meng_2020_hierarchical_topic_mining_via_joint_spherical_tree_and_text_embedding

Year

2020

Author(s)

Meng, Yu and Zhang, Yunyi and Huang, Jiaxin and Zhang, Yu and Zhang, Chao and Han, Jiawei

Title

Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding

Venue

KDD

Topic labeling

Manual

Focus

Primary

Type of contribution

Novel approach

Underlying technique

Manual labeling

Topic labeling parameters

Dataset 1
(NYT)

Nr of super-categories in the hierarchy: 8
Nr of sub-categories in the hierarchy: 12

Dataset 2
(arXiv)
Nr of super-categories in the hierarchy: 3
Nr of sub-categories in the hierarchy: 29

Table 1: Dataset statistics.

Corpus	# super-categories	# sub-categories	# documents
NYT	8	12	89,768
arXiv	3	29	230,105

Label generation

Topic labels act a **starting point** to define the structure of the topic hierarchy. Hierarchical Topic Mining takes only a topic hierarchy described by category names as user guidance, and aims to retrieve a set of coherent and representative terms under each category to help users comprehend his/her interested topics.

In this context:

- Gold-standard label (hierarchies) taken from datasets and used to evaluate the model.
- Generally dependent on end user when the model is utilised outside of this work.

Motivation

Contrary to the traditional method of assigning labels to generated topics, this work proposes a weekly supervised model to generate topic (hierarchies) starting from a (user provided) label hierarchy and a corpus of documents.

Topic modeling

(Guided / Weekly supervised) Hierarchical topic discovery (joint trees with text embeddings method)

Topic modeling parameters

- Embedding dimension p: 100
- Local context window size h: 5
- Number of representative terms to retrieve per category K: 5
- Learning rate α: 0.025 initially with linear decay

Nr. of topics

Dataset 1

(NYT): 30

Dataset 2

(arXiv): 32

Label

Single or multi-word label provided by the user or contained in the testing dataset.

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Topic hierarchies

Dataset: News and Scientific Literature

Problem statement

Proposing a Hierarchical Topic Mining task to guide the hierarchical topic discovery process with minimal user supervision, which takes a category tree described by category names only, and aims to mine a set of representative terms for each category from a text corpus to help a user comprehend topics.
Proposing a novel joint tree and text embedding method along with a principled optimization procedure that allows simultaneous modeling of the category tree structure and the corpus generative process in the spherical space for effective category-representative term discovery.

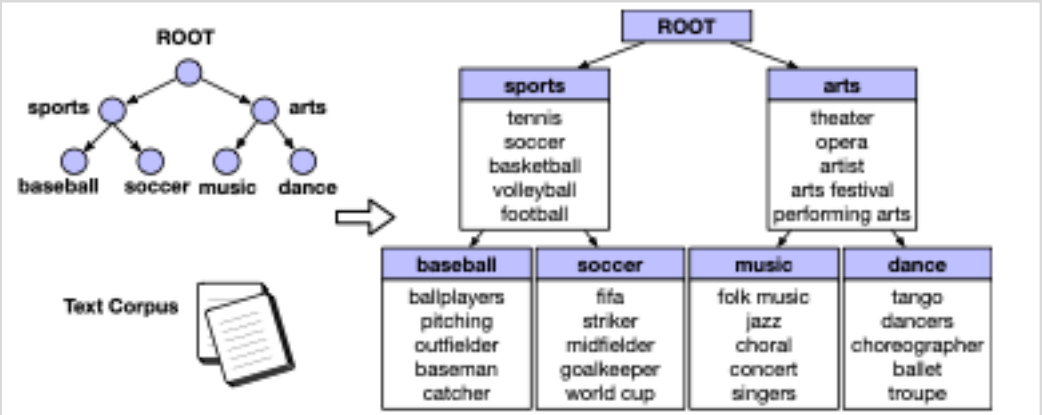

Figure 1: An example of Hierarchical Topic Mining. We aim to retrieve a set of representative terms from a given corpus for each category in a user-provided hierarchy.

Corpus

Dataset 1

Origin: New York Times

Nr. of documents: 89,768

Details: The New York Times annotated corpus (NYT)

Dataset 2

Origin: arXiv

Nr. of documents: 230,105

Details: arXiv paper abstracts

Both datasets are annotated with ground-truth category hierarchies

Document

Dataset 1

(NYT)
A single news article from The New York Times

Dataset 2
(arXiv)
A single publication hosted in the arXiv database
Data crawled from https://arxiv.org/.

Pre-processing

Infrequent words that appear less than 5 times are discarded
AutoPhrase is used to extract quality phrases, which are treated as single words during embedding training.

```
@inproceedings{meng_2020_hierarchical_topic_mining_via_joint_spherical_tree_and
_text_embedding,
author = {Meng, Yu and Zhang, Yunyi and Huang, Jiaxin and Zhang, Yu and Zhang,
Chao and Han, Jiawei},
title = {Hierarchical Topic Mining via Joint Spherical Tree and Text
Embedding},
year = {2020},
isbn = {9781450379984},
publisher = {Association for Computing Machinery},
address = {New York, NY, USA},
url = {https://doi.org/10.1145/3394486.3403242},
doi = {10.1145/3394486.3403242},
abstract = {Mining a set of meaningful topics organized into a hierarchy is
intuitively appealing since topic correlations are ubiquitous in massive text
corpora. To account for potential hierarchical topic structures, hierarchical
topic models generalize flat topic models by incorporating latent topic
hierarchies into their generative modeling process. However, due to their
purely unsupervised nature, the learned topic hierarchy often deviates from
users' particular needs or interests. To guide the hierarchical topic discovery
process with minimal user supervision, we propose a new task, Hierarchical
Topic Mining, which takes a category tree described by category names only, and
aims to mine a set of representative terms for each category from a text corpus
```

```
to help a user comprehend his/her interested topics. We develop a novel joint tree and text embedding method along with a principled optimization procedure that allows simultaneous modeling of the category tree structure and the corpus generative process in the spherical space for effective category-representative term discovery. Our comprehensive experiments show that our model, named JoSH, mines a high-quality set of hierarchical topics with high efficiency and benefits weakly-supervised hierarchical text classification tasks.},
booktitle = {Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &amp;amp; Data Mining},
pages = {1908–1917},
numpages = {10},
keywords = {text embedding, topic hierarchy, tree embedding, topic mining},
location = {Virtual Event, CA, USA},
series = {KDD '20}
}
```

#Thesis/Papers/Initial