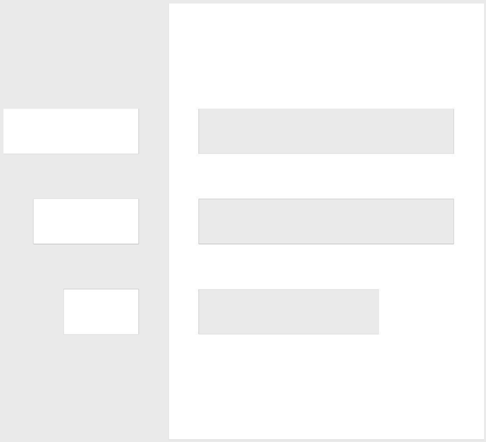
IPRally

Semi-Supervised Learning Methods for Patent Classification Using Search-Optimized Graph-Based Representations

Ekaterina Kotliarova and Sebastian Björkqvist

PatentSemTech’24

Goal



➔ Perform patent classification <u>quickly</u> in a <u>low-data environment</u>

- **Quickly** = training a model takes < 10 seconds, inference takes < 2 seconds.
- **Low-data environment** = Classification should work well with 10-20 samples per class.

Why?

IPRally

1

Manual labeling of patent documents is time-consuming.

Thus there is often only a small number of labeled documents available.

2

The classes might not map to existing IPC and CPC classes.

This means a custom classifier is required.

3

Enabling fast training of classifiers allows user to do quick iteration.

The model can be used to aid to label more data.

Our approach - PatentSemTech’23

1. Utilize pre-trained embeddings, optimized for patent search, as input for classifier.
2. Allow user to provide labels for a small set of documents, according to their own taxonomy.
3. Train a light-weight classifier using the embeddings and the given labels.
4. Allow the user to label more data utilizing the initial classifier, and repeat the training.

PatentSemTech’23 paper

New addition to our approach

IPRally

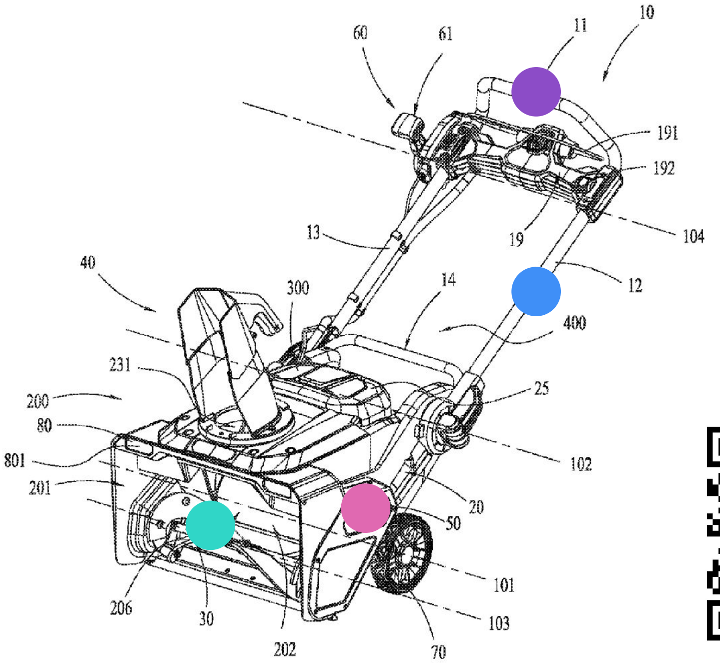
1. Utilize pre-trained embeddings, optimized for patent search, as input for classifier.
2. Allow user to provide labels for a small set of documents, according to their own taxonomy.
3. **Augment user-provided training data with other patent documents in a semi-supervised manner.**
4. Train a light-weight classifier using the embeddings and the given labels.
5. Allow the user to label more data utilizing the initial classifier, and repeat the training.

Patents as graphs

IPRally

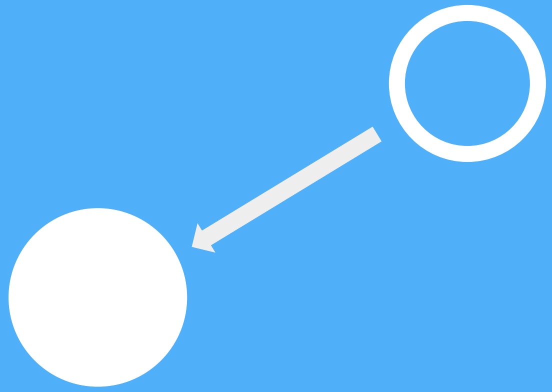
A snowthrower comprising a motor, an auger driven by the motor to rotate, a handle device for a user to operate, an auger housing for containing the auger and a frame for connecting the handle device and the auger housing, wherein the auger housing is made of at least two different materials.

- snowthrower
  - motor
  - handle device
    - **handle device** for a user to operate
  - auger housing
    - auger
      - **auger** driven by **motor** to rotate
    - at least two different materials
  - frame
    - **frame** for connecting **handle device** and **auger housing**



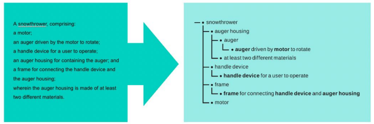
SIRIP’23 paper

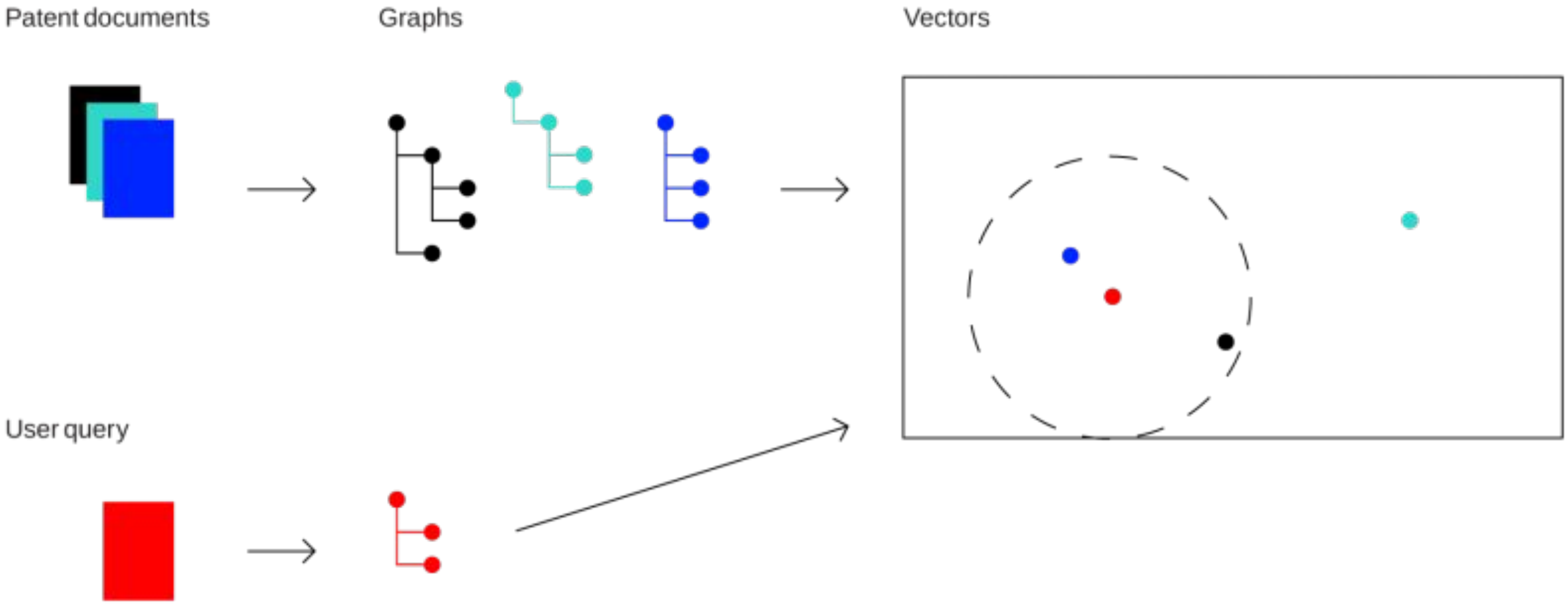
Training a model for search



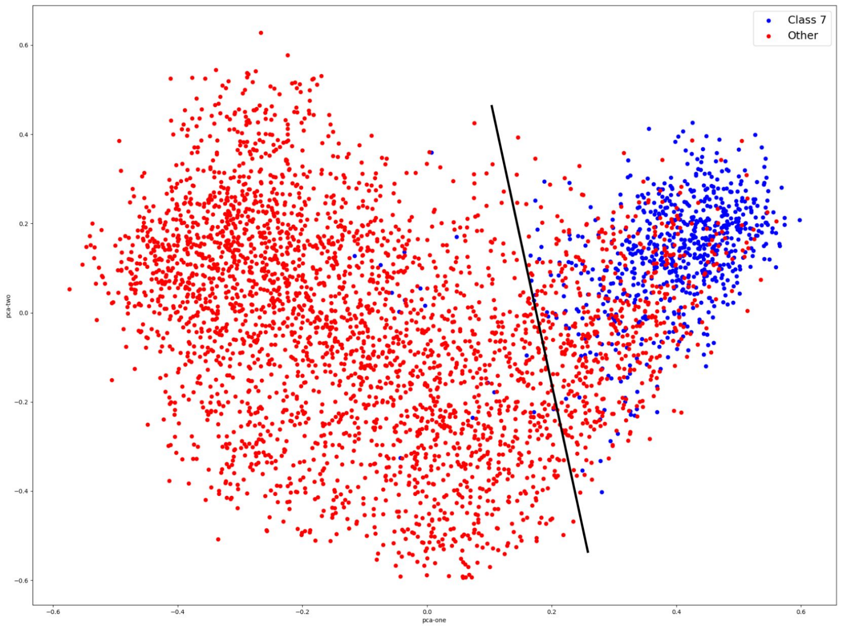
- We frame patent search as a metric learning problem.
- A graph neural network used to embed graphs to vectors.
- The model is trained in a supervised manner using patent office examiner citations.

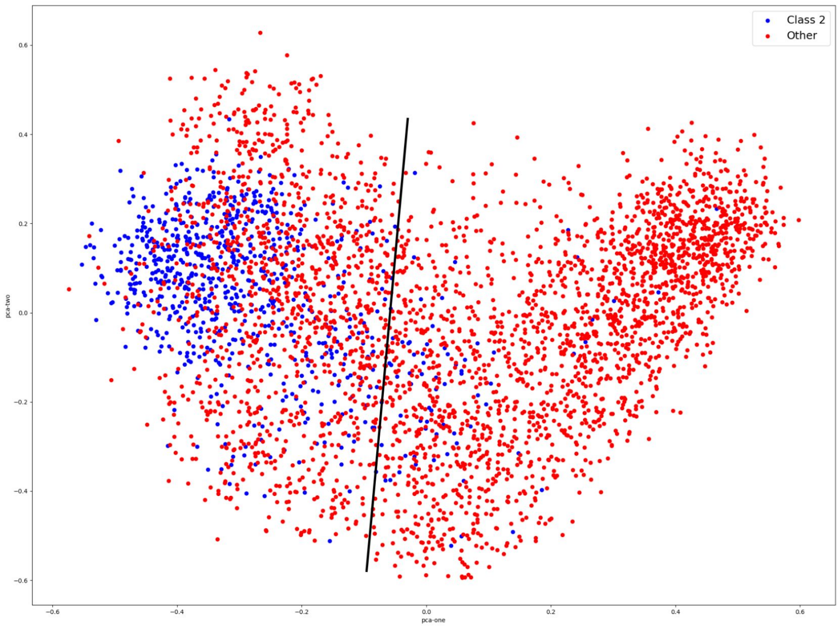
Deployment of search engine



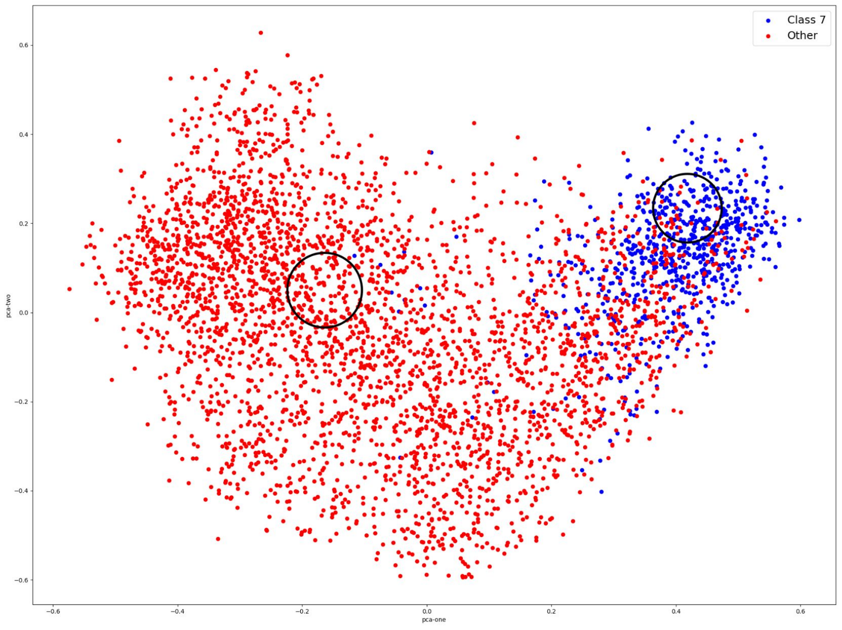


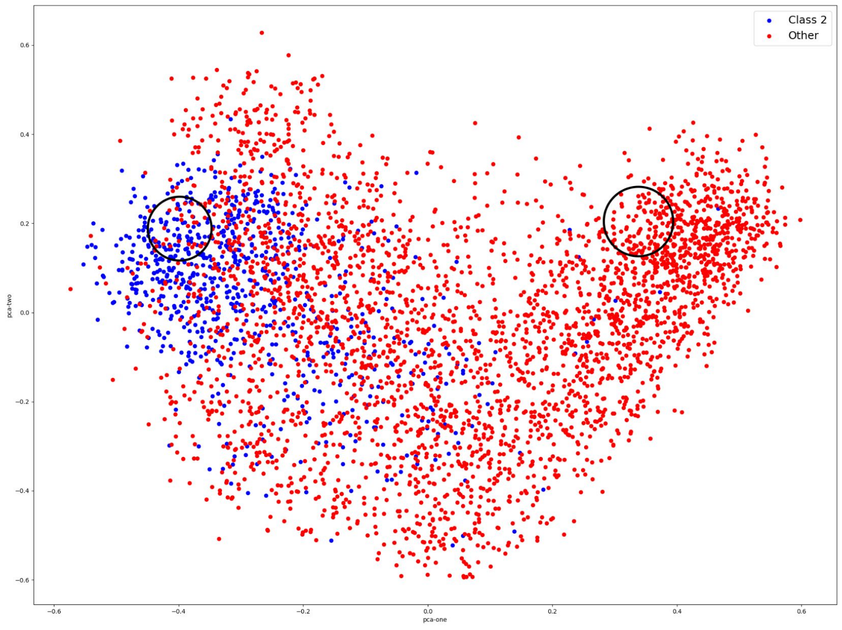
Training the classifiers - logistic regression





Training the classifiers - kNN





What if there's not enough samples?

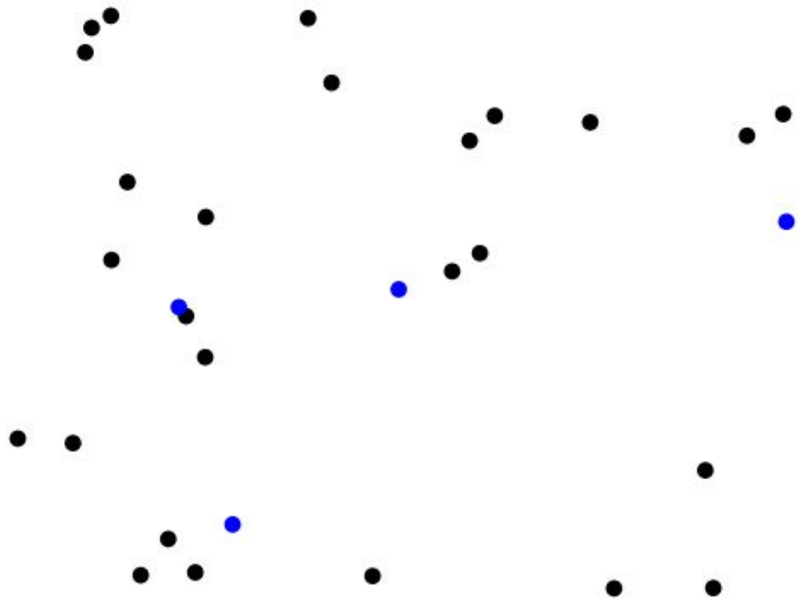
IPRally

- We have a great search engine for patents that will, given one document, find technically similar documents.
- If we have only a small amount of training data samples, can we use the search model to generate more samples to use in training?

Acquiring more labeled samples

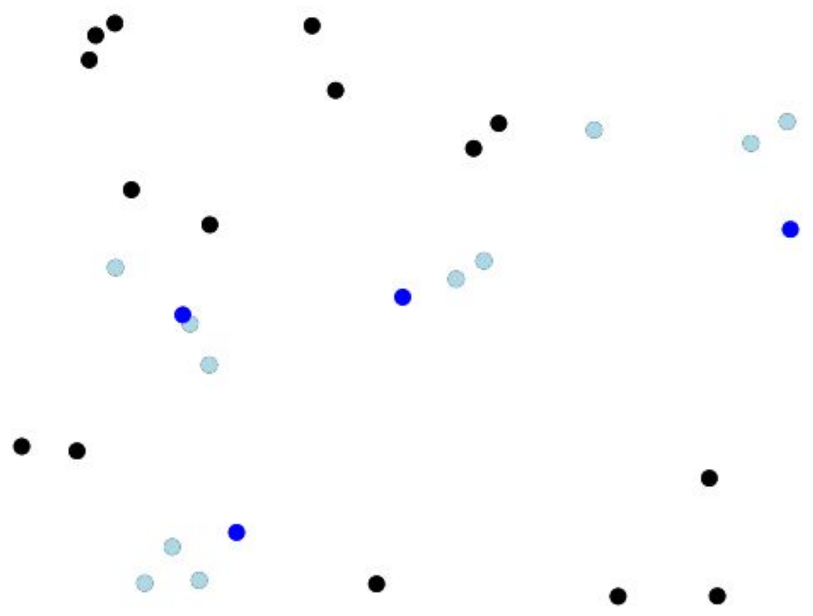
IPRally

Only a few labelled samples are available ⇨ The resulting classifier will be inaccurate.



Acquiring more labeled samples

Label more samples by selecting the nearest neighbors of existing samples!



Why does this work?

IPRally

1. The search engine is accurate ⇨ the nearest neighbors are technically similar to the original samples.
2. There is a large number of publications available ⇨ close neighbors are almost always available.

Evaluation

IPRally

- **Public datasets**
  - Quantum computing (Qubit) dataset (binary)
    - ~ 1400 unique patent families
    - Harris et al, WPI 61 (2020)
  - Cannabinoid edibles dataset (binary)
    - ~ 1600 unique patent families
    - https://github.com/swh/classification-gold-standard/

Evaluation

IPRally

➔ **Proprietary datasets**

- Mechanical engineering dataset (multi-label)
  - ~ 4700 unique patent families
  - 10 unique labels
- Chemical dataset (multi-label)
  - ~ 1300 unique patent families
  - 5 unique labels

Results - public datasets, logistic regression

IPRally

%	Cannabinoid edibles					Quantum computing				
	Semi-supervised		Up-sampling			Semi-supervised		Up-sampling		
	k=5	k=10	k=5	k=10	Baseline	k=5	k=10	k=5	k=10	Baseline
0.5	0.52	**0.59**	0.40	0.32	0.29	**0.68**	0.54	0.33	0.31	0.37
1	0.63	**0.64**	0.55	0.49	0.50	**0.64**	0.60	0.46	0.43	0.54
3	**0.77**	0.76	0.75	0.73	0.74	**0.81**	0.80	0.75	0.73	0.79
5	0.81	0.81	**0.81**	0.80	0.76	**0.84**	0.83	0.83	0.81	0.83
10	**0.85**	0.85	0.85	0.84	0.79	0.85	**0.87**	0.86	0.85	0.86
20	0.86	0.84	**0.87**	0.86	0.85	0.87	**0.87**	0.86	0.86	0.85
30	0.86	0.87	**0.88**	0.87	0.88	0.87	0.87	**0.88**	0.87	0.87
50	0.88	**0.88**	0.88	0.88	0.88	0.89	**0.89**	0.87	0.88	0.87
100	0.88	0.88	0.88	**0.89**	0.88	0.89	0.89	0.88	**0.90**	0.88

Metric: F1 score

Results - proprietary datasets, logistic regr.

IPRally

%	Mechanical engineering					Chemical				
	Semi-supervised		Up-sampling		Baseline	Semi-supervised		Up-sampling		Baseline
	k=5	k=10	k=5	k=10		k=5	k=10	k=5	k=10	
0.5	0.48	**0.48**	0.43	0.35	0.41	**0.21**	0.16	0.06	0.03	0.05
1	0.59	**0.60**	0.56	0.55	0.46	**0.34**	0.33	0.31	0.21	0.32
3	0.69	**0.70**	0.69	0.68	0.62	**0.45**	0.44	0.42	0.34	0.39
5	0.72	**0.73**	0.72	0.71	0.68	0.48	**0.49**	0.47	0.45	0.44
10	**0.74**	0.74	0.73	0.73	0.72	0.55	0.55	**0.55**	0.53	0.51
20	**0.75**	0.75	0.75	0.74	0.74	**0.59**	0.59	0.59	0.56	0.57
30	**0.76**	0.75	0.75	0.75	0.75	**0.62**	0.61	0.61	0.60	0.60
50	**0.76**	0.76	0.75	0.75	0.76	**0.63**	0.62	0.60	0.60	0.61
100	0.76	0.76	**0.76**	0.76	0.77	0.65	0.63	**0.65**	0.63	0.65

Metric: F1 score

Results in low-data environment - k=5, logistic regression

IPRally

- **On Qubit dataset**
  - Using **1% of original data** (9 samples) gets **71% of performance,** vs 61% without adding additional samples

- **On mech. eng. dataset**
  - Using **1% of original data** (37 samples) gets **76% of performance**, vs 60% without adding additional samples

Conclusion

IPRally

1

Search-optimized graph-based document representations capture enough information to use as input for classification.

2

The training data set can be augmented by searching for similar documents and labeling them according to the nearest neighbor.

3

This method is very well suited for low-data case, allowing training good classifiers with just some tens of samples.

Thank you!

sebastian@iprally.com

ipralley.com