Multilayer Networks for Scientific Field Classification

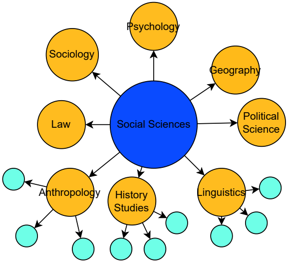
Gialitsis Nikolaos

SciNoBo: A Hierarchical Multi-Label Classifier of Scientific Publications
Gialitsis Nikos, Kotitsas Sotiris, Papageorgiou Haris
https://arxiv.org/abs/2204.00880

ngialitsis@iit.demokritos.gr

May 4, 2022

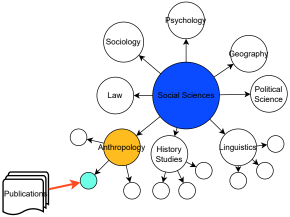
Field of Science Classification:What?



How to classify a publication according to a Field of Science (FoS) taxonomy?

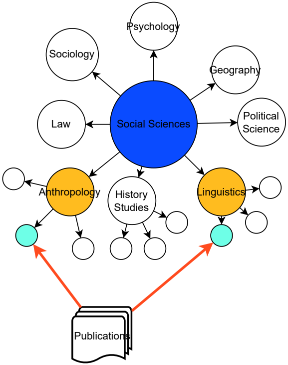
- All Science Journal Classification (ASJC) System
- International Classification of Diseases(ICD)
- Frascati Manual Classification
- Medical Subject Headings (MeSH)
- Web of Science (WoS) Categories and Subject Areas
- Scopus Subject Areas
- European Science Vocabulary (EuroSciVoc)
- Microsoft Academic Graph (MAG) Concepts

Field of Science Classification: Disciplinary



Publications may belong to a single path of the taxonomy

Field of Science Classification: Multidisciplinary



...or to multiple paths

Field of Science Classification: Why?

- Scientific Search Engines
- Recommendation Systems
- Monitoring of Research Output
- Normalization of Bibliometric Indicators
- Input to Machine Learning Models

Field of Science Classification: Disciplinary:How

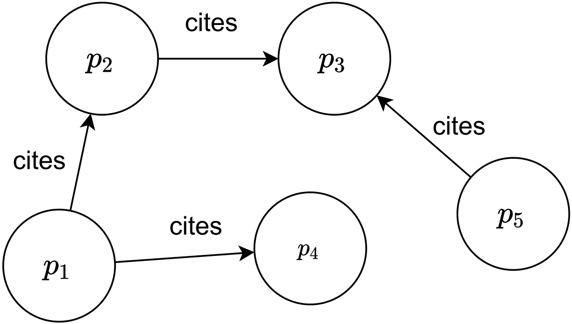
For classification, we can utilize:

- content: title, abstract, fulltext
- textual metadata
  - citation context
  - author-supplied keywords
  - affiliation names
  - venue names
- citations, references

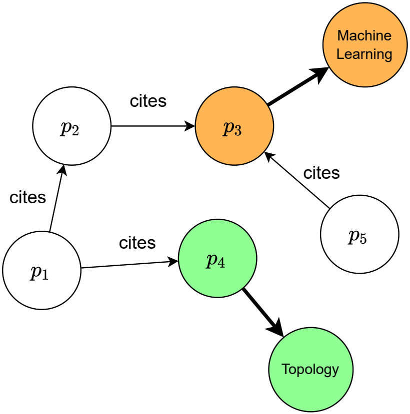
Limitations of Text-based Approaches

- Textual content is not always available
- Overlapping vocabulary among different classes
- Concept drift
- Discourse norms in specific fields
- Multilinguality

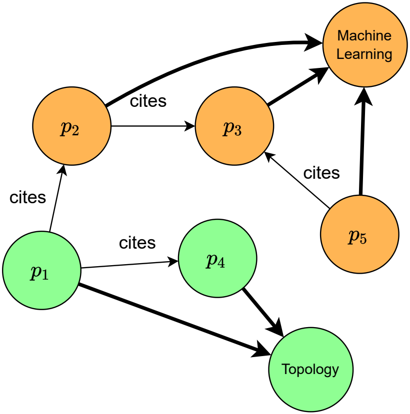
Utilizing Citation Networks



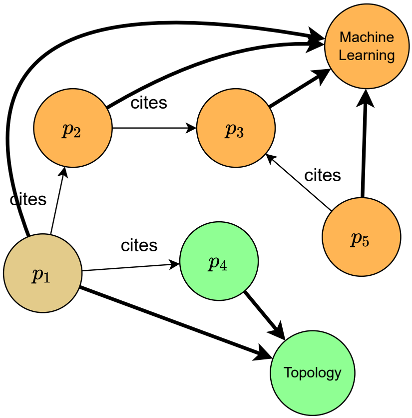
Label Propagation in Citation Networks - Round 0



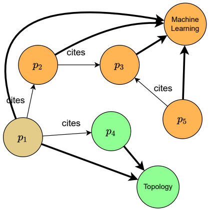
Label Propagation in Citation Networks - Round 1



Label Propagation in Citation Networks - Round 2

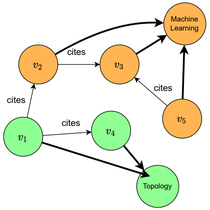


Limitation of Citation Networks



- Annotations are often not available at the publication level
- Volume and Growth Rate of Publications $\Rightarrow$ Constant Updates Needed
- All citations are weighted equally

A More Sane Approach: Venue-Venue Networks



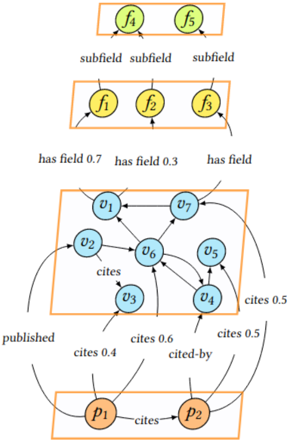
- a venue v_1 cites v_2 if v_1 has pubished works that cite works in v_2
- much more compact than publication-publication network
- venue deduplication can be performed to merge different instances of the same conference under one node

Using Venues as Intermediates

- Annotations are more easily curated at the Journal/Conference level (Venue-level)
- Multiple Venue classifications are available
  - ScienceMetrix Journal Classification

Main idea: Unite publications, venues and scientific fields in one network

Multilayer Networks



- Layer 1: $\{(p_i, p_j) : p_i \text{ cites } p_j\}$
- Layer 2: $\{(p, v) : p \text{ published in } v\}$
- Layer 3: $\{(v_i, v_j)\} : v_i$'s publications cite v_j publications
- Layer 4: $\{(v, f)) :$ publications in v are related to scientific field f
- Layer 5+: hieararchical relationships among fields

Label Propagation

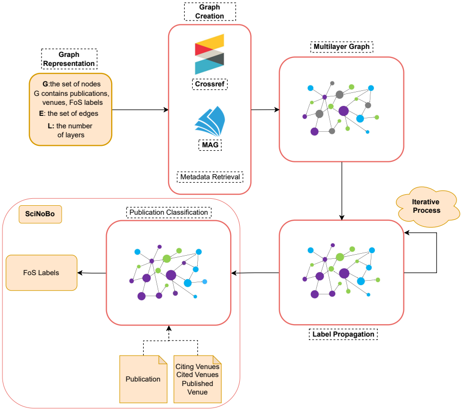
- Exploit network structure to link:
  - **venues** to **fields** through **venue-venue** relationships
  - **publications** to **fields** through **publication-publication** and **publication-venue** relationships
- similar intuition as in nearest neighbor classification

Inference Strategies in the Multilayer Network

After the venue-venue layer has been populated, we can infer the labels of individual publications in 3 main ways:

1. assign labels of publishing venues
2. assign labels of referencing venues
3. assign labels of citing venues + reference venues

Overview of Proposed System (SciNoBo)



Gialitsis et.al - SciNoBo: A Hierarchical Multi-Label Classifier of Scientific Publications

Field of Science Taxonomy

Labels according to the Field of Science Taxonomy

- Level 1 & Level 2 from OECD/FORD
- Level 3 extension through Science Metric labels

Levels of FoS	**Number of Labels**
Level 1	6
Level 2	42
Level 3	174

Graph Creation (1)

- Multilayer Graph populated by exploiting Crossref and Microsoft Academic Graph
  - time span 2016-2021: publications + references
  - retrieve publications' venue(s)
  - insert publication nodes and venue nodes
- create directed weighted edges between venues
  - weight (v_i, v_j): number of publications in v_i citing publications in v_j
  - direction: citing/cited
- normalize all outgoing edge weights to sum up to 1
- deduplicate venues by mapping full names to abbreviations

Graph Creation (2)

- Insert scientific field nodes from the Field of Science Taxonomy
- A small portion of venues have known L3 labels (ScienceMetrix classification)
- add edges between venue nodes and L3 scientific field nodes

Label Propagation in the Multilayer Network

Propagation ⇒ most venues are assigned to fields of science after two rounds.

Nodes	Edges	Pre Label Propagation Lvl1-Lvl2-Lvl3			Post Label Propagation Lvl1-Lvl2-Lvl3		
94752	3112953	N/A	84	14889	32049	32049	32324

Dataset for Evaluation

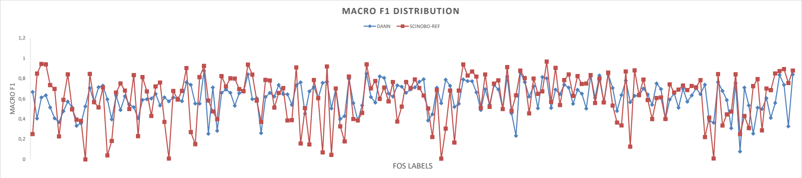
- Dataset: collect samples from Journals annotated by ScienceMetrix
  - 500 publications per Journal and 500 publications per Scientific Field
  - sampled from Microsoft Academic Graph

Statistics	Train Set	Development Set	Test Set	Total
Number of Instances	1687826	120282	120307	1928415

Table: Statistics of the dataset used for training and evaluating our methods.

Comparison to a Deep Attentive Neural Network

We compared against the Deep Attentive Neural Network architecture (DANN) proposed by Kandimala et.al for single-labeled classification trained on abstracts.



Experimental Results

To cater for multidisciplinarity, we present results with two settings:

1. TOP-1: Output only the most probable Field of Science label
2. TOP-2: Output the two most probable Field of Science labels

MODELS	MACRO-F1	WEIGHTED-MACRO-F1	MICRO-F1
DANN-TOP-1	0,4465 ± 0,003	0,5873 ± 0,002	0,4563 ± 0,002
DANN-TOP-2	0,6007 ± 0,008	0,7307 ± 0,002	0,6154 ± 0,003
SciNoBo-CITREF-TOP-1	0,4627 ± 0,0	0,4900 ± 0,0	0,4900 ± 0,0
SciNoBo-CITREF-TOP-2	0.6000 ± 0,0	0,6208 ± 0,0	0,6177 ± 0,0
SciNoBo-REF-TOP-1	0,4781 ± 0,0	0,5000 ± 0,0	0,5022 ± 0,0
SciNoBo-REF-TOP-2	0,6190 ± 0,0	0,6277 ± 0,0	0,6205 ± 0,0
SciNoBo-PUB-TOP-1	**0,7303 ± 0,0**	**0,7309 ± 0,0**	**0,7503 ± 0,0**
SciNoBo-PUB-TOP-2	**0,8200 ± 0,0**	**0,8223 ± 0,0**	**0,8270 ± 0,0**

Conclusions

We presented SciNoBo for a multi-label hierarchical classification. SciNoBo :

- classifies with minimal metadata
- takes into account context through citing relationships
- accounts for multidisciplinarity as more than one labels can be assigned
- is competitive even against a deep-learning method

Work in Progress

There are plans to:

- create human curated publication-level dataset to allow for more accurate evaluations
- extend the proposed Taxonomy to Levels 4,5 and 6 utilizing community detection and neural topic modeling
- replace label propagation with message passing in a Graph Neural Network
- replace MAG with OpenAlex as it has been discontinued

Questions

Thank you. Let's discuss.