim, Title)의 **TF-IDF** 값을 구한 뒤 전체 데이터 set에 merge
3. 전체 데이터를 7:3으로 Training, Test set 분리
4. SMOTE 기법을 활용하여 Training set의 유효 특허 데이터를 Garbage 특허 데이터의 개수와 동일하게 생성
5. 머신러닝 모델 학습

- 3차 실험 결과

	precision	recall	f1-score
0	1.00	0.96	0.98
1	0.19	0.71	0.30
accuracy		0.95	

Decision Tree

	precision	recall	f1-score
0	1.00	1.00	1.00
1	1.00	0.86	0.92
accuracy		0.99	

Random Forest Classifier

Text 데이터 추가 시
recall 향상

실험 결과 및 결론

모든 실험에 대한 결과 정리와 결론, 개선사항

data	model	accuracy	precision	recall
only preprocessed data (without text data (abstract, claim, title))	Logistic Regression	0.98	0.00	0.00
	SVM	0.98	0.00	0.00
	Decision Tree	0.99	0.67	0.29
	Random Forest	0.99	1.00	**0.57**
preprocessed data with network data (without text) + SMOTE	Logistic Regression	0.87	0.02	0.14
	SVM	0.70	0.02	0.43
	Decision Tree	0.98	0.36	0.57
	Random Forest	1.00	1.00	**0.71**
preprocessed data with network data and text data(TF-IDF) + SMOTE	Decision Tree	0.95	0.19	0.71
	Random Forest	0.99	1.00	**0.86**

- 결론

1. 데이터 imbalance의 문제는 **SMOTE 기법**을 사용하여 해결
2. 특허의 서지적 데이터 뿐 아니라 **네트워크 데이터, text 데이터**를 학습 데이터에 포함하면 예측 모델의 recall 점수가 향상
3. 머신러닝 모델 중 **tree 계열의 모델** 성능이 우수

- 개선방향

1. 학습 속도 단축을 위해 데이터 set의 size를 줄이는 작업 필요

   ex) embedding, 차원축소 등
2. 머신러닝 모델 뿐만 아니라 딥러닝 모델 확장 가능
3. 기계번역 모델을 활용하면 US 특허 뿐만 아니라 KR, JP, EP 등 여러 국가의 유효 특허 분류기로 확장 가능

감사합니다

이 성과는 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임
(No. NRF-2015R1C1A1A01056185).