Towards Enhanced Topic Discovery on Semantic Maps for Biomedical Literature Exploration

Bin Choi*
Yale-NUS College

Brian Ondov†
Department of Biomedical Informatics and Data Science
Yale University

Huan He‡
Department of Biomedical Informatics and Data Science
Yale University

Hua Xu§
Department of Biomedical Informatics and Data Science
Yale University

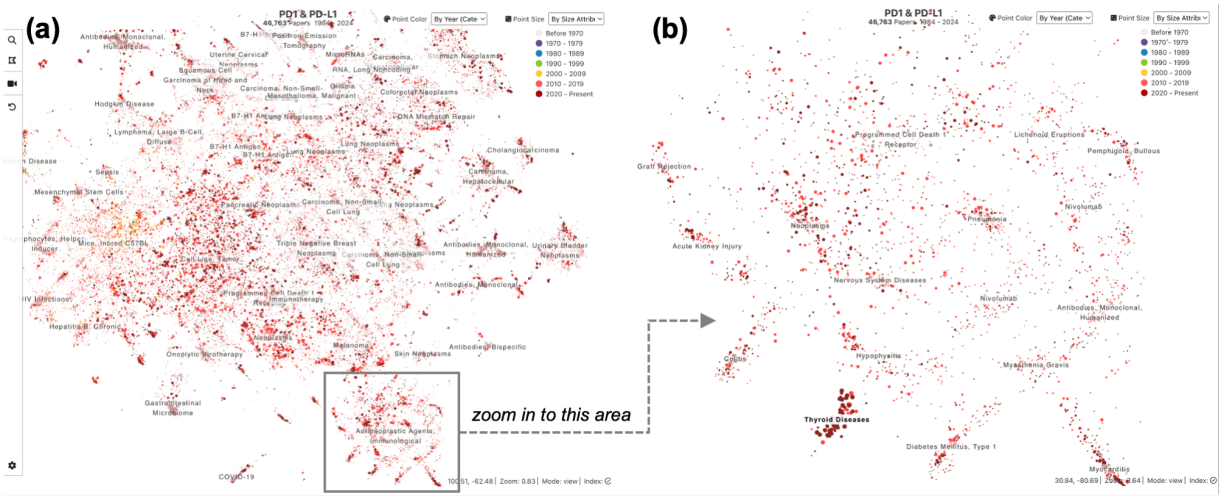

Figure 1: The screenshot of the labels generated for the semantic map, showing (a) the topic labels of top-level clusters on the overall map, and (b) the detailed topic labels of a zoomed-in area.

Abstract

The rapid growth of biomedical research has led to an overwhelming volume of literature, making it challenging for researchers to efficiently explore and analyze. While existing tools provide an overview of semantic maps and publication distributions, further refinement is needed to reveal fine-grained nuances and hierarchical topics. To address this, we propose a novel method for hierarchical topic modeling and label generation on 2D semantic maps. Our approach consists of three steps. First, we apply density-based hierarchical clustering using HDBSCAN to construct a topic tree. Second, we employ a novel tree-based TF-IDF method to refine topic representation using MeSH terms, capturing both general and local topic distinctions. Finally, we optimize label positioning using a centroid-based method to enhance visualization.

Index Terms: Topic modeling, semantic map, tree-based TF-IDF.

1 Introduction

The rapid growth of biomedical research has resulted in an exponential increase in the volume of publications, making it essential for researchers to utilize advanced tools to efficiently explore and analyze this vast body of literature [3]. In recent years, literature-mining tools have gained significant attention, helping researchers extract valuable insights and information from large collections of publications. Among these tools, the application of large language models (LLMs) and text embedding techniques has become increasingly pivotal [6]. By converting complex scientific texts into unified high-dimensional vectors, LLMs enable the use of machine learning and data mining techniques to facilitate data analysis and visualization. However, it is still challenging to reveal insights and nuances from large-scale literatures and their high-dimensional vectors, particularly in clustering and topic representation, which are crucial for understanding the topical landscape. In an effective semantic map, closely located embeddings should be semantically related, enabling users to grasp the thematic structure of the map. While existing methods such as BERTopic [4] and Top2Vec [1] offer embedding-based topic modeling, they tend to operate at a fixed level of generality and may not fully capture the fine-grained nuances and hierarchical structures present in large-scale literature collections. To address this, we propose a novel methodology that can capture both general and specific topics across different levels of semantic maps for biomedical publications, allowing for a more interactive and detailed exploration of the topics in clusters and sub-clusters.

2 Methodology

Our proposed method takes a 2D semantic map as input, which is generated from a collection of publications using LLM embedding and dimensional reduction techniques. In this semantic map, each publication is mapped to a point on the map using its 2D embeddings, which are generated using transformer-based language model [2] (BAAI/bge-small-en-v1.5) and t-distributed stochastic neighbor embedding (tSNE) [7] to maintain meaningful spatial relationships. Then, our method conducts three phases to visualize the topic labels for this semantic map: (1) clustering, (2) topic representation, and (3) label positioning and data file generation. The details of each phase are as follows.

*e-mail: binchoi@u.yale-nus.edu.sg
†e-mail: brian.ondov@yale.edu
‡e-mail: huan.he@yale.edu
§e-mail: hua.xu@yale.edu

2.1 First phase: density-based hierarchical clustering

In this phase, we identify groups of semantically similar papers by analyzing their proximity in the semantic space. This process is conducted in a hierarchical manner, beginning with fine-grained clusters, and expanding to identify more general clusters. To capture the hierarchical structure of the semantic map, we propose using a topic tree, which organizes and represent topics as a tree structure within the semantic space. Each node in the topic tree represents a cluster of publications. Child nodes represent sub-clusters, reflecting subtopics within the broader context of the parent cluster. In this bottom-up process, we leverage a density-based clustering algorithm, HDBSCAN [5], which does not require pre-specifying the number of clusters and excels in handling outliers, to assign relevant documents to clusters.

Initially, we use heuristically defined parameters to identify fine-grained clusters. Then, the algorithm's parameters are incrementally adjusted to identify larger, coarser-grained clusters. As each new, higher-level of clusters are identified, we assign pre-identified sub-clusters to new 'parent' clusters based on membership conditions, such as an overlap threshold. Some fine-grained clusters may not map to a parent cluster and remain as root nodes, representing the highest-level view of their topic. This process continues until an appropriate clustering for the highest-level view of the semantic space is achieved.

2.2 Second phase: topic representation based on MeSH terms, c-TF-IDF, and tree-based TF-IDF

As MeSH terms offer comprehensive coverage of biomedical topics and are organized hierarchically, we use them for topic representation of clusters. Through the first phase, we have identified clusters of publications that represent the semantic/topics at varying levels of granularity. At the highest level, we can determine the major topics present on the semantic map through what BERTopic coins as class-based TF-IDF (c-TF-IDF) [4]. In our adaptation, we compile the MeSH terms for each document within a cluster and compare them with those of other clusters to identify the most relevant and representative MeSH term for each cluster.

While this approach works well for representing topics at the root nodes of the topic forest, it struggles to distinguish local differences among sibling sub-clusters further down the tree, often resulting in non-discriminative topics. To address this, we propose a novel tree-based TF-IDF procedure that applies c-TF-IDF only within sibling (or relative) sub-clusters. This method more effectively identifies local differences and nuances among sibling subclusters, as confirmed by our experiments.

2.3 Third phase: label positioning and data file generation

In the final phase, we focus on positioning the generated topic labels on the 2D semantic map and producing structured data outputs to support further analysis and visualization. After determining the representative topic for each cluster from the second phase, we employ a combination of centroid-based and hierarchical methods for label placement. Each topic label is initially positioned at the centroid of its corresponding cluster to reflect the geometric center of the documents within that cluster, and we assigned a zoom level to each level. Once label positions are decided, we generate a JSON format data file for rendering the labeled semantic map. The file includes the cluster membership of each document, the hierarchical structure of clusters, and the final positions of the topic labels.

3 Discussion and Future Work

To validate our method, we implemented the above algorithms in Python and developed a prototype based on our existing visualization system using three.js and WebGL techniques. As shown in Figure 1, we generated topic labels on a collection of 46,763 papers related programmed cell death 1 (PD1) and visualized them on the semantic map of the paper collection (Fig. 1a). While users zoom in on this semantic map, labels of fine-grained clusters will appear (Fig. 1b). During the development of our method, we demonstrated our prototype to domain experts of immunotherapy and got positive feedback. They commented that the labels are intuitive, but some labels are visually overlapped. In addition, some clusters share the same labels despite having different topics.

As our next step, we plan to optimize our positioning algorithm by adopting density-aware method to adjust the label positions based on the local density and proximity of neighboring clusters. We are also going to improve the topic summarization using large language models.

Acknowledgments

The authors wish to thank A, B, and C. This work was supported in part by a grant from XYZ.

REFERENCES

[1] D. Angelov. Top2Vec: Distributed Representations of Topics. *arXiv preprint arXiv:2008.09470*, Aug 2020. doi: 10.48550/arXiv.2008.09470 1

[2] BAAI. bge-small-en-v1.5. https://huggingface.co/BAAI/bge-small-en-v1.5, 2023. Hugging Face. 1

[3] R. González-Márquez, L. Schmidt, B. M. Schmidt, P. Berens, and D. Kobak. The landscape of biomedical research. *Patterns (N Y)*, 5(6):100968, Apr 2024. doi: 10.1016/j.patter.2024.100968 1

[4] M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, Mar 2022. doi: 10.48550/arXiv.2203.05794 1, 2

[5] L. McInnes, J. Healy, and S. Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017. doi: 10.21105/joss.00205 2

[6] E. Simon, K. Swanson, and J. Zou. Language models for biological research: a primer. *Nat Methods*, 21:1422–1429, 2024. doi: 10.1038/s41592-024-02354-y 1

[7] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *J Mach Learn Res*, 9:2579–2605, 2008. doi: 10.1007/s10994-008-5078-3 1