zhang_2018_taxogen_unsupervised_topic_taxonomy_construction_by_adaptive_term_embedding_and_clustering

Year

2018

Author(s)

Zhang, Chao and Tao, Fangbo and Chen, Xiusi and Shen, Jiaming and Jiang, Meng and Sadler, Brian and Vanni, Michelle and Han, Jiawei

Title

TaxoGen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering

Venue

KDD

Topic labeling

Fully automated

Focus

Secondary

Type of contribution

Novel approach

Underlying technique

TF-IDF combined with topic taxonomy analysis

Topic labeling parameters

\

Label generation

A topic label is selected by identifying the **most representative term** belonging to the topic (which in this context is part of a taxonomy).
A representative term for a topic should appear frequently in it but not in the sibling topics (i.e. the other topics belonging to the same level in the taxonomy).

Term representativeness is measured using the documents that belong to the topic. The TF-IDF scheme is used to obtain the documents belonging to each topic. With these documents, the following two factors are considered for computing the representativeness of a term t for topic Sk :

- **Popularity**: A representative term for Sk should appear frequently in the documents of Sk .
- **Concentration**: A representative term for Sk should be much more relevant to Sk compared to the sibling topics of Sk

To combine the above two factors, it is noted that they should have conjunctive conditions, namely a representative term should be both popular and concentrated for Sk . Thus, the representativeness of term t for topic Sk is defined as:

$$r(t, S_k) = \sqrt{pop(t, S_k) \cdot con(t, S_k)}$$

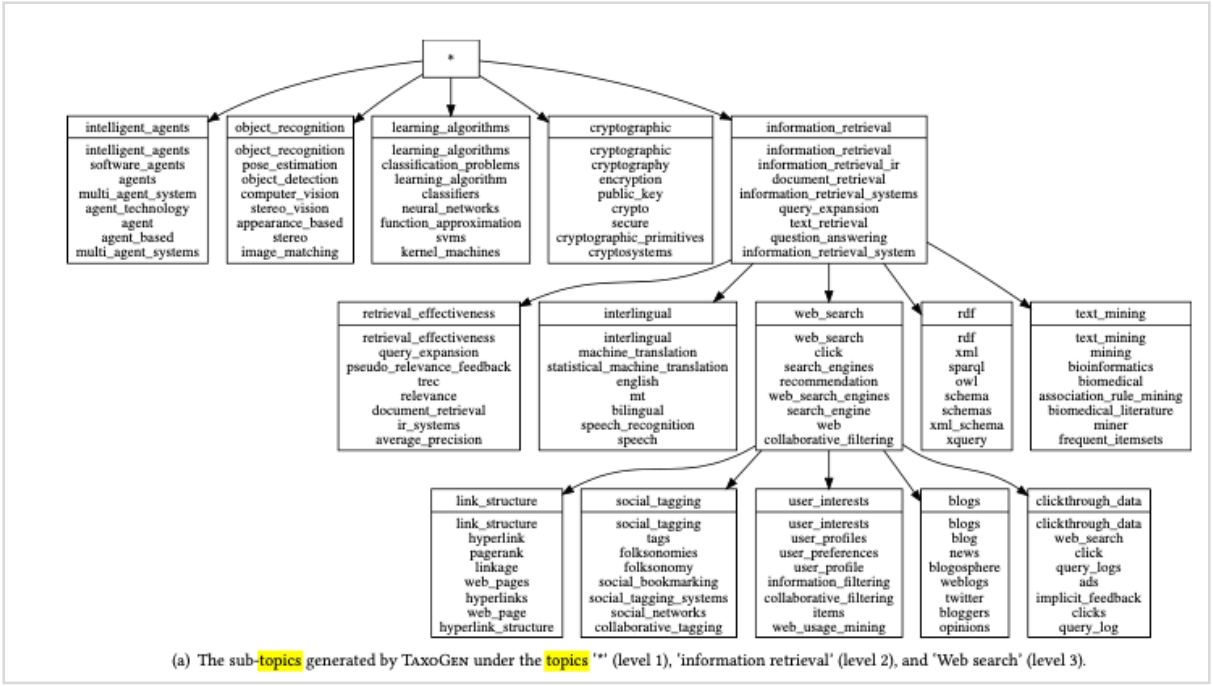
where pop(t,Sk) and con(t,Sk) are the popularity and concentration scores of t for Sk .
Let Dk denote the documents belonging to Sk , we define pop (t , Sk) as the normalised frequency of t in Dk:

$$pop(t, S_k) = \frac{\log(tf(t, \mathcal{D}_k) + 1)}{\log tf(\mathcal{D}_k)},$$

where tf(t, Dk) is number of occurrences of term t in Dk , and tf(Dk) is the total number of tokens in Dk.
To compute the concentration score, we first form a pseudo document Dk for each sub-topic Sk by concatenating all the documents in Dk.
Then we define the concentration of term t on Sk based on its relevance to the pseudo document Dk :

$$con(t, S_k) = \frac{\exp(rel(t, D_k))}{1 + \sum_{1 \le j \le K} \exp(rel(t, D_j))},$$

where rel(p,Dk) is the Okapi BM25 relevance of term t to the pseudo document Dk .



(a) The sub-topics generated by TaxoGen under the topics '*' (level 1), 'information retrieval' (level 2), and 'Web search' (level 3).

Motivation

Simplify the navigation of a topic taxonomy by highlighting the most representative term in each topic.

Topic modeling

Unsupervised topic taxonomy construction method by

Unsupervised (recursive) hierarchical clustering (with term embeddings)

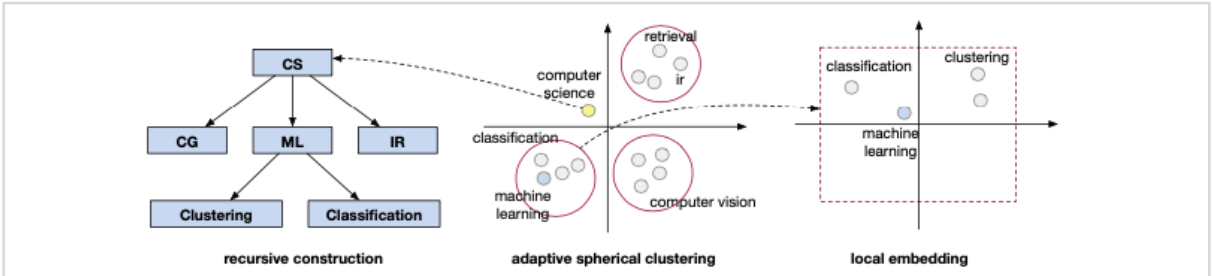


Figure 2: An overview of TaxoGen. It uses term embeddings to construct the taxonomy in a top-down manner, with two novel components for ensuring the quality of the resursive process: (1) an adaptive clustering module that allocates terms to proper topic nodes; and (2) a local embedding module for learning term embeddings on topic-relevant documents.

Topic modeling parameters

Number K for splitting a coarse topic: 5

Representativeness threshold δ for identifying general terms: 0.25 for DBLP and 0.15 for SP

Nr. of topics

Four level taxonomy on DBLP, each parent topic is split into five child topics.
Three-level taxonomy on SP.

Label

Single or multi word label (spaces in multi-word labels are replaced with underscores).

Label selection

The (single) most representative term in the topic (see “label generation”) for more details.

Label quality evaluation

No formal quality evaluation.
The only relevant statement is: “We find those labels are of good quality and precisely summarise the major research areas covered by the DBLP corpus. The only minor flaw for the five labels is ‘object recognition’, which is too specific for the computer vision area. The reason is probably be- cause the term ‘object recognition’ is too popular in the titles of computer vision papers, thus attracting the centre of the spherical cluster towards itself. ”

Assessors

\

Domain

Paper: Topic taxonomies
Dataset: Computer science research

Problem statement

Proposing an unsupervised method for constructing topic taxonomies, wherein every node represents a conceptual topic and is defined as a cluster of semantically coherent concept terms.
The proposed method uses term embeddings and hierarchical clustering to construct a topic taxonomy in a recursive fashion and it consists of

- A clustering module for allocating terms to proper levels when splitting a coarse topic into fine-grained ones
- An embedding module for learning term embeddings that maintain strong discriminative power at different levels of the taxonomy.

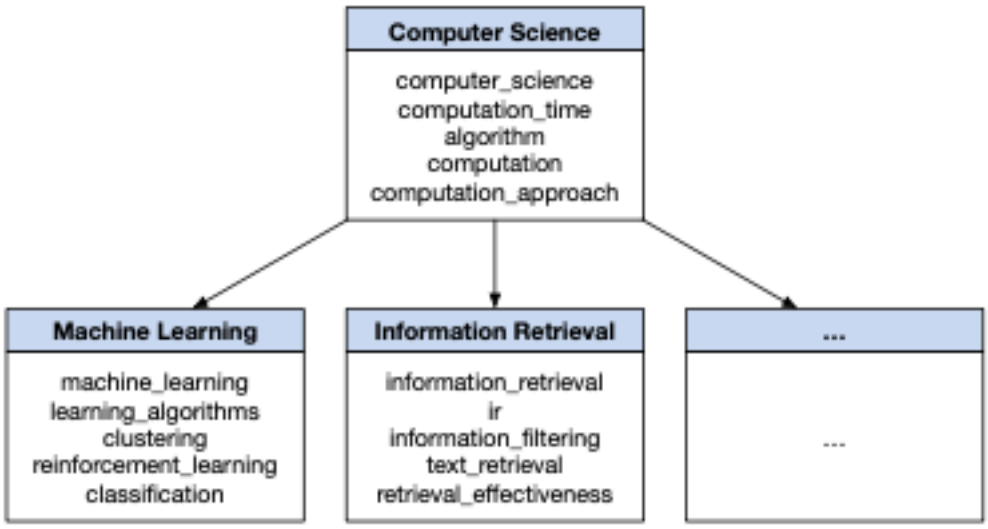


Figure 1: An example topic taxonomy. Each node is a cluster of semantically coherent concept terms representing a conceptual topic.

Corpus

Dataset 1

Origin: DBLP

Nr. of documents: 1,889,656

Details: Titles of computer science papers from the areas of information retrieval, computer vision, robotics, security & network, and machine learning.

Dataset 2

Origin:

Nr. of documents: 94,476

Details: Paper abstracts from the area of signal processing

Document

Dataset 1

Paper title

Dataset 2

Paper abstract

Pre-processing

Dataset 1

An existing NP chunker is applied on paper titles to extract all the noun phrases and then remove infrequent ones to form the term set, resulting in 13,345 distinct terms.

Dataset 2

All the noun phrases are extracted from the abstracts to form the term set and obtain 6,982 different terms.

```
@inproceedings{zhang_2018_taxogen_unsupervised_topic_taxonomy_construction_by_adaptive_term_embedding_and_clustering,
author = {Zhang, Chao and Tao, Fangbo and Chen, Xiusi and Shen, Jiaming and Jiang, Meng and Sadler, Brian and Vanni, Michelle and Han, Jiawei},
title = {TaxoGen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering},
year = {2018},
isbn = {9781450355520},
publisher = {Association for Computing Machinery},
address = {New York, NY, USA},
url = {https://doi.org/10.1145/3219819.3220064},
doi = {10.1145/3219819.3220064},
abstract = {Taxonomy construction is not only a fundamental task for semantic analysis of text corpora, but also an important step for applications such as information filtering, recommendation, and Web search. Existing pattern-based methods extract hypernym-hyponym term pairs and then organize these pairs into a taxonomy. However, by considering each term as an independent concept node, they overlook the topical proximity and the semantic correlations among terms. In this paper, we propose a method for constructing topic taxonomies, wherein every node represents a conceptual topic and is defined as a cluster of semantically coherent concept terms. Our method, TaxoGen, uses term embeddings and hierarchical clustering to construct a topic taxonomy in a recursive fashion. To ensure the quality of the recursive process, it consists of: (1) an adaptive spherical clustering module for allocating terms to proper levels when splitting a coarse topic into fine-grained ones; (2) a local embedding module
```

```
for learning term embeddings that maintain strong discriminative power at
different levels of the taxonomy. Our experiments on two real datasets
demonstrate the effectiveness of TaxoGen compared with baseline methods.},
booktitle = {Proceedings of the 24th ACM SIGKDD International Conference on
Knowledge Discovery &amp;amp; Data Mining},
pages = {2701–2709},
numpages = {9},
keywords = {text mining, taxonomy construction, word embedding},
location = {London, United Kingdom},
series = {KDD '18}
}
```

#Thesis/Papers/Initial