Hierarchical topic tree: A hybrid model comprising network analysis and density peak search

Mengjia Wu[1] and Yi Zhang[2]

[1] *Mengjia.Wu@student.uts.edu.au*
Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney (Australia)

[2] *Yi.Zhang@uts.edu.au*
Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney (Australia)

Abstract
Topic hierarchies can help researchers to develop a quick and concise understanding of the main themes and concepts in a field of interest. This is especially useful for newcomers to a field or those with a passing need for basic knowledge of a research landscape. Yet, despite a plethora of studies into hierarchical topic identification, there still lacks a model that is comprehensive enough or adaptive enough to extract the topics from a corpus, deal with the concepts shared by multiple topics, arrange the topics in a hierarchy, and give each topic an appropriate name. Hence, this paper presents a one-stop framework for generating fully-conceptualized hierarchical topic trees. First, we generate a co-occurrence network based on key terms extracted from a corpus of documents. Then a density peak search algorithm is developed and applied to identify the core topic terms, which are subsequently used as topic labels. An overlapping community allocation algorithm follows to detect topics and possible overlaps between them. Lastly, the density peak search and overlapping community allocation algorithms run recursively to structure the topics into a hierarchical tree. The feasibility, reliability, and extensibility of the proposed framework are demonstrated through a case study on the field of computer science.

Introduction
The last decades have witnessed a great accumulation of scientific documents, resulting in information overload for researchers. Aiming to improve this situation, a substantial number of bibliometric studies on topic extraction, knowledge mining, and text analytics have been undertaken, each looking for efficient ways to extract information from textual data and concise ways of presenting the knowledge found (Ba et al., 2019; Qian et al., 2020; Song et al., 2016; Wu et al., 2020; Zhang et al., 2018; Zhang et al., 2017). What many of those studies have shown is that, one, organizing research topics into curated hierarchical structures is an excellent way of quickly conveying a great deal of knowledge about the composition of a research field to those who are unfamiliar with it, and, two, constructing these arrangements is nontrivial and highly challenging. While very broad overviews of a field are not particularly difficult to generate, creating interactive topic maps that show fields at different and especially fine levels of granularity and disentangling the rising complexities of inter-/multi- disciplinary studies is another story altogether. In fact, all but the most rudimentary techniques still rely heavily on expert knowledge.

That said, advancements in natural language processing (NLP) are reducing this dependence, with methods capable of automatically identifying and stratifying the thematic concepts found in a dataset of literature. Among these methods, hierarchical latent Dirichlet allocation (hLDA) (Blei et al., 2010) is especially well-known. However, there are a couple of aspects of hLDA that could be improved. These include sometimes weak associations between the generated parent and child topics; and internal unigram incoherence within topics (Qian et al., 2020; Xu et al., 2018); a propensity to represent each topic as a conglomeration of unigrams and probabilities; and a tendency to label topics with appropriate names, which reduces the interpretability of the results. There are also alternative approaches of building topic hierarchies, such as taxonomy identification (Shang et al., 2020), ontology construction (Wong et al., 2012),

and knowledge graphs (Yang et al., 2017). But, despite substantial efforts to the contrary, these techniques inevitably suffer from either an excessive number of parameters that need to be fine-tuned and/or issues with creating clean partitions between topics. Hard clustering algorithms like K-means or non-negative matrix factorization (Qian et al., 2020; Zhang et al., 2018), which most of these techniques are based on, struggle to find clear divisions between topics with high levels of overlap, convergence, or interactivity – characteristics that typify the process of scientific development.

Aiming to solve these issues, we propose a novel framework called *Hierarchical Topic Tree* (HTT) that operates in a recursive manner to reveal the topic hierarchies within a set of documents. The framework comprises a term co-occurrence network and two algorithms: DPS, a density peak search algorithm, modified to work with networks; and OCA, an overlapping community allocation algorithm. We assume that every topic consists of a core term, which becomes the topic's label, and a set of affiliated terms. Applying the density peak search algorithm to a term co-occurrence network reveals the density peak terms that meet some specific criteria for being used as a topic's label. The terms associated with every core topic term, i.e., the affiliated terms, are then determined and partitioned by the overlapping community allocation algorithm, which means terms can be assigned to multiple topics. These two steps are run recursively on partitioned subnetworks to identify deeper hierarchies in the term co-occurrence network until no core topic terms (topic labels) are found. To demonstrate the practical workings of the HTT framework, we conducted a case study on 6,267 academic articles published in the expansive field of computer science. The final results show a tree with six main branches and 120 sub-branches in a complex, but cohesive, hierarchical structure. The three main contributions our work makes include: 1) a density peak search algorithm that identifies and labels the topics in a corpus; 2) a community allocation algorithm that recognizes topic overlaps, which may indicate knowledge convergence; and 3) a model that requires two hyperparameters – a density threshold and an overlap threshold, which makes the process of tuning parameters easy and the model adaptable to a variety of cases.

The rest of this paper is organized as follows. The Related Works section next gives a brief review of the work on topic hierarchy identification. The Methodology section sets out the details of our proposed methodology. The Case Study section follows, presenting the data, results, and empirical insights derived from the computer science domain case. We then wrap our study with a conclusion, the study's limitations, and future directions of research.

Related Works

Hierarchies are instinctive, basal structures to humans that naturally aid our sensemaking of scientific knowledge composition. Blei pioneers the automation of topic hierarchy identification by developing the two perhaps most renowned algorithms in identifying topic hierarchies – the Chinese restaurant process (CRP) (Blei et al., 2004) and hierarchical latent Dirichlet allocation (hLDA) (Blei et al., 2010). However, the efficacy of the hLDA model largely depends on the pre-processing quality and may generate unsatisfactory results otherwise (Qian et al., 2020; Xu et al., 2018). The latter works pay efforts to modify topic hierarchy identification from different perspectives, including introducing the idea of recursive hierarchy detection (Wang et al., 2013), involving distance-dependent discrepancies for the CRP (Song et al., 2016), adding external ancillary information (Shang et al., 2020; Wang et al., 2015; Xu et al., 2018), and using alternative topic partition method like non-negative matrix factorization (Qian et al., 2020). But those studies either suffer from the need for a pre-defined tree structure or the lack of a labeling strategy. In practical terms, hierarchical structures vary hugely from discipline to discipline, especially for disciplines of vastly different forms, such as biomedicine versus artificial intelligence. As for the topic labeling strategy, most bibliometric approaches constitute topics as a set of semantically similar terms or records (Colavizza & Franceschet,

2016; Hou et al., 2018; Porter et al., 2020). Similarly, in mainstream topic modeling approaches, a topic is not represented with one all-encompassing label but, rather, as a bag of words or phrases and their corresponding probabilities. With both approaches, one still has to manually dive into the specific words, phrases, or even documents to infer the broad subject matter of the topic and decide on a name.

Density peak clustering was first proposed in *Science* by Rodriguez and Laio (2014). It is based on the premise is that the center of a cluster is more densely packed than the surrounding regions and that areas of high density tend to be relatively far apart. As a clustering method, density peak search has proven to be very fast and quite accurate. Compared to traditional K-means or density-based clustering algorithms like DBSCAN, density peak searches identify cluster centroids purely based on the one characteristic of density. There are no additional parameters and no multiple iterations, which means the clustering process is extremely efficient and highly robust to parameter selection. Du et al. (2016) have since improved this method by using average K-nearest neighbor (KNN) density to emphasize the importance of local density instead of the circle radius approach used originally. This notion of density accords with the characteristics a topic label should have in that a highly representative topic label will be strongly connected to its affiliated terms but as different as possible from other topic labels. This parallel motivated our idea to automatically name topics through a KNN-modified density peak-based clustering algorithm.

Methods

Concept definitions and problem formulation

Definitions of the main concepts referred to in the methodology are as follows.

- Topic term: Nominal phrases extracted from scientific documents via a series of natural language processing and cleaning steps.
- Topic: A set of topic terms with their corresponding probabilities headed by a core topic term. Term overlaps under the same parent topic are allowed for different topics.
- Hierarchical topic tree (HTT): HTT is both the name of our methodology framework and the final output. As an output, an HTT is a tree-structure that consists of topic nodes residing on different layers of a tree, as illustrated in Figure 1. The length from the root node to the nodes on the deepest layer is called the tree depth. A higher layer topic is a parent topic, and its connected topics in lower layers are called child topics. Child topics under the same parent topic are siblings. The associations between a parent and child topic are assumed to be stronger than the associations between siblings.

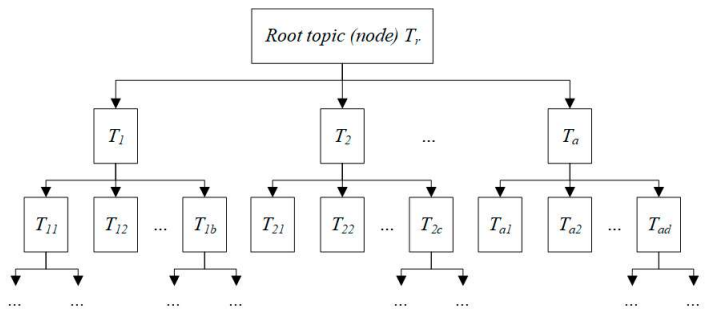

Figure 1. An example HTT

Problem formulation: The study aims to: 1) identify research topics with different granularities and construct a topic tree automatically from a collection of scientific documents; 2) label every topic with an appropriate name; and 3) detect topic overlaps.
HTT accomplishes these goals through the following steps.

Term clumping and network construction

The process begins by extracting topic terms from a corpus of documents. This is done with VantagePoint[1] and a term clumping process (Zhang et al., 2014). With the extracted terms in hand, the next step is to construct a weighted co-occurrence network of topic terms, denoted as $G = (V,\ E)$. V is the set of nodes representing the extracted topic terms, and E is the set of edges representing term co-occurrence. The graph is formulated according to the following equation:

$$w_{V_iV_j(i\neq j)} = \begin{cases} \dfrac{1}{CF(V_i, V_j)} & if\ V_i\ and\ V_j\ co-occur\ in\ at\ least\ one\ document \\ 0 & otherwise \end{cases}$$

where $w_{V_iV_j(i\neq j)}$ is the edge weight of $E_{V_iV_j(i\neq j)}$ and $CF(V_i, V_j)$ is the co-occurrence frequency of V_i and V_j.

Density peak search (DPS)

This algorithm is designed to identify core terms for topic labels. The primary concern when applying density peak clustering to network data is finding appropriate proxies for the distance and density measurements. Bai et al. (2017) use r-step topological distance as a proxy. However, this strategy necessitates a redundancy parameter r and a weighted parameter t, both of which need to be fine-tuned and both of which reduce the model's adaptability. Therefore, we opted to develop a new distance proxy, although still based on the topological distance between nodes:

$$d_{V_iV_j} = \begin{cases} w_{V_iV_j} & if\ V_i\ and\ V_j\ are\ connected \\ SPL_{V_iV_j} & if\ \ V_i\ and\ V_j\ are\ unconnected\ a\ path\ exits\ between\ them \\ \text{NA} & if\ no\ path\ exists\ between\ V_i\ and\ V_j \end{cases}$$

where $SPL_{V_iV_j}$ is the length of the shortest path from node V_i to V_j.

Generally, the co-occurrence network of high-frequency terms is fully connected, which means there will be at least one path from V_i to V_j. Hence, using the proposed new distance proxy, the kernel local KNN density and distance to the nearest denser point of every term can be calculated as:

$$\rho_{V_i} = \exp\left(-\frac{1}{K}\sum_{j\in KNN(V_j)} d(V_i, V_j)^2\right)$$

[1] More details could be found at www.vantagepoint.com.

$$\delta_{V_i} = \begin{cases} \max\limits_{V_j}(d_{V_i V_j}) & if\ \rho_{V_i} = \max(\rho_{V_i}) \\ \min\limits_{V_j \in V_{\rho_{V_j} > \rho_{V_i}}} (d_{V_i V_j}) & otherwise \end{cases}$$

In the few cases where the co-occurrence network includes several unconnected components, we will generate a virtual root node for the final HTT. Then each component will be processed separately as a branch of the virtual root node.

The original DPC algorithm identifies the cluster centroids with higher values of ρ and δ by observing the $\rho - \delta$ plot. However, when applying this algorithm to a real-world dataset, the boundaries of centroids and other terms are not always that clear. Therefore, in HTT, these selection criteria are quantitative. V_c denotes the potential centroids of all the communities, and the criteria for selecting the final centroids are formulated as follows:

1) Density peak: The selected centroids should be density peaks, denoted as:

$$\rho_{V_c} = \max_{V_i \in KNN(V_c)} \rho_{V_i}$$

in which $KNN(V_c)$ denotes the K-nearnest neighbor nodes of V_c.

2) Sparsity: To guarantee the identified centroids are sparse to each other, we set the node's distance to its parent node as a quantitative minimum threshold, which also indicates the associations of child nodes are weaker than the associations with their common parent node. This criterion is expressed as follows:

$$\delta_{V_c} > d_{V_r V_c}$$

in which V_r denotes the parent node of V_c.

Initially, there is no root node to measure whether a node meets Criterion 2). Hence, we will only use Criterion 1) to identify root nodes. If only one node meets criterion 1), it will automatically become the root node. Otherwise, a virtual root node is generated, and the n identified nodes would become children to the virtual root.

Overlapping community allocation (OCA)

The next step is to distinguish overlapping topics between communities and ensure they are given proper multiple assignments. Thus, every node is assigned a probability vector $p_{V_i} = \{p_{i,1}, p_{i,2}, p_{i,3}, \dots, p_{i,n}\}$, which reflects the probabilities that V_i belongs to core terms identified. Specifically, the probability that node V_i belongs to a community (topic) with the core term V_c is calculated as follows:

$$p_{i,c} = \frac{\min\limits_{V_j \in V} d(V_j, V_c)}{d(V_i, V_c)}$$

In disjoint community allocation, node V_i will be exclusively allocated to its closest centroid c if $c = argmax_t\{p_{i,t}, t = 1,2,3, \dots, n\}$. However, our aim is to allocate a node to more than one potential community with high probabilities. Hence, we employ an overlap threshold σ to decide multiple communities the node V_i could belong to. The rule applied is that if $\frac{p_{i,t}}{p_{i,c}} > \sigma$,

node V_i will be assigned to both community t and c. The output of this step is n overlapping communities with their assigned terms and probabilities.

Recursive hierarchy detection

The previous steps partition the network into n subnetworks, with each subnetwork comprising a core topic term and representing a sibling topic on the second layer. To extend the hierarchy into deeper layers, new subcommunities are detected by recursively applying the modified DPS and OCA algorithms to the partitioned subnetworks. When partitioning the parent networks into subnetworks, terms that belong to more than one topic, i.e., community overlaps, are excluded. This is because our approach aims at revealing hierarchies that exclusively belong to the parent topic. The recursive loop ends when no further core topic terms are detected in any subnetwork or the term number in the subnetwork is less than K.

The output of this step is the final HTT, with each node represented by a core topic term and linked to a set of terms. Topic overlaps containing terms shared by sibling topics are detected as well. This recursive process is illustrated in Figure 2, where each color represents a different stratum in the hierarchy. From top to bottom, the HTT has a root topic and one or multiple layers of topics generated by the iterations of DPS and OSA algorithms. Topics generated in the same iteration are siblings to each other and share a mutual parent topic.

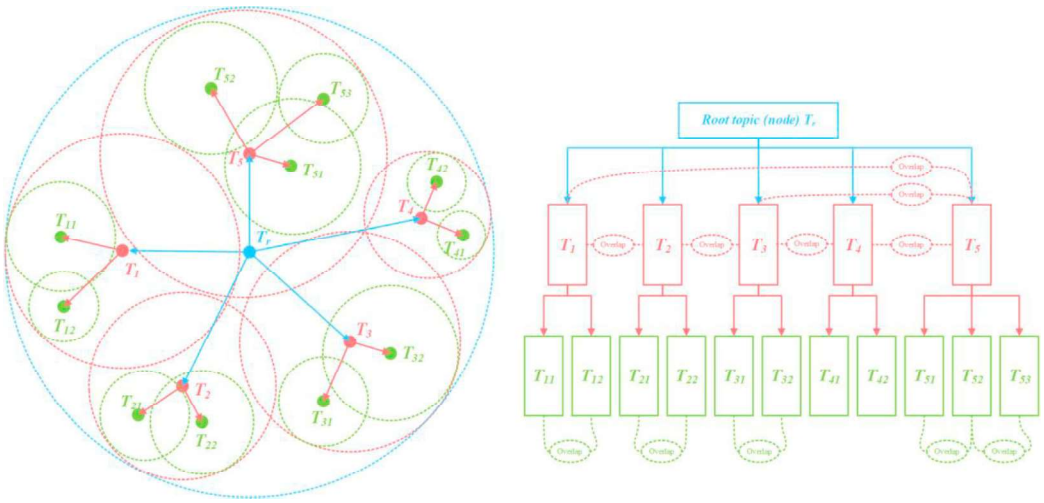

Figure 2. The recursive process of hierarchy construction

Methodology evaluation

According to criteria from previous studies, a well-curated hieratical topic structure should meet at least two characteristics: semantically coherent topics and high-quality parent-child topic relationships (Qian et al., 2020; Shang et al., 2020; Xu et al., 2018). Hence, we designed two indicators - topic coherence and parent-child association index (PCAI) to quantify the two characteristics. Additionally, we calculate the weight loss ratio of network edges to measure the information loss in the HTT process. Please note that the topics mentioned in this section contain overlapping terms, the association strength between two topics means the total sum of the edge weight's reciprocal of the pairwise terms from the two topics, the internal topic association of a topic strength refers to the total sum of the edge weight's reciprocal of pairwise terms from the topic itself.

- Topic coherence: Previous studies employ pointwise mutual information (PWI) to measure the topic coherence, but we consider it does not provide an intuitive and universal measure of topic coherence because its value range is -∞ to +∞ and its values vary hugely in multiple studies (Qian et al., 2020; Wang et al., 2013; Xu et al., 2018). Hence, in the current study, we measure the coherence of a topic via calculating the proportion of its total internal association strength against its total association strength with itself and its siblings.

$$Coherence_{T_i} = \frac{1}{|T_i|} \sum_{V_m \in T_i} \frac{\sum_{V_n \in T_i, m \neq n} CF(V_m V_n)}{\sum_{T_j \in children(parent(T_i))} \sum_{V_k \in T_j} CF(V_m V_k)}$$

- Parent-child association index (PCAI): This indicator is only applied to parent nodes in the final HTT (including the virtual root node if it exists). For every parent node, the PCAI equals the ratio of the total pairwise association strength among its children topics over the total association strength of itself and all children topics subtracted by 1.

$$PCAI_{T_i} = 1 - \frac{\sum_{T_m, T_n \in children(T_i), m \neq n} \sum_{V_p \in T_m^K, V_q \in T_n^K} CF(V_p, V_q)}{\sum_{T_j \in children(T_i)} \sum_{V_x \in T_j^K, V_y \in T_i^K} CF(V_x, V_y)}$$

- Information loss index: This index measures the overall information loss when the co-occurrence network being transformed into a hierarchical tree structure. The smaller value of information loss reflects the model's better performance of retaining information.

$$Information\ loss\ index_{T_i} = \frac{\sum_{T_m, T_n \in children(T_i), m \neq n} \sum_{V_p \in T_m, V_q \in T_n} CF(V_p, V_q)}{\sum_{V_x \in T_i, V_y \in T_i, x \neq y} CF(V_x, V_y)}$$

Case Study: The hierarchy of research topics in computer science

To demonstrate the methodology, we conducted a case study on the field of computer science, decomposing its many and varied research interests into topic hierarchies.

The corpus comprised 6,267 highly-cited papers published between 2010 and 2021 retrieved from the Web of Science (WoS) Core Collection database spanning the mainstream research topics regarding this domain. WoS is a well-curated multidisciplinary database with 74.8 million scientific publications from over 21,100 journals. Category information is assigned to every journal, and articles with the top 1% of citations received per field are flagged[2]. The search strategy used to assemble the corpus was as follows:

(WC = "Computer Science") AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article), Refined by: ESI Top Papers: (Highly Cited in Field), IC Timespan=2010-2021
WC: Web of Science Category;

Data pre-processing

Before applying our methods to the dataset, we ran VantagePoint's natural language processing (NLP) function to extract the raw words and phrases from the titles and abstracts. We then executed a term clumping process that removes noise and consolidates synonyms to arrive at a final list of topic terms. From this list, we selected terms with a frequency greater than 2. The

[2] https://clarivate.com/webofsciencegroup/solutions/essential-science-indicators/

stepwise cleaning results are given in Table 1. The final output was a term co-occurrence network consisting of 2,134 terms.

Table 1. Stepwise cleaning results

Step	*Description*	*# Terms*
1	Raw terms retrieved with NLP	132,846
2	Consolidated terms with the same stem, e.g., "information system" and "information systems"	116,898
3	Removed spelling variations, removed terms starting/ending with non-alphabetic characters, e.g., "Step 1" or "1.5 m/s", removed meaningless terms, e.g., pronouns, prepositions, and conjunctions	114,459
4	Removed general single-word terms, e.g., "information" *	96,245
5	Consolidated synonyms based on expert knowledge, e.g., "co-word analysis" and "word co-occurrence analysis"	84,828
6	Eliminated all terms occurring less than 5 times	2,134

*Note: Given that most single-word terms take on additional context when used in multi-word phrases, e.g., "information" vs. "information systems", we opted to remove generic single-word terms. Further, some multi-word terms were consolidated into a single-word form in Step 2 (e.g., "classification method" became "classification"). Non-general single-word terms were retained.

Parameter selection

Before generating the HTT, we selected appropriate values for the KNN density parameter K and the overlap threshold σ. Optimal values of K were determined through a sensitivity analysis by monitoring the number of initially identified core topic terms against K. The corresponding plot is presented in Figure 3.

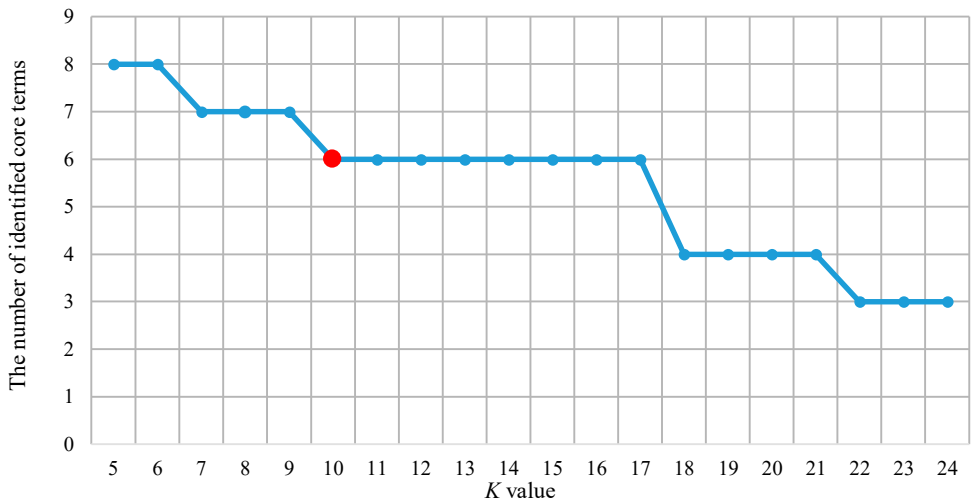

Figure 3. The plot of K against the number of identified core topic terms

HTT returned six initial core topic terms at every setting of K between 10 and 17. Therefore, to detect as many topics as possible, we set K to 10 and the overlap threshold σ to 0.8.

Tree generation

With the co-occurrence network as input to the DPS and OCA algorithms, the graph was recursively partitioned into subnetworks of topics in different layers, and the overlaps between topics were evaluated and assigned accordingly. The algorithms stopped at the eighth iteration, yielding a nine-level HTT of computer science research. Figures 4 and 5 illustrate the HTT and detailed terms in topics and their overlaps, respectively.

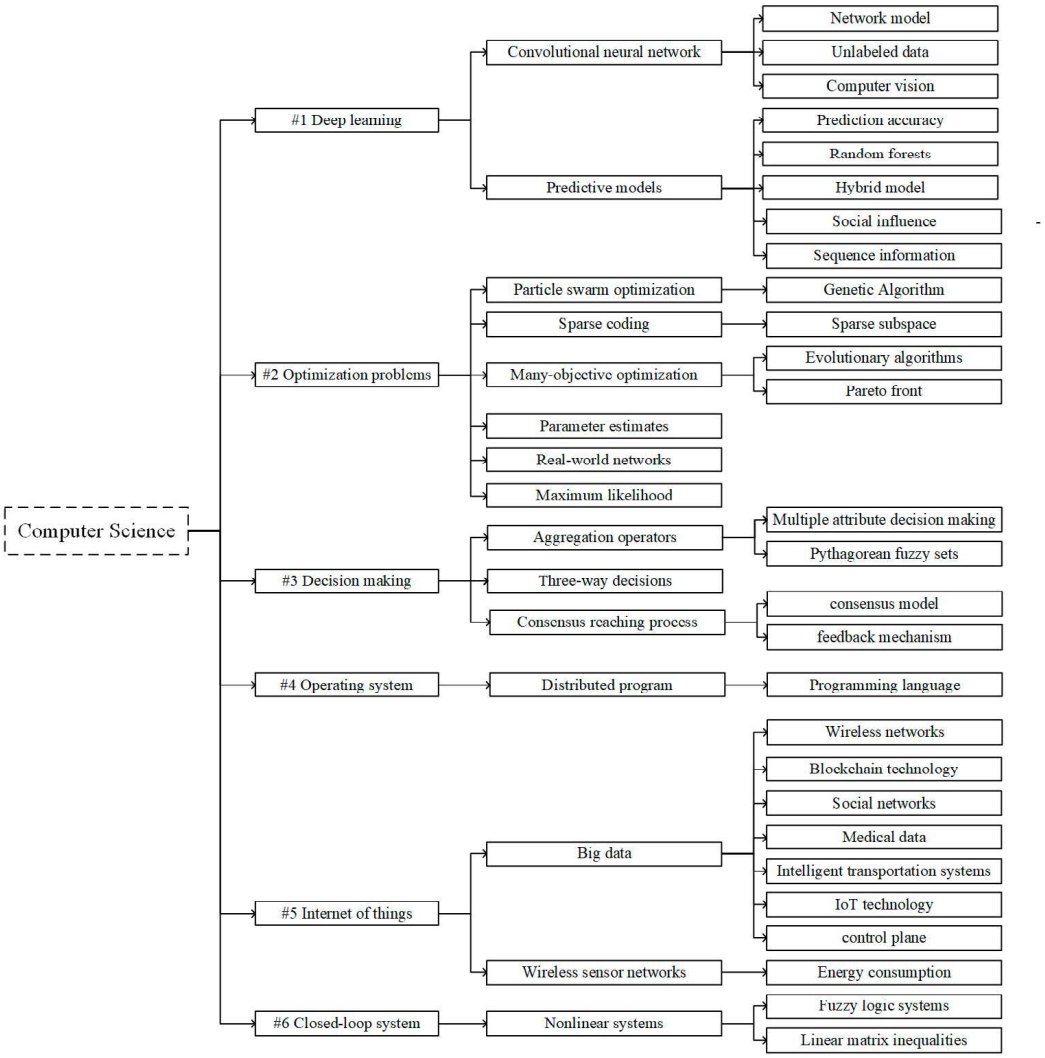

Figure 4. The HTT for computer science[3]

[3] Constraints on the page size limit the tree to its top three layers. The full HTT is available at https://github.com/IntelligentBibliometrics/HTT.

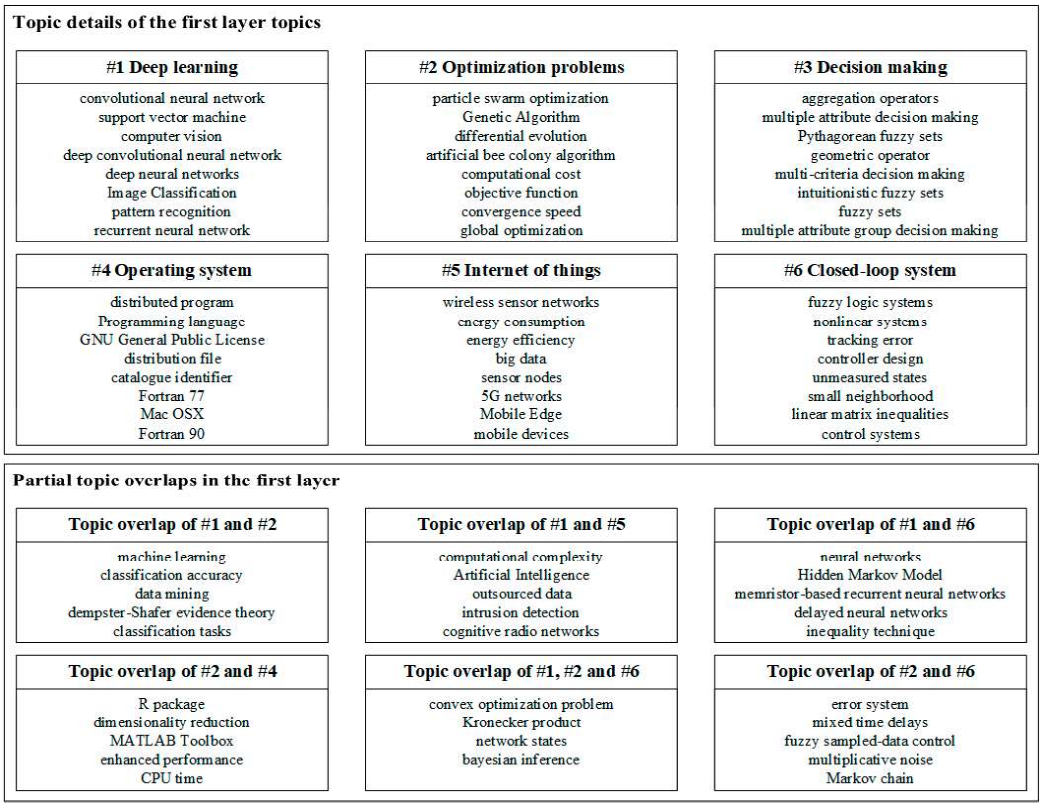

Figure 5. The topic details and partial topic overlaps in computer science

Evaluation and discussion

To evaluate the performance of HTT in this case, we calculated the average topic coherence, PCAI, and information loss of the final HTT, with their values presented as 0.619, 0.847, and 6%, respectively. The high PCAI value indicates our methods yield solid and reliable relationships between parent and their corresponding child topics. The low average information loss index suggests that the HTT evenly retains more than 93% of the information in every hierarchy construction process. The topic coherence is above 0.6, which is acceptable in partitioning the tangling research topics in the computer science domain that includes many multi-disciplinary interactions and knowledge convergence.

In Figure 4, the six topics in the first tier reflect six relatively separate research directions, which result from the idea of DPS that each core label should be topologically distant from each other. Simple observation confirms that the selected label terms with high density are also representatives of the terms they lead. Drilling down into each of the six initial parents, *#1 Deep learning* branches off into topics that pertain to various neural network techniques, such as convolutional neural networks and recurrent neural networks, and onwards to the tasks they are used to solve in the real world, e.g., computer vision, image classification, etc. The lower branch of this topic groups the models and metrics associated with deep learning, such as random forest and prediction accuracy. *#2 Optimization problem[s]* spans the different techniques, algorithms, and research objects associated with optimization and its sub-problems. *#3 Decision making* captures the models, strategies, sub-problems relevant to decision intelligence and its processes. #4 *Operating system[s]* groups the research topics surrounding computing architectures and software, which is a fundamental aspect of computer science. #5 *The Internet*

of Things (IoT) connects big data and sensor technology with its many spheres of application. Last, #6 *Closed-loop system[s]* leads the branch of topics concerning the convergence of computer science with engineering and control systems.

We also generated insights into cross-direction convergence from the topic overlaps in Figure 5. The overlapping terms between #1 and #2 include "data mining", "classification accuracy", and "classification tasks", which are universal concepts for both deep learning and optimization studies. Overlapping terms of #2 and #4 describe two programming tools (R, MATLAB) and computer performance (enhanced performance, CPU time). This overlap indicates a direction of solving optimization problems using computer operating system-based applications. Likewise, the other overlapping terms all indicate different kinds of topic convergence. Intriguingly, "machine learning" was also assigned to this overlapping section. Conventionally, deep learning would be regarded as a sub-topic of machine learning; however, the two terms are close neighbors in this term co-occurrence network, and "deep learning" has a higher KNN density. What this reflects is that deep learning has overshadowed its precursor technologies to become the more dominant research focus. Interestingly, this outcome raises questions over the temporal associations users attach to hierarchies and how the HTT framework prioritizes attention over evolution. This is a question we leave to future study.

Conclusions

This paper presents an end-to-end framework called HTT for identifying topic hierarchies from a co-occurrence network. The methodology combines density peak search and overlapping community allocation to provide a solution that extracts the topics from a corpus, identifies topic overlaps, arranges the topics in a hierarchy, and gives each topic an appropriately descriptive name. In HTT, the core term to each topic in a co-occurrence network, to be used as its label, is determined by the term's density peak characteristics, while overlapping community allocation detects overlaps among different topics. Recursive implementation of these two algorithms generates a hierarchical topic tree. A case study on the topic hierarchies in computer science demonstrates the feasibility and reliability of the proposed methodology.

In future studies, we plan several improvements to the HTT framework. These include: 1) Automatic parameter tuning: To further improve the adaptability of the methodology, we plan to change the DPS and OCA algorithms into nonparametric functions. Then, optimal values of K and σ could be selected automatically via a maximum entropy model or other approaches. 2) Leveraging additional forms of similarity: Co-occurrence networks are a classical input in bibliometric approaches, but they have also been criticized for their tendency to include too many irrelevant keyword pairs. Other forms of similarity, or combinations of similarities, such as semantic similarity based on topological distance, may prove to be a more effective proxy for the density peak search process. We plan to test these ideas in a future study. 3) HTT for streaming data. We also intend to build a variant of HTT that considers the temporal relationship between topics and how the research topics evolve over time.

Acknowledgments

This work is supported by the Australian Research Council under Discovery Early Career Researcher Award DE190100994.

References

Ba, Z., Cao, Y., Mao, J., & Li, G. (2019). A hierarchical approach to analyzing knowledge integration between two fields—a case study on medical informatics and computer science. *Scientometrics, 119*(3), 1455-1486.

Bai, X., Yang, P., & Shi, X. (2017). An overlapping community detection algorithm based on density peaks. *Neurocomputing, 226*, 7-15.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM), 57*(2), 1-30.
Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems, 16*(16), 17-24.
Colavizza, G., & Franceschet, M. (2016). Clustering citation histories in the Physical Review. *Journal of Informetrics, 10*(4), 1037-1051.
Du, M., Ding, S., & Jia, H. (2016). Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems, 99*, 135-145.
Hou, J., Yang, X., & Chen, C. (2018). Emerging trends and new developments in information science: A document co-citation analysis (2009–2016). *Scientometrics, 115*(2), 869-892.
Porter, A. L., Zhang, Y., Huang, Y., & Wu, M. (2020). Tracking and Mining the COVID-19 Research Literature. *Frontiers in Research Metrics and Analytics, 5*, 12.
Qian, Y., Liu, Y., & Sheng, Q. Z. (2020). Understanding hierarchical structural evolution in a scientific discipline: A case study of artificial intelligence. *Journal of Informetrics, 14*(3), 101047.
Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science, 344*(6191), 1492-1496.
Shang, J., Zhang, X., Liu, L., Li, S., & Han, J. (2020). *Nettaxo: Automated topic taxonomy construction from text-rich network.* Paper presented at the Proceedings of the Web Conference 2020.
Song, J., Huang, Y., Qi, X., Li, Y., Li, F., Fu, K., et al. (2016). Discovering hierarchical topic evolution in time‐stamped documents. *Journal of the Association for Information Science and Technology, 67*(4), 915-927.
Wang, C., Danilevsky, M., Desai, N., Zhang, Y., Nguyen, P., Taula, T., et al. (2013). *A phrase mining framework for recursive construction of a topical hierarchy.* Paper presented at the Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
Wang, C., Liu, J., Desai, N., Danilevsky, M., & Han, J. (2015). Constructing topical hierarchies in heterogeneous information networks. *Knowledge and Information Systems, 44*(3), 529-558.
Wong, W., Liu, W., & Bennamoun, M. (2012). Ontology learning from text: A look back and into the future. *ACM computing surveys (CSUR), 44*(4), 1-36.
Wu, M., Zhang, Y., Zhang, G., & Lu, J. (2020). Exploring the genetic basis of diseases through a heterogeneous bibliometric network: A methodology and case study. *Technological Forecasting and Social Change, 164*, 120513.
Xu, Y., Yin, J., Huang, J., & Yin, Y. (2018). Hierarchical topic modeling with automatic knowledge mining. *Expert Systems with Applications, 103*, 106-117.
Yang, S., Zou, L., Wang, Z., Yan, J., & Wen, J.-R. (2017). *Efficiently answering technical questions—a knowledge graph approach.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.
Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., et al. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*(4), 1099-1117.
Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85*, 26-39.
Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology, 68*(8), 1925-1939.