UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

AN INTELLIGENT BIBLIOMETRIC SYSTEM FOR KNOWLEDGE ASSOCIATION AND HIERARCHY DISCOVERY

by

Mengjia Wu

A Thesis Submitted
in Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2023

Certificate of Original Authorship

I, Mengjia Wu, declare that this thesis is submitted in fulfilment of the requirements for the award of the Doctor of Philosophy degree, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed
Signature: prior to publication.

Date: 31/05/2023

ABSTRACT

AN INTELLIGENT BIBLIOMETRIC SYSTEM FOR KNOWLEDGE ASSOCIATION AND HIERARCHY DISCOVERY

by

Mengjia Wu

Unravelling the intricate knowledge patterns and uncovering the underlying intelligence concealed within scientific literature constitutes a persistent objective for the bibliometric and data mining research communities. The rapid proliferation of scientific publications, coupled with the increasing prevalence of cross-/multi-/inter-disciplinary collaborations and the expanding scope of knowledge, pose continuous challenges for scholars seeking to remain abreast of the latest advancements and attain a comprehensive comprehension of their respective domains. Present knowledge mining methodologies encounter difficulties in flexibly accommodating diverse emerging demands, often necessitating prior expertise from domain specialists to achieve effective analysis, thereby impeding their practicality in real-world knowledge analysis tasks.

Aiming to contribute more adaptive and feasible knowledge mining approaches, this thesis incorporates bibliometric and management theories, data mining and natural language processing techniques (i.e., intelligent bibliometrics) to construct an intelligent bibliometric system for 1) knowledge association analysis and inference and 2) knowledge hierarchy extraction and characterisation. The system consists of two methodologies, with the scientific literature corpora as the input and bioentity rankings, bioentity association predictions and topic hierarchy visualisations as the output. The first methodology is a heterogeneous bioentity analysis methodology (HBAM), which focuses on the biomedical domain and provides a literature-based knowledge discovery approach that ranks extracted bioentities and predicts undiscovered bioentity associations. This methodology leverages bioentities' heterogeneity and latent semantic similarities to facilitate more comprehensive bioentity ranking and more accurate entity association prediction. The second methodology focuses on knowledge hierarchies and develops two hierarchical topic tree (HTT) models to extract and visualise topic hierarchies from scientific literature data adaptively. The two models can generate consistent research topics and solid parent-child topic relationships, with the latter refined as parameter-free and has better adaptivity. Lastly, the constructed intelligent bibliometric system integrates the proposed methods and a work pipeline, a Python-developed graphical user interface is then developed to provide an accessible for non-technical background users to conduct customised analysis. Academic researchers, policymakers, and entrepreneurs in certain domains can benefit from the system's ability to uncover knowledge associations and profile knowledge hierarchies for informed decision-making.

Dissertation directed by Dr Yi Zhang and Distinguished Professor Jie Lu.
School of Computer Science, Faculty of Engineering and Information Technology

Acknowledgements

I would like to express my sincere appreciation to my supervisor, Dr. *Yi Zhang*, and co-supervisor, Prof. *Jie Lu*, for giving me the opportunity to embark on this memorable PhD journey and for their guidance and support every step of the way. I still remember the day when Dr. *Yi Zhang* picked me up from the Sydney airport, marking the beginning of this incredible story. Looking back, I thoroughly enjoyed my PhD studies under his supervision. As a supervisor, he shared valuable research experience, provided substantial and detailed advice on my papers, and gave me plenty of opportunities to communicate and collaborate with researchers from around the world. Moreover, he always respected my research ideas and my preference for exploring technical issues. He was always eager to listen to my opinions and foster a positive collaborative environment. I feel honoured to be his first PhD student and sincerely hope that together, we can further strengthen our research team in the future. Prof. *Jie Lu* has also had a significant influence on me. Her expertise, patience, and passion for academic work have served as a role model for me. Every seminar or speech she delivered provided me with valuable insights that sparked new research ideas. During our internal workshops, she patiently taught us how to improve our presentation skills. Additionally, when the new year rolled around, she invited students to gather at her place for a party, making me feel at home despite being far away from my family. My thanks to my supervisors come from the bottom of my heart, as they have truly made my PhD journey complete.

I would also like to extend my gratitude to all members of DeSI, with special thanks to Ms. *Keqiuyin Li*, Dr. *Yiliao Song*, Dr. *Feng Liu*, Mr. *Zhaoqing Liu*, Ms. *Kun Wang*, Mr. *Tianyu Liu*, Mr. *Ming Zhou*, and Mr. *Guangzhi Ma*. Their academic suggestions, companionship, and the joy and fun they brought to my PhD journey are truly appreciated.

I gratefully acknowledge the funding provided by UTS that supported me in completing my research and this thesis. I would also like to express my gratitude to Ms. *Jemima Moore* for her invaluable assistance in proofreading and refining written English in most of my papers.

Lastly, I want to express my deepest and most heartfelt gratitude to my beloved family. Being separated from them for nearly four years has been challenging for each member. However, their unwavering support, regular video calls, and constant cheerful and empowering encouragement have carried me through difficult times and kept me resilient in the face of every obstacle. I would like to extend special thanks to my mother and my elder sister. My mother, as a single parent, raised me and instilled in me the strength to confront any challenges, including the difficulties I encountered in my research. My elder sister, the epitome of independence and fortitude, helped me overcome the pain of my parent's separation and consistently reminds me of how proud she is of me. I would not have come this far without their unwavering encouragement and belief in me. These two remarkable women in my life consistently exemplify the indomitable spirit of females, and their support and love are truly priceless treasures to me.

Mengjia Wu
Sydney, Australia, 2023.

List of Publications

I published 17 journal papers and 7 conference papers during my candidature. 2 journal papers are currently under review. Among the published papers, 11 are published in JCR Q1/ERA A-ranking venues. I am the first author of 8 papers, and 5 papers are published with my supervisor as the first author and me as the second author.

Journal Papers Published

J-1. **Wu, M.**, Zhang, Y., Zhang, G., and Lu, J. (2021). Exploring the genetic basis of diseases through a heterogeneous bibliometric network: A methodology and case study. *Technological Forecasting and Social Change*, 164, 120513. (ERA A, JCR Q1)

J-2. **Wu, M.**, Kozanoglu, D. C., Min, C., and Zhang, Y. (2021). Unravelling the capabilities that enable digital transformation: A data-driven methodology and the case of artificial intelligence. *Advanced Engineering Informatics*, 50, 101368. (JCR Q1)

J-3. **Wu, M.**, Zhang, Y., Grosser, M., Tipper, S., Venter, D., Lin, H., and Lu, J. (2021). Profiling Covid-19 genetic research: A data-driven study utilizing intelligent bibliometrics. *Frontiers in Research Metrics and Analytics*, 6, 683212.

J-4. Zhang, Y., **Wu, M.**, Tian, G. Y., Zhang, G., and Lu, J. (2021). Ethics and privacy of artificial intelligence: Understandings from bibliometrics. *Knowledge-Based Systems*, 222, 106994. (JCR Q1)

J-5. Zhang, Y., **Wu, M.**, Miao, W., Huang, L., and Lu, J. (2021). Bi-layer network analytics: A methodology for characterizing emerging general-purpose technologies. *Journal of Informetrics*, 15(4), 101202. (ERA A, JCR Q1)

J-6. Zhang, Y., **Wu, M.**, Hu, Z., Ward, R., Zhang, X., and Porter, A. (2021). Profiling and predicting the problem-solving patterns in China's research systems: A methodology of intelligent bibliometrics and empirical insights. *Quantitative Science Studies*, 2(1), 409-432.

J-7. Zhang, Y., **Wu, M.**, Lin, H., Tipper, S., Grosser, M., Zhang, G., and Lu, J. (2020). Framework of computational intelligence-enhanced knowledge base construction: Methodology and a case of gene-related cardiovascular disease. *International Journal of Computational Intelligence Systems.*

J-8. Zhang, Y., Cai, X., Fry, C. V., **Wu, M.**, and Wagner, C. S. (2021). Topic evolution, disruption and resilience in early COVID-19 research. *Scientometrics*, 126(5), 4225-4253. (ERA A, JCR Q1)

J-9. Cetindamar, D., Kitto, K., **Wu, M.**, Zhang, Y., Abedin, B., and Knight, S. (2022). Explicating AI literacy of employees at digital workplaces. *IEEE Transactions on Engineering Management.* (JCR Q1)

J-10. Kajikawa, Y., Mejia, C., **Wu, M.**, and Zhang, Y. (2022). Academic landscape of Technological Forecasting and Social Change through citation network and topic analyses. *Technological Forecasting and Social Change*, 182, 121877. (ERA A, JCR Q1)

J-11. Mejia, C., **Wu, M.**, Zhang, Y., and Kajikawa, Y. (2021). Exploring topics in bibliometric research through citation networks and semantic analysis. *Frontiers in Research Metrics and Analytics*, 6.

J-12. Porter, A. L., Zhang, Y., Huang, Y., and **Wu, M.** (2020). Tracking and mining the COVID-19 research literature. *Frontiers in Research Metrics and Analytics*, 5, 594060.

J-13. Li, X., Yao, Q., Tang, X., Li, Q., and **Wu, M.** (2020). How to investigate the historical roots and evolution of research fields in China? A case study on iMetrics using RootCite. *Scientometrics*, 125(2), 1253-1274. (ERA A, JCR Q1)

J-14. Grosser, M., Lin, H., **Wu, M.**, Zhang, Y., Tipper, S., Venter, D., Lu, J., and

Dos Remedios, C. G. (2022). A bibliometric review of peripartum cardiomyopathy compared to other cardiomyopathies using artificial intelligence and machine learning. *Biophysical Reviews*, 1-21.

J-15. Huang, Y., Zhang, Y., **Wu, M.**, Porter, A., and Barrangou, R. (2021). Determination of Factors Driving the Genome Editing Field in the CRISPR Era Using Bibliometrics. *The CRISPR Journal*, 4(5), 728-738. (JCR Q1)

J-16. Alsolbi, I., **Wu, M.**, Zhang, Y., Joshi, S., Sharma, M., Tafavogh, S., Sinha, A and Prasad, M. (2022). Different approaches of bibliometric analysis for data analytics applications in non-profit organisations. *Journal of Smart Environments and Green Computing.*

J-17 Zhang, Y., **Wu, M.**, and Lu, J. (2022). Stepping beyond your comfort zone: Diffusion-based network analytics for knowledge trajectory recommendation. *Journal of the Association for Information Science and Technology*, https://doi.org/10.1002/asi.24754. (EAR A*, JCR Q1)

Journal Papers Submitted

J-18. **Wu, M.**, Zhang, Y., Zhang, G., and Lu, J. (2021). Hierarchical topic tree: A topic hierarchy profiling approach incorporating k-shell decomposition and community detection. *Journal of the Association for Information Science and Technology*, under review. (EAR A*, JCR Q1)

J-19. **Wu, M.**, Zhang, Y., Markley, M., Cassidy, C., Newman, N., and Porter, A. (2022). Covid-19 Knowledge Deconstruction and Retrieval: An Intelligent Bibliometric Solution, *Scientometrics*, accepted. (ERA A, JCR Q1)

Conference Papers

C-1. **Wu, M.**, Zhang, Y., and Li, X., (2022). Exploring Associations within Disease-Gene Pairs: Bibliometrics, Word Embedding, and Network Analytics, *Proceedings of the 2022 Portland International Conference on Management of Engineering and Technology (PICMET)*, 1-7. (ERA A)

C-2. **Wu, M.**, and Zhang, Y. (2020). Intelligent bibliometrics for discovering the associations between genes and diseases: Methodology and case study. *Proceedings of Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2020)*, 8-15.

C-3. **Wu, M.**, and Zhang, Y. (2021). Hierarchical topic tree: A hybrid model comprising network analysis and density peak search. *Proceedings of the 18th International Conference on Scientometrics and Informetrics Conference*, 1241-1252. **(Winner of the ISSI student travel award)**

C-4. **Wu, M.**, Zhang, Y., Lu, J., Lin, H., and Grosser, M. (2020). Recommending scientific collaborators: Bibliometric networks for medical research entities. *Proceedings of the 14th International FLINS Conference (FLINS 2020)*, 480-487.

C-5. **Wu, M.**, Zhang, Y., Markley, M., Cassidy, C., Newman, N., and Porter, A. (2022). Covid-19 Knowledge Deconstruction and Retrieval: Solutions of Intelligent Bibliometrics, *Proceedings of Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2020)*, 92-103.

C-6. Zhang, Y., **Wu, M.**, Wang, X., and Chen, H. (2022). Navigating the Tradeoffs between Independence and Collaboration: A Network Analytic Method and Case Study, *Proceedings of the 2022 Portland International Conference on Management of Engineering and Technology (PICMET)*, 1-7. (ERA A)

C-7. Alsolbi, I., **Wu, M.**, Zhang, Y., Tafavogh, S., Sinha, A., and Prasad, M. (2022). Data Analytics Research in Nonprofit Organisations: A Bibliometric Analysis. *Proceedings of Pattern Recognition and Data Analysis with Applications*, 751-763.

Contents

List of Figures

Abbreviation

NLP - Natural Language Processing

LBD - Literature-based Discovery

ACM - Association for Computing Machinery

ICD - International Classification of Diseases

HBAM - Heterogeneous Bioentity Analysis Methodology

HTT - Hierarchical Topic Tree

COVID-19 - Coronavirus Disease 2019

AI - Artificial Intelligence

OAG - Open Academic Graph

DWPI - Derwent World Patent Index

GUI - Graphical User Interface

KNN - K-nearest Neighbours

PPI - Protein-protein Interaction

SAO - Subject-action-object

TRIZ - Theory of Inventive Problem Solving

TF-IDF - Term Frequency-Inverse Document Frequency

CRP - Chinese Restaurant Process

hLDA - Hierarchical Latent Dirichlet Allocation

DBSCAN - Density-based Spatial Clustering of Applications with Noise

AF - Atrial Fibrillation

GWAS - Genome-wide Association Studies

DC - Degree Centrality

CC - Closeness Centrality

BC - Betweenness Centrality

IR - Intersection Ratio

RA - Resource Allocation

SERA - Semantics-enhanced Resource Allocation

NCBI - National Center for Biotechnology Information

NIH - National Institute of Health

NLM - National Library of Health

OMIM - Online Mendelian Inheritance in Man

KEGG - Kyoto Encyclopedia of Genes and Genomes

GDA - Gene-disease Association Score

VDA - Variant-disease Association Score

TP - True Positive

FN - False Negative

WRA - Weighted Resource Allocation

LPI - Local Path Index

RWR - Random Walk with Restart

DT - Digital Transformation

CS - Computer Science

DPC - Density Peak Clustering

DPS - Density Peak Search

SEP - Scientific Evolutionary Pathways

OCA - Overlapping Community Allocation

PCAI - Parent-child Association Index

DPCI - Derwent Patent Citation Index

PWI - Pointwise Mutual Information

ICT - Information and Communications Technology

KSI - K-shell Index

LPA - Label Propagation Algorithm

aLPA - Asynchronous Label Propagation Algorithm

SLPA - Semi-synchronous Label Propagation Algorithm

IS - Information Science

MAG - Microsoft Academic Graph

TC - Topic Coherence

PCTA - Parent-Child Topic Association

STE - Sibling Topic Exclusiveness

ICU - Intensive Care Unit

JCR - Journal Citation Report

PCA - Principal Component Analysis

PCD - Principal Component Decomposition

Chapter 1

Introduction

1.1 Background

Scientific literature serves as the fundamental repository of human knowledge in contemporary sciences. Since scientific knowledge is primarily presented in unstructured text, reading has traditionally been the principal method for researchers and the general public to stay abreast of scientific advancements and emerging knowledge. However, the exponential growth of literature and the rapid advancement of data analytic techniques have brought about significant transformations. These changes have presented two major implications: 1) scholars face considerable challenges in managing the overwhelming volume of research papers, and 2) novel opportunities have arisen for the bibliometric and computer science research communities. In recent decades, the advent of natural language processing (NLP), machine learning, and network analytics techniques has empowered us to harness big data from scientific literature for knowledge extraction and discovery (Tang et al., 2008; Chen et al., 2021; Wang et al., 2019a; Sinha et al., 2015). Although the ever-expanding depth and breadth of knowledge, coupled with ongoing knowledge interactions, disruptions, and recombination (Dan and Chieh, 2008; Kaplan and Vakili, 2015), pose challenges in achieving this objective; scholars have endeavoured to address specific tasks aimed at unravelling knowledge composition and development patterns, which can subsequently be applied in downstream research and applications. Among these tasks, knowledge association discovery and hierarchy extraction have emerged as particularly significant and complex endeavours.

Knowledge association discovery aims to identify and establish connections and relationships between structured knowledge (Sun et al., 2020b). Despite some knowledge association tasks being performed on knowledge bases and graphs (Hao et al., 2019; Sun et al., 2020c; Guo et al., 2019), another crucial area of research, known as literature-based discovery (LBD), focuses on leveraging scientific literature data to uncover knowledge associations (Pyysalo et al., 2019; Crichton et al., 2020). LBD entails the process of inferring novel, credible, and informative knowledge by explicitly or implicitly associating two or more disparate concepts found in the literature (Bruza and Weeber, 2008). As scientific knowledge is primarily conveyed through unstructured textual formats, the practical approach involves extracting entities as knowledge units (Ding et al., 2013) and utilising these entities and their relationships for further analysis. It is worth noting that most knowledge association tasks are tailored to specific domains, with the biomedical field standing out as a prominent domain of interest due to the extensive knowledge yet to be discovered and the immense potential value of new findings in this area (Al-Aamri et al., 2019; Ding et al., 2013; Shang et al., 2014). Therefore, this thesis focuses specifically on the biomedicine field. The extracted entities derived from scientific texts in biomedicine are referred to as bioentities, encompassing diseases, chemicals, genes, and other relevant entities (Kim et al., 2004).

Hierarchy is another significant and intricate aspect of knowledge organisation and advancement. Bernstein (2000) articulates that hierarchy is a common nature of knowledge structure and describes natural sciences as behaving “explicit, coherent, systematically principled and hierarchical organisation of knowledge”. Real-world hierarchical knowledge structures exist in a broad range of knowledge domains, e.g., the Association for Computing Machinery (ACM) Computing Classification System[1]

[1] https://dl.acm.org/ccs

in computer science, the International Classification of Diseases (ICD) in medicine[2], and the library classification system in information science[3], etc. Those observations in real-world instances indicate that hierarchy is an innate structure rooted in knowledge development and has been widely accepted by academic communities (Ba et al., 2019; Qian et al., 2020; Song et al., 2016; Xu et al., 2018). Furthermore, knowledge hierarchy can help domain newcomers and stakeholders quickly comprehend the knowledge components of a research field and benefit various downstream applications, including knowledge recommendation and inference (Gao et al., 2019; Yang et al., 2017; Dinneen et al., 2018). However, manually curated knowledge hierarchy systems are not available in every segmented knowledge domain, especially for the new emerging fields. Hence, data-driven approaches that can build such hierarchies automatically are in urgent demand and of great practical value.

Despite the presence of methodologies for biomedical knowledge association discovery and hierarchy extraction from scientific literature data, these current approaches are subject to common limitations.

- First, existing biomedical knowledge association discovery methods are designed for a specific research case (for example, a specific gene or disease) and rely on expertise interpretation or prior knowledge, which are not generalised for a broad range of applications;

- Second, most biomedical knowledge association discovery methods concentrate on one single entity type and ignore the interactions between heterogeneous categories of entities;

- Third, current biomedical knowledge association inference methods ignore the semantic relationships between bioentities, which can be used as a valuable

[2] https://www.who.int/standards/classifications/classification-of-diseases

[3] https://www.oclc.org/en/dewey.html

feature in knowledge association inference;

- Fourth, existing knowledge hierarchy extraction techniques inevitably suffer from an excessive number of parameters, including the hierarchy depth and the number of topics, that need pre-defined or decided by prior knowledge, making them less adaptable for cases from different domains.

To address the above concerns, this thesis constructs an intelligent bibliometric system for biomedical knowledge association analysis and hierarchy extraction. The constructed system consists of existing work as follows:

The heterogeneous bioentity analysis methodology (HBAM) develops a work pipeline for processing biomedical literature data, sorting heterogeneous bioentities and predicting bioentity associations. It incorporates a heterogeneous entity network construction procedure, a non-dominated sorting genetic algorithm-based scoring scheme, a bioentity2vec training model and a semantics-enhanced link prediction method to rank bioentity importance/specificity and predict unobserved emerging entity associations. The semantic similarity between bioentities improves the performance of link prediction tasks and facilitates more accurate predictions validated by experimental and empirical evidence;

The hierarchical topic tree (HTT)-I model provides a feasible and handy approach for extracting topic hierarchies by inputting a term co-occurrence network. It exploits the idea of density peak clustering to identify term nodes with high density and relatively long distances from other high-density nodes as community centroids. An overlapping community allocation algorithm then applies to complete the community partition. This process iterates until no density peak nodes can be found. The values of three evaluation indicators on the HTT-I model demonstrate that it can generate consistent topics, solid parent-child topic associations with reasonable information loss;

The HTT-II model presents a refined version of HTT that fits a broader range of network inputs with different degrees of clustering tendency. Still using the term co-occurrence network as the input, the HTT-II model adopts k-shell decomposition and the Louvain algorithm to partition parent and child layers of terms and terms belonging to different topics. Compared with the HTT-I model, the HTT-II model is parameter-free and embraces a different design to partition terms into parent and child topics. This design can better retain coupling knowledge and differentiate terms in parent and child topics. The results from the comparison experiment demonstrate that the HTT-II model can generate consistent topics, solid parent-child topic associations and exclusive sibling topics;

Apart from the methodological contributions, this thesis covers multiple empirical studies to validate the performance and practical effectiveness of the proposed methods, with the case foci covering disciplines of biomedicine (investigations on genetic factors of atrial fibrillation and COVID-19), management (digital transformation conceptualisation and AI ethical issues identification), and information sciences (profiling research landscapes in the computer science and information science disciplines). Data sources of the empirical studies include academic papers from the Web of Science (WoS)[4], PubMed databases[5], Open Academic Graph (OAG)[6] and patents from the Derwent World Patent Index (DWPI)[7]. The results derived from multiple empirical studies generate insights into 1) research frontiers and foundations in relevant academic research fields and 2) strategic management and decision-making in relevant industrial sectors.

[4] https://www.webofscience.com

[5] https://pubmed.ncbi.nlm.nih.gov/

[6] https://www.aminer.cn/oag-2-1

[7] https://clarivate.com/products/ip-intelligence/ip-data-and-apis/derwent-world-patents-index/

1.2 Research Aim and Objectives

The overall aim of this thesis is;

to develop an intelligent bibliometric system for knowledge association discovery and knowledge hierarchy characterisation.

This thesis blends bibliometrics, management theories, NLP and network analytic techniques to develop an intelligent bibliometric system that can infer knowledge association and represent knowledge hierarchy from scientific literature data. To achieve this aim, we focus on three concrete research questions:

Question 1: Is there a practical way to analyse and predict biomedical knowledge association from scientific literature data?

Question 2: Is there a feasible way to extract and characterise knowledge hierarchy from scientific literature data?

Question 3: Can an information system be constructed to analyse entity associations and extract knowledge hierarchy automatically?

To answer the above questions, the objectives of this thesis are to:

i. Research Objective 1 (RO1): Establish a heterogeneous bioentity analysis methodology for knowledge association analysis and prediction.

   LBD is an efficient and cost-effective approach to uncovering and inferring unknown biomedical entity associations. However, existing biomedical knowledge association studies are not adaptive enough for varying cases and overlook bioentity semantic relationships. Aiming to fill those gaps, the first research objective of this thesis is to develop an LBD methodology to support bioentity association analysis and prediction. Specifically, the proposed methodology is capable of 1) integrating heterogeneous categories of bioentities to conduct the comprehensive analysis, 2) adapting to a broad range of research cases, and 3) leveraging the semantic features of biomedical entities to realise more accurate bioentity association prediction.

ii. Research Objective 2 (RO2): Develop an adaptive hierarchical topic extraction model to identify research topics from scientific literature data and uncover the hidden hierarchical knowledge structures.

   Hierarchy is a born characteristic of knowledge in its formation and development. Scientific studies and discoveries expand newborn knowledge fields in breadth and depth simultaneously to shape the hierarchical knowledge system, which helps scholars understand the knowledge landscape, identify knowledge frontiers, and detect cross-domain research opportunities. Considering that most current hierarchical topic extraction modes require excessive parameters to decide or fine-tune manually, the second objective of this thesis is to develop a hierarchical topic extraction model that can 1) automatically extract research themes/topics and organise knowledge hierarchy from scientific literature data, and 2) adaptively fit different real-world inputs and generate decision-making insights into real-world cases.

iii. Research Objective 3 (RO3): Construct an intelligent bibliometric system that realises knowledge association analysis, prediction and knowledge hierarchy extraction.

   With the proposed methods from Objectives 1 and 2, it remains a challenge for non-technical background users to implement the developed functions and obtain insights into their interests. Hence, our third objective is to integrate the proposed methods into a systematic workflow and develop an accessible graphic user interface (GUI) to access the designed functionalities, assisting users in running the workflow on their customised issues of interest.

1.3 Research Significance

1.3.1 Theoretical significance

Previous knowledge association discovery and hierarchy extraction studies develop isolated methods or models targeting a specific research case or domain. This thesis provides a one-stop system that integrates the literature data pre-processing steps, bioentity analysis and knowledge hierarchy extraction. The theoretical significance of this thesis includes the following:

First, this thesis proposes a heterogeneous bioentity analysis methodology to facilitate automatic bioentity extraction, sorting and association prediction from biomedical literature. We devise a new heterogeneous network-based bioentity sorting scheme in the methodology to quantify the importance/specificity of ranked bioentities to the target entity. The ranking scheme comprehensively balances influence from four categories of bioentities and integrates multiple network influence indicators via multiple-objective optimisation.

Second, this thesis develops a semantic-enhanced link prediction approach in the HBAM to predict the undiscovered associations between bioentities. Semantic similarities derived from the context of the bioentities are incorporated in the link prediction score calculation. The empirical experiments demonstrate that semantics can improve prediction accuracy.

Third, the HTT-I model devises an adaptable way to extract topic hierarchies from term co-occurrence networks. The model exploits the idea of density peak clustering to identify network community centroids, which does not require the mandatory inputs of a pre-defined number of topics or topic tree depth as many current approaches do. This characteristic of the proposed model makes it a feasible solution for research cases with little prior knowledge available.

Fourth, the HTT-II model further improves the adaptivity of this topic hierarchy extraction method. It incorporates the k-shell decomposition and Louvain community detection methods to automatically partition parent-child layer topics and segment different topic directions recursively. The HTT-II model is parameter-free and fits various networks with varying degrees of clustering tendency, especially for data input from relatively narrow knowledge domains in which knowledge is highly coupling and tangling.

Last, an intelligent bibliometric system and its graphical user interface (GUI) are developed to provide accessible functions proposed above to provide a one-stop platform for knowledge association analysis, hierarchy extraction, and further bibliometrics-derived data analysing functions.

1.3.2 Practical significance

The practical significance can be summarised from three perspectives:

First, one of the practical uses of the proposed HBAM is to depict a disease's gene importance-specificity map and predict emerging gene-disease associations. As reflected by the rise of precision medicine, taking individual genetic variability into account for personal healthcare services is a frontier trend in modern medical research. Awareness of a disease's genetic bases can contribute much to better risk assessment, diagnostics, and treatment strategies. The proposed methodology exploits scientific literature data to capture known associations between bioentities and diseases and further incorporates semantics-enhanced link prediction to identify undiscovered emerging disease-gene associations. This part of the work can provide a low-cost and efficient tool for scientific researchers and clinical doctors to realise a target disease's genetic factor analysis quickly.

Regarding the HTT models, topic hierarchies primarily assist stakeholders in quickly comprehending the knowledge components of a research field of interest. Beyond this, it can help academic researchers, policymakers, and entrepreneurs make more informed decisions. For example, the topic hierarchies of a specific target discipline could empower individual researchers to better grasp the frontiers of research in that field, supporting them to access more relevant literature via hierarchy-based document retrieval. Additionally, creating topic hierarchies for multiple disciplines may help policymakers map research resource distributions across different domains or help to justify their funding allocation strategies. Topic hierarchies for emerging subjects or technologies, like COVID-19 treatments or electric vehicles, could help companies to chart major research pathways or be used to inform more reasonable business strategies.

Furthermore, this thesis constructs a comprehensive one-stop system and develops a python-based GUI to analyse knowledge association and extract knowledge hierarchy from scientific literature data. The system integrates the proposed functions into a systematic workflow, and the GUI enables non-technical background users to access the proposed functions efficiently and perform customised analysis.

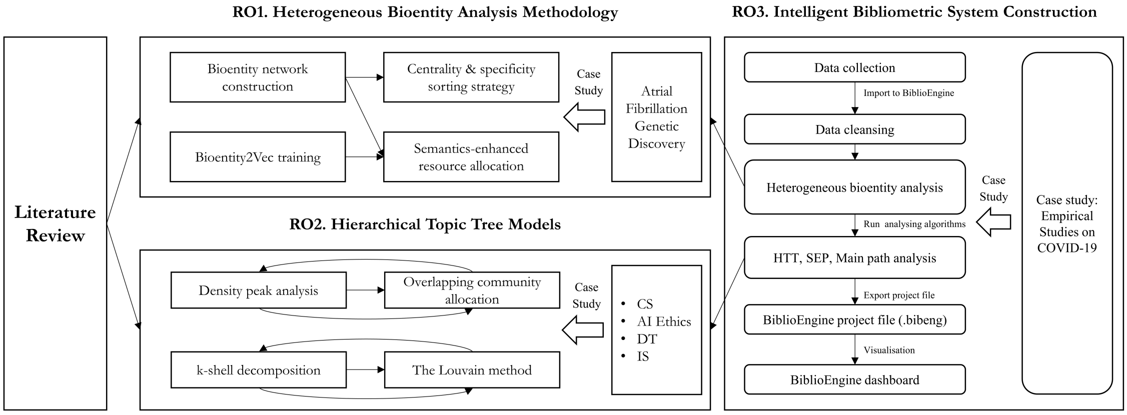

RO1. Heterogeneous Bioentity Analysis Methodology
RO3. Intelligent Bibliometric System Construction
Literature Review
Bioentity network construction
Centrality & specificity sorting strategy
Bioentity2Vec training
Semantics-enhanced resource allocation
Case Study
Atrial Fibrillation Genetic Discovery
RO2. Hierarchical Topic Tree Models
Density peak analysis
Overlapping community allocation
k-shell decomposition
The Louvain method
Case Study
• CS
• AI Ethics
• DT
• IS
Data collection
Import to BiblioEngine
Data cleansing
Heterogeneous bioentity analysis
Run analysing algorithms
HTT, SEP, Main path analysis
Export project file
BiblioEngine project file (.bibeng)
Visualisation
BiblioEngine dashboard
Case Study
Case study: Empirical Studies on COVID-19

Figure 1.1 : Research methodology

1.4 Research Methodology and Process

Our research methodology framework is illustrated in Figure 1.1. We first identified knowledge association and hierarchy discovery as our research interests. Then we conducted a literature review and reviewed relevant studies from a critical perspective. The review process helped us identify limitations in current studies and helped us formulate the research questions. To address those limitations and propose an integrated intelligent bibliometric system, we constructed three modules corresponding to our research objectives: A heterogeneous bioentity analysis methodology that can generate bioentity importance ranking and infer association prediction, two hierarchical topic tree models that can adaptively extract topic hierarchies, and an intelligent bibliometric system to integrate all the proposed functions and provide a systematic scientific literature analysis pipeline. As indicated in Figure 1.1, all the modules are linked and can be integrated, with methods proposed in each module validated by at least one empirical case study.

1.5 Thesis Organisation

This thesis contains seven chapters, and they are organised as follows. The structure of this thesis is shown in Figure 1.2.

- *Chapter 1*: This chapter introduces the research background, questions, objectives, significance, methodology, and structure of this thesis.

- *Chapter 2*: This chapter presents a literature review of relevant studies, including network analytics in bibliometrics, literature-based discovery studies, and existing flat/hierarchical topic extraction techniques.

- *Chapter 3*: This chapter proposes a literature-based discovery methodology for bioentity association analysis and prediction in the biomedical domain.

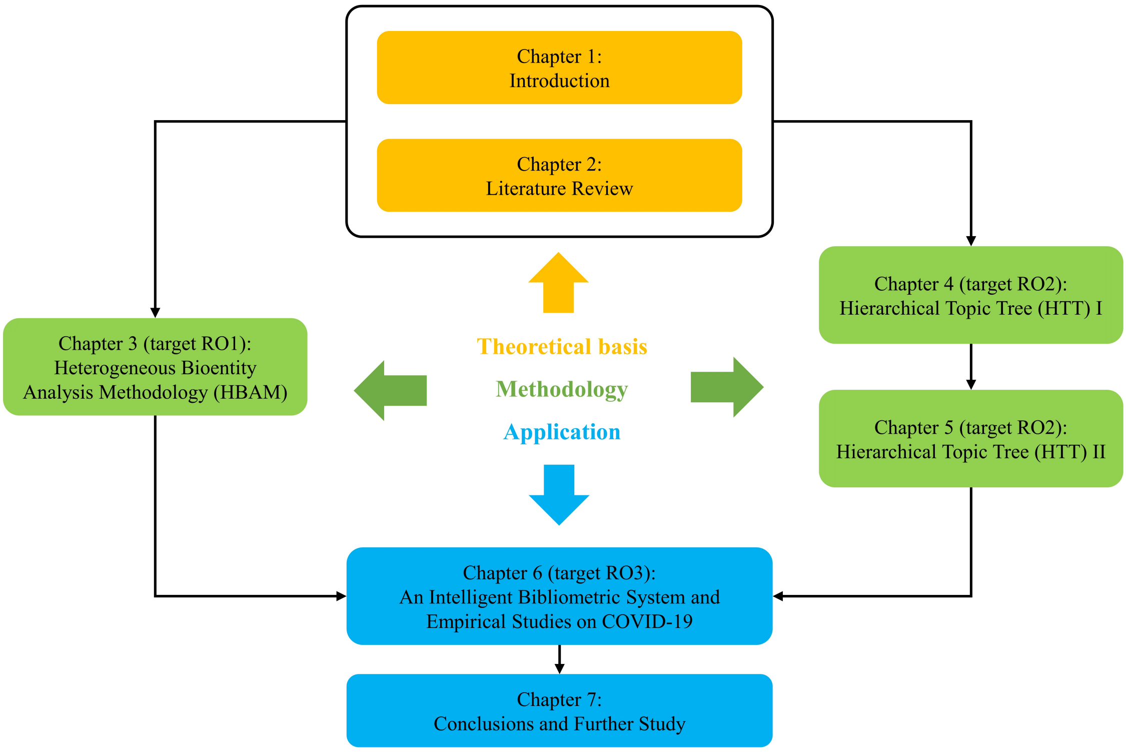


Figure 1.2 : Thesis structure

The methodology incorporates a heterogeneous entity network construction procedure, a non-dominated sorting genetic algorithm-based scoring scheme, a bioentity2vec training model and a semantics-enhanced link prediction method to rank bioentity importance/specificity and predict unobserved emerging bioentity associations. The pilot studies related to this chapter are published in the *Portland International Conference on Management of Engineering and Technology 2022* as **C-1** and *Extraction and Evaluation of Knowledge Entities from Scientific Documents 2022* as **C-2**; The journal paper related to this chapter is published in *Technological Forecasting and Social Change* as **J-1**.

- *Chapter 4*: This chapter presents the first version of the HTT model to identify topic structures from term co-occurrence networks. Using the term co-occurrence network as the input, the proposed model exploits the ideas of k-nearest neighbour (KNN) density and density peak clustering to identify term nodes with high density and relatively far distance to other high-density nodes. The identifying process recursively runs on the partitioned network to detect term groups, their overlaps and parent-child relationships, constructing the finalised topic tree structure. Compared with existing hierarchical topic extraction models, the proposed method demonstrates high adaptivity with fewer parameters to be decided. The practical effectiveness of the proposed model is validated by case studies on profiling the research landscape in the computer science domain, conceptualising the definition of digital transformation and uncovering emerging AI ethical debating themes. The related work of this methodology is published in *the 18th International Conference on Scientometrics and Informetrics Conference* as **C-3**. The applied research studies of the early version of the proposed model were published in *Advanced Engineering Informatics* as **J-2** and *Knowledge-based Systems* as **J-4**.

- *Chapter 5*: This chapter raises the refined non-parametric version of the HTT model that is more adaptable for term co-occurrence networks with different degrees of clustering tendency. The refined model incorporates k-shell decomposition and the Louvain community detection methods to group scientific term nodes as topics and extracts the hierarchical structure of topics based on the core-periphery and community characteristics of term co-occurrence networks. Compared with the HTT-I model, it is parameter-free and can adaptively and automatically generate topic hierarchy results to fit the given input network. The theoretical effectiveness of the proposed model is validated by a comparative analysis with five baseline approaches; Its practical value is endorsed by case studies depicting research segmentation in information sciences. The related work of this methodology is now under review in the *Journal of the Association for Information Science and Technology* as

Publication **J-18**.

- *Chapter 6*: This chapter introduces an intelligent bibliometric system that integrates the proposed methods and BiblioEngine, a Python-developed GUI for intelligent bibliometric analysis. Further, it demonstrates the practical value of the system through two empirical studies on COVID-19. Results through the HBAM analysis highlight core genes and a group of candidate novel genes that play a vital role in COVID-19. The HTT results profile an overall research landscape of COVID-19 research progress and further uncover the knowledge foundation for COVID-19 vaccination studies. The GUI development and applications pilot study was published in *International Journal of Computational Intelligence Systems* as **J-7**. One of the empirical studies of COVID-19 literature is published in *Frontiers in Research Metrics and Analytics* as **J-3**; The other is currently accepted by *Scientometrics* as J-19.

- *Chapter 7*: A summary of the thesis contents and its contributions are given in the final chapter. Further study recommendations are presented as well.

Chapter 2

Literature Review

This chapter reviews relevant studies to this thesis. Section 2.1 details the application of network analytics in bibliometrics, including the relevant theories, data sources, algorithms and research trends. Section 2.2 presents literature-based discovery and network medicine studies that uncover biomedical knowledge associations. Section 2.3 introduces topic extraction methods in bibliometrics and recent application studies. The following summary of two primary modelling methods, flat and hierarchical schemes, gives the methodological pathways of topic extraction techniques. Section 2.4 wraps this chapter with the limitations of existing studies.

2.1 Network Analytics in Bibliometrics

Modern bibliometrics can be traced back to the observations of Derek Price on the patterns of scientific activities (Price, 1986). Early definitions of bibliometrics emphasise "the application of mathematics and statistical methods to books and other media of communication (Pritchard, 1969; Price, 1986)", involving indicators such as citation/co-citation statistics, word co-occurrence, and co-authorships (Zhang et al., 2017c). It methodologically highlights the quantitative analysis of scientific literature and other relevant data sources. The increasing diversity of available data sources rapidly extends the scope of bibliometric data from books to a wide range of information resources in science, technology and innovation, such as research articles, patents, and academic proposals, as well as to social media data (e.g., Facebook, Twitter) (Zhang et al., 2013). Information technologies, especially artificial intelligence (AI) techniques, further strengthen the capabilities of bibliometrics in analysing large-scale data with enhanced efficiency, effectiveness, and robustness. Example pilot studies in this direction spearhead a cross-disciplinary approach that develops computational models incorporating bibliometric indicators with AI techniques, which we call intelligent bibliometrics (Zhang et al., 2020b; Wu et al., 2021c; Zhang et al., 2017c, 2018b).

Network format data is increasingly attracting research interest in a broad range of research fields, with intriguing scientific phenomena uncovered from analysis of biological, social, textual and many other networks. Since the early 2000s, research interests in complex network analysis have rapidly expanded from applied physics to various domains (Borgatti et al., 2009; Palla et al., 2005). Now more widely known as social network analysis, network analytics was a relative latecomer to bibliometrics. Initially, network analytics in bibliometrics was exploited to investigate research collaborations and disciplinary interactions through bibliographic couplings (Yan and Ding, 2009; Yang et al., 2010). However, once network analytics began to be combined with citation networks, co-citation networks, and co-authorship networks, attention from the bibliometric community increased dramatically. Understanding the topological structures of these networks has provided insights into a large volume of open research topics, such as collaboration and citation patterns (Ding, 2011; Liu et al., 2005). More recently, the introduction of word co-occurrence networks and natural language processing (NLP) techniques have provided more advanced angles to discover knowledge structures and identify research domains (Ravikumar et al., 2015; Zhang et al., 2012). Algorithms for community detection, link prediction, random walks, and others are also lending novel tools to increase the scope of traditional techniques and to undertake new types of analysis – for example, recommending potential collaborators (Yan and Guns, 2014; Huang et al., 2018b), discovering technological opportunities (Park and Yoon, 2018), and detecting/predicting emerging topics and technologies (Érdi et al., 2013; Huang et al., 2018a).

Previous studies incorporating bibliometric network analytics have 1) used topological indicators, such as centrality, to identify critical nodes and determine their actual meanings –e.g., influential researchers in a co-authorship network (Li et al., 2013; Yan and Ding, 2009); 2) used topology-based approaches, such as community detection and link prediction, to recognise specific behaviours, relationships, and patterns, e.g., collaborations (Yan and Guns, 2014), disciplinary interactions (Huang et al., 2020b), or problem-solving patterns (Zhang et al., 2021b); 3) connected bibliometric networks with a broad scenario of innovation paradigms, e.g., technology roadmaps (Jeong et al., 2021) and technology opportunity analysis (Park and Yoon, 2018; Ren and Zhao, 2021). Very few studies use heterogeneous bibliometric networks, but two are worth highlighting. Aiming to understand the collaborative/citing patterns of academic researchers, Ding (2011) applied an approach to a citation network and a co-authorship network that incorporated topic models with a random walk approach. Compared to typical network analytics, this work creatively embedded topic models with two bibliometric indicators. The other study is our adventure in applying heterogeneous bibliometric networks for measuring emerging general-purpose technologies (Zhang et al., 2021c).

The methodology of network medicine expands the application of network analytics within the biomedical field. Its predominant research paradigm involves constructing a network composed of biomedical entities and employing network analytic approaches to investigate the interactions among these entities (Barabási et al., 2011). Over time, numerous biomedical data sets and interaction networks have been curated, encompassing protein-protein interaction (PPI) networks (RN52), metabolic networks (Lawson et al., 2017; Schellenberger et al., 2010), regulatory networks (Clemente-Casares et al., 2016; Newburger and Bulyk, 2009), RNA networks (Anastasiadou et al., 2018), gene co-expression networks (Van Dam et al., 2018), among others. In addition to biological networks, there also exist self-constructed networks, such as entity similarity networks (Ravindra et al., 2020) and entity co-occurrence networks (Xu et al., 2020). There are certain trade-offs associated with using biological networks versus self-constructed networks for downstream analysis. Self-constructed networks have the capacity to integrate diverse data sources and offer more comprehensive relationships between biomedical entities. For example, these networks can incorporate information from biomedical databases (Piñero et al., 2016; Kanehisa and Goto, 2000), literature (Zhang et al., 2021c), and clinical trials. However, it should be noted that the edges in self-constructed networks may not entirely represent explicit biological relationships and could potentially contain some noise. For instance, co-occurrence networks may inadvertently convert negative associations between two entities that appear in the same context into links, thereby introducing a mixture of positive and a few negative links.

Further, research has been undertaken to unlock the knowledge within these established networks using various approaches. As the representative studies of this stream, Lei and Ruan (2013) proposed a topological similarity-based method to reduce the sparsity of a protein-protein interaction network, reconstructing a more condensed network for genetic analysis with better computational efficiency and more accurate predictions. Ganegoda et al. (2014) developed a method for constructing tissue-specific gene networks from a whole disease-gene network and applied a path-based similarity measurement to validate its usefulness. Valdeolivas et al. (2019) constructed a heterogeneous network containing diseases, genes, and proteins as entities and further implemented a random walk on the network to infer disease-gene interactions. However, despite all these successful explorations of fundamental knowledge bases, only a narrow slice of the possible biomedical entities is covered in each study. Plus, the results do not include very recent discoveries, and the economic cost and human effort to establish and continue maintaining the data sets that drive these solutions is enormous.

In recent studies, COVID-19 stands out as a unique task because the unprecedented amount of emerging knowledge it brings is closely related to the established knowledge foundation and rapidly reshaping a new knowledge structure. Hence, identifying the links between "new" and "old" knowledge becomes a significant task in COVID-19 knowledge profiling and retrieval. Along with the rapid accumulation of COVID-19 studies, bibliometricians have started analysing relevant literature to follow the latest research progress. The early-stage bibliometric analysis presents descriptive analyses of country-level research productivity (Chahrour et al., 2020), supporting sources (Nasab and Rahim, 2020), collaborating dynamics (Cai et al., 2021; Fry et al., 2020), and citing patterns (Hossain, 2020; Kousha and Thelwall, 2020). Apart from these efforts to measure research activity, uncovering new knowledge from the rapidly accumulating literature, i.e., literature-based discovery, is becoming a more critical task as such insights can support research and clinical and policy decisions (Hristovski et al., 2005; Swanson, 1986; Wu et al., 2021c). Following the literature-based discovery stream, Pourhatami et al. (2021) adopt co-word analysis to identify past coronavirus-related topics, pointing out promising research gaps in antibody-virus interactions, emerging infectious diseases, and coronavirus detection methods. Yu et al. (2021) apply entity metrics on an entity network extracted from the literature, highlighting ACE-2 and C-reactive protein as significant biomarkers and chemicals in diagnosing and treating COVID-19. Similar findings were reported by Wu et al. (2021b) through network analysis on biomedical entities extracted from COVID-19 literature, with more significant biomarkers, drugs, and complications identified. Ebadi et al. (2021) applied machine learning approaches to different COVID-19 publication sources and compared the highlights and differences in research topics. These literature-based discovery studies provide substantial evidence of explicit and implicit knowledge associations from extant research and insights that inspire deeper explorations in the future.

2.2 Literature-based Discovery

Literature-based discovery (LBD) is a workflow of inferring novel, credible and informative knowledge by associating two or more disparate literature concepts explicitly or implicitly (Bruza and Weeber, 2008). Swanson (1986) first employed this workflow and discovered the plausible association between fish oil and Raynaud syndrome treatment. This research concretes the potential of LBD in real-world applications. It arouses LBD's popularity in a variety of biomedical research issues: the exploration of potential treatments for diseases (Kostoff and Briggs, 2008), uncovering new therapeutic uses for existing drugs (Ding et al., 2013), revealing adverse drug effects (Shang et al., 2014) and inferring gene association for diseases (Al-Aamri et al., 2019), etc. The LBD expands its application into other domains such as discovering new problems for electric vehicles (Vicente-Gomila, 2014), water purification (Kostoff et al., 2008), climate change (Marsi et al., 2014) and robotics (Ittipanuvat et al., 2014), etc. Even though LBD's usefulness has been proven in interdisciplinary research (Small, 2010), most LBD approaches are still problem-derived and domain-specific. From a technical perspective, core methodologies and computational algorithms adopted in LBD research are continuously evolving. Statistical distribution models and co-occurrence analysis are the primitive LBD approaches used to quantify the concepts and relationships with the intention of knowledge inference (Gordon et al., 2002; Lindsay and Gordon, 1999; Petrič et al., 2009).

As indicated by a large volume of existing studies, one of the most potential applications that benefit from blending bibliometrics with network analytics is to uncover entity associations in the biomedical field. Literature-based discovery provides a more widely accessible pathway to explore the genetic basis for disease in the biomedical domain (Zhang et al., 2018c, 2016). One of the earliest attempts to discover genetic knowledge from the literature by Stapley and Benoit (1999) was to extract terms from the articles based on a dictionary and use those terms in conjunction with a set of rules to construct a gene co-occurrence network. Jenssen et al. (2001) further validated the usefulness of co-occurring patterns by visualising a global human genome network comprising millions of articles from three large-scale data sets of biomedical literature. The resulting network revealed significant, meaningful associations between co-occurring gene names at the document level. Adamic et al. (2002) applied statistical analysis to disease-gene co-occurrences using the binomial distribution to ink genes with diseases through text analysis. They scored the relevance of genes and diseases to discover novel associations, shedding light on multiple biomedical entity analyses.

Though straightforward, those pioneer approaches still face multiple limitations, such as neglecting concepts' contextual information, the need for concept disambiguation and the lack of knowledge representation and visualisation. Different efforts have been addressed in the following decades to improve this situation. Incorporating structured knowledge bases/ontologies facilitates semantic augmentation to fulfil the linkage information of concepts (Baker and Hemminger, 2010; Cameron et al., 2013; Lever et al., 2018; Preiss et al., 2015). The involvement of machine learning methods lifts concept extraction and data pre-processing in LBD to an upper level. For example, supervised learning and natural language processing unprecedentedly improve term extraction, disambiguation and consolidation (Mallory et al., 2016; Song et al., 2015; Wei et al., 2019). Unsupervised approaches, such as clustering algorithms, upgrade the concept profiling level from terms to topics (Zhang et al., 2018b). The adoption of network/graph theory provides a novel framework for literature knowledge representation and visualisation, along with those global (e.g., network completeness, diameter), local (e.g., network path or modularity-based communities) and node individual indicators (e.g., node properties such as centrality measures) measuring the literature knowledge structure and algorithms like link prediction inferring emerging knowledge associations (Al-Aamri et al., 2019; Crichton et al., 2018; Kastrin et al., 2016). Besides, other knowledge discovery models like subject-action-object (SAO) (Tsourikov et al., 2000) and the Theory of Inventive Problem Solving (TRIZ) (Savransky, 2000) are also frequently used knowledge representation paradigms for LBD research.

Natural language processing techniques have further improved the efficiency and accuracy of biomedical entity extraction (Habibi et al., 2017; Mallory et al., 2016; Pletscher-Frankild et al., 2015; Wei et al., 2013). Garten et al. (2010) employed a text mining-based extractor to perform sentence-level drug and gene co-occurrence analysis. The results show the superiority of using a text-derived network over a manually-curated network of drug-gene relationships to make predictions. Özgür et al. (2008) established a disease-specific gene interaction network and used network centrality measures to infer genes with potential links to prostate cancer and already-known seed genes. Al-Aamri et al. (2019) developed an approach based on network centrality, where the classifier is trained using a bootstrapping method. As a result, the model was able to parse the entire human genome. Some studies also leverage the semantic similarity between entities. For example, Coulet et al. (2010) built semantic networks of pharmacogenomic entities based on text data and inferred their interactions. Schlicker et al. (2010) improved gene prioritisation accuracy by involving semantic similarities generated from an ontology of genetic terms. With all the indicators in a basket, Heo et al. (2019) combined entity co-occurrence with word embedding techniques to produce a comprehensive index to measure the relationships between entities related to Alzheimer's disease.

2.3 Topic Extraction in Bibliometrics

Topic extraction is the process of mining and labelling topics from documents to represent the major themes or concepts from the document content, in which citation and textual features are heavily involved (Zhang et al., 2018b; Velden et al., 2017; Blei, 2012). As one of the primary content analysis methods, topic extraction is of significant interest to the bibliometric community. The extracted topics are represented by either a sub-collection of scientific literature or a set of scientific terms that hold recognised capabilities in knowledge interpretation and exploration, e.g., profiling research disciplines and technological areas (Zhang et al., 2016, 2017a; Ravikumar et al., 2015), identifying latent relationships (Zhang et al., 2021b, 2017c; Guo et al., 2016), and predicting potential future changes in either collaborative patterns or research interests (Huang et al., 2018b; Yan and Guns, 2014; Zhang et al., 2018c). Further analysis of those topics can help clarify cross-/inter-/multi-disciplinary interactions or predict future emerging research topics/interests (Zhang et al., 2017a, 2018c).

The benefits and value of topic extraction in profiling the knowledge landscape and facilitating knowledge discovery can be observed from numerous topic analysis case studies (Begelman et al., 2006; Kajikawa et al., 2022; Mejia et al., 2021). Scholars cluster semantically similar text (e.g., a collection of documents or similar terms) as topics and develop topic analysis approaches with different emphases, including topic identification (Small et al., 2014), tracking (Zhang et al., 2017c), and visualisation (Huang et al., 2014). The following section reviews two different topic extraction schemes that adopt different views in profiling extracted topic structures.

2.3.1 Flat topic extraction approaches

From the traditional bibliometric perspective, researchers have exploited various bibliometric indicators to identify topics hiding in the literature, including (1) Co-word analysis (Wartena and Brussee, 2008): assuming that words (referring to a broad definition including words, phrases or entities etc.) co-occurring in the same context (e.g., sentence, paragraph, document or keywords) tend to associate with the same theme, word co-occurrence is a plausible indicator to identify research topics; (2) Citation analysis (Colavizza and Franceschet, 2016; Hou et al., 2018): Citation relationships reflect the directed knowledge flow from one to another, citation analysis-based topic extraction mainly works under the assumption that the citing and cited articles share similar research topics, bibliographic coupling (Li et al., 2017) and co-citation (Shiau et al., 2017) analysis are also two prevailing approaches in research topic analysis.

While there are diverse choices available for constructing data sources for topic extraction, there are distinct differences among these classic bibliographic data inputs. For instance, (Yan and Ding, 2012) compared the similarities among six scholarly networks, namely citation networks, co-citation networks, bibliographic coupling networks, co-word networks, co-authorship networks, and topical networks. Their findings revealed that co-word networks exhibited the highest similarity to topical networks compared to the other four alternatives. This can be attributed to the fact that both co-word networks and topical networks primarily focus on the research content of publications rather than the flow of citations, which may contain substantial interdisciplinary or cross-disciplinary interactions.

Generally, those collected indicators will then be grouped to generate research topics via clustering algorithms (K-means, hierarchical agglomerative clustering or fuzzy c-means, etc.). Words, phrases or papers with high similarity are organised together, indicating separated research topics divided by their data characteristics. Apart from clustering algorithms, the community detection method is also an option for generating topics based on those indicators. Such algorithms group topologically similar keywords/documents as topics based on their connectivity in the keyword/term co-occurrence or citation networks(Huang et al., 2018a; Waltman and Van Eck, 2013). In more recent works, incorporating community detection with word embedding techniques has led to novel solutions for knowledge representation and topic extraction (Zhang et al., 2018b).

The blooming of text mining and NLP techniques enlarge the scope of topic extraction sources to unstructured textual data like titles, abstracts and full texts. Topic models represented by Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003) dominate the topic modelling methodologies, it leads the popularity of topic extraction for the following years with various mutations and derivations (Jelodar et al., 2019; Suominen and Toivanen, 2016; Yau et al., 2014). Topic modelling provides a complementary approach for free text analysis in bibliometric research. The combination of topic modelling with the traditional co-occurrence-based method is also attracting attention (Shams and Baraani-Dastjerdi, 2017).

2.3.2 Hierarchical topic extraction approaches

Hierarchies are instinctive, basal structures to humans that naturally aid our sense-making of scientific knowledge composition. Hierarchically organised research topics could provide a fine-grained structure for the target knowledge system. The last decade has witnessed the rapid growth of scientific literature and the increasing challenges facing researchers in their attempts to quickly and precisely retrieve knowledge from massive bodies of literature. With concisely profiling knowledge structures as their aim, many studies have shown that organising 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 (Ba et al., 2019; Qian et al., 2020; Xu et al., 2018). But they also show that constructing these topic hierarchies is nontrivial and highly challenging (Song et al., 2016). While broad and flat overviews of a field are not particularly difficult to generate, creating science maps that show fields at different granularity and disentangling the rising complexities of inter-/multi- disciplinary studies is another story altogether (Borner, 2015; Kay et al., 2014; Leydesdorff and Rafols, 2009).

There have been numerous attempts to identify topic hierarchies from scientific documents, such as text-based approaches like hyponym detection (Ponzetto and Strube, 2007; Seitner et al., 2016), hierarchical topic modelling (Blei et al., 2010), term embedding and clustering (Zhang et al., 2018b), and network-based approaches, including community detection (Shang et al., 2020; Wang et al., 2015a) and k-shell decomposition (Xiao et al., 2016). Nevertheless, the currently existing approaches always seem to have an adaptivity issue in that such techniques inevitably suffer from an excessive number of parameters that need to be determined. Specifically, existing clustering algorithms used in the studies above, like K-means, non-negative matrix factorization, and topic modelling (Shang et al., 2020; Qian et al., 2020; Xu et al., 2018; Zhang et al., 2018a), need to manually specify the appropriate number of topics or the topic hierarchy depth based on prior knowledge or expertise. This issue results in the fact that applying the method to a new field will require the involvement of domain experts, which significantly adds to usage costs.

Blei et al. (2004) pioneer the automation of topic hierarchy identification by developing the two perhaps most renowned algorithms in identifying topic hierarchies: The Chinese restaurant process (CRP) and hierarchical latent Dirichlet allocation (hLDA) (Blei et al., 2010). Those algorithms eliminate the mandatory input of topic numbers and theoretically enable infinite topic detection. However, in real-world applications, 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); Those models will not perform well if a volume of meaningless words is allocated in the higher layer topics. 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., 2015a; Xu et al., 2018), and using alternative topic partition methods like non-negative matrix factorisation (Qian et al., 2020). But those studies either suffer from the need for a pre-defined tree structure or the lack of a labelling strategy. In practical terms, hierarchical structures vary hugely from discipline to discipline, especially for fields of vastly different forms, such as biomedicine versus artificial intelligence. As for the topic labelling strategy, most bibliometric approaches constitute topics as a set of semantically similar terms or records (Colavizza and Franceschet, 2016; Hou et al., 2018; Porter et al., 2020).

With the heuristic hierarchical topic modelling algorithm, Wang et al. (2013) developed an algorithm for recursively construing a hierarchy of topics from a document set. Song et al. (2016) propose a hierarchical topic evolution model based on the distance-dependent CRP, which is a mutation of CRP. Xu et al. (2018) modified the hierarchical model by involving prior knowledge. A more recent work Qian et al. (2020) utilised the document-term matrix to detect the hierarchical topic structure by non-negative matrix factorisation. Their work pre-defined a three-layer design for the artificial intelligence topic tree. A general limitation exposed in their studies is the decision of dendrogram depth (i.e., the layer number of the hierarchical tree) needs to be defined manually: In realistic cases, the hierarchical structure varies hugely from discipline to discipline, three layers may fit relatively young domain like artificial intelligence, but in some areas like biomedical domain, if we have a look at the International Classification of Disease (ICD)[1] or the Medical Subject Headings[2], the tree structure could be much deeper and more complicated.

Network analytics provides another research trajectory for hierarchical topic extraction. Clauset et al. (2008) proposed that real-world networks often exhibit natural hierarchical structures that can reveal multilevel patterns. Typical examples can be observed in ecological, biomedical and social science networks, for instance, eco-

[1] https://www.who.int/standards/classifications/classification-of-diseases

[2] https://www.nlm.nih.gov/mesh

logical niches in food webs (Endrédi et al., 2018), phylogenetic trees in species evolutionary networks (Schaub and Peel, 2020), and organisation hierarchies in business management (Josephs et al., 2022). In the bibliometric field, leveraging hierarchical network structures to reveal scientific knowledge landscape and research intelligence is also a valuable research trajectory (Ba et al., 2019; Palla et al., 2015; Xiao et al., 2016). The hierarchy structure in networks enlightens our model design; Clauset et al. (2008) created a random dendrogram network to simulate the accurate graph by maximum likelihood estimation; they assumed that each node has a probability of p to connect with other nodes. Still, this model will not fit an existing real-world network because edges in that network are not randomly linked.

To uncover such important hierarchical characteristics of networks, research works have attempted to represent and characterise them from statistical or topological perspectives. The current approaches to discovering complex network hierarchy exploit the random walk, stochastic block modelling and k-shell decomposition.Rosvall and Bergstrom (2011) applied the random walk method to measure network flow and identified the hierarchical clustering of networks by optimising the shortest multilevel random walkers. Peixoto (2014) constructed a nested community generative model to derive the multi-scale network hierarchy. Lyzinski et al. (2016) modelled real-world networks with the hierarchical stochastic block model and recursively run community detection to detect the structural similarity of multilevel community subgraphs. K-shell decomposition is a heuristic method that decomposes a network into multiple shells of subgraphs from the connectivity from dense to sparse. Among the current network hierarchy profiling approaches, k-shell decomposition has a wide range of applications due to its robustness and algorithmic simplicity (Fang et al., 2017; Lin et al., 2021; Xiao et al., 2016). Previous bibliometric studies have examined the effectiveness of revealing hierarchical knowledge structures and landscapes (Ba et al., 2019; Xiao et al., 2016), yet some issues remain when generalising it to different cases.

2.4 Limitations of Previous Studies

Contemporary methods of biomedical LBD have some significant shortcomings. The main problem is that most approaches are designed with a singular focus on a specific disease, resulting in very few generalised models available. Additionally, there is a tendency toward the imbalance between quantitative approaches and expert knowledge. Too often, getting good results relies on substantial human intervention and prior knowledge, which can be hard to access. These constraints are particularly problematic for rare diseases and diseases where multiple genes may contribute to a condition. Apart from that, most existing approaches omit semantic features of bioentities that could be valuable in association analysis and prediction.

Regarding the existing topic extraction methods, both flat and hierarchical topic extraction approaches are mainly based on clustering or classification algorithms, which require the input of specific parameters like the pre-defined number of topics, pre-defined hierarchy depth, or both. This may result in two significant limitations: 1) It will harm the method's adaptivity to different cases, in which extra expert prior knowledge will be needed for deciding those parameters and 2) An inappropriate selection of those parameters may harm the topic extraction performance.

Chapter 3

Heterogeneous Bioentity Analysis Methodology

3.1 Introduction

In modern medicine, deciphering the genetic basis of diseases plays a vital role in their diagnosis, treatment, and prevention. However, for most disorders and abnormalities, it is not yet known whether genes, gene mutations, genetic variations, etc., play a pathogenetic role (Cookson et al., 2009; Goldstein, 2009). The high costs of genetic linkage analysis (Ott, 1999) and genome-wide association studies (GWAS) (Bush and Moore, 2012) have spawned an urgent need to prioritise candidate factors for further investigation. Researchers have established medical ontologies and curated molecular networks in past decades to analyse and infer molecular interactions for diseases based on accumulated experimental and clinical experience. Although these curated knowledge bases provide structured data for genetic insights into diseases, their use is still limited for 1) knowledge bases primarily focusing on a single category of bioentity and 2) the high cost of establishing and maintaining these knowledge bases.

Advanced text mining techniques combined with a fast-growing body of rich biomedical texts may provide an accessible and economically-viable pathway to solving those issues (Opap and Mulder, 2017) via literature-based knowledge discovery. Techniques such as co-occurrence analysis (Cohen et al., 2005), meta-analysis (Wang et al., 2017), centrality measurement (Al-Aamri et al., 2019), text mining (Mallory et al., 2016), and machine learning (Kim et al., 2017) have broadly enabled scientific literature as a valuable data source of exploring the genetic basis of various diseases.

Still, contemporary methods of scientific literature data analysis have some significant shortcomings. The main problem is that most approaches are designed with a singular focus on a specific disease, resulting in very few generalised models available. Additionally, there is a tendency toward the imbalance between quantitative approaches and expert knowledge (Al-Aamri et al., 2019). Too often, getting good results relies on substantial human intervention and prior knowledge, which can be hard to access. These constraints are particularly problematic for rare diseases and diseases where multiple genes may contribute to a condition.

Aiming to address these concerns, this chapter proposes a generalised and adaptable bibliometric methodology for investigating the bioentity association of target diseases. Four categories of bioentities are considered in this methodology to provide a comprehensive analysis: diseases, chemicals, genes, and genetic variations. The methodology is data-driven and does not require human intervention, guaranteeing its adaptability to different cases. Further, the proposed methodology exploits text semantics to refine the weighted link prediction approach and has the predicting capacity to infer likely associations that have yet to be identified.

The main components of this methodology include a heterogeneous bibliometric network, a Bioentity2Vec model, a suite of network analytics indicators, and a link prediction algorithm. The nodes of the bibliometric network represent the four types of bioentities, and the edges represent sentence-level co-occurrences between nodes. The Bioentity2Vec model follows the algorithmic design of Word2Vec (Mikolov et al., 2013), where all the entities are embedded as vector representations and then used to generate an adjacency matrix of pairwise semantic similarities. The network indicators include a series of centrality measures that characterise the importance of each entity, plus a novel indicator called intersection ratio that measures the specificity of an entity to the disease under study. The link prediction algorithm incorporates the semantic similarity of entities to modify the resource allocation algorithm and achieves better performance.

To validate our processes and demonstrate the effectiveness of the proposed framework, we conducted a case study on a corpus of 54,219 academic papers related to atrial fibrillation (AF). In a comparison test with a divided dataset designed to provide ground truth, our link prediction method identified 74% of the factor associations that would come to emerge spanning genes and genetic variants. In the same test on data up to 2020, we discovered strong evidence for five potential undiscovered associations and mediocre evidence for another five.

There are several novel aspects of this work. First, the literature-based method in our work does not rely on prior biomedical knowledge. Further, our methodology uses heterogeneous networks to analyse bioentity interactions. Lastly, the combination of contextual semantics and topological similarity enhanced with link prediction is a powerful new technique with broad applications in information science.

The rest of this chapter is organised as follows. Section 3.2 presents the details of the proposed methodology. The case study on AF appears in Section 3.3, along with the results. Section 3.4 wraps up the study with a discussion and conclusions.

3.2 Heterogeneous Bioentity Analysis Methodology (HBAM)

The methodology of this chapter is illustrated in Figure 3.1. The five blocks of the methodology include bioentity extraction, heterogeneous network construction, Bioentity2Vec training, core entity identification, and semantic-enhanced link prediction.

3.2.1 Bioentity extraction and heterogeneous network construction

As mentioned, the methodology covers four types of bioentities: Diseases, chemicals, genes, and genetic variations. 1) Diseases include disorders, symptoms, risk factors, and complications. 2) Chemicals cover chemical elements, clinical medica-

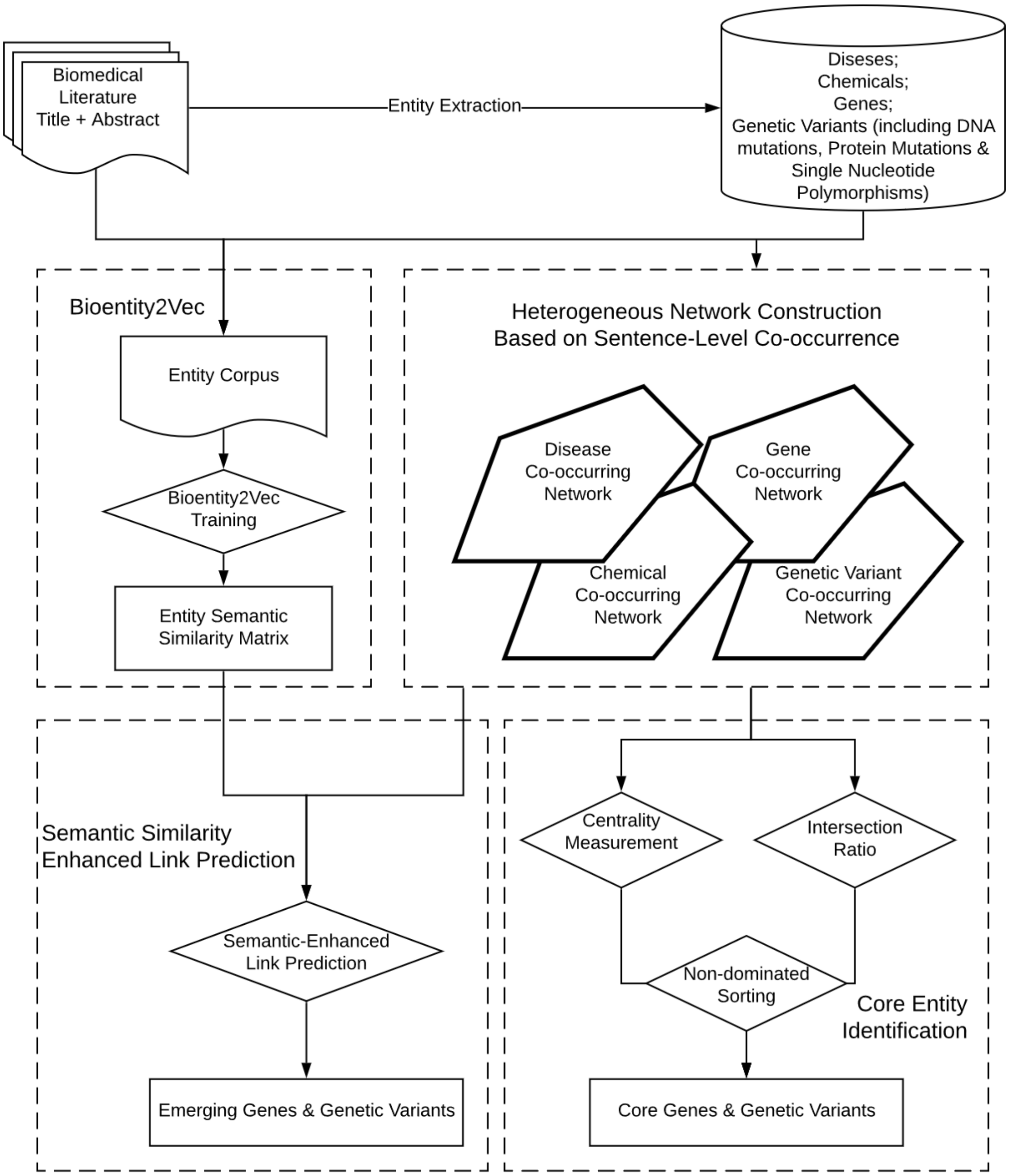

Biomedical Literature Title + Abstract
Entity Extraction
Diseses;
Chemicals;
Genes;
Genetic Variants (including DNA mutations, Protein Mutations & Single Nucleotide Polymorphisms)
Bioentity2Vec
Entity Corpus
Bioentity2Vec Training
Entity Semantic Similarity Matrix
Heterogeneous Network Construction Based on Sentence-Level Co-occurrence
Disease Co-occurring Network
Gene Co-occurring Network
Chemical Co-occurring Network
Genetic Variant Co-occurring Network
Semantic Similarity Enhanced Link Prediction
Semantic-Enhanced Link Prediction
Emerging Genes & Genetic Variants
Centrality Measurement
Intersection Ratio
Non-dominated Sorting
Core Entity Identification
Core Genes & Genetic Variants

Figure 3.1 : Research framework of the HBAM

tions, and other compounds. 3) Genes are the basic unit of heredity, occupying a fixed position on the chromosome. 4) Genetic variants include DNA mutations (i.e., a permanent change in a DNA sequence), protein mutations (i.e., proteins encoded with a mutated gene) and single nucleotide polymorphisms (SNPs) (i.e., normal variations of a single nucleotide in a gene sequence) (Arias et al., 1991).

Heterogeneous networks refer to networks that incorporate multiple categories of nodes and edges. Compared with homogeneous networks, leveraging heterogeneous for network analysis can integrate and fuse information from various sources and domains and further enable more accurate context-aware (the relationships and characteristics of entities) analysis. These four entities are represented as nodes in the weighted heterogeneous network. Working under the hypothesis that sentence-level co-occurrences indicate a stronger association between pairwise entities than document-level co-occurrence, the edges reflect co-occurrence frequency at the sentence level. The weights are derived from an adjacency matrix A, in which V_i^m is the mth node in the ith category and

$$A_{V_i^m V_j^n} = \begin{cases} CF(V_i^m, V_j^n) & \text{if } V_i^m \text{ and } V_j^n \text{ co-occur in a sentence} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

$CF(V_i^m, V_j^n)$ is the sentence co-occurrence frequency between V_i^m and V_j^n.

This network can also be denoted as a graph representation:

$$G = (V_K, E_{K(K+1)/2}) \tag{3.2}$$

where V is the set of K entity categories and E is the set of $K(K+1)/2$ types of edges connecting the different categories of nodes. An illustration of the network is provided in Figure 3.2.

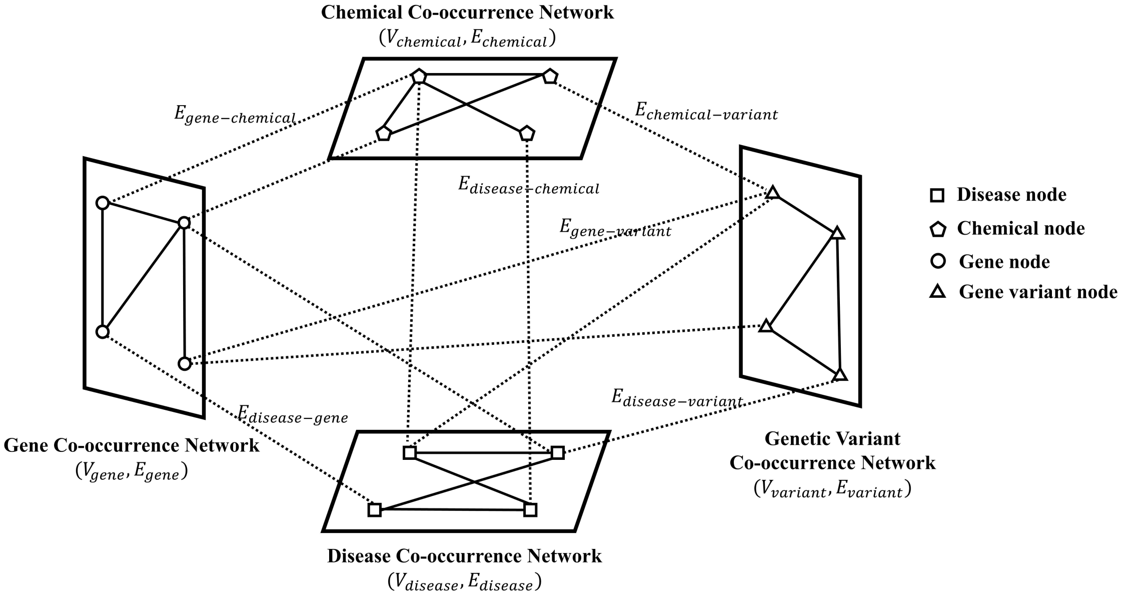


Figure 3.2 : Illustration of the heterogeneous network

3.2.2 Bioentity2Vec modelling

Sparked by the idea of the well-regarded Word2Vec natural language model (Mikolov et al., 2013), our semantic similarity measures are taken from a context-based perspective using a model we developed called Bioentity2Vec. Like Word2Vec, Bioentity2Vec converts bioentities into vectors by projecting one-hot representations into a lower dimension while mainly preserving the semantic meaning of the content. In our case, the bioentities are treated as words, and those words placed in sequence constitute the training corpus. We selected Skip-Gram as our training algorithm since it better fits small datasets. A summary of the Skip-Gram training process follows.

Given an entity $E(i)$ in a corpus, the probabilities of other entities in a certain window size w are predicted based on the given central entity $E(i)$ (Rong, 2014). The global objective is to maximize the average conditional probability for all windows in the corpus, which is formulated as:

$$LF = \frac{1}{n}\sum_{i=1}^{n}(\sum_{-w\leq j\leq w, i\neq 0}\log_2 P(E(i+j)E(i))) \tag{3.3}$$

The first step is to calculate the pairwise similarity of entities via cosine similarity and then generate a semantic adjacency similarity matrix $S_{V_i^m V_j^n}$:

$$S_{V_i^m V_j^n} = \cos(v_{V_i^m}, v_{V_j^n}) = \frac{v_{V_i^m} v_{V_j^n}}{|v_{V_i^m}||v_{V_j^n}|} \tag{3.4}$$

where $v_{V_i^m}$ is the corresponding vector of entity node V_i^m. Applying this formula to all entity pairs produces a pairwise adjacency matrix $S_{V_i^m V_j^n}$ of semantic similarity for all entities.

3.2.3 Network analytic measures

Centrality measures

Centrality measures comprise three indicators, degree centrality, closeness centrality, and betweenness centrality (Freeman et al., 1979; Zhang et al., 2021c), reflecting the node's capacity to aggregate, disseminate, and transfer information across a network. All three centrality measures have been proven efficient in revealing key nodes in biomedical networks (Al-Aamri et al., 2019). Their formal definitions are given below.

Degree Centrality (DC): This indicator measures the direct influence of a node on other nodes by calculating the proportion of its degree. An entity with a high value of degree centrality indicates that it has direct interactions with many other entities. It is calculated as:

$$DC(V_i^m) = \frac{\sum_{j=1}^{K}\sum_{n=1}^{|V_j|} A_{V_i^m V_j^n}}{|V_K| - 1} \tag{3.5}$$

where $|V_K|$ is the number of all K categories of nodes in the network and $|V_j|$ is the node number in the jth category.

Closeness Centrality (CC): This indicates a node's topological distance from all other nodes in the network, reflecting the global impact of a node towards all other nodes within the network. It is calculated as follows:

$$CC(V_i^m) = \frac{|V_K| - 1}{\sum_{j=1}^{K} \sum_{n=1}^{|V_j|} d_{V_i^m V_j^n}} \tag{3.6}$$

where $d_{V_i^m V_j^n}$ is the topological distance from node V_i^m to node V_j^n.

Betweenness Centrality (BC): This indicator measures a node's capability of connecting any other two nodes. In a network, a high betweenness centrality indicates that the node has a solid potential to be a crucial connector or transmitter. It is calculated by the sum of possibilities that any shortest paths connecting two other nodes go through the target node:

$$BC(V_i^m) = \frac{2 \sum_{x,y=1}^{K} \sum_{a=1}^{|V_x|} \sum_{b=1}^{|V_y|} \frac{\sigma(V_x^a V_y^b)_{V_i^m}}{\sigma(V_x^a V_y^b)}}{(|V_K| - 1)(|V_K| - 2)} \tag{3.7}$$

where $\sigma(V_x^a V_y^b)$ is the number of all shortest paths from node V_x^a to V_y^b and $\sigma(V_x^a V_y^b)_{V_i^m}$ is the number of these paths that pass through node V_i^m.

Intersection ratio

The intersection ratio is an indicator we designed to distinguish entities specifically associated with the target disease. While all centrality indicators reflect some aspect of a node's significance in the global network, some entities with high centrality measures may not be associated with the target disease at an exceptionally high level. Those entities are usually general terms representing fundamental chemicals or genetic factors related to a relatively broad range of conditions. Thus, we aim to distinguish the general entities from those highly relevant to the target disease. To this end, we develop an indicator based on the Jaccard coefficient (Wartena et al., 2010) and call it the intersection ratio. This indicator reflects an entity's specificity as the rate of a node's interaction with the target disease over all other diseases:

$$IR(V_i^m) = \frac{w(V_i^m, V_{disease}^t)}{\sum_{a=1}^{|V_disease|} w(V_i^m, V_{disease}^a)} \tag{3.8}$$

where $V_{disease}^t$ represents the node of the target disease, and $w(V_i^m, V_{disease}^t)$ refers to the weight of the edge connecting V_i^m and $V_{disease}^t$.

Traditionally, bibliometrics-based indicators are combined by specific strategies (e.g., entropy) into a unique value or are pairwise visualised based on diverse actual requirements (Zhang et al., 2017b). However, we aim to build a general methodology; Therefore, we introduced the non-dominated sorting algorithm to rank the entities based on a combination of the four metrics. Technically, non-dominated sorting is a multi-objective optimization procedure that compares samples containing multiple objectives or dimensions and ranks them according to their "dominance" over each other (Yuan et al., 2014). An entity A would dominate an entity B if A was better than B according to at least one of the four indicators but was no worse than B in any of the others. Once sorted, the items are divided into several consecutive Pareto fronts according to their domination counts. For example, if entity A is better than entity B in all four measures, it will be assigned to the dominant Pareto front. With top ranks in all metrics, the entities on this front have the strongest associations with the disease under study.

The pseudo-code for the non-dominated sorting algorithm is shown in Algorithm 1. This set of measurements outputs four lists of entities related to the target disease, i.e., core diseases, chemicals, genes, and genetic variations ranked in non-dominated order.

Algorithm 1: Non-dominated sorting algorithm

```
for $V_i$ in $V_K$ do
    for node $V_i^m$ in $V_i$ do
        $Domination[V_i^m] = 0$;
        for node $V_i^n$ in category $i$ do
            if $\forall x \in [1, d]$, $M_x(V_i^m) \geq M_x(V_i^n)$ $(m \neq n)$ then
                $Domination[V_i^m] += 1$;
                // $M_d(V_i^m)$ refers to the $d$th dimensional
                    measurement of $V_i^m$
            end
        end
    end
end
```

Semantic similarity-enhanced link prediction

Link prediction describes approaches that estimate the probability of particular links emerging in a network in the future (Liben-Nowell and Kleinberg, 2007). The results from our pilot study show that, of all the neighbour-based comparison methods, resource allocation (RA) (Zhou et al., 2009) is the most accurate (Zhang et al., 2021b,c). The original RA algorithm follows the assumption that every node in a network has one unit of a resource, and a common neighbour to two nodes will act as a transmitter, evenly distributing its resource to the connected nodes. The RA index of an unconnected pair of nodes is the sum of all resources obtained from all the neighbours common to the two nodes. In simple terms, it reflects the potential for a direct link emerging between the nodes. The higher the value, the greater the possibility is for a future link.

Inspired by Lü and Zhou (2010), who developed a weighted version of this algorithm, we conjecture that assessing the semantic similarity between two nodes and using that to weight to the RA index will increase the link prediction accuracy. Hence, we incorporated an additional procedure into the algorithm that involves the semantic matrix of bioentities generated by the Bioentity2Vec model. Thus, the final refined RA index is calculated as:

$$MRA_{V_i^m V_j^n} = S_{V_i^m V_j^n} \sum_{V^t \in \Gamma(V_i^m) \bigcap \Gamma(V_j^n)} \frac{CF(V_i^m, V^t)|S_{V_i^m V^t}| + CF(V^t, V_j^n)|S_{V^t V_j^n}|}{\sum_{V_k \in \Gamma(V^t)} CF(V^k, V^t) S_{V_k V^t}} \tag{3.9}$$

where V_i^m is the target disease, and V_j^n belongs to the set of genetic factors that have never before co-occurred with the target disease.

Applying the modified link prediction approach in a pairwise manner (V_i^m, V_j^n) generates the final output, which is a ranked list of genetic factors and their corresponding modified RA index scores. The assumption underlying the prediction that a genetic factor is associated with a disease is: if the target disease node V_i^m and the genetic factor node V_j^n do not co-occur, but they share at least one common neighbour, then they have the potential to be directly associated. The modified RA score tells us how strong that potential is. The common neighbour could be any one of the four entities. For example, they may both be associated with another genetic factor, or they may both be reactive to the same chemical.

3.3 Case Study: Knowledge Association Analysis for Atrial Fibrillation

Atrial fibrillation (AF) is one of the most common forms of cardiac arrhythmia. The disease progress of AF is closely related to atrial size and the extent of atrial fibrosis, both of which are affected by genetic factors. Although several gene groups and genetic mutations have been linked to AF, clinical evidence and mechanistic explanations are still far from sufficient to begin integrating our knowledge of these genetic risk factors into clinical practice (Feghaly et al., 2018). For these reasons, exploring the associations between genes and AF as our case study can assess the proposed method and have practical significance for advancing research frontiers in the AF area.

3.3.1 Data collection

PubMed is the largest global biomedical literature database, comprising more than 30 million citations across the MEDLINE database, life science journals, and other online book resources. We used the term “atrial fibrillation” with a MeSH search strategy limited to the “species” human across PubMed titles to guarantee precise AF-related search results. No restrictions were placed on the publication date. In all, 54,219 records were retrieved from the following search query:

”(”Atrial Fibrillation”[Mesh] AND Humans[Mesh])”

Search Date: 28 April 2020

3.3.2 Bioentity extraction and network construction

The high error rate is a common challenge in gene name recognition tasks. To lessen this problem, we assembled the extractor’s vocabulary list by combining terms from three different biomedical dictionaries:

- Medical Subject Headings (MeSH)[1] is a medical thesaurus provided by PubMed that contains the standardised concepts of diseases and chemicals.

[1] https://www.ncbi.nlm.nih.gov/mesh/

- NCBI *Homo Sapiens* Gene Dictionary[2] provided by the United States National Institute of Health (NIH), covers the known genes of Homo sapiens.

- dbSNP database[3] is a register of known sequence variants in the human genome, established in 1999. It contains the discovered DNA mutations, protein mutations and SNPs. Each SNP record is associated with a unique SNP ID.

We selected Pubtator[4] as the extractor. Pubtator is a deep learning-based bioentity extraction tool developed by the National Library of Medicine (NLM) (Wei et al., 2019). It can automatically extract categorised biomedical concepts from the titles and abstracts of PubMed articles.

The extraction process resulted in 577,809 raw biomedical concepts with accompanying text locations and unique identifiers. The concepts included diseases, chemicals, genes, DNA and protein mutations, SNPs, and species. We excluded the species concepts since our focus is on humans and restricted the genetic factors to the scope of the human genome using dbSNP. We then mapped every concept to its corresponding dictionary, removing noisy concepts (see Step 1, Table 3.1) and consolidating all synonyms (Step 2, Table 3.1). After these two steps, 6,318 unique bioentities remained. We further excluded 480 concepts that did not co-occur with any other concept (i.e., isolated nodes) to result in a final set of 5,838 entities. The stepwise pre-processing tallies are given in Table 3.1.

The co-occurrence network construction process revealed 48,988 edges reflecting sentence-level co-occurrence across the 5,838 nodes of the network. Among the four types of entities, there can be ten types of edges; their counts are provided in Table 3.2.

[2] ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/

[3] https://www.ncbi.nlm.nih.gov/snp/

[4] https://www.ncbi.nlm.nih.gov/research/pubtator/api.html

Table 3.1 : Stepwise results of the pre-processing procedure

	Raw	**Step 1**	**Cleaned**	**Step 2**	**All nodes**	**Del.**	**Nodes**
Disease	440,610	Removed noisy concepts like "cardioembolic", "JAGS", "nonvitamin", etc. that could not be mapped to MeSH	434,198	MeSH	2,239	-199	2,040
Chemical	104,702		101,512	MeSH	2,187	-183	2,004
Gene	31,209	Exclude genes that do not belong to Homo Sapiens	26,948	NCBI Gene	1,506	-93	1,413
Genetic variant							
- DNA mutation	223	Removed variants with unclear loci (i.e., could not be mapped to an SNP ID)	161	dbSNP	386	-5	381
- Protein mutation	770		555				
- SNP	925		217				
Total	**577,809**	-	**563,235**		**6,318**	**-180**	**5,838**

Table 3.2 : Counts of the different types of edges

	Disease (2,040)	Chemical (2,004)	Gene (1,413)	Genetic variant (215)
Disease (2,040)	19,181	10,977	5,318	469
Chemical (2,004)	10,977	5,248	2,463	123
Gene (1,413)	5,318	2,463	3,477	654
Genetic variant (215)	469	123	654	495

Table 3.3 : Centrality measures and intersection ratio statistics

		Disease	Chemical	Gene	Genetic variant
Degree centrality	Max.	0.668	0.106	0.050	0.006
	Min.	0.668	0.106	0.050	0.006
	Min.	0.668	0.106	0.050	0.006
	Avg.	0.668	0.106	0.050	0.006
Closeness centrality	Max.	0.739	0.493	0.471	0.433
	Min.	0.0001	0.0001	0.0001	0.0001
	Avg.	0.397	0.381	0.382	0.378
	Std.	0.063	0.068	0.085	0.074
Betweenness centrality	Max.	0.630	0.020	0.005	0.002
	Min.	0	0	0	0
	Avg.	0.0005	0.0001	0	0
	Std.	0.014	0.0007	0.0003	0.0001
Intersection ratio	Max.	1	1	1	1
	Min.	0	0	0	0
	Avg.	0.276	0.364	0.459	0.555
	Std.	0.280	0.363	0.377	0.443

3.3.3 Identifying core bioentities associated with AF

Core entities (i.e., highly-relevant entities) have high values on degree, closeness and betweenness centrality measures, plus a high intersection ratio value. We calculated these metrics for all 5,838 entities. A summary of the pertinent statistics by entity type is provided in Table 3.3.

Core genes

Following the steps described in the methodology, we began with gene nodes to apply the non-dominated sorting algorithm to the three centrality measures and normalise the domination counts to reflect their global importance. We then juxta-

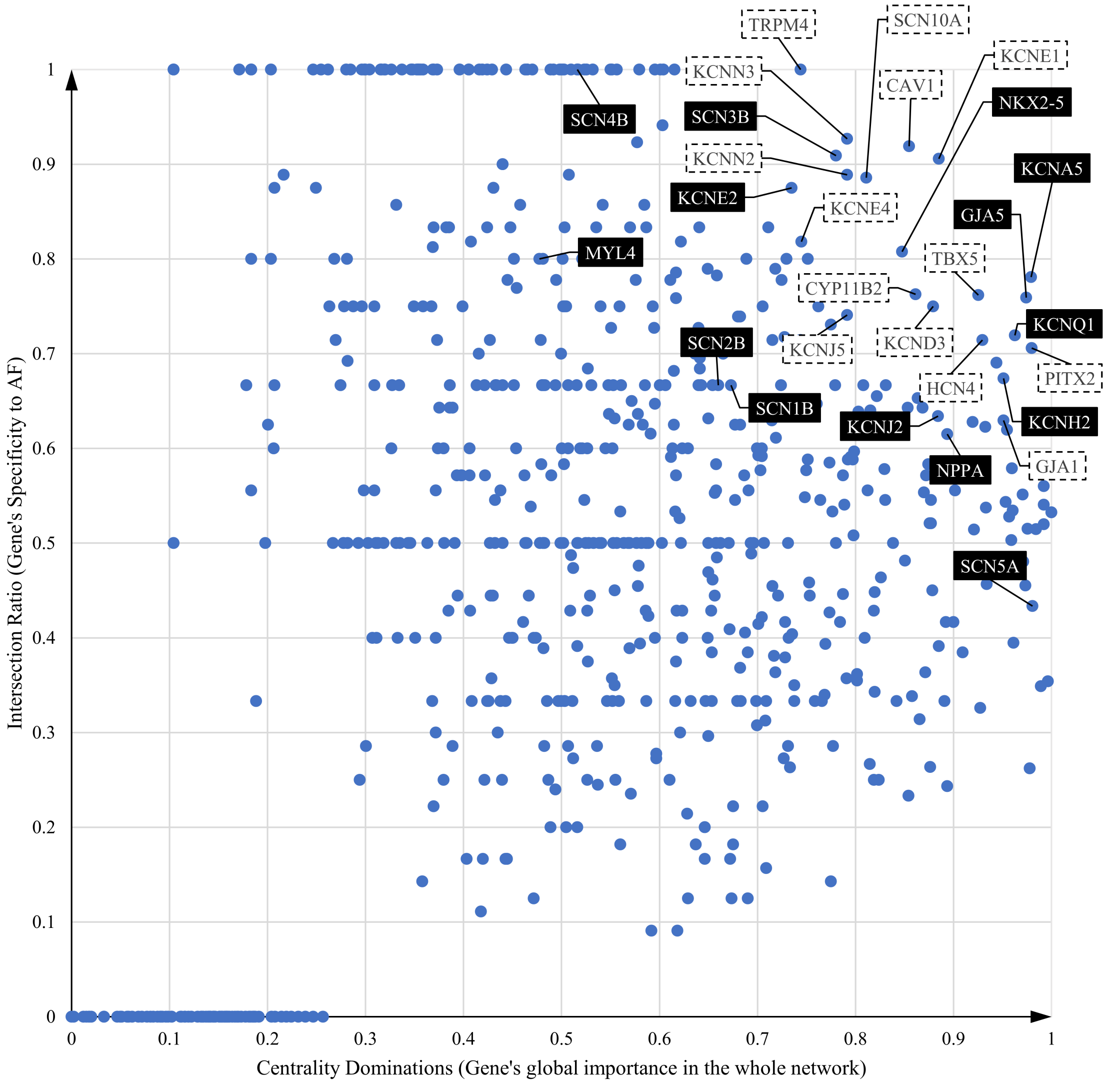


Figure 3.3 : Gene map in an importance-specificity coordinate system

posed this against normalised intersection ratios, which reflect the gene's specificity to AF. This produced a 2-D gene scatter map of importance v.s. specificity, as shown in Figure 3.3. Global importance is plotted on the X-axis and specificity is plotted on the Y-axis. The genes of most concern to us are those in the top right corner – i.e., the genes with both high centrality domination and a high intersection ratio, which means they are not only essential but also specific to AF. However, part of the purpose of this case study is to evaluate this framework. Hence, we corroborated these results with a manual review of three biomedical knowledge bases: Online Mendelian Inheritance in Man (OMIM)[5], the Kyoto Encyclopedia of Genes and Genomes (KEGG)[6], and the Genetics Home Reference-NIH[7]. Throughout the investigation, we divided the core genes into two loci: seed genes – genes with known functions associated with the incidence of AF; and suspected correlated genes – genes with unknown functions that are possibly related to AF but yet to be explored):

i. **Seed genes (in black boxes)**: Nine genes with direct associations to AF are documented in knowledge bases. According to OMIM, most of the different subtypes of AF[8] are caused by mutations or variations in these nine genes. The noted gene and subtype correlations are as follows: *KCNQ1*-AF subtype 3, *KCNE2*-AF subtype 4, *KCNA5*-AF subtype 7, *KCNJ2*-AF subtype 9, *NPPA*-AF subtype 6, *GJA5*-AF subtype 11, and *SCN3B*-AF subtype 17. Although OMIM does not explicitly state associations with particular AF subtypes for the other two genes in this group, *KCNH2* and *NKX2–5*, the Genetic Home Reference-NIH lists them as significant genes in AF's progression.

[5] More information could be found at https://www.omim.org/

[6] More information could be found at https://www.genome.jp/kegg/

[7] More information could be found at https://ghr.nlm.nih.gov/

[8] More information can be found at https://omim.org/entry/608583

In examining the significance of those genes from the literature, we found that Sinner et al. (2008) identified a positive correlation between mutation *K897T* in *KCNH2* and a higher incidence of AF. Similarly, Xie et al. (2013) associated the *NKX2–5* loss-of-function mutations *p.N19D* and *p.F186S* with AF via a cohort study on 136 patients with idiopathic atrial fibrillation. It is worth highlighting that our framework placed all seed genes prominently; 36%, i.e., *SCN5A*, *MYL4*, *SCN1B*, *SCN2B*, had either lower IR or centrality domination, pushing them out of the top right corner toward the left or bottom.

ii. **Suspected gene loci (in white boxes)**: Fourteen genes in the list are frequently studied because their mutations or variations are statistically proven to be associated with AF. Hence, they are suspected genetic factors, but the underlying mechanisms as to why are less understood than with the seed genes. The 14 genes are *KCNE1*, *KCNN2*, *KCNN3*, *KCNJ5*, *KCND3*, *CAV1*, *SCN10A*, *TBX5*, *PITX2*, *ZFHX3*, *GJA1*, *HCN4*, *CYP11B2*, and *TRPM4*. The literature review revealed the following mutation/variation associations: *G25V* and *G60D* in *KCNE1* (Olesen et al., 2012), *rs337711* in *KCCN2* and *rs75190942* in *KCNJ5* (Christophersen et al., 2017), *rs13376333* to *KCNN3* (Ellinor et al., 2010), *rs12044963* in *KCND3*, *rs11773845* in *CAV1*, *rs6790396* in *SCN10A*, *rs883079* in *TBX5*, *rs2129977* in *PITX2*, *rs2359171* in *ZFHX3*, *rs13191450* in *GJA1*, *rs74022964* in *HCN4* (Roselli et al., 2020), *T-344C* in *CYP11B2* (Li et al., 2012).Düzen et al. (2017) found that *TRM4* expression was significantly upregulated in leukocytes of non-valvular AF patients.

Other core entities

We then applied non-dominated sorting to the other three entity categories and generated the corresponding core entity lists. The top 20 diseases, chemicals, genes, and genetic variants are given in Table 3.4. To evaluate the quality of the sorted genetic factors, we compared our results with data from the authoritative disease-gene association discovery database DisGeNET (Piñero et al., 2016). DisGeNET integrates data from various sources, including curated knowledge bases, modelled data, inferred data, and the literature[9]. Users can also rank the associations between diseases and genetic factors according to several provided metrics. We chose the gene-disease association score (GDA) and variant-disease association score (VDA)[10] as the best comparison to our results.

In the disease category, terms in normal type are the most common physiological or pathological phenomenon relevant to the presence or treatment of AF. The terms in italics are symptoms, complications, and risk factors. Awareness of these concepts is critical to understanding the treatment of AF. One noticeable term in the list is gastroesophageal reflux, frequently reported in AF patients. However, judging from current research progress, any association between the two is still inconclusive (Huang et al., 2019). Further studies to supplement the literature may reveal gastroesophageal reflux has underlying significance to this research area.

In the chemicals list, terms in roman are treatments, while terms in italics are critical receptors and ion channels in the pathogenesis of AF. Caffeine and Omega-3 fatty acids are two noticeable chemicals on the list. Over years of research, the association between caffeine and AF has intriguingly been reversed from a risk factor (Curatolo and Robertson, 1983) to one with potential preventive benefits (Abdelfattah et al., 2018). The inconsistency of these results warrants further research to provide clear evidence on the issue. The same turnabout is true of Omega-3. Once lauded as a health supplement to reduce cardiovascular disease (Abdelhamid et al., 2018), Sheikh et al. (2019) now report that Omega-3 might increase the incidence of AF. The controversy has yet to be settled. These two results show that the proposed

[9] More information can be found at https://www.disgenet.org/dbinfo

[10] More information about the metrics could be found at https://www.disgenet.org/dbinfo

Table 3.4 : The top 20 bioentities of each category

Rank	Disease	Chemical	Gene		Genetic variant	
			Symbol	**DisGeNET Ranking**†	**SNP ID**	**DisGeNET Ranking**
#1	*Fibrosis*	Acetylcholine	*KCNA5*	Gene: 3/73	rs2200733	AF: 1/584
#2	Atrial Remodeling	AVE0118	*GJA5*	AF: 4/939	rs13376333	AF: 5/584
#3	Atrial Flutter	*Ryanodine*	PITX2	AF: 2/939	rs2108622	Variant: 6/20
#4	Arrhythmias, Cardiac	Diltiazem	KCNE1	Gene: 10/95	rs1805127	Variant: 1/17
#5	Myocardial Stunning	Propafenone	*KCNQ1*	Gene: 6/281	rs699	Variant: 21/134
#6	*Inflammation*	Sotalol	TBX5	AF: 7/939	*rs3789678*	(Zhao et al., 2015)
#7	Tachycardia, Supraventricular	**Caffeine**	*KCNH2*	AF: 1/939	rs6795970	Variant: 5/13
#8	Atrial Premature Complexes	Isoproterenol	ZFHX3	AF: 5/939	rs3807989	AF: 4/584
#9	Mitral Valve Stenosis	Quinidine	CAV1	AF: 8/939	rs2106261	AF: 3/584
#10	*Stroke*	Verapamil	HCN4	AF: 9/939	rs10033464	AF: 2/584
#11	Heterotaxy Syndrome	*Potassium*	*NKX2-5*	AF: 12/939	rs17042171	AF: 12/584
#12	**Gastroesophageal Reflux**	*Magnesium*	CYP11B2	AF: 119/939	rs7193343	AF: 6/584
#13	Sick Sinus Syndrome	Procainamide	KCND3	Gene: 4/94	rs7164883	AF: 7/584
#14	*Stenosis, Pulmonary Vein*	Digoxin	SCN10A	AF: 10/939	rs6584555	AF: 21/584
#15	*Thromboembolism*	*Calcium*	KCNN2	Gene: 1/18	**rs121912507**	Variant: 3/4
#16	*Sleep Apnea, Obstructive*	Flecainide	KCNN3	AF: 6/939	rs1805120	Variant: 1/3
#17	*Rheumatic Diseases*	*Ibutilide*	*SCN3B*	Gene: 8/33	rs120074192	Variant: 1/10
#18	Atrioventricular Block	**Fatty Acids, Omega-3**	*NPPA*	Gene: 12/217	rs3903239	AF: 16/584
#19	Heart Disease	*Adenosine Triphosphate*	CRP	AF: 181/939	rs1152591	AF: 14/584
#20	Venous Thromboembolism	*Sodium*	*KCNJ2*	AF: 11/939	rs10824026	AF: 11/584

† *"AF" refers to the gene's ranking in the list of AF-associated genes, indicating the gene's importance to AF. "Gene" refers to AF's ranking in the gene's list of associated diseases, indicating the gene's specificity to AF. The same rule applies to genetic variant ranking.*

Note: the entities in regular, italic font and bold font respectively represent different types of core entities, please refer to the following explanations for details.

method can identify debated chemicals for further exploration.

To validate the identified genetic variants, we used ClinVar (Landrum et al., 2016), SNPedia (Cariaso and Lennon, 2012), rankings from DisGeNET, and complementary evidence from the literature. The variants in italics are variants of seed genes. We found most of the factors identified have known associations with AF, indicating their importance or specificity to AF. One exception is *rs3789678*, which does not appear in DisGeNET but is noted as having a significant association with AF in the literature (Zhao et al., 2015). Additionally, there was a notable mutation *rs121912507*, which refers to the *G628* gene transfer in *KCNH2*. This SNP is not directly related to the occurrence of AF but is an adenovirus-mediated transgene expression that could be used as an effective gene therapy to prevent postoperative AF.

From this analysis, we are confident in concluding that our approach can identify, with relatively good accuracy, a list of bioentities strongly related to a given target disease. Compared to traditional approaches, such as term frequency or TF-IDF value-based sorting algorithms, this strategy produces a list of relevant, specific, and frontier entities that are not biased by the popularity of research topics.

3.3.4 Link prediction validation

Before running the link prediction algorithm, we validated its usefulness on rolled-back data. The experiment was designed as follows:

We divided the dataset into five-year brackets and constructed a network for each. k AF-linked genes or SNPs identified in the last five years were used as the true-labelled samples in the test set. We then tested our semantics-enhanced version of the RA algorithm (SERA) along with four other methods on the remaining data and compared the results. The link predictions were output as a mixed list of genes, and SNPs were ranked according to their RA index scores. Any gene or SNP predicted in the top n that was also in the true label set was counted as a true positive (TP) and a false negative (FN) otherwise. n is a threshold we initially set to k. The four methods chosen for comparison were:

i. RA: The original version of the resource allocation approach (Zhou et al., 2009).

ii. WRA (Lü and Zhou, 2010): The weighted version of the resource allocation algorithm. The assumption of this algorithm is the same as RA, but the diffusing rate is measured as a weight ratio instead of as a proportion of degree centrality.

iii. LPI (Lü et al., 2009): Local path index is a local similarity-based index calculated by the weighted sum of the number of paths of lengths two and three. We used the default settings of 1 and 0.01 respectively for the paths of lengths two and three.

iv. RWR (Tong et al., 2008): Random walk with restart is based on a global similarity measurement for pairwise nodes. The idea of the random walk is that a particle starts from a seed node and randomly jumps to a connected node with a probability of p. The ultimate probability of the particle reaching a target node after a certain number of iterations is the possibility of forming a direct link between the seed and target node (Lovász, 1993).

To fairly evaluate the performance of all algorithms, we used the top n hit rate as the assessment metric, defined as:

$$\text{Top } n \text{ hit rate} = \frac{TP}{TP + FN} \tag{3.10}$$

TP is the correct results in the top n predictions, and FN is the number of other samples in k true-labelled samples but not in top n predictions.

Table 3.5 : Experimental results

	RWR	LPI	RA	WRA	**SERA (proposed)**
Top k hit rate	0.183	0.132	0.205	0.212	**0.283**
Top 100 hit rate	0.445	0.181	0.436	0.392	**0.502**
Top 200 hit rate	0.621	0.576	0.714	0.632	**0.742**

The outcomes are provided in Table 3.5. SERA had a better hit rate than the other four baselines, validating its effectiveness. The top 200 predictions covered 74% of the genetic factor associations that would appear in the next five years, according to the true label set. This promising result demonstrates that our strategy can substantially reduce the heavy workload of manually seeking new candidate factors. However, the low hit rate for the complete list of predictions (top k) is less than optimal. We can conclude four reasons for this result: 1) Our experiment effectively simulates data streaming over time, which is a strict standard from the validation perspective. 2) Our approach is purely data-driven without any human intervention or supervision. 3) The network is co-occurrence-based and cannot distinguish between positive and negative associations, such as “A is not associated with B”. 4) The community’s awareness of AF molecular mechanisms is still at a relatively early stage, and not all possible discoveries were made in the last five years. Therefore, it is reasonable to assume that some predicted associations may exist but require more time to uncover.

3.3.5 Predicting the future emerging genetic factors

The top 15 genes and SNPs from the link prediction procedure with the entire network are in Table 3.6. The results were validated against the DisGeNET database and the literature. Detailed explanations of the identified evidence are given below.

Table 3.6 : Predicted disease-gene associations and evidence

Rank	Gene/Variant	DisGeNET[1]	P/N[2]	Evidence[3]	Evidence level[4]
#1	Gene: BGLAP	–	+	The activity of the encoded protein of BGLAP can be reduced by Warfarin, a treatment for AF (Sato et al., 2010; Yamagishi, 2019)	B
#2	SNP:rs4762	–	+	rs4762 was identified as not significantly associated with AF by Zhao et al. (2015); Kuken et al. (2020) found the opposite.	B
#3	SNP: rs337711	0.700	+	Bentzen et al. (2020) and Christophersen et al. (2017) find no significant association; Wang et al. (2018) found the opposite.	A
#4	SNP: rs11264280	0.700	+	Wang et al. (2018) found no significant association; Pan et al. (2020) found the opposite.	A
#5	Gene: HP	–	+	Eryd et al. (2011) found no significant association between HP's protein product and the incidence of AF.	C
#6	Gene: PKP2	0.400	+	Yeung et al. (2019) found patients with mutations in desmosomal PKP2 have smaller atria. Alhassani et al. (2018) found large PKP2 deletion is related to lone AF.	A
#7	Gene: DUOX2	–	–	No clear findings	C
#8	Gene: OLR1	0.010	+	The activity of OCR1 increased by atrial modelling during AF (Bukowska et al., 2008)	A
#9	SNP: rs3097	–	–	Zhao et al. (2015) found no significant association.	C
#10	Gene: MGP	–	+	Warfarin treatment inhibits this gene expression.	B
#11	Gene: S100A6	-	+	The encoded protein of S100A6 combined with RTEN is a potential biomarker of AF (Doulamis et al., 2019)	B
#12	Gene: IL3	–	–	No clear findings	C
#13	Gene: IL20	–	–	No clear findings	C
#14	Gene: TFF3	–	+	The encoded protein of TFF3 combined with P3NP is a potential biomarker of AF (Doulamis et al., 2019)	B
#15	Gene: NOX4	0.020	+	NOX4 may be a gene therapy for ibrutinib-induced AF (Chen et al., 2019; Yang et al., 2020a)	A

[1] *The association score with AF retrieved from DisGeNET.*

[2] *The overall evaluation of evidence identified from the literature, ”+” indicates positive and ”−” indicates negative.*

[3] *Sourced from LitVar, SNPedia &/or Literature Retrieval.*

[4] *A: positive evidence found both in DisGeNET and the literature; B: positive evidence found only in literature; C: neither the literature nor DisGeNET provide evidence of an association.*

- #1 *BGLAP* and #10 *MGP*: These are associations to treatments for AF as opposed to AF's cause. Sato et al. (2010) and Yamagishi (2019) discovered that long-term use of Warfarin, a regular treatment used in non-rheumatic atrial fibrillation, significantly reduces the activity of osteocalcin (BGLAP) which protects bones from fracture and inhibits the effectiveness of matrix Gla-protein (MGP) in preventing vascular calcification. Those discoveries have led to alternative recommended treatments for AF patients with a high risk of fracture or vascular calcification, such as non-vitamin K oral anticoagulants (NOACs).

- #2 *rs4762*: Despite Zhao et al. (2015) finding that this mutation is not significantly associated with AF, Kuken et al. (2020) recently published a study showing that *AGTT 174 M* (*rs4762*) is associated with the occurrence of AF in the Han and Uyghur ethnic groups in Xinjiang, China. These conflicting results suggest that this SNP may perform differently for different ethnicities.

- #3 *rs337711*: A large-scale genome-wide association study identified a correlation between *rs337711* and AF with a significant statistical P-value (Christophersen et al., 2017). However, another experiment based on vitro electrophysiology analysis and animal models failed to capture the association of this SNP with any atrial or ventricular changes in *KCNN2* mRNA expression (Bentzen et al., 2020). These contrasting results may lead to further exploration of the molecular mechanism of this SNP.

- #4 *rs11264280*: Wang et al. (2018) conducted a clinical case study in the Chinese Han population but did not identify any significant correlation between *rs11264280* and AF. However, a later Mendelian randomisation study conducted by Pan et al. (2020) indicates a notable correlation between this SNP and AF at a P-value of 3.07×10^{-79}, which is far smaller than the universal conspicuous level. Again, the conflicting results may suggest that this SNP performs differently for different ethnicity.

- #5 *HP*: Eryd et al. (2011) conducted a cohort study to identify the association of haptoglobin (HP) level with AF and identified an insignificant level of correlation.

- #6 *PKP2*: Several studies mention a potential association between the gene *PKP2* and AF. Bourfiss et al. (2016) discovered that mutation in desmosomal *PKP2* could result in a significantly smaller atrial size in AF patients, which suggests a different arrhythmogenic mechanism of AF. Alhassani et al. (2018) reported a family case with large pathogenic *PKP2* deletion, resulting in cardiac arrhythmias, including persistent lone AF. These researchers also claim that AF occurring as genetic ventricular cardiomyopathy could be a secondary phenotype of a common underlying genetic variant.

- #9 *rs3907*: The associations between AF and rs3907 have barely been investigated at the current stage. According to the minimal existing evidence (Zhao et al., 2015), this SNP is not significantly correlated with AF.

- #11 *S100A6* and #14 *TFF-3*: A few studies have been conducted on the potential associations between these two genes and AF. The limited evidence suggests that the genetic combinations of *TFF-3* and *P3NP*, *S100A6* and *RETN* may be biomarkers for AF (Doulamis et al., 2019).

- #15 *NOX4*: Mounting evidence is revealing an association between *NOX4* and AF. Chen et al. (2019) evaluated the mediation of *CD44*/*NOX4* signals in atrial tachycardia-induced oxidative stress and Ca2+-handling abnormalities, providing a possible explanation for the onset/progression of AF. Yang et al. (2020a) identified elevated expressions of *NOX4* in an ibrutinib-induced AF mice group and proposed inhibiting *NOX* as a potential novel AF therapy for ibrutinib-induced AF.

- We could not find supporting evidence for the other four predictions. However, in conversations with several domain experts, we were advised that the field is in a relatively early stage of research progress, and more time is needed to examine such associations. These empirical insights can provide a clear direction for future research undertakings.

3.4 Summary

Considering individual genetic variability in contemporary medical research has become a research frontier. The recognition of disease-specific genetic foundations has proven to be invaluable for enhancing risk assessment, diagnostics, and treatment strategies. To address this, our framework leverages scientific literature data to identify known associations between bioentities and diseases, integrating these findings with network analytics and link prediction techniques. In this modified link prediction algorithm, we utilize the Bioentity2Vec model to obtain semantic similarities between entities. Although pre-trained language models, such as BERT and GloVe, have exhibited impressive performance in various downstream tasks, including text classification and named entity recognition, their effectiveness often relies on fine-tuning with labelled datasets tailored to specific domains. However, in the context of the HBAM framework, the lack of labelled datasets poses a challenge to the fine-tuning process. As suggested by Reimers and Gurevych (2019), the inherent design of BERT makes it less suitable for text-unsupervised tasks such as semantic similarity search or clustering, which aligns with the nature of our HBAM research problem. Consequently, we have opted to employ the bioentity2vec model on a limited dataset to derive contextual embeddings for further investigation. The results of a comprehensive case study indicate that our strategy offers promising potential as a solution for entity association prediction and recommendation.

Our empirical study focused on atrial fibrillation. The results of the case analysis presented some critical bioentities associated with AF. However, they also revealed controversial findings, such as the association between AF and gastroesophageal reflux, Omega-3 fatty acids, and caffeine. Therefore, from one perspective, the framework can be seen as a tool for generating a data-driven, bird's eye view of cardiovascular research. From another perspective, it is a decision support system that produces insights into prior research that may need to be re-examined or pointers toward future research that is likely to prove fruitful.

Chapter 4

Hierarchical Topic Tree - I Model

4.1 Introduction

The last decades have witnessed a significant 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). What many of those studies have shown is that organising 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. While very broad overviews of a field are not particularly difficult to generate, creating interactive topic maps that show fields at different levels of granularity and disentangling the rising complexities of inter-/multi- disciplinary studies is another story altogether. To our knowledge, current 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 literature dataset. Among these methods, hierarchical latent Dirichlet allocation (hLDA) (Blei et al., 2010) is especially well-known. However, a couple of aspects of hLDA could be improved. These include occasional weak associations between the generated parent and child topics; 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 to building topic hierarchies, such as taxonomy identification (Shang et al., 2018), 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 an excessive number of parameters that need to be fine-tuned or issues with creating clean partitions between topics. Current hard clustering algorithms like K-means or non-negative matrix factorization (Qian et al., 2020; Zhang et al., 2018b), 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 hierarchical topic extraction model called Hierarchical Topic Tree - I(HTT-I). The model 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 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 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 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-I framework, we conducted three case studies on computer science, AI ethics and digital transformation literature datasets. The main contributions our work makes include: 1) a density peak search algorithm that identifies and labels the topics in a corpus; 2) an overlapping community allocation algorithm that recognises 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 chapter is organised as follows. Section 4.2 sets out the details of our proposed methodology. Sections 4.3, 4.4, 4.5 follow, presenting the data, results, and empirical insights derived from three case studies. We then wrap up our study with a conclusion, the study's limitations, and future research directions.

4.2 HTT-I Methodology

4.2.1 Concept definitions and problem formulation

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

i. Topic term: Nominal words and phrases extracted from scientific literature textual data. The terms can come from data sources or be extracted from titles, abstracts, or full texts via NLP and cleaning steps.

ii. 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.

iii. Hierarchical topic tree - I (HTT-I): HTT-I is both the name of our proposed method and the final output. As an output, an HTT-I is a tree structure consisting of topic nodes residing on different tree layers, as illustrated in Figure 4.1. 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

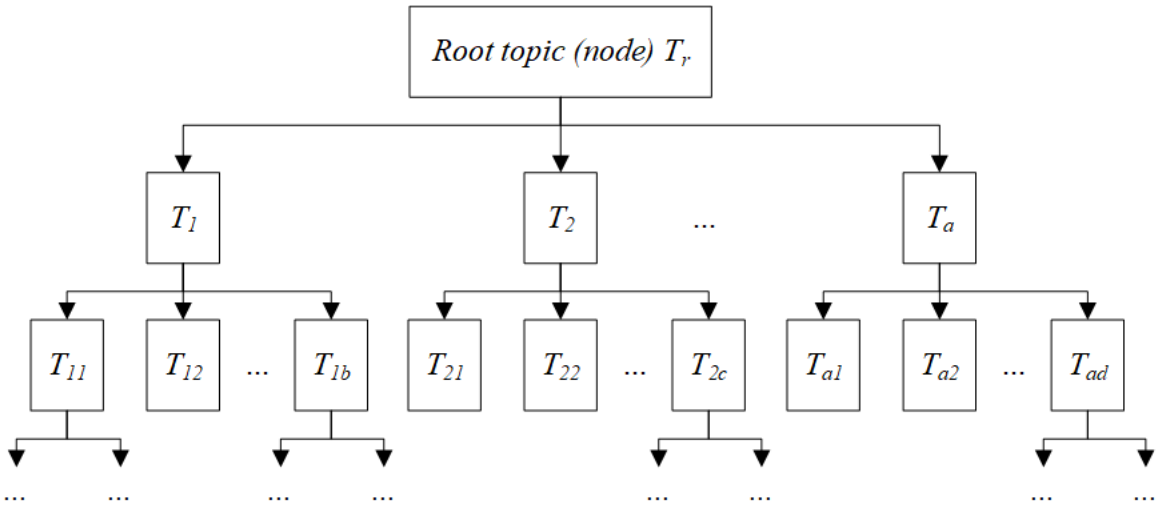


Figure 4.1 : Illustration of an HTT-I example

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 associations between siblings.

iv. Problem formulation: The methodology aims to: 1) identify research topics with different granularity 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.

4.2.2 Data pre-processing 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., 2014b) or any other term extraction workflow. With the extracted terms, the next step is constructing 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

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

set of edges representing term co-occurrence. The graph is formulated according to the following equation:

$$w_{V_iV_j,i\neq j} = \begin{cases} \frac{1}{CF(V_i,V_j)} & \text{if } V_i \text{ and } V_j \text{ co-occur in at least one document} \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

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

4.2.3 Density peak search (DPS)

Density peak clustering was first proposed by a paper on *Science* by Rodriguez and Laio (2014). It is based on the premise that the centre 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 one-off clustering method, the density peak clustering method is more straightforward and computationally efficient than traditional K-means or density-based clustering algorithms like DBSCAN. There are no additional parameters and multiple iterations, meaning the clustering process is highly robust to parameter selection. Du et al. (2016) have since improved this method by using average K-nearest neighbour (KNN) density to emphasise the importance of local density instead of the original circle radius approach. 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 related terms but as different as possible from other topic labels. This parallel motivated our idea to name topics through a KNN-modified density peak-based clustering algorithm automatically.

This algorithm is also designed to identify core terms for topic labels. When applying density peak clustering to network data, the primary concern 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 reduce the model's adaptability. Therefore, we opted to develop a new distance proxy, although still based on the topological distance between nodes:

$$d_{V_iV_j,i\neq j} = \begin{cases} w_{V_iV_j,i\neq j} & \text{if } V_i \text{ and } V_j \text{ are connected} \\ SPL_{V_iV_j,i\neq j} & \text{if } V_i \text{ and } V_j \text{ are unconnected but a path exists between them} \\ NA & \text{if no path exists between } V_i \text{ and } V_j \end{cases} \tag{4.2}$$

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:

$$\rho_{V_i} = exp(-\frac{1}{K}\sum_{j\in KNN(V_j)} d^2_{V_iV_j}) \tag{4.3}$$

$$\delta_{V_i} = \begin{cases} \max\limits_{V_j} d(V_i, V_j) & \text{if } V_i \text{ and } V_j \text{ co-occur in at least one document} \\ \min\limits_{V_j \in V_{\rho_{V_j} > \rho_{V_i}}} d(V_i, V_j) & \text{otherwise} \end{cases} \tag{4.4}$$

In the few cases where the co-occurrence network includes several unconnected components, we will generate a virtual root node for the final HTT-I. 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-I, 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:

i. Density peak: The selected centroids should be density peaks, denoted as:

$$\rho_{V_c} = \max_{V_i \in KNN(V_c)} \rho_{V_i} \tag{4.5}$$

where $KNN(V_c)$ denotes the K-nearnest neighbour nodes of V_c.

ii. Centroid 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:

$$\sigma_{V_c} > d_{V_r V_c} \tag{4.6}$$

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 will be generated, and the n identified nodes will become children of the virtual root.

4.2.4 Overlapping community allocation (OCA)

The next step is to distinguish overlapping topics between communities and ensure they are given multiple proper assignments. Thus, every node is assigned a probability vector $p_{V_i} = \{p_{i,1}, p_{i,2}, p_{i,3}, \ldots, p_{i,n}\}$, 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:

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

4.2.5 Recursive hierarchy detection

The previous steps partition the network into n subnetworks, with each subnetwork comprising a core topic term and a set of affiliated terms. 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 number of terms in the subnetwork is less than K.

The output of this step is the finalised 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 4.2, where each colour represents a different stratum in the hierarchy. From top to bottom, the HTT-I has a root topic and one or multiple layers of topics generated by the iterations of DPS and OCA algorithms. Topics generated in the same iteration are siblings to each other and share a mutual parent topic.

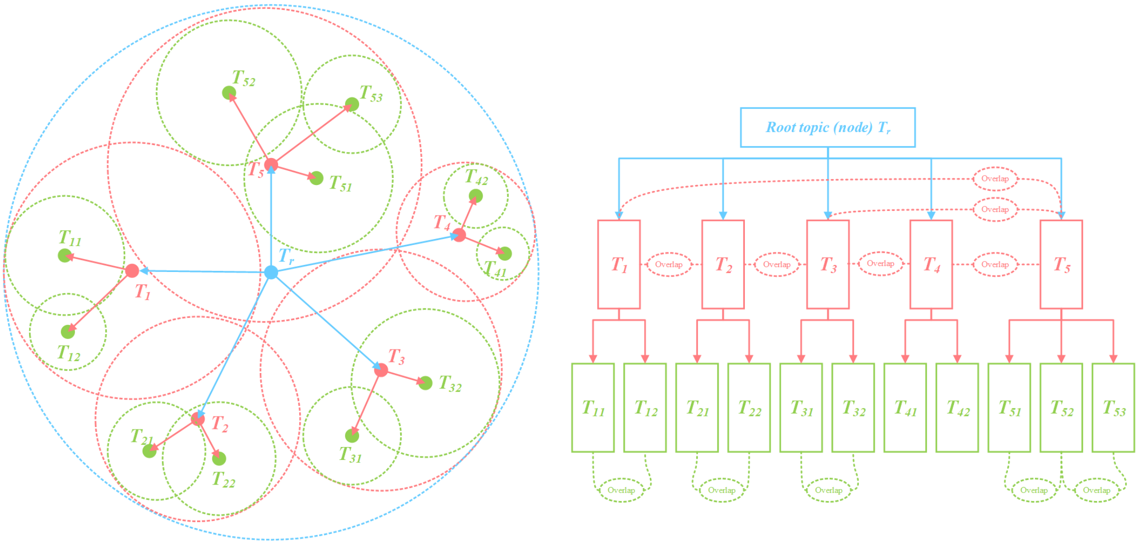


Figure 4.2 : The recursive process of hierarchy construction

4.2.6 Evaluation indicators

According to criteria from previous studies, a well-curated hierarchical 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 topic association (PCTA) to quantify the two characteristics. Additionally, we calculated the weight loss ratio of network edges to measure the information loss in the HTT-I process. Please note that the topics mentioned in this section contain overlapping terms. The association strength between two topics means the sum of the edge weight's reciprocal of the pairwise terms from the two topics, and the internal topic association of a topic strength refers to the sum of the edge weight's reciprocal of pairwise terms from the topic itself.

i. 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 $-\infty$

to $+\infty$ and its values vary hugely in multiple studies (Qian et al., 2020; Xu et al., 2018; Wang et al., 2013). Hence, in this study, we measure the coherence of a topic by calculating the proportion of its total internal association strength against its total association strength with itself and its siblings. The calculating formula is as follows:

$$Coherence_{T_i} = \frac{1}{|T_i|}\sum_{V_M\in T_i}\frac{\sum_{V_n\in T_i, m\neq n} CF(V_m, V_n)}{\sum_{T_j\in children(parent(T_i))}\sum_{V_k\in T_j} CF(V_m, V_k)} \tag{4.7}$$

ii. Parent-child topic association (PCTA): This indicator is only applied to parent nodes in the final HTT-I (including the virtual root node if it exists). For every parent node, the PCTA 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. It proxies how strongly the parent topics are associated with their corresponding child topics.

$$PCTA_{T_i} = 1 - \frac{\sum_{T_m,T_n\in children(T_i),m\neq n}\sum_{V_p\in T_m^K, V_q\in T_n^K} CF(V_p, V_q)}{\sum_{T_j\in children(T_i)}\sum_{V_x\in T_j^K, V_y\in T_i^K} CF(V_x, V_y)} \tag{4.8}$$

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

$$\text{Information loss index}_{T_i} = \frac{\sum_{T_m,T_n\in children(T_i),m\neq n}\sum_{V_p\in T_m, V_q\ inT_n} CF(V_p, V_q)}{\sum_{V_x\in T_i, V_y\in T_i, x\neq y} CF(V_x, V_y)} \tag{4.9}$$

4.3 Case Study I: Topic Hierarchies in the Computer Science Discipline

4.3.1 Data collection and pre-processing

To demonstrate the proposed methodology, we conducted a case study on the field of computer science, decomposing its many and varied research interests into topic hierarchies. The corpus in this case study comprised 6,267 highly-cited papers published between 2010 and 2021 retrieved from the Web of Science (WoS) core collection database. 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.

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 5. The stepwise cleaning results are given in Table 4.1. The final output was a term co-occurrence network consisting of 2,134 terms.

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

Table 4.1 : Step-wise results of the pre-processing procedure

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 and 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 five 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.

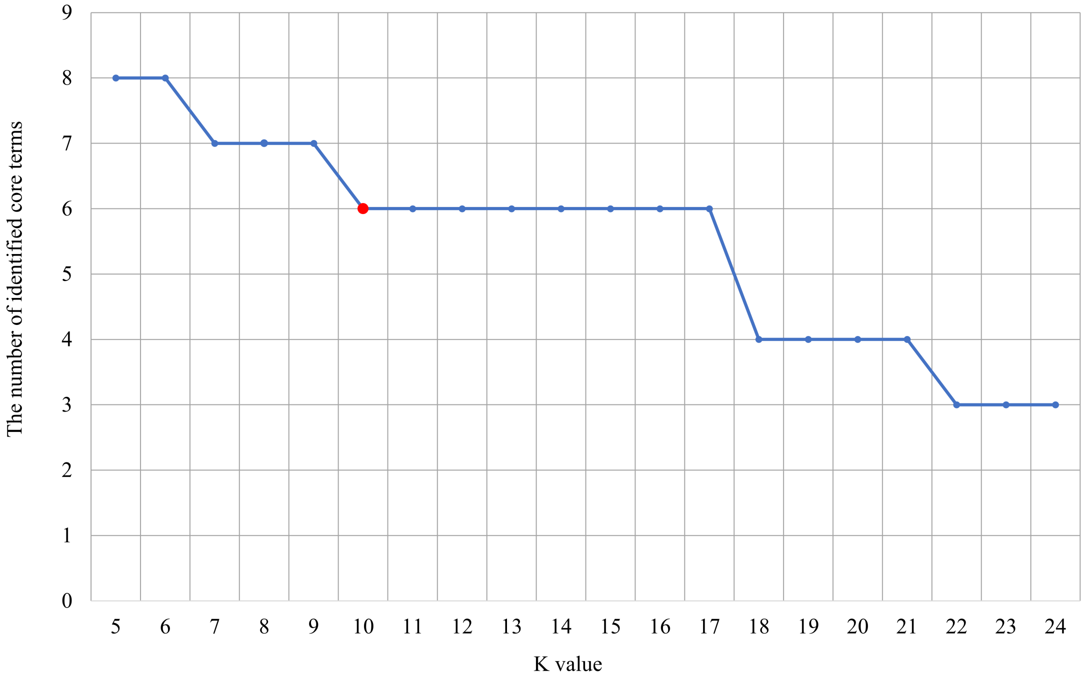


Figure 4.3 : K against the number of identified core topic terms plot

4.3.2 HTT-I result and interpretation

Before generating the HTT-I, we selected appropriate values for the KNN density parameter K and the overlap threshold σ. Optimal values of K were determined through sensitivity analysis by monitoring the number of initially identified core topic terms against K. The corresponding plot is presented in Figure 4.3.

The first round of tests 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.

With the term co-occurrence network as input to the DPS and OCA algorithms, the graph was recursively partitioned into subnetworks of topics in different layers. The overlaps between topics were evaluated and assigned accordingly. The algorithms stopped at the eighth iteration, yielding a nine-level HTT-I of computer science research. Figures 4.4[3] and 4.5 illustrate the HTT-I result and detailed terms in topics and their overlaps, respectively.

To evaluate the performance of HTT-I in this case, we calculated the average topic coherence, PCTA, and information loss of the final HTT-I, with their values as 0.619, 0.847, and 6%, respectively. The high PCTA 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-I 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 4.4, 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 the others. Simple observation confirms that the selected label terms with high density also represent the terms they lead. Drilling down into each of the six initial parents, #1 *Deep learning* branches off into topics that pertain to convolutional neural networks and then onwards to the relevant tasks they are used to solve, e.g., *computer vision*, *image segmentation*, etc. The lower branch of this topic groups the models and metrics associated with *deep learning*, such as *random forests* and *prediction accuracy.* #2 *Optimisation* problems span different techniques, algorithms, and research objects associated with optimisation and its sub-problems. #3 *Decision-making* captures the models, strategies, and sub-problems relevant to decision intelligence and its processes. #4 *Operating systems* group the research topics surrounding computing architectures and software, a fun-

[3] Constraints on the page size limit the tree to its top three layers. The total HTT-I result is available at HTT-Ips://github.com/IntelligentBibliometrics/HTT-I/blob/main/CS%20case%20results.png

damental aspect of computer science. #5 The *Internet of Things (IoT)* connects big data and sensor technology with its many application spheres. Last, #6 *Closed-loop systems* 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 4.5. The overlapping terms between #1 and #2 include *data mining*, *classification accuracy*, and *classification tasks*, which are universal concepts for both deep learning and optimisation studies. Overlapping terms of topics #2 and #4 describe two programming tools (*R*, *MATLAB*) and computer performance (*enhanced performance*, *CPU time*). This overlap indicates a direction of solving optimisation 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 neighbours in this term co-occurrence network, and *deep learning* has a higher KNN density. This reflects that *deep learning* has overshadowed its precursor technologies to become the more dominant research focus.

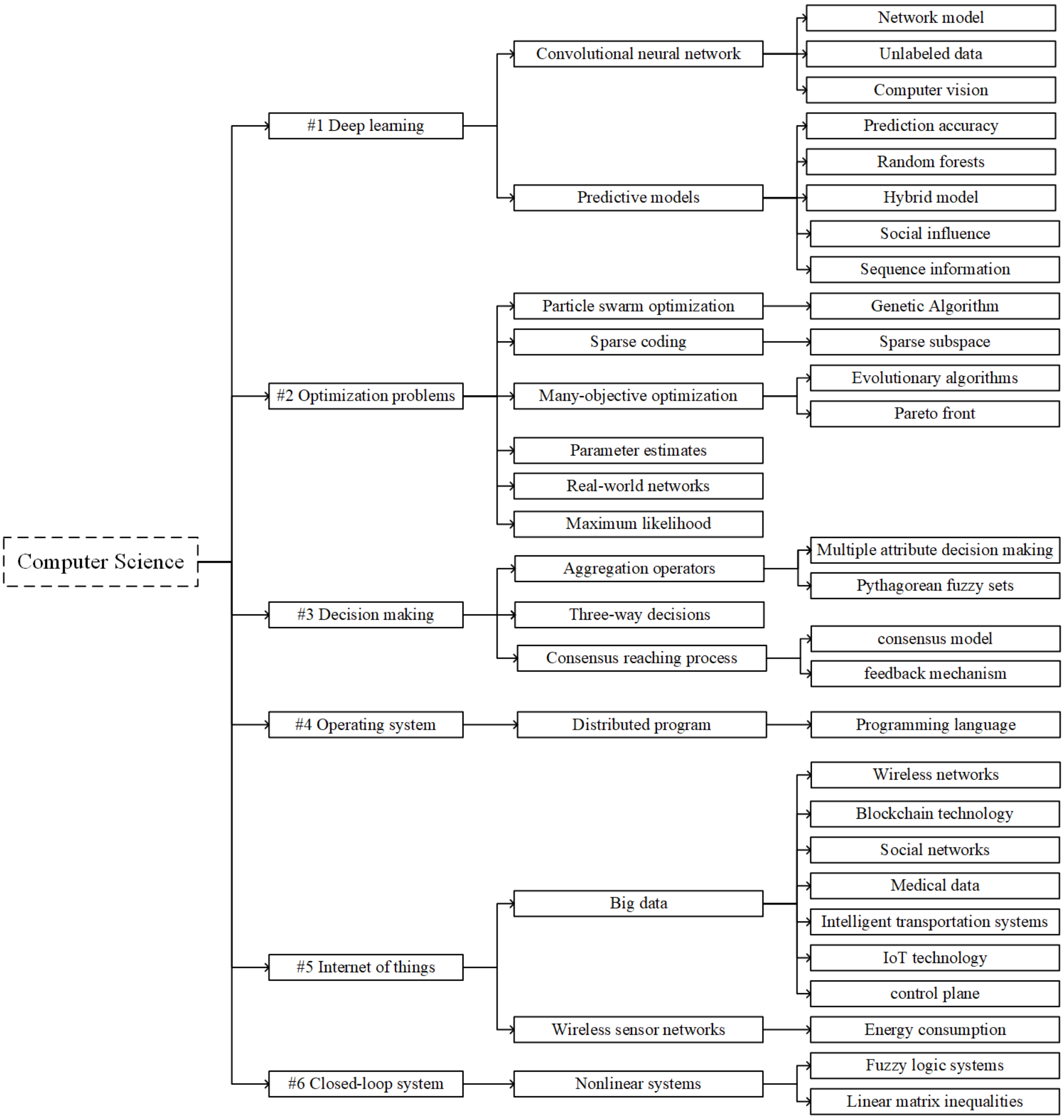
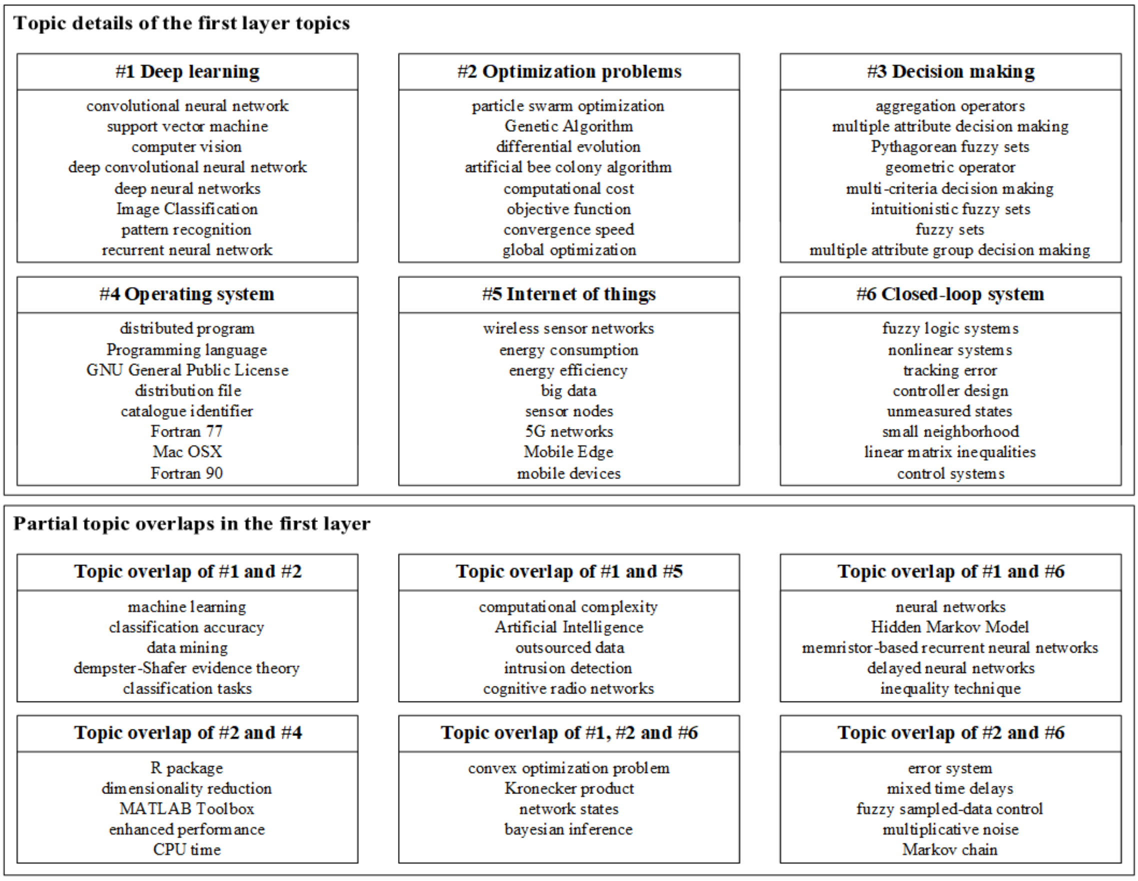


Figure 4.4 : The HTT-I result for computer science



Topic details of the first layer topics

#1 Deep learning

convolutional neural network
support vector machine
computer vision
deep convolutional neural network
deep neural networks
Image Classification
pattern recognition
recurrent neural network

#2 Optimization problems

particle swarm optimization
Genetic Algorithm
differential evolution
artificial bee colony algorithm
computational cost
objective function
convergence speed
global optimization

#3 Decision making

aggregation operators
multiple attribute decision making
Pythagorean fuzzy sets
geometric operator
multi-criteria decision making
intuitionistic fuzzy sets
fuzzy sets
multiple attribute group decision making

#4 Operating system

distributed program
Programming language
GNU General Public License
distribution file
catalogue identifier
Fortran 77
Mac OSX
Fortran 90

#5 Internet of things

wireless sensor networks
energy consumption
energy efficiency
big data
sensor nodes
5G networks
Mobile Edge
mobile devices

#6 Closed-loop system

fuzzy logic systems
nonlinear systems
tracking error
controller design
unmeasured states
small neighborhood
linear matrix inequalities
control systems

Partial topic overlaps in the first layer

Topic overlap of #1 and #2

machine learning
classification accuracy
data mining
dempster-Shafer evidence theory
classification tasks

Topic overlap of #1 and #5

computational complexity
Artificial Intelligence
outsourced data
intrusion detection
cognitive radio networks

Topic overlap of #1 and #6

neural networks
Hidden Markov Model
memristor-based recurrent neural networks
delayed neural networks
inequality technique

Topic overlap of #2 and #4

R package
dimensionality reduction
MATLAB Toolbox
enhanced performance
CPU time

Topic overlap of #1, #2 and #6

convex optimization problem
Kronecker product
network states
bayesian inference

Topic overlap of #2 and #6

error system
mixed time delays
fuzzy sampled-data control
multiplicative noise
Markov chain

Figure 4.5 : Topic details and partial topic overlaps in computer science

4.4 Case Study II: Reveal the Topic Hierarchies in AI Ethics Research

A pandora's box of artificial intelligence (AI) has been opened and these disruptive technologies are transforming the daily lives of human beings in relation to new ways of thinking and behavioural patterns with enhanced capabilities and efficiency. There are many examples of AI applications in use today, such as smart homes (Harper, 2006), smart farming (Walter et al., 2017), precision medicine (Collins and Varmus, 2015) and healthcare surveillance systems (Hossain et al., 2020). The ethical and privacy issues surrounding the use of AI have been a topic of growing interest among diverse communities. For example, the general public has expressed concern about the impact of the increased use of robots on unemployment and inequality (Bossmann, 2016), social scientists have raised deep privacy concerns related to surveillance systems (Müller, 2020), and limited regulation of social media has raised debate with technical giants on the abuse of private data. Despite these concerns, the AI community stands behind the efficiency and robustness of their AI models. There is an urgent need to guide the research community to understand these ethical and privacy challenges.

To address these concerns, this case study reports on bibliometric research to comprehensively profile the key ethical and privacy issues discussed in the research articles and to trace how such issues have changed over the past few decades. We integrated a set of intelligent bibliometric approaches within a framework for diverse analyses. With specific foci in topic analysis, we initially retrieved terms from the combined titles and abstracts of collected articles and used a term clumping process (Zhang et al., 2014b) to remove noisy terms and consolidate technical synonyms. We answered the questions about the topical landscape using the approach of HTT-I. We anticipate that the empirical insights identified in this study will motivate the

AI community to extensively and comprehensively discuss the ethical and privacy issues surrounding AI and will guide the implementation of AI in line with an ethical framework.

4.4.1 Data collection and pre-processing

The Web of Science (WoS), owned by Clarivate, is a well-recognised integrative platform of bibliometric data sources. Of these, the WoS All Databases covers all the WoS's subscribed resources which we used as our primary data source when considering AI ethics as an emerging topic covering both natural sciences and social sciences. Its major debates exist not only in journal articles but also in a wide range of resources (e.g., conference proceedings and other types of research publications). Our special interest is in the ethical issues surrounding AI at both the macro and micro levels. Thus, topic analyses would focus on the WoS All Databases. In addition, since the WoS Core Collection database provides a curated form of full bibliographical information (e.g., author affiliations, countries/regions, and forward and backward citations), we particularly focused on an analysis of the key entities that contribute to the research on AI ethics and the interactions between these entities. Comparably, the WoS All Database covers a relatively ”full” collection of various types of articles in WoS, with a priority on data coverage, but the WoS Core Collection only contains journal articles collected in selective indexes (e.g., Science Citation Index), highlighting the quality of its data collection. In other words, the WoS Core Collection is a subset of the WoS All Database, with a filtered data collection. The search process ended up with 4,375 articles. The search strategy is given below:

TS = ((”artificial intelligence” OR ”big data”) AND (”disinform*” OR ”ethic*”
OR ”crimin*” OR ”moneti*” OR ”data control*” OR ”implicit trust*” OR
”addiction*” OR ”contestab*” OR ”moral*” OR ”digit* transparen*” OR

"algorithm* transparen*" OR "accountabilit*" OR "liabilit*" OR "fairness*"))

Data source: WoS All Databases

4.4.2 HTT-I result and interpretation

In this section, we applied an HTT-I topic analysis to the collected dataset from the WoS All Databases. We initially retrieved 93,364 terms from the combined titles and abstracts of the 4,375 articles and conducted a term clumping process (Zhang et al., 2014b) to remove noise and consolidate the technical synonyms, reducing the total number of terms to 52,054. Then, we used the 2,163 terms appearing in more than two articles as the core set of terms to generate the HTT-I result shown in Figure 4.6.

Figure 4.6 enhances the understanding of the details of AI ethical issues, especially the connections between specific AI techniques and moral concerns. Among its 71 nodes, the HTT-I result lists 27 AI techniques (e.g., *machine learning*) and AI-driven applications, devices, and products (e.g., *robots* and *autonomous vehicles*), 28 ethical topics (e.g., *fairness* and *discrimination*), and 16 societal topics (most of them about medical and healthcare issues). The four main branches of this HTT-I result represent four significant issues relating to AI ethics, that is, #1 *AI techniques and potential ethical issues*, #2 *technological and political implications of AI ethics*, #3 *data privacy*, and #4 *privacy in healthcare*. We discuss these four issues in detail:

- #1 *AI techniques and potential ethical issues*: Figure 4.6 reveals the key AI techniques that may raise ethical concerns, such as *machine learning* (including *deep learning*, *computer vision*, *neural networks*, *natural language processing*, etc.), *ontologies*, *communication technologies*, and *neuroscience*. *Machine learning*, one of the key areas in AI, shares close connections with almost all AI techniques and thus attracts the most attention in this HTT-I and is con-

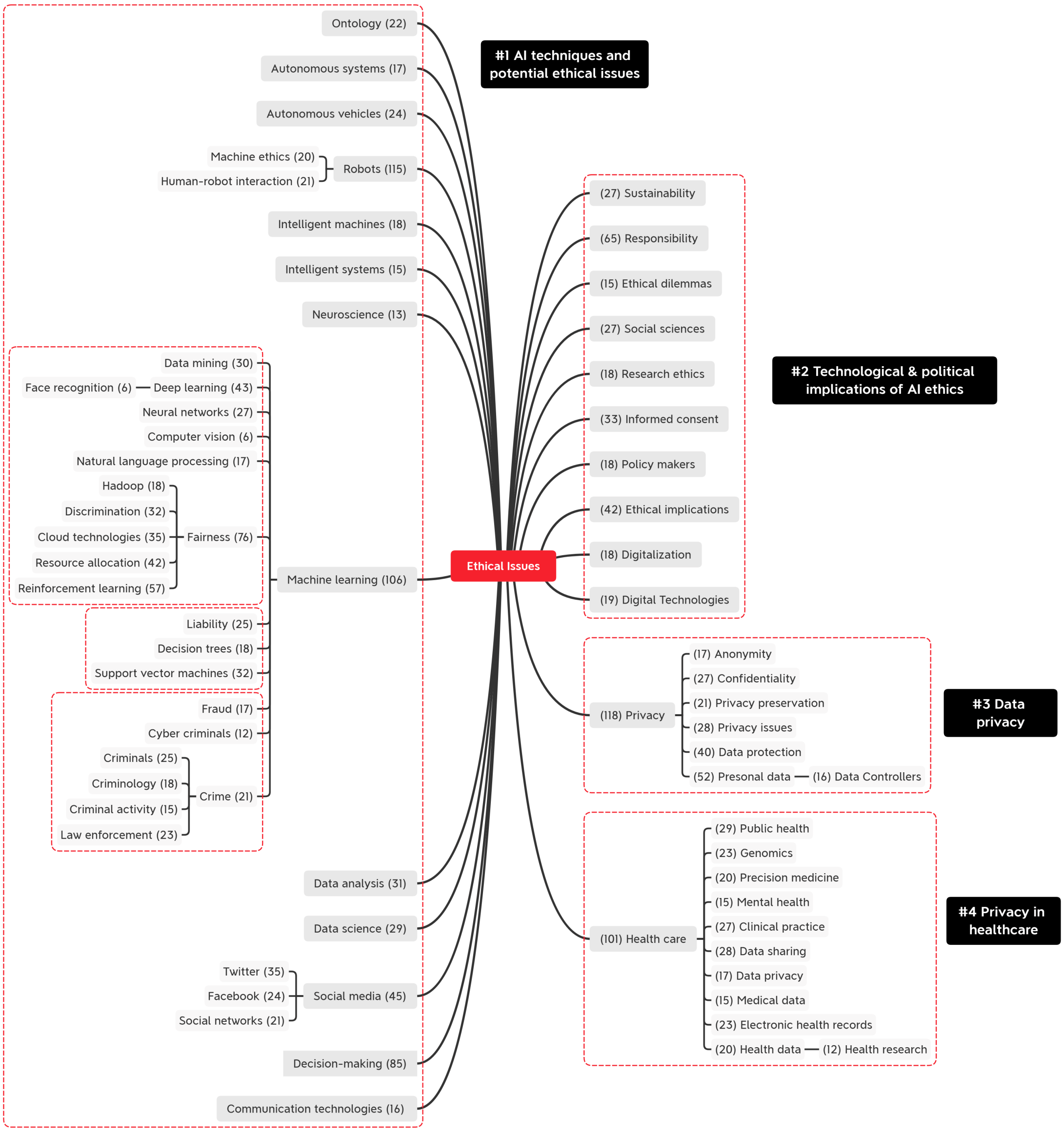


Figure 4.6 : HTT-I result of AI ethics research

nected with all ethical issues, such as *fairness*, *discrimination*, *liability*, *fraud*, and *criminals*. It is easy to explain these cases. For example, applying AI models to make decisions entails a justiciable ”right to a well-calibrated machine decision” (Kalluri, 2021; Huq, 2020), AI-driven fraud in social media, political elections, and financial markets (e.g., fake videos and identifications manipulated by AI techniques, such as image processing and face recognition) has become a major concern (King et al., 2020). How to validate AI recommendations with human knowledge in actual cases, such as clinical practice, is challenging both the AI community and the receptivity of the general public (Price et al., 2019). A brand-new topic of brain-computer interfaces is attracting increasing attention from the public, and ethical issues (such as privacy) and related regulations are appearing in public reading materials.

- #2 *Technological and political implications of AI ethics*: As an extension of the ethical issues in #1, #2 further extends AI's influence from ethics to the broad society through specific technological and political implications, such as *sustainability*, *responsibility*, and *digitalisation*. From the perspective of a complex ecosystem, these societal reactions could be the resilient progress of an ecosystem responding to disruptions introduced by AI techniques and their resulting ethical issues (Zhang et al., 2021a).

- #3 *Data privacy* and #4 *Privacy in healthcare*: #3 and #4 are a specific case of AI ethics. The big data boom initially activated the public's concerns about *data privacy*, where the illegal exposure of *personal data*, particularly those linked with social media, occurred. Furthermore, while analysing health data (e.g., electronic health records), including clinical trials and gene sequencing data, provides evidence for *precision medicine*, privacy concerns in medical and healthcare sectors then become not only a societal issue but also a threat to national strategies and the sustainability and balance of nature (Webber et al., 2015).

4.4.3 Case summary

In the HTT-I result, key AI techniques such as machine learning, data analysis, robots and intelligent systems, and cloud technologies generate concerns about the ethical issues relating to AI. Fairness and discrimination are critical concerns because AI models are applied in decision support in diverse scenarios. Data privacy, particularly in the healthcare and medical sectors, is a cause of increasing problems. Cybercrime and fraudulent behaviour are particularly concerning in the absence of appropriate support from the law and regulations. Machine ethics are mainly related to robots, autonomous cars, and intelligent machines, highlighting a balance between machine consciousness and human rights.

4.5 Case Study III: Digital Transformation Conceptualisation

Digital transformation (DT) has become an emerging phenomenon in strategic information research and industrial business practice. At a macro level, society is experiencing profound changes due to the explosion in digital technology across various industries. At a micro level, organisations build their digital capabilities and take advantage of new digital technologies to realise innovations and create business value. Yet, despite the vast interest in DT, there are still some significant research gaps in this domain.

First, most conceptual definitions of DT are based on qualitative analyses, such as expert judgments or a literature review (Reis et al., 2018; Vial, 2019). Data-driven quantitative analysis has rarely been used to characterise DT or the capabilities that enable it. This was one of Vial (2019)'s most urgent and vital calls to researchers.

Second, many studies point out the significant role of dynamic and technological capabilities in an organisation's development (Bharadwaj, 2000; Chae et al., 2014; Eisenhardt and Martin, 2000; Helfat et al., 2009; Teece, 2007; Warner and Wäger, 2019). However, almost no studies bring them together to understand their role in enabling DT. Lastly, when it comes to technological capabilities, most studies on DT are general. They talk non-specifically about the realm of digital technologies and not the technologies themselves, which makes it difficult to generate specific theoretical and practical implications from the results.

Hence, this case study aims to fill the three critical gaps: the lack of quantitative analysis on a definition of DT, the missing link between DT and the capabilities that enable it, and the lack of attention to understanding the capabilities needed to leverage specific digital technologies. To fill the third gap, this case study presents a case study focusing on one emerging technology, artificial intelligence (AI), to exemplify the specific capabilities needed to leverage AI successfully in a DT journey. More specifically, the overall goal of this case study could be concluded as answering the three following case research questions:

- Case research question 1 (CRQ1): What is the definition of DT from a bibliometric perspective?
- Case research question 2 (CRQ2): What capabilities enable DT?
- Case research question 3 (CRQ3): What are the AI capabilities enabling DT?

This case study is to leverage bibliometrics to seek quantitative evidence of precisely what DT is and the capabilities needed to enable it successfully. The analysis framework devised for this study integrates a topic-tracking method called scientific evolutionary pathways (SEP) (Zhang et al., 2017c) with a novel method of identifying topic hierarchies named hierarchical topic tree (HTT-I). By incorporating these two methods with network analytics and a literature review, each of our three research questions can be answered empirically rather than subjectively.

More specifically, we collected 10,179 scientific articles from the Web of Science and 9,454 patents relating to AI from the Derwent Patent Citation Index. Through SEP analysis on the collected dataset of scientific papers, we identified the evolutionary patterns of research topics in the DT literature, enabling us to distil a general definition of DT (CRQ1). To unravel the specific capabilities enabling DT (CRQ2), we applied HTT-I coupled with a literature review of specific papers on digital capabilities to arrive at a comprehensive categorisation of the resources and competencies needed to successfully undertake a DT journey, with the specific capabilities from 39 core papers classified as dynamic capabilities, technological capabilities, platform capabilities, and other capabilities. Lastly, taking AI as our focus technology, we applied an HTT-I analysis to a corpus of patents on AI and conceptualised a four-level model to guide AI-enabled DT (CRQ3). From bottom to top, the model progressively presents data collection and transmission capabilities, bridging capabilities, algorithm capabilities, and application capabilities required for companies to leverage AI in their digital transformation process.

4.5.1 Data collection and pre-processing

We chose two data sources as our corpus: Academic papers from the Web of Science (WoS) core collection and patents from the Derwent Patent Citation Index (DPCI). The WoS is a well-curated multidisciplinary database with 74.8 million scientific publications from over 21,100 journals, while the DPCI contains 39 million patent citations covering all technologies. The following search strategy returned 10,179 articles related to DT from WoS:

TS = ("digit* transfor*" OR "digitisation*" OR "digitisation*" OR
"digitalisation*" OR "digitalisation*" OR "digit* capabilit*" OR "digit*

platform*" OR "digit* tech*" OR "digit* innova*" OR "digit* competence*" OR "digit* mind*" OR "digit* activit*" OR "digit* practice*" OR "digit* manag*")
AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article)
Timespan=2010-2020
Search date: 24 September 2020

These papers constitute Dataset 1. From these data, we additionally prepared a second, more focused dataset of 913 papers published in journals our experts deemed to be of "high quality". The list of the journal titles is provided in Appendix A. Dataset 3 was assembled to support the case study. It comprises 9,454 patents relating to AI drawn from the DPCI. Given the general concept of AI covers such a broad range of areas, we narrowed our search to only patent titles that contained "artificial intelligen*") with the following search strategy:

TI = ("artificial intelligen*") AND IP=(G06* OR H04* OR H01* OR G11* OR G10* OR G01* OR G02* OR H05* OR H02* OR H03* OR G09* OR G05* OR A63* OR G08* OR G03* OR B60* OR G07* OR F24* OR A61* OR B65* OR B23* OR B81* OR B25* OR C08* OR A45* OR B01* OR C25* OR C09* OR B64* OR C23* OR F16* OR A44* OR C12* OR B32* OR C03* OR B62* OR F04* OR B29* OR B41* OR B24* OR F25* OR F28* OR E04* OR F21* OR G12* OR G04* OR G16* OR C01* OR B66* OR C07* OR B22* OR A47* OR A01* OR B82* OR B05* OR C22*)
Search date: 06 November 2020

In summary, the three datasets are:

- Dataset 1: 10,179 scientific papers related to DT retrieved from the WoS.
- Dataset 2: A subset of Dataset 1 comprising 913 articles published in high-quality journals.

Table 4.2 : Step-wise results of the term clumping process

Step	Description	#Terms	
1	Raw terms retrieved with NLP	253,162	24,203
2	Consolidated terms with the same stem, e.g., “information system” and “information systems”	220,812	20,530
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	199,410	18,398
4	Removed general single-word terms, e.g., “information” *	174,880	15,281
5	Filtered technological terms suggested by experts	-	-
6	Consolidated synonyms based on expert knowledge, e.g., “co-word analysis” and “word co-occurrence analysis”	164,433	14,918
7	Eliminated all but the top 5000 most frequently occurring terms	5,000	5,000
8	Eliminated all terms occurring only once	-	-

- Dataset 3: 9,454 AI patents from the DPCI.

Before using the data for HTT-I analyses, we applied VantagePoint’s natural language processing (NLP) function to convert the datasets into a dictionary of raw words and phrases. We then executed a term clumping process (Zhang et al., 2014b) that removes noise and consolidates synonyms to arrive at a final list of topic terms. From this list, we selected the 5000 terms with the highest frequency from datasets 1 and 2 and those with a frequency greater than one from Dataset 3 for further analysis. The step-wise results are given in Table 4.2.

4.5.2 SEP results and interpretation

The SEP analysis helped us to define DT from a bibliometric perspective in answer to CRQ1. By running the SEP algorithm on processed Dataset 1, we generated the DT research SEP in Figure 4.7. It traces the changing focus of academic research over the last decade. Each node represents a topic, and each edge indicates predecessor-descendent relationships between two topics. The colours indicate the

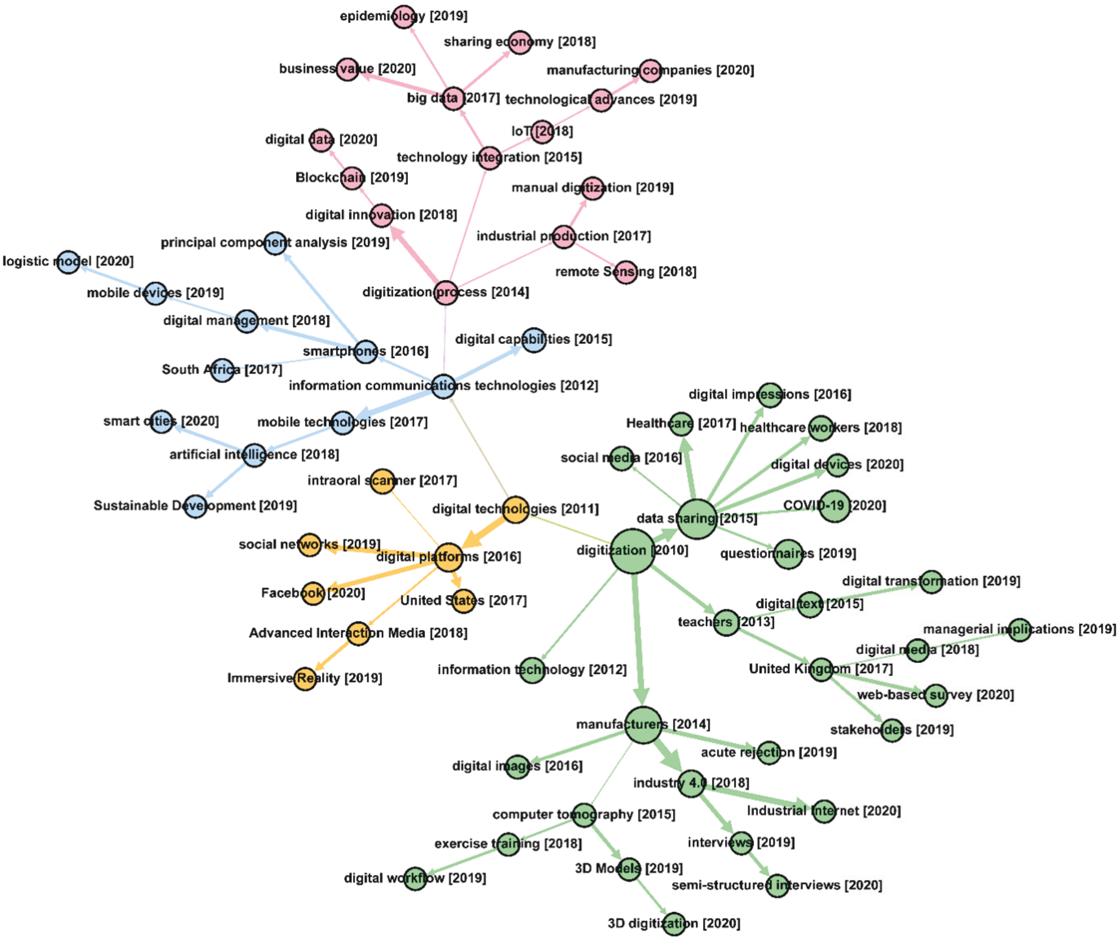


Figure 4.7 : The SEP within DT research

four topic communities detected by Gephi[4].

The green community (#I) encompasses the fundamental concepts of DT research at a macro level. The orange community (#II) indicates the initial development of digital technologies researched in DT and highlights the heavy involvement of interactive technologies. The blue community (#III) marks communications technologies (CTs) as a prominent technology related to DT and uncovers other relevant emerging technologies based on CTs, like mobile CTs and AI. Lastly, the pink community (#IV) denotes digitisation processes and the transition from theory into

[4] HTT-Ips://gephi.org/

practice.

#I *digitisation (green)*: This community can be considered a birds-eye view of the spectrum of research into DT. Many of the significant evolution emanate from the industries that need or benefit from DT, such as healthcare, education, and manufacturing (relevant topics: healthcare [2017], COVID-19 [2020], teachers [2013], and manufacturers [2014]). The milestones along the main pathways include changes to the fabric of industry itself (Industry 4.0 [2018], industrial internet [2020]), digitising objects (digital text [2015], digital media [2018], digital images [2016], 3D digitisation [2020]); and research methods (questionnaires [2019], interviews [2019], web-based surveys [2020], semi-structured interviews [2020]). Of the four communities, this one has the most comprehensive scope.

#II *Digital technologies (orange)*: Derived from Community #I, this community reveals the first offshoots of DT – the technologies developed. The topics include digital platforms, social networks, advanced interaction media, and immersive reality (Hein et al., 2019; Butler et al., 2020; Zhukov et al., 2018), primarily interactive and user-engaged. This community also acts as a bridge to Community #3, where information communication technologies (ICTs) emerge.

#III *ICTs (blue)*: Community #III encompasses smartphones [2016], mobile technologies [2017] and mobile devices [2018], highlighting the developing trends in mobile communications (Neirotti and Pesce, 2019). It also includes the emerging topics of data analytics, artificial intelligence, and sustainability (artificial intelligence [2018], principal component analysis [2019], logistics models [2020], sustainable development [2019], and smart cities [2020]) (Yang et al., 2020b; Brock and Von Wangenheim, 2019; Tumelero et al., 2019).

#IV *digitisation processes (pink)*: While Community #I shows the theoretical beginnings of DT, Community #IV shows the practical outcomes. Many of the topics in this community are either key enablers to DT (big data [2017], IoT [2018], remote sensing [2018], Blockchain [2019], etc.) or digitisation solutions (technology integration [2015], the sharing economy [2018], manual digitisation [2019]), which guides companies to realise a successful DT (Nicolescu et al., 2018; Bayer et al., 2020).

Overall, this SEP analysis reveals how the research into DT has evolved. The four different topic communities reveal the "entity" (#I), the "technologies" (#II and #III) and the "significant change" that has occurred as a result (#IV), providing quantitative evidence of how the theoretical foundations of DT can become a reality. These four communities accord with Vial (2019)'s definition that "DT is a process that aims to improve an entity by triggering significant changes to its properties through combinations of information, computing, communication, and connectivity technologies".

4.5.3 AI research papers - HTT-I result and interpretation

The HTT-I result reveals topic relationships from a cross-sectional and vertical perspective. In Figure 4.8, the *digital technologies* topic is the most dominant node. Therefore, both figures indicate that the development of digital technology is a critical enabler of DT. *Digital platforms* and *information technology* are the other two topic nodes in the tree without parent nodes, which means these two topics are equivalently important with digital technologies to DT. The remaining nodes are divided into ten topic clusters. We named each cluster after the node with the most substantial connection to its parent.

On the right side of the tree, there are seven topic clusters subordinate to the digital technologies root node. These seven clusters represent either specific digital technologies or the business implications of those technologies. The top four are *ICT (#1)*, *social networks (#2)*, *AI (#3)*, and *IoT (#4)* – all digital technologies. From a

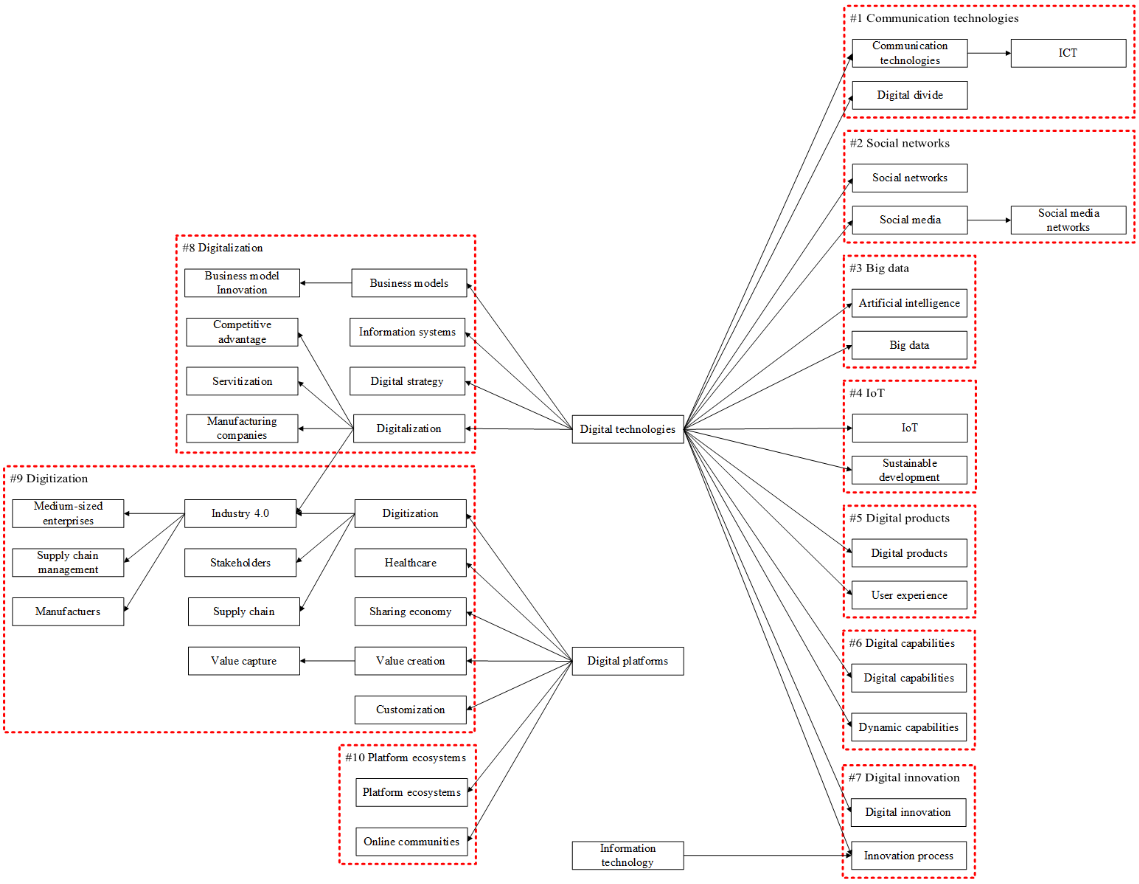


Figure 4.8 : The HTT-I result discovered in DT research

complete reading of some of the papers in these clusters, we find that disruptiveness is a characteristic shared by all topics (Zhukov et al., 2018; Bayer et al., 2020; Young et al., 2019). By disruption, we mean they can shake up industries, trigger the development of new business models, and segment markets in new ways (Danneels, 2004). The papers in these clusters articulate how the technologies they discuss can enable DT (Neirotti and Pesce, 2019; Brock and Von Wangenheim, 2019; Nicolescu et al., 2018; Bayer et al., 2020; Ardolino et al., 2018; Hartley and Sawaya, 2019; Wang et al., 2015b), and they present empirical evidence to prove it (Butler et al., 2020; Yang et al., 2020b; Chae, 2019). Thus, the social implication of the technologies in this cluster is also a research focus. For instance, the explosion and imbalance of ICT development are claimed to be one of the causes of the digital divide – a prominent social problem in implementing digital technologies (Afshar Ali et al., 2020; Srivastava and Shainesh, 2015), while IoT is recognised to have the potential to promote sustainable development by industry (Tumelero et al., 2019; Yang et al., 2020b).

The remaining three of the seven are *digital products (#5)*, *digital capabilities (#6)*, and *digital innovation (#7)*. These reflect the business implications of digital technologies.

The articles on *digital products #5* typically aim to promote the development or improvement of digital products based on digital technologies (Henfridsson et al., 2014; Øiestad and Bugge, 2014). Here, user experience is usually the key evaluation indicator (Shin, 2019). The focus in *digital capabilities #6* is on the capabilities required by organisations and individuals in the DT process, which are divided into digital (Gurbaxani and Dunkle, 2019; Pagoropoulos et al., 2017), dynamic (Demeter et al., 2020; Jantunen et al., 2018; Karimi and Walter, 2015; Freitas et al., 2020), and their combination (Fernandes et al., 2017; Antonucci et al., 2020). A few studies have attempted to shed light on the connotation of some capabilities. However, there is still no guidance in the form of widely-accepted and broad categories of the specific capabilities required for a successful DT. To further explore this question, we conducted a literature review, the results of which are summarised in the following subsection. The articles grouped under digital innovation #7 discuss how to realise digital innovation using digital technologies (Nylén and Holmström, 2015; Pershina et al., 2019; Trabucchi and Buganza, 2019). Here, information technology is also a parent of innovation processes. This indicates that information technology is a universal research topic but has a specific power to drive digital and business innovation (Candi and Beltagui, 2019; Trantopoulos et al., 2017).

The last subordinate cluster of digital technologies is digitalisation (#8) at the left top. The articles in this cluster discuss several prevalent issues for company management and industry governance in the DT process, such as alternative and innovative business models (Li, 2020; Loebbecke and Picot, 2015), digital strategy scheduling (Bharadwaj et al., 2013; Correani et al., 2020), the development and enhancement of business servitisation (Frishammar et al., 2019; Kohtamäki et al., 2020), and how to maintain one's competitive advantages through digitalisation (Black and van Esch, 2020; Ferreira et al., 2019). Another notable highlight in this cluster is the frequent mention of manufacturing as a representative industry experiencing digitalisation (Björkdahl, 2020; Pessot et al., 2020).

The next parent, *digital platforms*, is linked to two subordinate clusters: *digitisation (#9)* and *platform ecosystems (#10)*. Papers in the *digitisation #9* cluster claim the significance of digital platforms in realising business value (Hein et al., 2019; Alaimo and Kallinikos, 2017; Helfat and Raubitschek, 2018) and accelerating the digitisation process (Karimi and Walter, 2015). Further, this cluster contains several branches highlighting the highly-relevant sectors of digitisation in practice, including supply chains (Garay-Rondero et al., 2019; Ghadge et al., 2020; Ivanov et al., 2019), manufacturing (Culot et al., 2020; Ghobakhloo and Fathi, 2019;

Horváth and Szabó, 2019; Zheng et al., 2019), and healthcare (Agarwal et al., 2010; Holeman and Kane, 2020). Papers discussing *platform ecosystems #10* explore the establishment, development, and implications of platform-based digital ecosystems (Ghazawneh and Henfridsson, 2015; Wang and Miller, 2020; Yablonsky, 2020).

Intriguingly, our algorithm placed the topic *Industry 4.0* at the convergence of two parent nodes, digitisation and digitalisation, which means that, at the macro-level, realising industry 4.0 requires both digital technologies and transformative business practice (Bienhaus and Haddud, 2018; Frank et al., 2019).

In summary, this topic tree quantitatively reflects the composition of research topics in the field of DT, highlighting that technologies and platforms are two essential enablers of the transformation process. The digital technologies most frequently studied to promote DT are ICT, social networks, AI, and IoT. Establishing digital platforms can empower DT by creating digital ecosystems and providing systematic business digitisation approaches (Wang et al., 2019b; Hein et al., 2019). Among all the clusters, digital capabilities #6 specifically gathers papers discussing capabilities that enable DT, which corresponds to our CRQ2.

4.5.4 Categorising the capabilities that enable DT

The findings outlined in SEP and HTT-I result highlight that no overall category has emerged from the exacted terms but rather only categories that encompass technologies and digitisation. From Figure 4.7, we see digitisation plays a substantial role. This finding confirms the emerging newness of DT and the lack of a coherent theory behind it. It also shows the importance of digitisation as a broad category of observing transformations taking place through digital technologies. But, more importantly, our findings highlight that DT is highly connected with digital technologies and platforms, as Figure 4.8 shows. Hence, we must revise Vial's definition of DT (Vial, 2019), proposing that DT be defined as a process that aims to improve an entity by triggering significant changes to its properties through combinations of digital capabilities, technologies, and platforms.

Given that DT is a change process, it is critical to understand what capabilities can help its development and management. The digitisation cluster in Figure 4.7 and the *digital capabilities #6* cluster in Figure 4.8 point us to the academic work on this subject. *Digital capabilities #6* represents 59 articles, which we read to determine whether they offer any clear and specific insights into the capabilities required for DT at a company level. Here, clear and specific means definitions are given, and constructs have been developed and used to measure them. Of the 59 papers, 31 satisfied these criteria and were included in the review. The other papers typically focused on society-level issues, such as measuring the Industry 4.0 readiness of manufacturing in the EU (Castelo-Branco et al., 2019), or on individual-level topics, such as how to assess digital skills in citizens (Hidalgo et al., 2020). These papers were discarded.

Both authors read and classified the 31 articles based on the capabilities discussed, as listed in Table 4.3. The capabilities fall into two broad categories: dynamic and technological. Dynamic capability is the more extensive and diverse of the two but mainly includes traditional and non-traditional variations of three key concepts: sensing, seizing, and transforming (Teece et al., 1997). One article, driven by a study of 208 innovations in the insurance industry, uses the term "transformative capabilities", referring to sector-specific capabilities, such as developing services that fulfil customer needs, exploiting data for risk assessment, and underwriting (Stoeckli et al., 2018). Conversely, any technological capabilities have, by and large, been neglected.

The closest call to a technological capability is the digital capabilities defined in just a few studies, as shown in Table 4.3. All of these refer to IT-related ca-

Table 4.3 : The 31 articles of capabilities enabling DT

Capability category	Capabilities	Source
Dynamic capabilities	Sensing, seizing, transforming	(Demeter et al., 2020; Jantunen et al., 2018; North et al., 2019; Stoeckli et al., 2018; Day and Schoemaker, 2016)
	Digital sensing, digital seizing, digital transforming	(Warner and Wäger, 2019)
	Absorptive capacity	(Demeter et al., 2020)
	Integrative capabilities	(Demeter et al., 2020; Helfat and Raubitschek, 2018; Lin et al., 2016)
	Relational capabilities	(Demeter et al., 2020; Lin et al., 2016; Sun et al., 2020a)
	Innovative capability	(Ferreira et al., 2019; Helfat and Raubitschek, 2018)
	Dynamic managerial capabilities	(Annosi et al., 2019; Li et al., 2018)
Technological capabilities	Digital capabilities	(Fernandes et al., 2017; Ardolino et al., 2018; Gurbaxani and Dunkle, 2019; Pagoropoulos et al., 2017; Levallet and Chan, 2018)
	Dynamic IT capabilities	(Li and Chan, 2019)
	Big data capabilities	(Dremel et al., 2017)
	Information analytics	(Park and Mithas, 2020)
	Relational and information processing capability	(Saldanha et al., 2017)
Platform capabilities	Platform capability	(Li and Chan, 2019; Karimi and Walter, 2015; Sun et al., 2020a)
	Platform utilisation capabilities	(Annosi et al., 2019; Li et al., 2018)
Others	Business process management capabilities	(Antonucci et al., 2020; Ukko et al., 2019)
	Project capabilities	(Lobo and Whyte, 2017)
	Organisational learning capabilities	(Tortorella et al., 2020)
	Customer service capabilities	(Setia et al., 2013)
	R&D capabilities	(Szalavetz, 2019)
	Production capabilities	(Ukko et al., 2019; Szalavetz, 2019)
	Knowledge management	(Muninger et al., 2019)
	Top management understanding	(Gurbaxani and Dunkle, 2019; Muninger et al., 2019; El Sawy et al., 2016)
	Networking and collaboration competences	(Muninger et al., 2019)

pabilities with diverse and hard-to-generalise definitions. For example, one study refers to digital capabilities as the combination of a flexible IT infrastructure and a well-developed information management capability (Levallet and Chan, 2018). Another considers digital capabilities as a company's capacity to utilise its available IT resources (Fernandes et al., 2017).

The study by Li and Chan (2019) presents an in-depth conceptual model developed for IT departments. This unique study offers companies three sets of capabilities to manage their IT: 1) dynamic digital platforms covering IT infrastructure functionality, flexibility, and integration capability; 2) dynamic IT management consisting of IT deployment, exploration, and exploitation; and 3) dynamic IT knowledge management based on knowledge creation, transfer, and retention.

In another study, Gurbaxani and Dunkle (2019) offer a framework for DT consisting of six themes: strategic vision, the culture of innovation, know-how and IP assets, digital capabilities, strategic alignment, and tech assets. The majority of these items are managerial, but digital capabilities and tech assets speak to technical capabilities. For example, digital capability refers to the availability of expertise at both the strategic and technical levels and the level of skill at hand to define and execute digital strategies. Tech assets cover big data, data mining and analysis/data analytics, mobile technologies, cloud computing, and internet and wireless communications. These are deemed sufficient technology assets to implement a strategic vision. However, no details are supplied regarding the expertise needed to use any specific tech asset. The study asks survey respondents to rate their company's position compared with rival companies.

The remaining digital capabilities mentioned are related to IT, such as a company's big data assets (Dremel et al., 2017), or its ability to undertake information analytics (Park and Mithas, 2020) or relational/information processing (Saldanha et al., 2017). Many studies focus on general technologies, but they do not discuss any capabilities associated with specific technologies. For example, Muninger et al. (2019) investigated the capabilities needed to use social media to generate innovation, finding three non-technical capabilities companies should build upon knowledge management, top management understanding, and networking and collaboration.

There are also inconsistencies in the broad categories of capabilities, which further complicates a general understanding of what is required for a successful DT. For example, the authors of one study refer to "managerial capability" (Annosi et al., 2019) when what they really mean is the level of technical knowledge managers have. Further, the managerial capability is measured by qualitatively ranking the managers' responses to questions such as "Do you have employees dedicated to the management and/or research of new digital technology for your farm?" (the study concerns agriculture). However, again, the paper provides no details of any specific technologies.

This exclusion of technological capabilities from studies presents an intriguing opportunity for future studies to explore. Even though it is speculative, we think the gap may exist for two main reasons. First, it might be challenging to find common technological capabilities considering the wide range of different technological features for each digital technology. Second, the literature is vague on the definition of technological capabilities. Most articles seem to rely on infrastructure or technological investments to indicate a company's technological capabilities (Li and Chan, 2019). But this approach ignores the importance of the soft side of technologies, particularly know-how and intellectual property rights.

To fill this knowledge gap in the literature, this study draws on the understanding of technology management as a set of capabilities (Cetindamar et al., 2016). Further, CRQ3 asks: What are the AI capabilities enabling DT? Through this question, we focus on one digital technology, AI, and assess it through the most widely used indicator of technological proficiency: patents. We used network analytics to deconstruct the key technical knowledge that might be associated with a company in this arena. The following section exemplifies how technological capabilities can be derived for different technologies using HTT-I analysis on AI patents.

4.5.5 AI patents - HTT-I result and interpretation

By feeding the technical terms extracted from Dataset 3 into our HTT-I algorithm, we generated the HTT-I in Figure 4.9. We then used this tree result to identify the technological composition, divergence, and convergence of patented AI techniques.

The HTT-I result partitions the research directions into seven clusters. This HTT-I's primary root node is labelled *sensor* with three linked topic clusters: *sensor technologies* #A; *transmission technologies* #B; *AI applications* #C, which owns *robot*, *cloud technology*, and *Internet of Things (IoT)* as three typical AI applications. Derived from cluster #C, *robot* is further broken down into *robotics* (#D), and *robot functions* (#E), the topic *cloud technology* leads a whole cluster with the same name, i.e., *cloud technology* #F. *Neural networks* (#G) is a relatively independent cluster parent node that converges with the node machine learning.

The status of *sensor technology* as the root node demonstrates its foundational role in AI. Our patent review also confirmed that sensors were used as a primary data collection module in most granted patents.

Sensor technologies (#A) consist of various sensors designed to capture different input signals, such as pressure, humidity, images, ultrasonic waves, temperature, infrared light, etc. *Transmission technologies* (#B) comprises information and communication technologies used in data transmission modules, such as *Bluetooth*, *wireless fidelity (WIFI)*, and *wireless communication.*

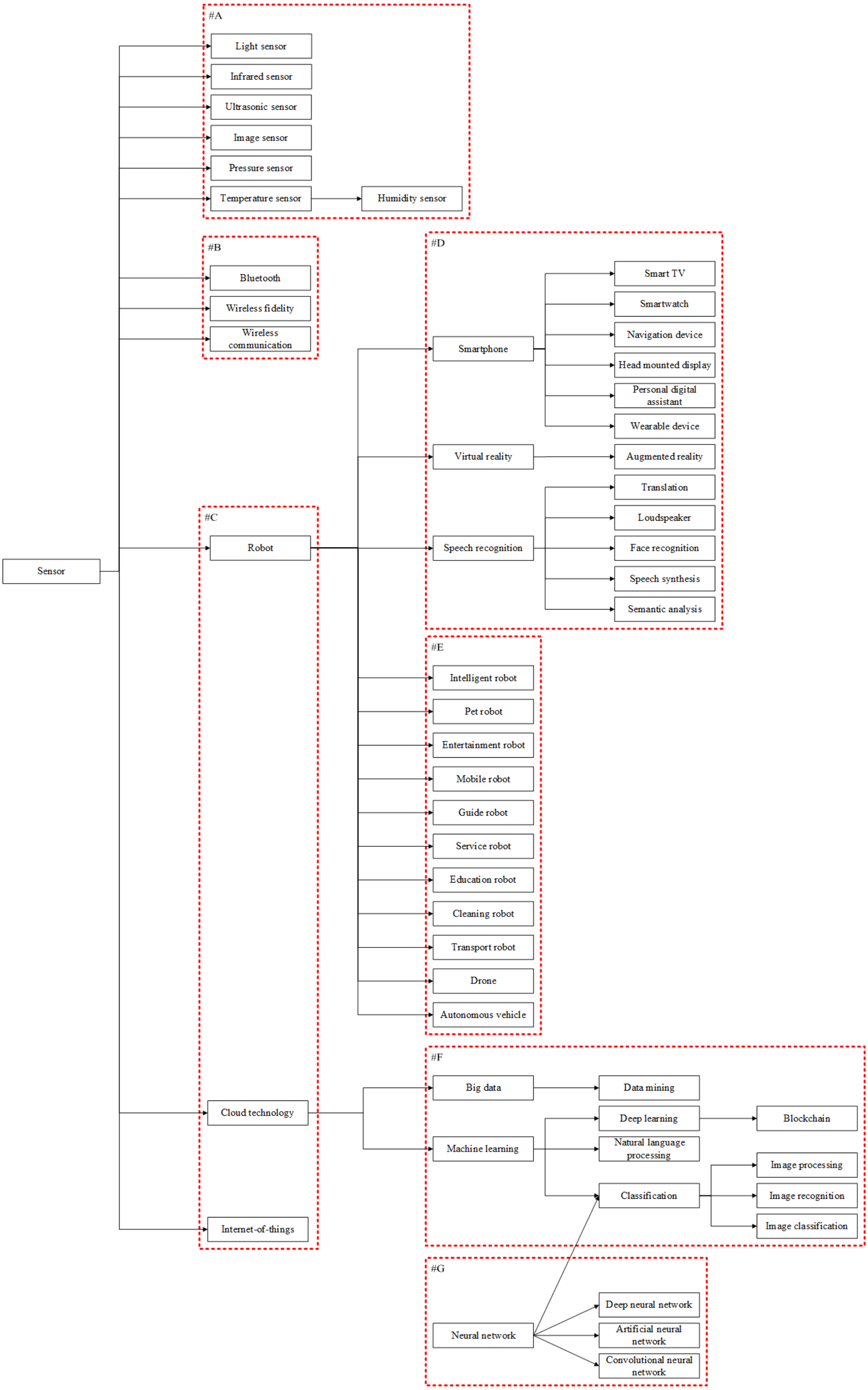


Figure 4.9 : The HTT-I result generated from the AI patents dataset

AI applications (#C) contains *robot*, *cloud technology* and *IoT* as three representative applications. *Robot* is the largest subordinate node and is further partitioned into *robotics* (#D) and *robot functions* (#E). *Robotics* #D is an interesting cluster housing many innovations that sit at the convergence of multiple other technologies, such as *speech recognition*, *face recognition*, *smartphone*, *wearable devices*, *virtual reality*, and *augmented reality*. *Robot functions* (#E) covers the robots built for various specific uses, including *drones*, *autonomous vehicles*, *pets*, *navigation*, *cleaning*, *service*, *education*, etc. *IoT* is another application in this cluster but it has no subordinate nodes. Our patent review reveals that current major patents in *IoT* tend to focus on intelligent hardware control. However, the terms describing pure hardware facilities (like rotating rods, supporting rods, etc.) were not included in the filtered technological terms. Our findings of the emergence of IoT and robotics in AI applications comply with the co-evolving patterns and convergences identified by Börner et al. (2020), whose citation analysis indicates that cross-citation between AI, IoT, and robotics has increased dramatically over the last decade.

Cloud technology (#F) is a unique AI application that bridges the connection between AI services and end-users. This cluster involves substantial specific AI algorithms and techniques. By referring to the relevant patents in this cluster, we found that AI algorithms and techniques always involve massive data processing and need cloud technology to provide a computing efficiency solution. From this perspective, *cloud technology* can be regarded as the prerequisite technology for the product realisation of AI. When diving into the subordinate nodes of *cloud technology*, we identify the following technological composition and changes:

- *Technological segmentation* (#F): *Machine learning* is segmented into *deep learning*, *natural language processing (NLP)*, and *classification*. *Classification* diverges into *image processing*, *image classification*, and *image recognition*, indicating that most classification tasks are related to image data.

- *Technological convergence* (#F and #G): *machine learning* and *neural networks* present a technological convergence to *classification.* This convergence indicates the incorporation of machine learning and neural network in improving the accuracy of classification tasks.

Based on Figure 4.8, we find AI capabilities can be classified into four levels:

i. Data collection and transmission: the capability to leverage technologies that collect data from the physical world or to transfer data within and between product modules. Sensors and ICTs are the representative technologies that realise such capabilities in AI inventions.

ii. Bridging: the ability to connect (disparate) end-users with AI products and services/products. Cloud technology is a crucial part of bridging capability since it plays an indispensable role in the deployment and large-scale implementation of advanced AI algorithms.

iii. Algorithms: the ability to use AI techniques and algorithms to perform specific business tasks. Typical examples include machine learning, deep learning, big data analysis, neural networks, etc.

iv. Applications: the capability to realise mature technological convergences between AI and other technologies to provide innovative products. IoT and robots, for example, are two mainstream applications of such a kind.

With this stratification, we can conceptualise the capabilities needed to leverage AI successfully within a DT process as the pyramid shown in Figure 4.9.

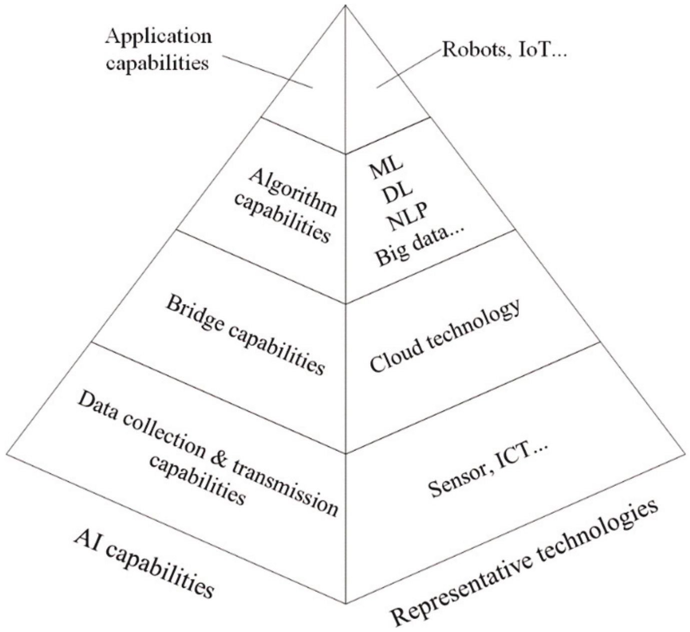


Figure 4.10 : The pyramid of AI capabilities

4.5.6 Case summary

DT is here to stay, and its revolutionary nature seems to speed up in parallel to making rapid changes in digital technologies (Schwab, 2017). Clarifying its definition, observing its evolution, and identifying its enablers could benefit the information systems and technology management disciplines. Without clear and empirically-validated definitions, the DT literature might remain in an adolescent development phase. In this study, we exploit methods of intelligent bibliometrics, including scientific evolutionary pathways, HTT-I analysis, and network analytics, to conduct a set of quantitative analyses as opposed to the qualitative analyses dominating the field. Through these approaches, we address three critical questions in DT:

- CRQ1: What is the definition of DT from a bibliometric perspective?

  The SEP analysis advances Vial's definition and solidifies those concepts with identified dynamic research communities. Additionally, the HTT-I analysis highlights several DT enablers, including digital technologies, digital platforms, and digital capabilities, based on which researchers can further extend DT's definition.

- CRQ2: What are the capabilities enabling DT?

  Our literature review, driven by the SEP and HTT-I analyses brings a broad range of digital capabilities together into a comprehensive set of categories, as given in Table 4.3. The studies reviewed highlight the key role of dynamic, technological, and platform capabilities in DT.

- CRQ3: What are the AI capabilities enabling DT?

  Having decomposed the hierarchical technologies from AI patents, we propose the pyramid model of four significant capabilities illustrated in Figure 4.10: data collection and transmission, bridging, algorithms, and applications.

This research benefits the DT field from both an academic and a practical perspective. From an academic standpoint, our study presents topic analyses that can help researchers understand the breadth and depth of DT research. It provides insights and clues to conduct a more in-depth analysis of certain research topics. From a practical perspective, the current literature is patchy and incomplete regarding an understanding of the capabilities needed for a successful DT. By bringing these diverse capabilities together, we make managers aware of some core resources and competencies that will likely prove helpful on their DT journeys.

Lastly, this case study sheds light on the definition of DT, its evolution, and the capabilities that enable it. It offers an approach to identifying technological capabilities for a specific digital technology, AI. Future studies might explore the generalisability of these techniques with empirical evidence of these capabilities with other technologies and more importantly in real-life applications and diverse contexts, such as different industries, countries, and organisation types. With enough such studies, it may be possible to compare capabilities in practice to determine the components and configurations that make DT successful more precisely.

4.6 Summary

In summary, this chapter presents an end-to-end framework called HTT-I 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-I, the core term to each topic in a co-occurrence network, to be used as its label, is determined by term density peak characteristics, while overlapping community allocation detects overlaps among different topics. The recursive implementation of these two algorithms generates a hierarchical topic tree. Case studies on the hierarchy of computer science, AI ethics and DT research topics demonstrate the proposed methodology's feasibility and reliability.

Chapter 5

Hierarchical Topic Tree - II Model

5.1 Introduction

In Chapter 4, we proposed HTT-I, the initial version of HTT that can extract topic hierarchies from term co-occurrence networks. However, the HTT-I model still has two parameters to be fine-tuned, which hinders its adaptivity and was listed as one of the future research improvements. To provide an adaptive and generalised way of constructing topic hierarchies in this chapter, we develop a network-based non-parametric hierarchical topic extraction model named Hierarchical Topic Tree - II (HTT-II) that can automatically construct the high-quality topic tree without the need for preset parameters. The model takes a co-term network as input and generates hierarchical term communities as topics on different layers. Inspired by the natural hierarchical structures of real-world networks and the influential node theory in social network analysis (Clauset et al., 2008; Kitsak et al., 2010; Sun et al., 2021), we employ k-shell decomposition to layer the nodes in a network as core or periphery according to their connectivity. Simultaneously, we apply the Louvain community detection method to partition nodes into communities. The densely-connected nodes in the core layer form different topics based on their given community labels, while the periphery nodes, along with their community information, will go into the next round for further partition. The two steps will recursively run until no community structure or core nodes can be found. The finalised output of the proposed method is a topic tree with every tree node (topic) formed by a group of term nodes. Originating from a virtual root topic, the tree diverges into multiple branches and ends with leaf topics.

A set of experiments and an empirical case study demonstrate the utility and effectiveness of our method. Comparisons to six baselines with multiple real-world co-term networks reveal that HTT-II analysis can construct a high-quality topic tree with high topic coherence, strong parent-child topic association and exclusive sibling topics. The case study, which focuses on 11,399 research papers in the field of information science, constructs an HTT-II result with five major branches and 144 topics. The HTT-II result uncovers five prominent research directions in the discipline: Data mining, bibliometrics, information seeking, information systems and ontology construction. These research streams are then broken down into more fine-grained research topics, as detailed in the case study.

The remainder of this chapter is organised as follows. Section 4.2 sets out the details of our proposed methodology. Section 3.3 follows, presenting the empirical study and demonstrating the effectiveness and practical value of HTT-II analysis. Section 4.6 wraps this chapter, concluding the study's limitations and future research directions.

5.2 HTT-II Methodology

5.2.1 k-shell decomposition

The k-core of a network is defined as the largest subnetwork in which each node has at least k edges. k-shell decomposition aims to assign each node a k-shell index that indicates the largest k-core the node exists in. The assigned k-shell indices partition nodes from high to low, reflecting the hierarchical structure of nodes from core to periphery in a network (Dorogovtsev et al., 2006). We give the pseudo-code of k-shell decomposition in Algorithm 2 and illustrate the k-shell decomposition process in Figure 5.1.

Algorithm 2: k-shell decomposition

```
Input  : $G = (V, E)$, where $V = \{1, 2, ..., n\}$ is the node set and
         $E = (i, j|i, j \in V)$ is the edge set.
Output: $KSI = \{ksi_1, ksi_2, ..., ksi_n\}$ // The $k$-shell index of every
    node
$l = 0, V^{'} = \emptyset$;
$G^{'} = G(V - V^{'})$;
while $V^{'} \neq V$ do
    for $i \in V - V^{'}$ do
        if $k_i^{'} == l$ then
            // $k_i^{'}$ is the degree of node in $G^{'}$
            $ksi_i = l, V^{'} \leftarrow$ add $i$ // Add $i$ to the $k$-shell index
                assigned nodes
        end
        $l+ = 1$
    end
end
```

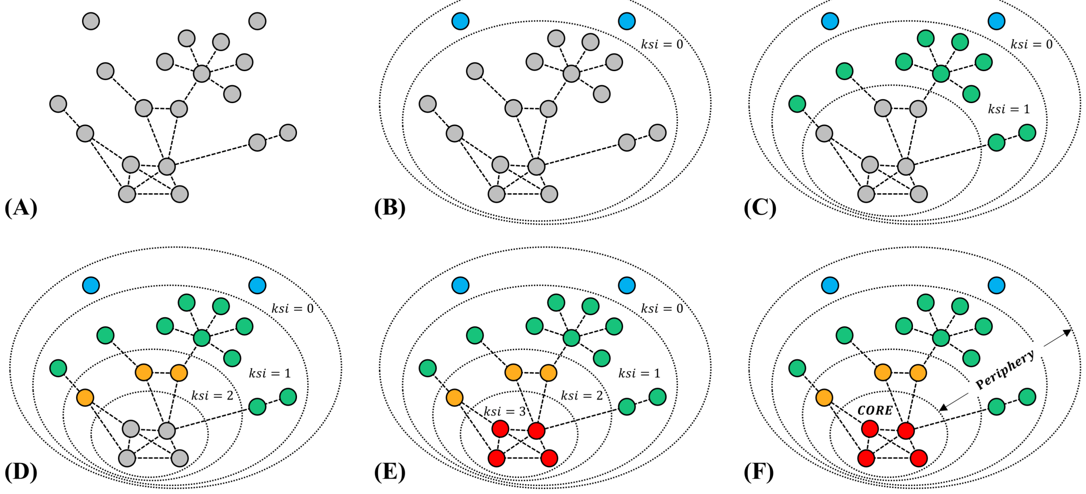


Figure 5.1 : Illustration of the k-shell decomposition method

Although the k-shell index is a robust and efficient measure for node influence and network hierarchy structures (Kitsak et al., 2010; Liu et al., 2015; Zhang et al., 2008), some issues remain when this measure is applied to reveal topic hierarchies hidden in co-term networks (Ba et al., 2019; Xiao et al., 2016). The first issue is that k-shell decomposition may yield many trivial and similar adjacent shells, especially when the network scale increases largely. In the previous practice, Carmi et al. (2007) reorganised the trivial shells by searching for transition shells according to the percolation theory. Xiao et al. (2016) compared the morphology of adjacent shells and manually reclassified similar shells as layers. However, such solutions may not work efficiently on large-scale networks, as hundreds or even thousands of shells can be generated, in which transition shells are much less noticeable. Besides, the manual decision can vary from subjective opinions and result in time cost aligning an appropriate decision. Another issue is that k-shell decomposition characterises the hierarchical network structure by a chain of node shells (node groups) from core to periphery nodes, which does not reflect the divisions of research topics on different granularity when it is applied to co-term networks; Substantial previous studies and practical experience have demonstrated the tree structure is more suitable for profiling hierarchical research landscapes and highlighting fine-grained topic segmentation (Qian et al., 2020; Shang et al., 2020; Yu et al., 2020; Zhu et al., 2019).

5.2.2 HTT-II conceptualisation

Aiming to address the two drawbacks of k-shell decomposition in revealing hierarchical knowledge structures, we devise an improved recursive algorithm named hierarchical topic tree - II (HTT-II) analysis. The proposed method addresses the above issues by incorporating the Louvain community detection method (Blondel et al., 2008) to break the innermost shells into multiple communities and organise the network hierarchy as a tree structure. As illustrated in Figure 5.2, our HTT-

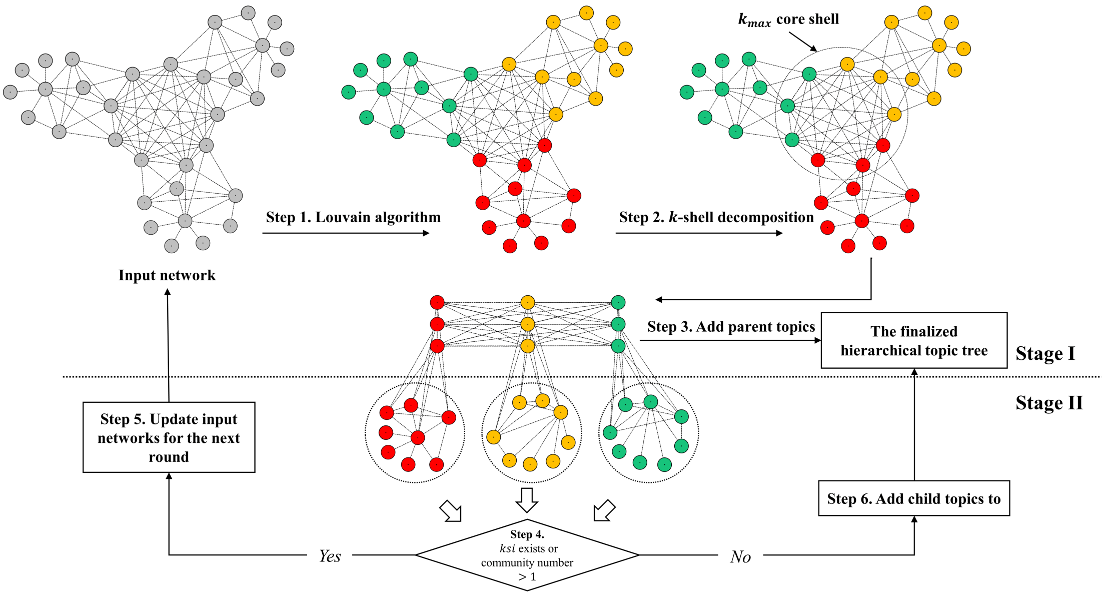

k_{max} core shell
Input network
Step 1. Louvain algorithm
Step 2. k-shell decomposition
Step 3. Add parent topics
The finalized hierarchical topic tree
Stage I
Stage II
Step 5. Update input networks for the next round
Step 6. Add child topics to
Step 4. ksi exists or community number > 1
Yes
No

Figure 5.2 : Research framework of the HTT-II model

II analysis consists of six steps: **Step 1**: Run the Louvain community detection algorithm to assign every term node a community label. **Step 2**: Apply k-shell decomposition to assign every term node a k-shell index (ksi) and identify term nodes in the k_{max} shell (k_{max} is the largest k-shell index). **Step 3**: Wrap every group of term nodes in both the k_{max} shell and the same community as a parent topic, then add the parent topics as tree nodes of the finalised hierarchical topic tree. **Step 4**: Judge the subnetworks formed by the rest of nodes in each community; If the k_{max} of the subnetwork exists or the subnetwork can be further partitioned to multiple communities, continue to **Step 5**, else go to **Step 6**. **Step 5**: Use the subnetworks as inputs to iterate steps 1-4. **Step 6**: Add the subnetworks as tree leaf nodes of the finalised hierarchical topic tree.

The design of HTT-II analysis ensures that: 1) Every topic derives from a k_{max} shell of term nodes that are strongly connected; hence there will be few trivial topics in the finalised tree result compared to the original k-shell decomposition approach. 2) The community detection process can partition nodes into multiple groups according to their topological features in each iteration, generating the branches to reflect research topic divisions on different levels in the finalised tree result. In most cases, our method design guarantees the generation of a tree structure with the automatically-decided number of nodes and layers representing the topic hierarchies of a co-term network.

5.2.3 HTT-II input and output

The input to HTT-II analysis is a co-term network, where the terms can be keywords (author-provided), pre-assigned topics (Shen et al., 2018), or terms extracted from a term extraction process (Zhang et al., 2014b). We retain raw co-occurrence semantic relationships because, although term embedding techniques have great merit, Shang et al. (2020) demonstrated that while term embedding clustering can roughly group terms from different domains, they do not perform well at distinguishing highly coupled terms in a specific field. A co-term network can be formally represented as $G = (V, E)$, where V represents the set of term nodes, and E represents the co-occurrence edges of the nodes in V.

The finalised HTT-II output is a tree structure, as illustrated in Figure 5.3. The tree originates from a virtual root topic that derives multiple branches and tree nodes (A, B, C, A1, A2, B1, B2, C1, and C2); Each tree node, composed by a term node group generated in steps 3 or 6, represents a topic. Each branch denotes a parent-child topic pair. The child topics derived from the same parent topic are siblings (A and B, B and C, A and C, A1 and A2, B1 and B2, C1 and C2).

5.2.4 HTT-II algorithmic details

From the algorithmic perspective, the HTT-II six steps can be concluded as a two-stage process. Stage I covers steps 1-3; It starts with the co-term network input and runs one-round community detection and k-shell decomposition algorithms to

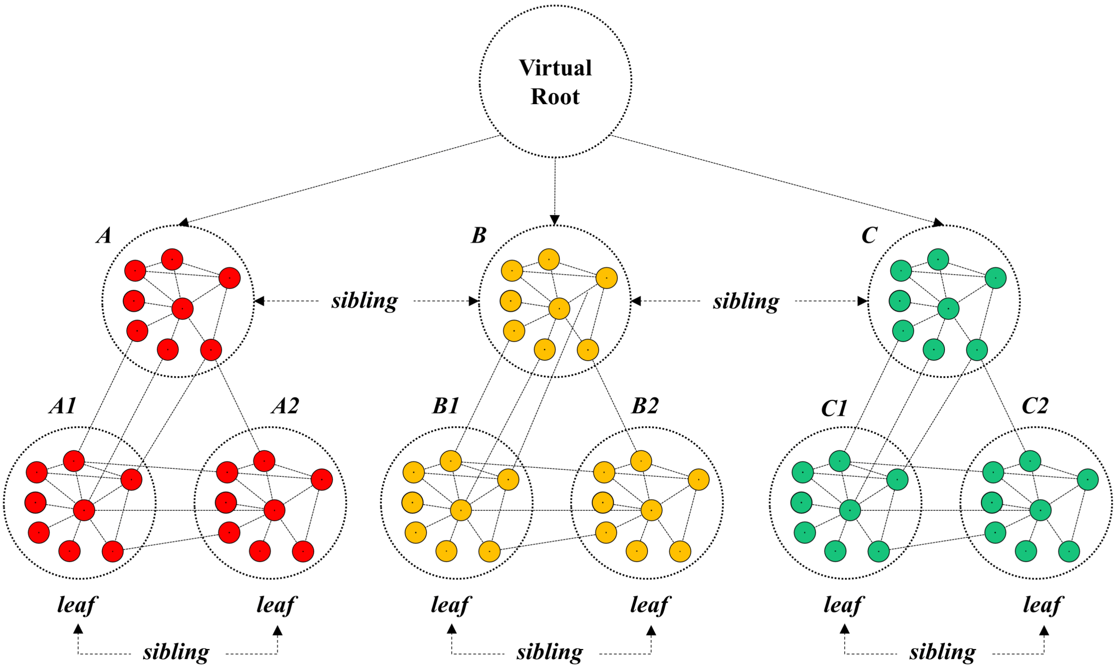


Figure 5.3 : Illustration of the HTT-II-generated tree structure

present a set of two-layer tree topics. Stage II runs steps in Stage I recursively to increase the tree depth incrementally; It will end until the stopping criterion is met.

Stage I: Network parent-child layer partition (Steps 1-3)

In this stage, we initially run the Louvain algorithm on the input network G and assign community labels for all nodes, representing the initial research direction segmentation characterised by different term node communities. Then, k-shell decomposition is applied to partition nodes into multiple shells. Together with the community information, the core nodes in the innermost shell will form different topics; The periphery nodes will retain their community information and go into Stage II. The pseudo-code of Stage I is given in Algorithm 3.

Algorithm 3: Network parent-child layer partition

```
Input  : $G = (V, E)$, where $V = \{1, 2, ..., n\}$ is the node set and
         $E = (i, j|i, j \in V)$ is the edge set.
Output: Two sets of node groups $T = \{\phi_1, \phi_2, ..., \phi_n\}$ that denotes the
         generated core topics and $\mathcal{L} = \{\ell_1, \ell_2, ..., \ell_n\}$ that denotes the
         periphery nodes of each core topic.
$T = \emptyset, \mathcal{L} = \emptyset$ // Initiate the layer and node set
$G^{'} = G(V - V^{'}) \leftarrow$ run community detection $\#C = \{1, ..., 2..., k, ..., n\}$
    // where $C_i$ denotes the community of $i$
$KSI = G(V, E) \leftarrow$ run algorithm 2 ;
for $i \in V$ do
    if $ksi_i == \max(KSI)$ then
        $\phi_{C_i} \leftarrow$ add $v$ // if the node is in the $max(KSI)$-core
            network
    else
        $\phi_{C_i} \leftarrow$ add $v$ // the periphery nodes in this community will
            go into the next layer
end
for $k \in C$ do
    if $|V_{\phi_k} > 2|$ and $|E_{\phi_k}| > 1$ then
        $T \leftarrow$ add $\phi_k$ ;
        $\mathcal{L} \leftarrow$ add $\ell_k$
end
```

Stage II: Recursive tree generation (Steps 4-6)

This stage constructs the topic tree by recursively running Step I on derived subnetworks and generating intensely connected topics and their parent-child relationships. The stopping criterion for each subnetwork is k_{max}=1, or the subnetwork cannot be partitioned into two or more communities. The stopping criterion was set up in Algorithm 2 to guarantee the algorithm can converge. The pseudo-code of Stage 2 is given in Algorithm 4.

Algorithm 4: Recursive tree generation

Input : $G = (V, E)$ and $\mathcal{R}$, $\mathcal{R}$ is the virtual root topic.

Output: The tree structure $\mathcal{H} = \{\phi_1, \phi_2, ..., \phi_n\}$;

The parent-child relationship mappings of topics

$\mathcal{M} = \{\phi_1 : \mathcal{R}, \phi_2 : \mathcal{R}, ..., \phi_n : \phi_k\}$, where $\phi_n : \phi_k$ denotes that ϕ_n is the child topic of ϕ_k.

```
$T, \mathcal{L} = G(V, E) \leftarrow$ run Algorithm 3 ;
$\mathcal{H} = \mathcal{H} \cup T, \mathcal{J} = \mathcal{J} \cup \mathcal{L}$ ;
for $\phi \in T$ do $\mathcal{M} \leftarrow$ add $\phi : \mathcal{R}$;
while $\mathcal{J}$ do
    $\mathcal{J} = \emptyset$ ;
    for $\ell \in \mathcal{J}$ do
        $G^{'} = G(\ell)$ ;
        $T^{'}, \mathcal{L}^{'} = G^{'} \leftarrow$ run Algorithm 3 ;
        $\mathcal{H} = \mathcal{H} \cap T^{'}, \mathcal{J} = \mathcal{J} \cap \mathcal{L}^{'}$ ;
        for $\phi \in T^{'}$ do
            $\mathcal{M} \leftarrow$ add $\phi : \phi_p$ // $\phi_p$ is the core topic generated in the last loop
        end
    end
end
```

5.3 Experiment

5.3.1 Experimental networks and baselines

To validate the effectiveness of our proposed method, we tested HTT-II and six baselines on three real-world co-term networks. The tested networks are generated from literature collections in the previous pilot studies, covering a broad range of research fields, including artificial intelligence (AI) ethics (Zhang et al., 2021d), bibliometrics (Mejia et al., 2021), and early COVID-19 research (Fry et al., 2020; Zhang et al., 2021a). We selected the three networks because they were from different disciplines, and the density of the three networks varies, representing different levels of knowledge interaction within the relevant research fields. The details of the test networks are given in Table 5.1.

Following the test networks, we selected two non-parametric community detection algorithms that can each be scaled to large networks and constructed six baselines. They are:

i. Asynchronous label propagation algorithm and k-shell decomposition (aLPA + k-shell): This baseline uses the same design of HTT-II but utilises aLPA (Raghavan et al., 2007) as the community detection algorithm. The aLPA algorithm initialises every node with a unique community label and uses randomisation and an asynchronous strategy to update the labels until every node

Table 5.1 : The information of three tested networks

	#Nodes	#Edges	Average degree	Density
N1 (AI ethics)	2,163	41,833	19.34	0.0179
N2 (Bibliometrics)	5,000	388,737	77.75	0.0311
N3 (COVID-19)	4,481	551,453	123.06	0.0549

adopts the label that most of its neighbours currently have.

ii. Semi-synchronous label propagation algorithm and k-shell decomposition (sLPA + k-shell): This baseline uses the same design of HTT-II but utilises sLPA (Cordasco and Gargano, 2010) as the community detection algorithm. sLPA is a refined version of aLPA. It adopts the same ending condition but introduces less randomisation to provide more stable results than the asynchronous model.

iii. Recursive Louvain method (rLouvain): This baseline uses the Louvain method (Blondel et al., 2008) to recursively partition the input network into hierarchical communities. The Louvain method starts from a singleton partition of nodes and then iteratively removes nodes and merges communities to maximise network modularity, an objective function that measures the overall community partition quality.

iv. Recursive aLPA (raLPA): This baseline uses the aLPA method to partition the input network into hierarchical communities recursively.

v. Recursive sLPA (rsLPA): This baseline uses the sLPA method to partition the input network into hierarchical communities recursively.

vi. K-shell decomposition (k-shell): This baseline uses the original k-shell decomposition method to partition nodes into hierarchical chain groups.

5.3.2 Evaluation indicators

To measure the quality of network tree structures generated by different approaches, we followed the evaluation criteria Shang et al. (2020) proposed and designed three quantitative indicators: Topic coherence, parent-child association, and sibling topic exclusiveness. Deeming each topic as a subnetwork of G, edges in E can exist within a topic and between any two topics (parent-child, sibling or non-connected topics in the generated topic tree). Based on this understanding, we define the three evaluating indicators as follows:

i. Topic coherence (TC): In a high-quality topic tree structure, the term nodes within the same topic should be densely connected; hence the network density is a straightforward measure. We calculate the weighted mean of all the topics to indicate the overall density of the generated tree structure:

ii. Parent-child topic association (PCTA): A parent topic should topologically have a substantial association with its child topics in the topic tree structure. We first calculate the mean link possibilities for each set of parent-child topic pairs to measure such association strength. Then, we exploit the weighted summation of the mean link possibilities for all the parent topics to measure the overall parent-child association strength of the topic tree. The calculating formula of this indicator is:

iii. Sibling topic exclusiveness (STE): Multiple sibling topics can derive from the same parent topic in the topic tree structure. However, the sibling topics are supposed to be as distinctive as possible to reflect sufficient divisions of different branches. Hence, for each topic, we calculate the ratio of its inner edges to its external edges with sibling topics. Following this, we exploit the weighted summation of this value for all topics to measure the overall sibling topic exclusiveness of the tree. The formula of this indicator is:

5.3.3 Experimental results

We applied the six baselines and our HTT-II analysis to all the networks and measured the quality of generated trees via the three evaluating indicators. The results are presented in Figure 5.4. The subfigures 5.4 (A)-(C), 5.4 (D)-(F), and 5.4

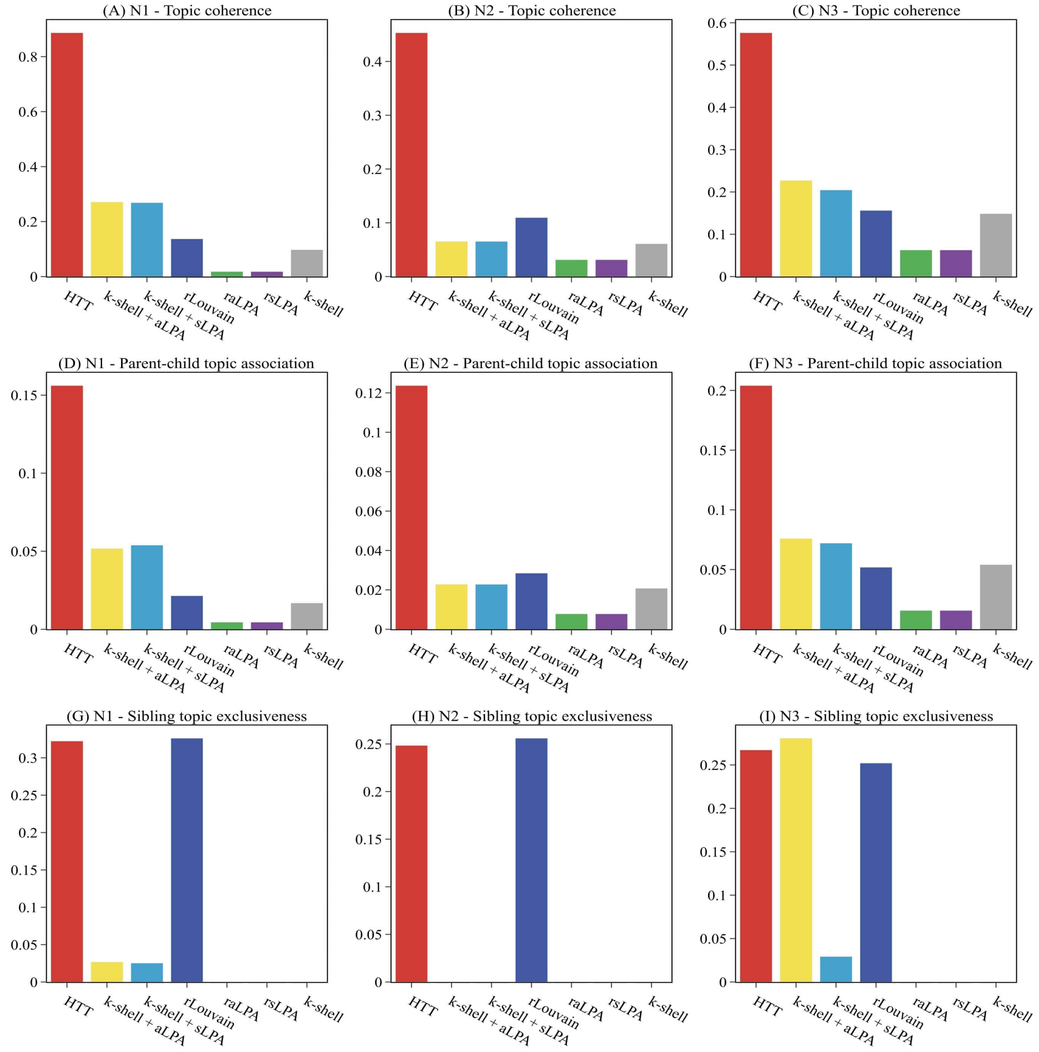


Figure 5.4 : The experimental results of seven methods

(G)-(I) respectively show HTT-II and the six baselines' TC, PCTA and STE values on the three test networks.

In Figures 5.4 (A)-(C), we observed that most TC values are significantly higher than the input network densities in Table 5.1, which means all the methods can transform sparse co-term networks into denser topic hierarchy representations. Among all the approaches, HTT-II significantly outperformed the six baselines on TC and PCTA in Figures 5.4 (A)-(F), indicating the effectiveness of incorporating the Louvain method and k-shell decomposition in generating coherent topics and consistent parent-child topic association. Focusing on the value variations of methods that use different community detection algorithms, the LPA-based methods generally performed less competitively than the Louvain method, resulting from the fuzzy boundaries between research topics formed by knowledge convergence and interactivity. According to a phenomenon frequently reported in many previous studies (Fiscarelli et al., 2019; Malhotra and Chug, 2021), LPA-based algorithms tend to merge smaller, less clear community structures into a single giant community. The original k-shell decomposition results show low TC values because it generated a substantial number of small and sparsely connected shells, which we introduced as one of its issues in the methodology section.

Figures 5.4 (G)-(I) present the STE values; The results illustrate that HTT-II still produced competitive results on three test networks, with rLouvain as a strong competitor. The ability of HTT-II to partition topics on different levels majorly comes from the Louvain community detection algorithm, which explains the similar performance of the two methods on this indicator. Besides, we observe some zero STE values in 5.4 (G)-(I). For k-shell decomposition, the zero values resulted from the fact that k-shell decomposition presented all the topic hierarchies as chains instead of tree structures, which did not yield any sibling topics in the results. The other zero values in 5.4 (G)-(I) were also a result of the chain structure issue. Although LPA-based methods were equipped with the capability to partition multiple topics on each level, they still generated only a single child topic or a few size-skewed sibling topics on the test networks as a result of the fuzzy topic boundaries.

We summarise three critical insights from the experimental results: 1) Incorporating the community detection methods into k-shell decomposition can help reveal more consistent and coherent hierarchical structures of networks. 2) The Louvain method outperforms the other two competitors in identifying topic boundaries from co-term networks. 3) Among the evaluating results of seven approaches, our proposed HTT-II can generate high-quality hierarchical topic results with highly coherent topics, strong parent-child associations and exclusive sibling topics.

5.4 Case Study: Topic Hierarchies in the Information Sciences Discipline

To demonstrate the practical use of HTT-II analysis, we conducted a case study profiling the hierarchy of research topics in the information sciences discipline. With the aid of the open data platform AMiner (Tang et al., 2008), we accessed the Microsoft Academic Graph (MAG) and collected 11,399 articles on information science (IS) from the nine most relevant journals (Hou et al., 2018): JASIST, Information Processing and Management, Journal of Informetrics, Information Research, Library and Information Science Research, Scientometrics, Research Evaluation, Journal of Documentation, and Journal of Information Science. MAG generates topic tags for each paper called fields of study (FoS) (Shen et al., 2018). We directly exploited these FoS tags to construct the co-term network containing 7,028 nodes and 137,088 edges. Then we run the HTT-II analysis on the constructed network and obtain a tree with 144 tree nodes; The topic details in the top three layers are profiled in Figure 5.5: Each circle represents a topic, and the size denotes the number of nodes contained in this topic. The different circle colours reflect the topic segmentation at the first layer.

Looking at the top-level topics, we see five prominent topic segmentation: 1) Data mining, 2) bibliometrics, 3) information seeking, 4) information systems, and 5) ontology. In the following discussion, we will dive into specific research papers within these five major topics and their child topics to interpret the results in detail. Note that the results may involve some computer science-based topics; However, our

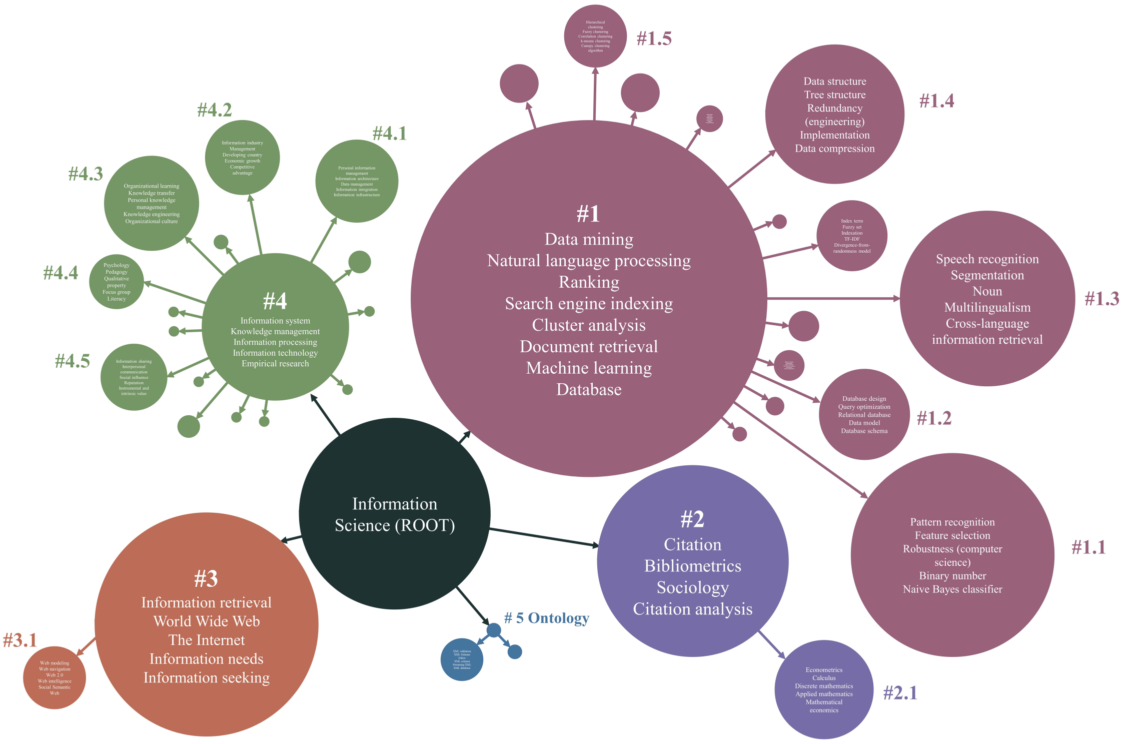


Figure 5.5 : The top three layers of the HTT-II result

interpretation is based on their applications in the information science domain, not their origins in the computer science discipline.

i. Data mining: Among the five major top-level branches, data mining is the largest. This branch indicates that data-driven approaches have become indispensable in IS research. From the term included in topic #1, we can glimpse multiple mainstream data mining approaches such as natural language processing (NLP), cluster analysis, machine learning, and database. The subordinate topics in this branch specify detailed technical directions of those approaches: including machine learning concepts and tools (#1.1), database design and utilisation (#1.2), NLP tasks and issues (#1.3), data engineering techniques (#1.4), and different clustering algorithms (#1.5). Diving into papers in each subordinate topic, #1.1 and #1.3 cover machine learning and NLP applications on a broad range of IS research tasks, including sentiment analysis (Huang et al., 2017; Melo et al., 2019; Onan and Korukoğlu, 2017), named entity extraction (Kholghi et al., 2017; Mao and Cui, 2018), and so on. Research papers in #1.2 consist of studies on database construction and database information retrieval performance and evaluation (Gu and Hwang, 2015; Yu et al., 2015). Relevant papers in #1.5 mostly contribute to the methodological innovations of clustering methods (Zhu et al., 2018) or they discuss the applications of clustering analysis given different research tasks, such as topic extraction (Fang et al., 2014; Zhang et al., 2018b), opinion mining (Hu et al., 2017; Pandey et al., 2017), and scientific behaviour patterns discovery (Xie et al., 2018).

ii. Bibliometrics: This topic highlights bibliometric studies as an important component of IS research. From the terms in the topic, we can observe that citation-related terms are frequently highlighted as citing is one of the essential scientific behaviours and the basic indicator for scientific literature analysis (Ding et al., 2009; Leydesdorff and Rafols, 2011; Zhai et al., 2018); The full list of terms also consists of several citation-derived subjects, including informetrics and altmetrics. Diving into subordinate topic #2.1, we find it intriguingly indicates two critical associated disciplines relevant to bibliometrics in IS research: mathematics and econometrics. Our investigation of the associated papers reveals that 1) mathematics and statistics are commonly applied in bibliometrics or derived research domains from identifying research activity patterns (Mir and Ausloos, 2018) and find correlations between dependent and independent variables (Abbasi et al., 2011; Thelwall, 2018b). 2) the appearance of econometrics is because the theoretical foundations of their research assumptions and the indicators developed in those bibliometric studies were derived from econometrics (Leydesdorff et al., 2019; Parolo et al., 2015;

Ruiz-Castillo and Costas, 2014). This connection indicates that transferring econometric methods to bibliometric research is a trending research paradigm.

iii. Information seeking: This topic is composed of studies on online information needs and information-seeking behaviours. Papers on this topic broadly throw light on developing or improving web information retrieval tools (Abdi et al., 2018; Fernández-Reyes et al., 2018; Song et al., 2019), exploring web-based information resources (Abad-García et al., 2018; Kousha et al., 2018; Thelwall, 2018a), proposing information recommendation algorithms (Cechinel et al., 2013; Pera and Ng, 2018), and evaluating information seeking behaviour and needs (Buchanan et al., 2019; Goyal et al., 2018; Ruthven et al., 2018). The subordinate topic #3.1 highlights the fine-grained research of web-based applications, such as web-based recommendation (Sisodia et al., 2017), web information extraction (Uçar et al., 2017), and web archiving (Dougherty and Meyer, 2014), etc.

iv. Information systems: This branch presents the qualitative research trajectory of information science and reflects its multidisciplinary nature. Studies within this branch broadly discuss issues in the information life cycle, including information creation, processing, utilisation, dissemination, and management in various social activities. The subordinate topics of this branch diverge into data management (#4.1), information industry development (#4.2), organisation knowledge management (#4.3), information education (#4.4), and information dissemination (#4.5). Diving into relevant papers in the biggest subordinate topic #4.3, we found they focus on discovering methods of producing, organising, using, storing, and sharing knowledge on the personal and organisational levels (Ahmad, 2018; La Bella et al., 2018; Shen et al., 2019).

v. Ontology: This branch is a relatively small first-layer interest group in line with information science studies. Papers in this branch majorly discuss the definition (Almeida, 2013), construction (Lubani et al., 2019; Lumsden et al., 2011) and utilisation (Browne et al., 2019; Rodríguez-García et al., 2019; Yeh et al., 2008) of ontology. The subordinate topic #5.1 highlights the prevalence of XML schemes in ontology construction (Aouadi et al., 2012; Hacherouf et al., 2015).

However, we also noticed that there are still a few coupled terms in the top-layer topics based on our experience in this field, such as the term "information technology" in topic #4, which can also make sense if it is partitioned into topic #3. This results from the fuzzy topic boundaries formed by knowledge convergence and interactivity – characteristics that typify scientific development. "Hard" community detection methods, which allocate each node into one single community, cannot well capture and represent such fuzzy characteristics. Despite our HTT-II performing best on the proposed indicators in the experiment, it still adopts a hard community detection algorithm and may result in coupled terms in topics. Hence, involving overlapping community detection methods to identify the possible overlaps between topics is a promising improvement. We will also include this issue in the Discussion section.

There are some limitations to our current study. First, HTT-II analysis is a method that reflects the knowledge components of a field. Yet, it does not generate the developing trending of topics along with time, which might be more significant and intriguing to scientists. Topic composition and hierarchies are constantly changing; Hence, we intend to build a variant of HTT-II that considers the temporal relationship between topics and how those research topics evolve. Second, HTT-II only focuses on textual data and exploits the semantic relationships from the research literature data; However, this may overlook more available data patterns, such as author collaborations, venue associations, co-citations, and citation coupling. These heterogeneous data sources have the potential to help HTT-II yield more accurate and explainable topic hierarchy results. In subsequent studies, we anticipate embedding more external information like author collaborations, publication types, and geographical locations to build a more sophisticated hierarchical topic model that incorporates exterior features. Last, as we stated at the end of the case study section, the hard partition of communities will inevitably result in coupled terms in different research topics. To generate more comprehendible and informative hierarchical topic results, we plan to equip HTT-II with overlapping community detection approaches and enable it to reveal topic overlaps.

5.5 Summary

This case study presents an adaptive and non-parametric method of identifying topic hierarchies from scientific documents called HTT-II analysis. The proposed method devises a recursive process incorporating the Louvain community detection and k-shell decomposition methods and provides a universal solution that automatically extracts topics and their hierarchical relationships. In HTT-II, k-shell decomposition and community detection parse the nodes to allocate them simultaneously in landscape and portrait directions of the tree. Nodes connected densely form different topics according to their community labels, with periphery nodes from the same community composing their child topics. Recursive implementation of this process generates the whole hierarchical structure of all identified topics until the topics are too small to be subdivided. As such, the number of topics and tree depth are also decided automatically. Experiments and a case study on real-world co-term networks demonstrate the theoretical and empirical effectiveness of HTT-II analysis.

Compared with the HTT-I model, there are a few methodological and practical differences worth highlighting:

- From a methodological standpoint, the HTT-I model utilizes a combination of K-nearest neighbours (KNN) and density peak methods, whereas the HTT-II model employs an alternative framework based on k-shell decomposition and Louvain community detection. Two primary distinctions between these models are evident: Firstly, HTT-I requires the specification of a parameter K to establish the KNN criterion, whereas HTT-II does not, rendering the latter model more adaptable. Secondly, while the HTT-I model possesses the ability to detect overlapped communities, the current HTT-II model does not support the identification of overlapping topics.

- The experimental results demonstrate that both the HTT-I and HTT-II models are capable of generating coherent topics and establishing robust parent-child topic relationships. However, the discrepancies observed in the generated trees indicate that the HTT-I model tends to capture topics at finer granularities, potentially overlooking significant links within the original network.

- From a practical perspective, it is advisable to apply the HTT-I model to networks composed of multi-disciplinary or inter-disciplinary knowledge, as it is capable of uncovering topic overlaps. Conversely, the HTT-II model is better suited for networks constructed with single-domain knowledge.

5.5.1 Technical implications

This method makes two main methodological contributions to the literature. Most importantly, HTT-II analysis provides an adaptive way of extracting topic hierarchies from scientific documents without human interference or prior expertise. The non-parametric characteristic of HTT-II is advantageous for 1) newcomers who know little about a targeted discipline and struggle to fine-tune a clustering model's parameters and 2) identifying topic hierarchies in a newly emerging research field with little knowledge background. Besides, constructing high-quality topic hierarchies can also benefit various downstream applications. For example, hierarchical topic structures can navigate efficient document retrieval in digital libraries (Dinneen et al., 2018); A built topic/item hierarchical structure can help extract multi-level features of authors/users to facilitate more accurate and explainable topic/item recommendations (Gao et al., 2019; Zhang et al., 2014a); Curated topic tree structures can serve as a hierarchical knowledge graph to empower resolving knowledge inference and question answering tasks (Yang et al., 2017).

5.5.2 Practical implications

The feasibility and utility of the proposed approach and its early versions have been proven to be useful in the pilot case studies, e.g., conceptualising the definition of digitalisation (Wu et al., 2021a) and AI literacy (Cetindamar et al., 2022), profiling research landscapes in AI ethics (Zhang et al., 2021d) and bibliometrics (Mejia et al., 2021). Those case studies empirically demonstrate the practical effectiveness of HTT analysis. Intuitively, topic hierarchies assist stakeholders in quickly comprehending the knowledge components of a research field of interest. Beyond this, it can help academic researchers, policymakers, and entrepreneurs make more informed decisions. For example, the topic hierarchies of a specific target discipline could empower individual researchers to better grasp the frontiers of research in that field, supporting them to access more relevant literature via hierarchy-based document retrieval. Additionally, creating topic hierarchies for multiple disciplines may help policymakers map research resource distributions across different domains or help to justify their funding allocation strategies. Topic hierarchies for emerging subjects or technologies, like COVID-19 treatments or electric vehicles, could help companies to chart major research pathways or be used to inform more reasonable business strategies.

Chapter 6

An Intelligent Bibliometric System and Empirical Studies on COVID-19

In this chapter, we introduce the constructed intelligent bibliometric system architecture, the developed GUI BiblioEngine for accessing the system, and two empirical COVID-19 case studies using this system. The system architecture integrates the proposed methods in Chapters 3, 4 and 5 and provides an accessible tool for non-technical background users to conduct customised case studies. To demonstrate the practical use of the system, we introduce two empirical studies on COVID-19 literature datasets to show how users can comprehensively leverage the analysing results to generate insights into real-world research issues.

6.1 Term explanation

The explanation of terms used in this section is given as follows:

- Bioentity ranking: This term refers to the process of ranking biomedical entities (including diseases, drugs, genes and genetic variations) extracted from literature according to their network importance and specificity to a target disease. The ranking can be entity category-separated or all category-mixed.
- Heterogeneous bioentity analysis: This term refers to the HBAM framework, which constructs the heterogeneous co-occurrence network of multiple categories of biomedical entities and contains ranking analysing and entity association inference methods.
- The Hierarchical Topic Tree (HTT): HTT is a network-driven methodology

designed to uncover research topics and their underlying hierarchical relationships. This method employs two models and has the capability to generate a tree-like structure where topics are organized in a hierarchical manner. These hierarchically-organized topics provide a comprehensive understanding of the interconnections and dependencies among different research areas.

- Scientific Evolutionary Pathways (SEP): SEP is a topic-tracking approach specifically designed to detect and trace research topics within a time-labelled document stream. This method effectively identifies and analyzes the evolution of research topics over time, allowing for the tracking of their progression, changes, and interrelationships within the scientific literature.

6.2 The Intelligent Bibliometric System Architecture

The architecture of the intelligent bibliometric system is illustrated in Figure 6.1. It consists of two modules and a work pipeline: Module A is designed to collect and pre-process data input; Module B is built up for conducting multiple intelligent bibliometric analyses; Module C illustrates the work pipeline to show how a customised case study is conducted. The interface of our constructed GUI (BiblioEngine) is given in Figure 6.2.

i. **Module A: Data input and pre-processing**

   This module is designed for loading and pre-processing raw data inputs. The inputs supported are 1) text/XML files downloaded from PubMed/PMC databases and 2) raw search strings. If the users input raw search strings, our GUI will access the PubMed E-utilities application programming interface (API)[1] to download the literature dataset automatically. However, this approach may face a data scalability issue due to PubMed API data regulations. Further, the

[1] https://www.ncbi.nlm.nih.gov/books/NBK25500/

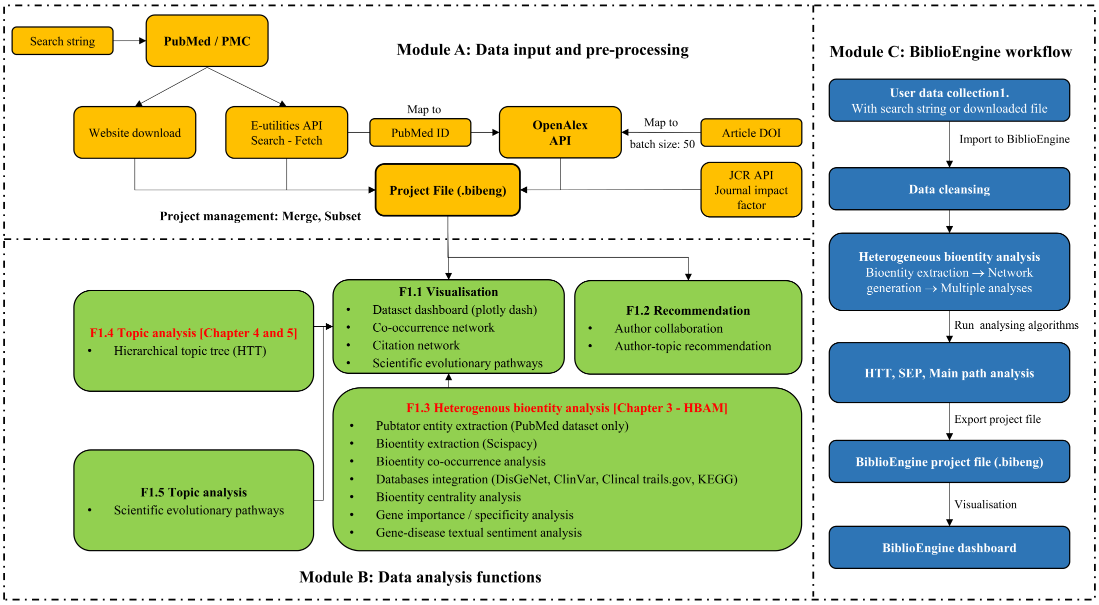

Search string
PubMed / PMC
Module A: Data input and pre-processing
Website download
E-utilities API
Search - Fetch
Map to
PubMed ID
OpenAlex
API
Map to
batch size: 50
Article DOI
JCR API
Journal impact
factor
Project File (.bibeng)
Project management: Merge, Subset
F1.1 Visualisation
Dataset dashboard (plotly dash)
Co-occurrence network
Citation network
Scientific evolutionary pathways
F1.2 Recommendation
Author collaboration
Author-topic recommendation
F1.4 Topic analysis [Chapter 4 and 5]
Hierarchical topic tree (HTT)
F1.3 Heterogenous bioentity analysis [Chapter 3 - HBAM]
Pubtator entity extraction (PubMed dataset only)
Bioentity extraction (Scispacy)
Bioentity co-occurrence analysis
Databases integration (DisGeNet, ClinVar, Clincal trails.gov, KEGG)
Bioentity centrality analysis
Gene importance / specificity analysis
Gene-disease textual sentiment analysis
F1.5 Topic analysis
Scientific evolutionary pathways
Module B: Data analysis functions
Module C: BiblioEngine workflow
User data collection1.
With search string or downloaded file
Import to BiblioEngine
Data cleansing
Heterogeneous bioentity analysis
Bioentity extraction → Network
generation → Multiple analyses
Run analysing algorithms
HTT, SEP, Main path analysis
Export project file
BiblioEngine project file (.bibeng)
Visualisation
BiblioEngine dashboard

Figure 6.1 : Intelligent bibliometric system architecture

GUI will apply multiple pre-processing steps to the imported dataset, including mapping the collected papers to the OpenAlex[2] database to curate disambiguated author/affiliation information and concept information, searching Journal Citation Report (JCR) API[3] to retrieve journal impact factor information.

ii. **Module B: Data analysis functions**

This module includes the HBAM proposed in Chapter 3 as the heterogeneous bioentity analysis block and the HTT models proposed in Chapters 4 and 5 as the Topic analysis - HTT block. The two function blocks will run automatically on the imported dataset and generate corresponding results in tabular formats or visualisations. In the current version of this system, we support both HTT models (I and II) for topic analysis. Apart from the two pro-

[2] https://openalex.org/

[3] https://jcr.clarivate.com/jcr/home

posed methods, we also integrate two methods: Scientific evolutionary pathways (SEP) analysis developed in our pilot study (Zhang et al., 2017c) and diffusion-based recommendation algorithm proposed in a more recent exploration (Zhang et al., 2022) to fulfil the work pipeline of intelligent bibliometrics. The primary visualisation tools we have integrated into the GUI dashboard are the Python Plotly[4] and gravis[5] packages and the JavaScript Vega visualisation package[6]. Another two individual visualisation tools we additionally employed are VoSViewer[7] and Gephi software[8].

iii. **Module C: Systematic analysis workflow**

This workflow illustrates how non-technical background users can access this system and conduct a customised analysis. It consists of six steps: 1) User data collection (via search string or downloaded file), 2) data cleansing, 3) heterogeneous bioentity analysis via HBAM, 4) HTT, SEP and citation main path analysis, 5) export to BiblioEngine project file (save as a persistent model), and 6) Dashboard visualisation.

6.3 BiblioEngine GUI Introduction

In this section, we will present our developed GUI named BiblioEngine for realising the intelligent bibliometric system. The screenshot of the main interface is in Figure 6.2.

[4] https://plotly.com/

[5] https://robert-haas.github.io/gravis-docs/

[6] https://vega.github.io/

[7] https://www.vosviewer.com/

[8] https://gephi.org/

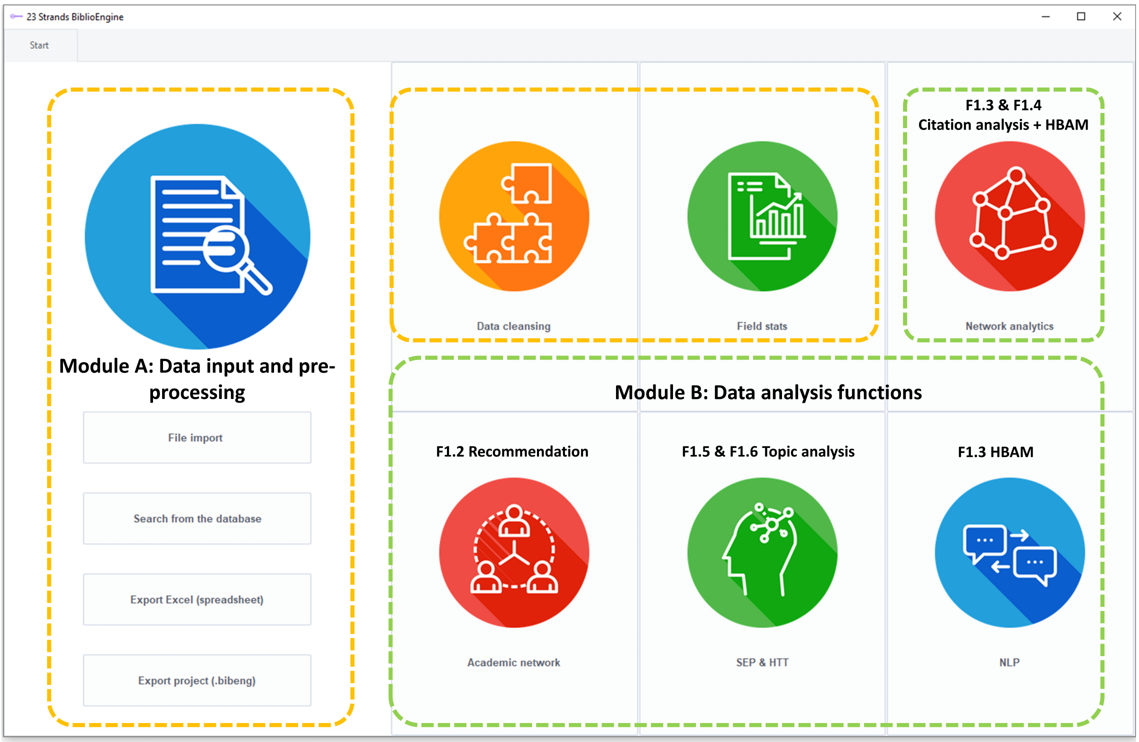


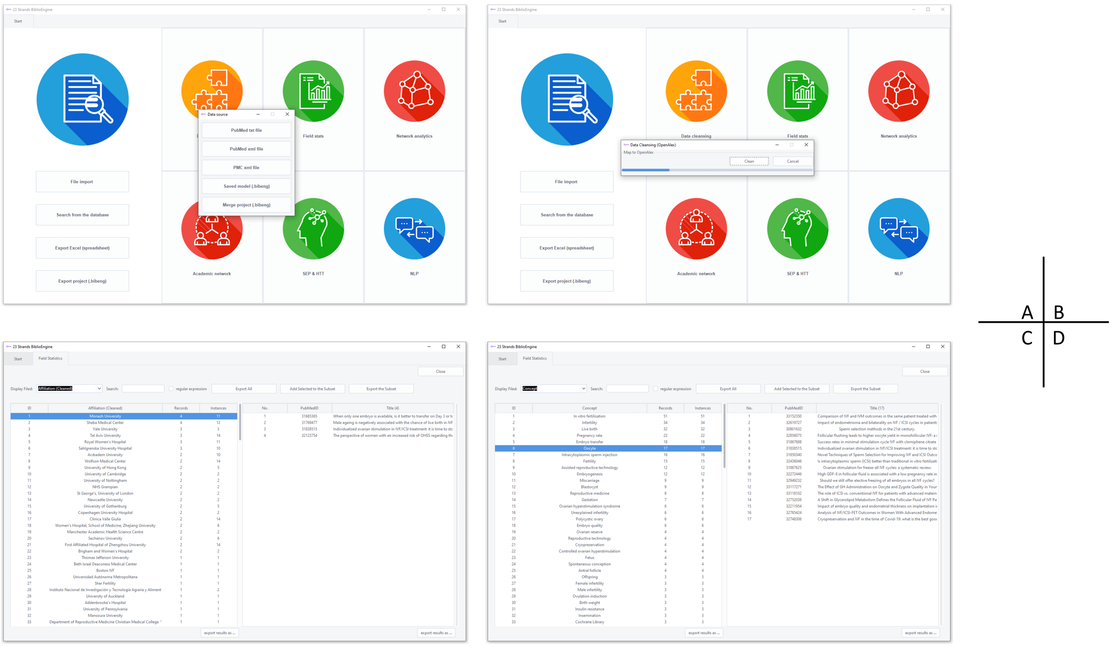
Figure 6.2 : The BiblioEngine GUI

6.3.1 Module A: Data input and pre-processing

The data import and pre-processing steps are shown in Figure 6.3. BiblioEngine supports inputs from PubMed/PMC database download and raw string search, which is listed in Figure 6.3A. Figure 6.3B is mapping the imported dataset to OpenAlex for retrieving disambiguated authors, affiliations, and concept terms (that can be used as our SEP/HTT input). Figures 6.3C and 6.3D respectively present the disambiguated affiliations and concept terms.

6.3.2 Module B: F1.1 Visualisation

For visualising the processed dataset and analysing results, we developed a Plotly-based dashboard as shown in Figure 6.4 to illustrate the year distribution, bibliographic indicator (including year, journal, authors, affiliations, extracted bioentity, etc.) rankings and a series of filters to narrow down the imported dataset.



Display Filed: Affiliation (Cleaned)

ID	Affiliation (Cleaned)	Records	Instances
1	Monash University	4	11
2	Sheba Medical Center	4	12
3	Yale University	3	3
4	Tel Aviv University	3	14
5	Royal Women's Hospital	3	11
6	Sahlgrenska University Hospital	3	10
7	Acıbadem University	2	10
8	Wolfson Medical Center	2	14
9	University of Hong Kong	2	5
10	University of Cambridge	2	2
11	University of Nottingham	2	2
12	NHS Grampian	2	2
13	St George's, University of London	2	2
14	Newcastle University	2	2
15	University of Gothenburg	2	5
16	Copenhagen University Hospital	2	2
17	Clinica Valle Giulia	2	14
18	Women's Hospital, School of Medicine, Zhejiang University	2	6
19	Manchester Academic Health Science Centre	2	2
20	Sechenov University	2	6
21	First Affiliated Hospital of Zhengzhou University	2	14
22	Brigham and Women's Hospital	2	2
23	Thomas Jefferson University	1	1
24	Beth Israel Deaconess Medical Center	1	1
25	Boston IVF	1	1
26	Universidad Autónoma Metropolitana	1	1
27	Sher Fertility	1	1
28	Instituto Nacional de Investigación y Tecnología Agraria y Aliment	1	2
29	University of Auckland	1	1
30	Addenbrooke's Hospital	1	1
31	University of Pennsylvania	1	1
32	Mansoura University	1	1
33	Department of Reproductive Medicine Christian Medical College '	1	1

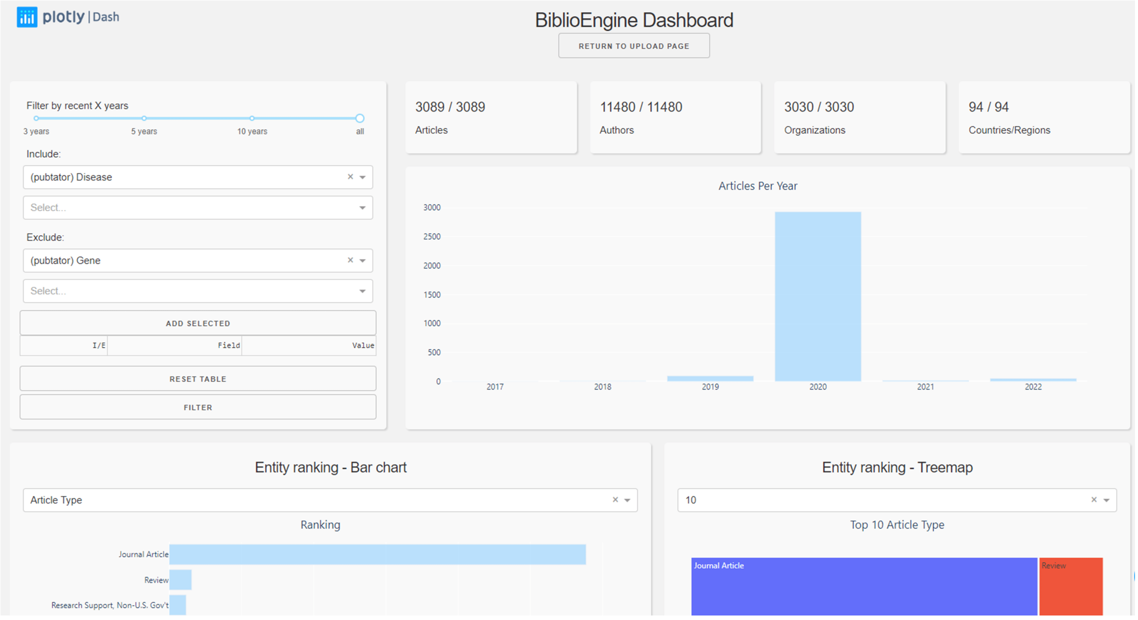
No.	PubMedID	Title (4)
1	31685365	When only one embryo is available, is it better to transfer on Day 3 or t
2	31769477	Male ageing is negatively associated with the chance of live birth in IVF
3	31838515	Individualized ovarian stimulation in IVF/ICSI treatment: it is time to sto
4	32123754	The perspective of women with an increased risk of OHSS regarding th

Display Filed: Concept

ID	Concept	Records	Instances
1	In vitro fertilisation	51	51
2	Infertility	34	34
3	Live birth	32	32
4	Pregnancy rate	22	22
5	Embryo transfer	18	18
6	Oocyte	17	17
7	Intracytoplasmic sperm injection	16	16
8	Fertility	15	15
9	Assisted reproductive technology	12	12
10	Embryogenesis	12	12
11	Miscarriage	9	9
12	Blastocyst	9	9
13	Reproductive medicine	8	8
14	Gestation	7	7
15	Ovarian hyperstimulation syndrome	6	6
16	Unexplained infertility	6	6
17	Polycystic ovary	6	6
18	Embryo quality	6	6
19	Ovarian reserve	4	4
20	Reproductive technology	4	4
21	Cryopreservation	4	4
22	Controlled ovarian hyperstimulation	4	4
23	Fetus	4	4
24	Spontaneous conception	4	4
25	Antral follicle	4	4
26	Offspring	3	3
27	Female infertility	3	3
28	Male infertility	3	3
29	Ovulation induction	3	3
30	Birth weight	3	3
31	Insulin resistance	3	3
32	Insemination	3	3
33	Cochrane Library	3	3

No.	PubMedID	Title (17)
1	33152350	Comparison of IVF and IVM outcomes in the same patient treated with
2	32619727	Impact of endometrioma and bilaterality on IVF / ICSI cycles in patients
3	30801632	Sperm selection methods in the 21st century.
4	32856073	Follicular flushing leads to higher oocyte yield in monofollicular IVF: a
5	31867688	Success rates in minimal stimulation cycle IVF with clomiphene citrate
6	31838515	Individualized ovarian stimulation in IVF/ICSI treatment: it is time to sto
7	31850340	Novel Techniques of Sperm Selection for Improving IVF and ICSI Outco
8	32436048	Is intracytoplasmic sperm (ICSI) better than traditional in vitro fertilizati
9	31867625	Ovarian stimulation for freeze-all IVF cycles: a systematic review.
10	32272446	High GDF-8 in follicular fluid is associated with a low pregnancy rate in
11	32949232	Should we still offer elective freezing of all embryos in all IVF cycles?
12	33117271	The Effect of GH Administration on Oocyte and Zygote Quality in Your
13	33118102	The role of ICSI vs. conventional IVF for patients with advanced matern
14	32752038	A Shift in Glycerolipid Metabolism Defines the Follicular Fluid of IVF Pa
15	32211954	Impact of embryo quality and endometrial thickness on implantation i
16	32765424	Analysis of IVF/ICSI-FET Outcomes in Women With Advanced Endome
17	32748306	Cryopreservation and IVF in the time of Covid-19: what is the best goo

Figure 6.3 : Module A: Data input and pre-processing



Include:
(pubtator) Disease
Select...

Exclude:
(pubtator) Gene
Select...

I/E	Field	Value

3089 / 3089 Articles

11480 / 11480 Authors

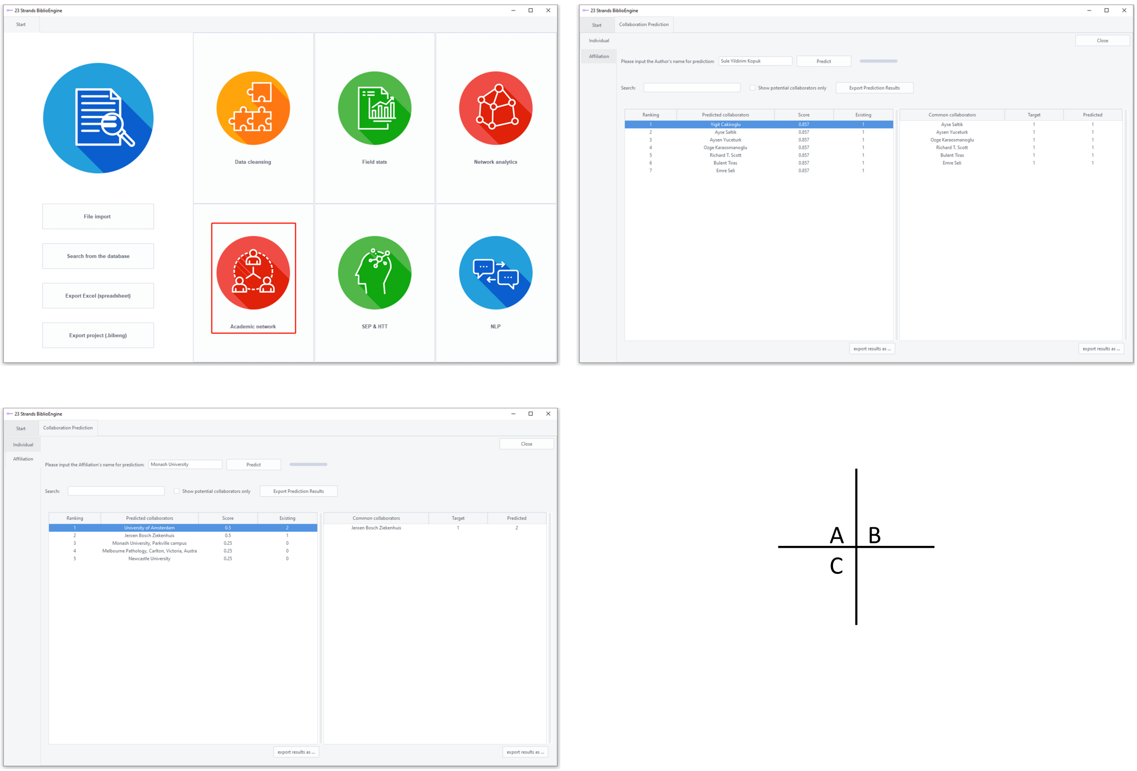
3030 / 3030 Organizations

94 / 94 Countries/Regions

Article Type

10

Figure 6.4 : Module B: F1.1 Visualisation



Please input the Author's name for prediction: Sule Yildirim Kopuk

Ranking	Predicted collaborators	Score	Existing
1	Yigit Cakiroglu	0.857	1
2	Ayse Saltik	0.857	1
3	Aysen Yuceturk	0.857	1
4	Ozge Karaosmanoglu	0.857	1
5	Richard T. Scott	0.857	1
6	Bulent Tiras	0.857	1
7	Emre Seli	0.857	1

Common collaborators	Target	Predicted
Ayse Saltik	1	1
Aysen Yuceturk	1	1
Ozge Karaosmanoglu	1	1
Richard T. Scott	1	1
Bulent Tiras	1	1
Emre Seli	1	1

Please input the Affiliation's name for prediction: Monash University

Ranking	Predicted collaborators	Score	Existing
1	University of Amsterdam	0.5	2
2	Jeroen Bosch Ziekenhuis	0.5	1
3	Monash University, Parkville campus	0.25	0
4	Melbourne Pathology, Carlton, Victoria, Austra	0.25	0
5	Newcastle University	0.25	0

Common collaborators	Target	Predicted
Jeroen Bosch Ziekenhuis	1	2

Figure 6.5 : Module B: F1.2 Recommendation

6.3.3 Module B: F1.2 Recommendation

This is the extra function we developed in (Zhang et al., 2022). It aims at recommending research topics and collaborators for scholars and organisations. Figure 6.5A, B and C show the function entrance and individual/organisational collaborator recommendation results.

6.3.4 Module B: F1.3 Heterogeneous bioentity analysis

Figure 6.6A show the function entrance for HBAM. Figures 6.6B, C and D illustrate the stepwise guidance of performing Pubtator extraction, network construction, indicator calculation and non-dominated ranking to the processed dataset. Figure 6.7 gives the visualisation results via functions we developed in the dashboard.

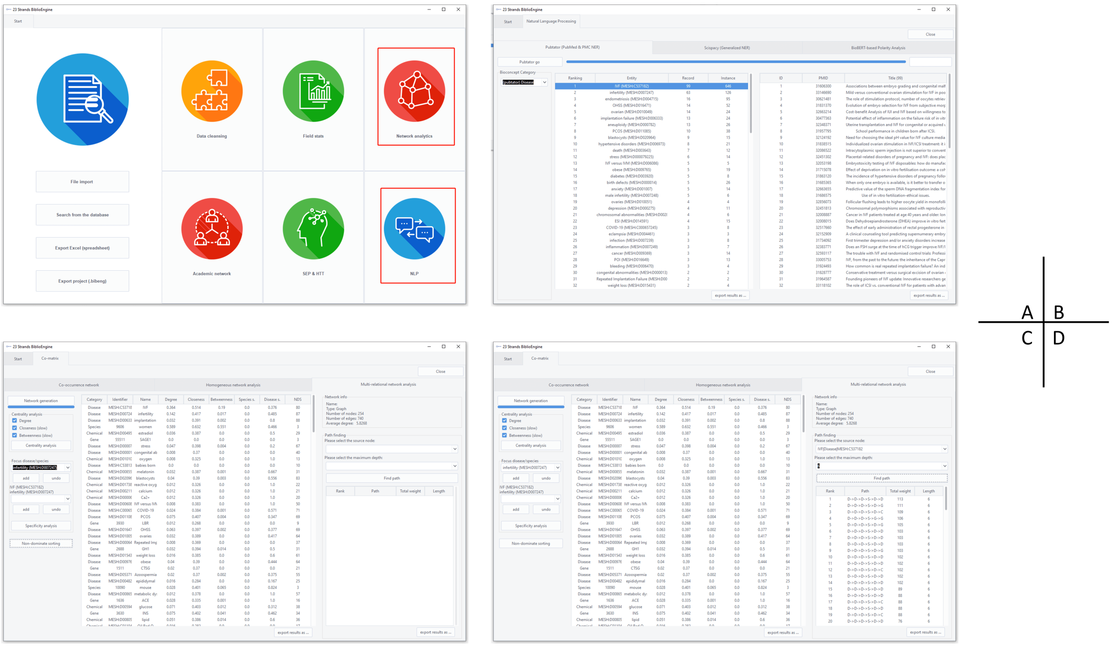
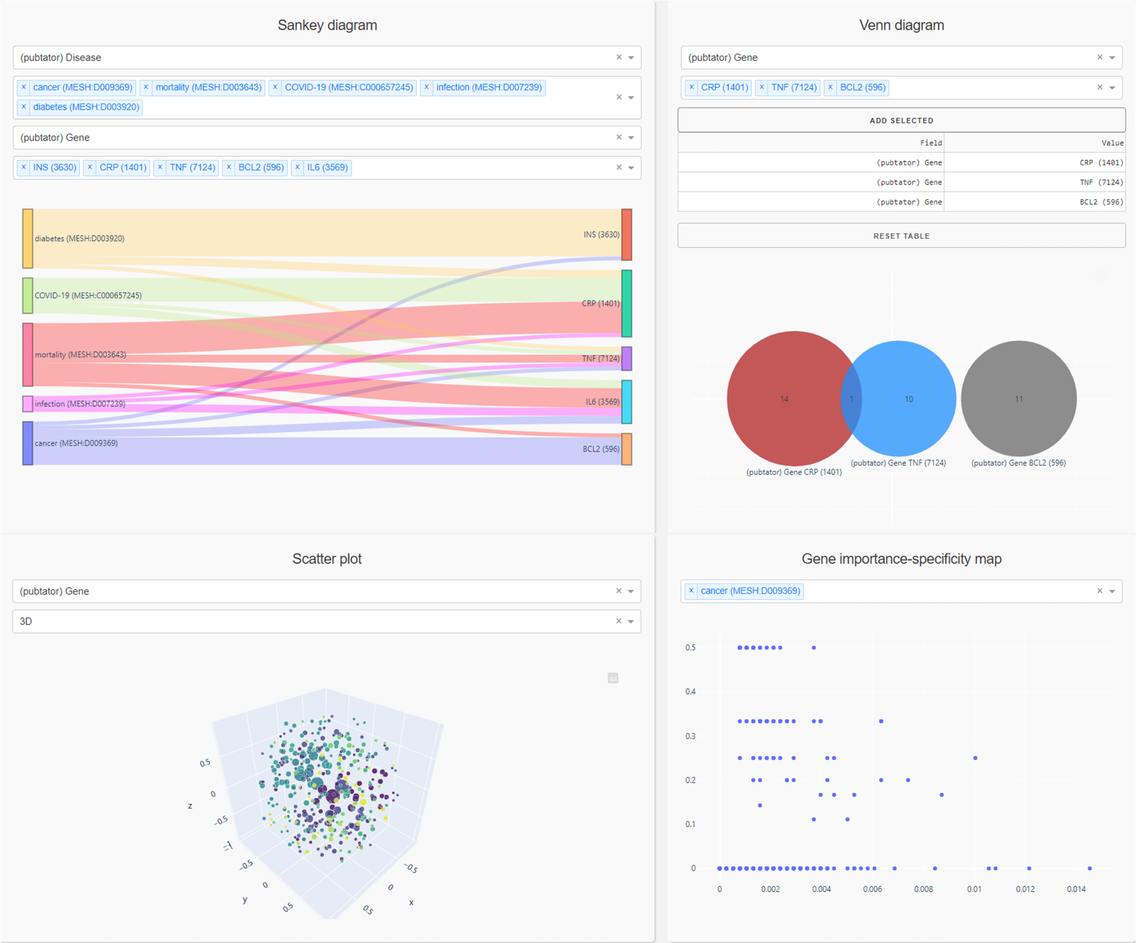


Figure 6.6 : Module B: F1.3 Heterogeneous bioentity analysis

6.3.5 Module B: F1.4 and F1.5 Topic analysis - SEP and HTT

Figure 6.8A show the function entrance for SEP and HTT analyses. Figure 6.8B is the operation for SEP and HTT analysis, with the input of either the MeSH terms affiliated with PubMed papers or concept terms retrieved from OpenAlex. The visualisations of the analysing results are given in Figures 6.8C and 6.8D. They are integrated into BiblioEngine and realised by the Python Gravis package and Vega visualisation grammar.



Sankey diagram

(pubtator) Disease

cancer (MESH:D009369) · mortality (MESH:D003643) · COVID-19 (MESH:C000657245) · infection (MESH:D007239) · diabetes (MESH:D003920)

(pubtator) Gene

INS (3630) · CRP (1401) · TNF (7124) · BCL2 (596) · IL6 (3569)

Venn diagram

(pubtator) Gene

CRP (1401) · TNF (7124) · BCL2 (596)

ADD SELECTED

Field	Value
(pubtator) Gene	CRP (1401)
(pubtator) Gene	TNF (7124)
(pubtator) Gene	BCL2 (596)

RESET TABLE

14
1
10
11
(pubtator) Gene CRP (1401)
(pubtator) Gene TNF (7124)
(pubtator) Gene BCL2 (596)

Scatter plot

(pubtator) Gene

3D

Gene importance-specificity map

cancer (MESH:D009369)

Figure 6.7 : Module B: F1.3 Heterogeneous bioentity analysis results visualisation

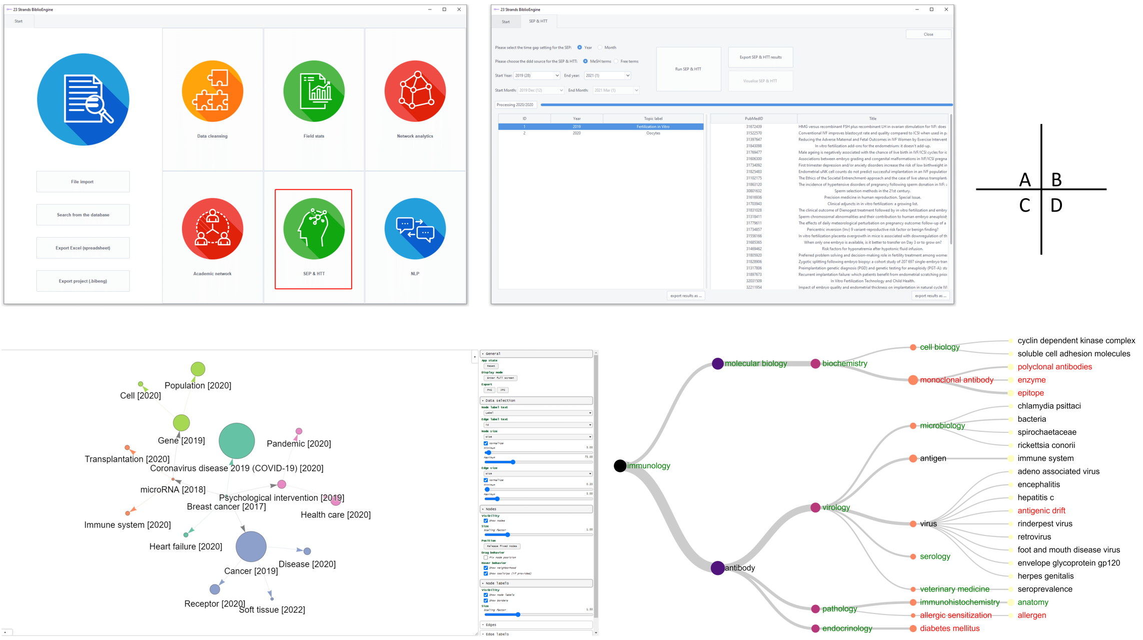


Figure 6.8 : Module B: F1.4 and F1.5 Topic analysis - SEP and HTT

6.4 The Comparison of BiblioEngine with Existing Bibliometric Tools

In contrast to existing bibliometric tools such as CiteSpace (Chen, 2006), VoSViewer (Van Eck and Waltman, 2010) and Bibliometrix (Aria and Cuccurullo, 2017), BiblioEngine boasts a range of distinctive features that set it apart in the field of literature mining. Notably, BiblioEngine excels in the algorithms to uncover entity-level knowledge, facilitate entity association prediction, and profile topic evolution and hierarchy. When compared to its counterparts, BiblioEngine offers several notable advantages:

- Enhanced entity-level knowledge: BiblioEngine excels in extracting and analysing entities within the literature, enabling a deeper understanding of the knowledge landscape. By delving into the granular details of entities, BiblioEngine enables researchers to gain valuable insights on a more fine-grained level.

- Topic extraction and profiling algorithms: BiblioEngine seamlessly integrates the HTT and SEP algorithms to identify topic evolution and hierarchy within time-labelled scholarly documents effectively. Extensively validated through numerous case studies, these algorithms have consistently demonstrated their ability to derive valuable real-world insights.

- Integrated visualisations and interactive dashboard: BiblioEngine offers an intuitive and user-friendly dashboard, ensuring ease of use for researchers of varying technical backgrounds. This platform serves as a convenient pipeline for users to apply entity extraction, heterogeneous network analysis, and topic analysis techniques to their collected literature data. Moreover, it provides sophisticated visualisation options to effectively present analysis outcomes. The platform offers visually appealing and customisable visual layouts, enhancing the clarity and interpretation of results.

6.5 COVID-19 Bioentity Association Analysis

In the following two empirical studies, we will focus on COVID-19, a global public health threat, to demonstrate how to use the proposed system in discovering bioentity associations and uncovering knowledge hierarchy. The two studies were conducted separately in our two research papers, but the relevant data analysis parts were performed using BiblioEngine.

During the time of this thesis being curated, the COVID-19 pandemic has developed into an unprecedented global crisis that impacts people's daily lives and healthcare services provision. To stop its spread and efficiently control it, the biomedical research community has responded proactively on multiple fronts, including in the field of genetic research. By deciphering the genetic mechanisms underlying the body's response to SARS-CoV-2 infection, researchers can better understand COVID-19 pathogenesis, diagnosis, treatment, and, potentially, prevention, including optimising vaccine development. In practical terms, the multiple COVID-19 genetic research efforts have resulted in substantial research publications (Chua et al., 2020; Pairo-Castineira et al., 2021; Shin et al., 2020; Wrobel et al., 2020). However, the downside of this productivity is that the quantity of COVID-19 literature proliferates and results in difficulties for researchers in comprehending this field's changing knowledge landscape, particularly regarding the emerging information on various genes involved in COVID-19 response.

Bibliometrics is a subject that deciphers the patterns generated by scientific endeavours by quantitatively tracking and measuring research activities. The intensive growth of COVID-19 publications has triggered considerable attention from the bibliometrics community. As part of those studies,Colavizza et al. (2021) tracked the topics in COVID-19 research, Chahrour et al. (2020) presented descriptive statistics of publication distribution, Fry et al. (2020) revealed the influence of COVID-19 on scientific activities such as international collaboration patterns, Zhang et al. (2021a) highlighted topical disruptions and resilience to the coronavirus research focus. Traditional bibliometric studies are conducted based on the statistical analysis of bibliographic information such as author entities, keywords and citations. Empowered with artificial intelligence (AI) techniques like text mining and network analytics, current bibliometrics has developed novel capabilities of excavating implicit knowledge and inferring potential knowledge associations from literature data, which could be described as intelligent bibliometrics (Zhang et al., 2020b). The case study presented here differs from other COVID-19 bibliometric studies in that it focuses specifically on COVID-19 genetic research. It utilises multiple traditional and intelligent bibliometric analyses to profile this emerging field's research landscape and address the following question: What specific genes are frequently highlighted, and which ones are emerging as relatively newly-described entities that may be potentially important in the COVID-19 genetic studies?

Previous efforts to answer these questions mostly consist of COVID-19 topic analysis (Colavizza et al., 2021; Pourhatami et al., 2021; Haghani and Bliemer, 2020; Zhang et al., 2021a), literature-based discovery studies (Wise et al., 2020; Wu et al., 2021b; Yu et al., 2021), and literature search tools (Chen et al., 2021; Trewartha et al., 2020). A common approach in current topic studies is to apply co-word clustering (Pourhatami et al., 2021) or topic modelling (Colavizza et al., 2021; Tran et al., 2020) to the COVID-19 literature. Such studies have helped to track newly emerging knowledge but have often overlooked the relationships between new evidence and previously established coronavirus knowledge. For example, what are the similarities and differences between the diagnosing criteria, treatments, and social responses for SARS and SARS-CoV-2? In such cases, established knowledge can effectively discover and synthesise new knowledge (Haghani and Bliemer, 2020; Haghani and Varamini, 2021; Hu et al., 2021; Petrosillo et al., 2020). In addition, current literature-based discovery studies are conducted on macro levels and do not focus on specific knowledge domains to discover targeted knowledge for people pursuing different interests (Wise et al., 2020; Yu et al., 2021). For example, the interests of virologists and pathologists lie in tracing the spike protein mutations of SARS-CoV-2 (Pairo-Castineira et al., 2021; Starr et al., 2021), while clinical doctors are eager to follow the latest progress in diagnosis and treatment (Felsenstein et al., 2020; Merrill et al., 2020). For this reason, combining topic analysis and literature-based discovery approaches is a promising way to fill these two gaps. Further, few of the available COVID-19 knowledge search tools provide visualisations or other efficient ways to assist users in understanding the retrieved results (Trewartha et al., 2020; Zhang et al., 2021a). A concise and appropriate visualisation could save a huge amount of time in finding the right papers to follow or in narrowing down their search scope. Aiming to fill these research gaps, we employed the heterogeneous bioentity analysis methodology that provides a systematic solution to answering the three cited research questions.

6.5.1 Data collection and pre-processing

PubMed is a public biomedical literature database developed by the National Library of Medicine (NIH) and comprises over 32 million medical articles and online books. Falagas et al. (2008) recommend PubMed as the optimal bibliometric database for medical and life sciences, which exactly coheres with the foci of our study. Its advantage in biomedical information retrieval is providing specialised functions like Medical Subject Heading (MeSH) search and biomedical filters (including the species filter we used in this work) to return precise results. While considering Web of Science (WoS) is also a well-recognised data source in traditional bibliometric studies, we compared the search results of WoS and PubMed using the same search string (the filter and MeSH terms excluded) and noted that 93% of our collected data are indexed by WoS, indicating a wide coverage of PubMed data. Given that circumstances, in our study, we only exploited the PubMed database as our data source.

Using the search strategy below, we collected a data set of the genetic research performed on COVID-19 and SARS-CoV-2 from PubMed:

(”COVID-19/genetics”[MeSH Terms] OR ((”genes”[MeSH Terms] OR ”genetics”[MeSH Terms] OR ”gene”[All Fields] OR ”genes”[All Fields] OR ”genome”[All Fields] OR ”genetics”[All Fields]) AND (”COVID-19”[All Fields] OR ”SARS-Cov-2”[All Fields]))) AND (humans [Filter])

Search date: 08/03/2021

The search yielded 5,632 records related to COVID-19 genetic research. We restricted the species to humans since our primary goal is only to explore the critical human genes that act in COVID-19 infection.

Table 6.1 : Stepwise results of the pre-processing procedure

	Raw	**Step 1**	**Cleaned**	**Step 2**	**Nodes**
Disease	31,974	Removed noisy concepts like "cardioembolic", "JAGS", "nonvitamin", etc. that could not be mapped to MeSH	31,963	MeSH	801
Chemical	4,494		3,724		678
Gene	11,211	Exclude genes that do not belong to Homo Sapiens	8,781	NCBI Gene	968
Genetic variant					
- DNA mutation	69		17		126
- Protein mutation	349	Removed variants with unclear loci (i.e., could not be mapped to an SNP ID)	91	dbSNP	
- SNP	104		104		
Total	**48,201**	-	**44,680**	-	**2,573**

6.5.2 Bioentity extraction and pre-processing

Following the heterogeneous bioentity analysis methodology (HBAM) proposed in Chapter 3, we first exploited Pubtator to extract bioentities from the 5,632 PubMed records, resulting in 48,201 raw biomedical concepts, including diseases, chemicals, genes and genetic variants. The cleaning process was then conducted to map every concept to the corresponding dictionary, remove noisy concepts (Step 1) and consolidate concepts (Step 2). 2,573 unique bio-entities remained after the cleaning steps with the stepwise results presented in Table 6.1. The 2,573 bi-entities were then used to construct the heterogeneous bioentity co-occurrence network.

The finalised network comprised 2,573 bioentity nodes and 31,848 co-occurrence edges. The counts of different types of nodes and edges are given in Table 6.2. The numbers in the parentheses are the node counts of the corresponding bioentities, and the numbers in the tabular cells are edge counts.

6.5.3 Bioentity ranking analysis

To begin with the bioentity analysis, we listed the top ten highly frequent bioentities and presented them in Table 6.3. The monthly changes of those entities

Table 6.2 : Counts of the different types of edges

	Disease (801)	Chemical (678)	Gene (968)	Genetic Variant (126)
Disease (801)	8,231	4,872	6,966	499
Chemical (678)	4,872	2,121	2,268	37
Gene (968)	6,966	2,268	5,692	385
Genetic Variant (126)	499	37	385	777

Table 6.3 : The top ten entities ranked by the raw frequency

Ranking	**Disease**	**Chemical**	**Gene**	**Genetic Variant**
1	Death	Oxygen	ACE2	rs2285666
2	Pneumonia	Hydroxychloroquine	TMPRSS2	rs12329760
3	Inflammation	remdesivir	IL6	rs4646116
4	Fever	Serine	CRP	rs11385942
5	Neoplasms	Chloroquine	TNF	rs12252
6	Respiratory Distress Syndrome, Adult	Lipids	CD4	rs1244687367
7	Cough	Azithromycin	ACE	rs143936283
8	Diabetes Mellitus	lopinavir-ritonavir drug combination	CD8A	rs73635825
9	Hypertension	Nitrogen	IFNG	rs8176746
10	Zoonoses	Aldosterone	FURIN	rs8176719

during the entire year 2021 are provided in Figures 6.9, 6.10 and 6.11. Genetic variation entities are not given due to the relatively small amount of data in this category. We only traced the frequency changes to January 2021 since the latter collection is incomplete due to publishing lags.

Generally, the frequencies of the top ten entities in each category keep increasing during the early pandemic. Predictably, *ACE2* is the top-mentioned gene with a noticeable frequency gap with the following genes. This is mainly because *ACE2* is the primary functional receptor for the SARS-CoV-2 virus in human cells and plays a core role in the virus infection process (Zheng et al., 2020). Regarding the comorbidities, apart from cough, respiratory distress syndrome and inflammation that could be regarded as symptoms or directly associated disease manifestations of COVID-19 infection, the comorbidities in COVID-19 genetic studies include neo-

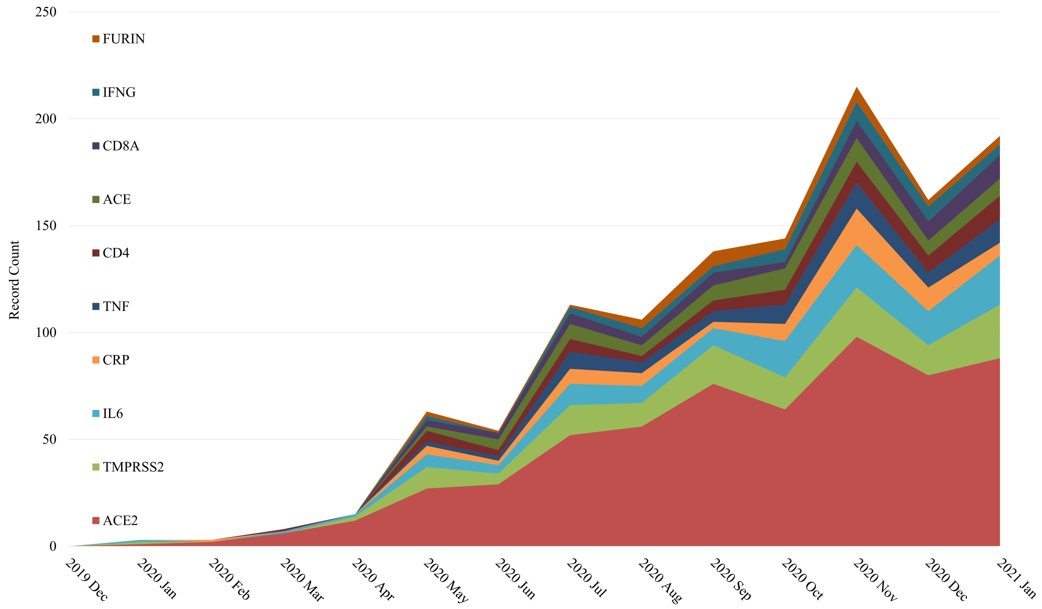


Figure 6.9 : Top ten gene frequency changes

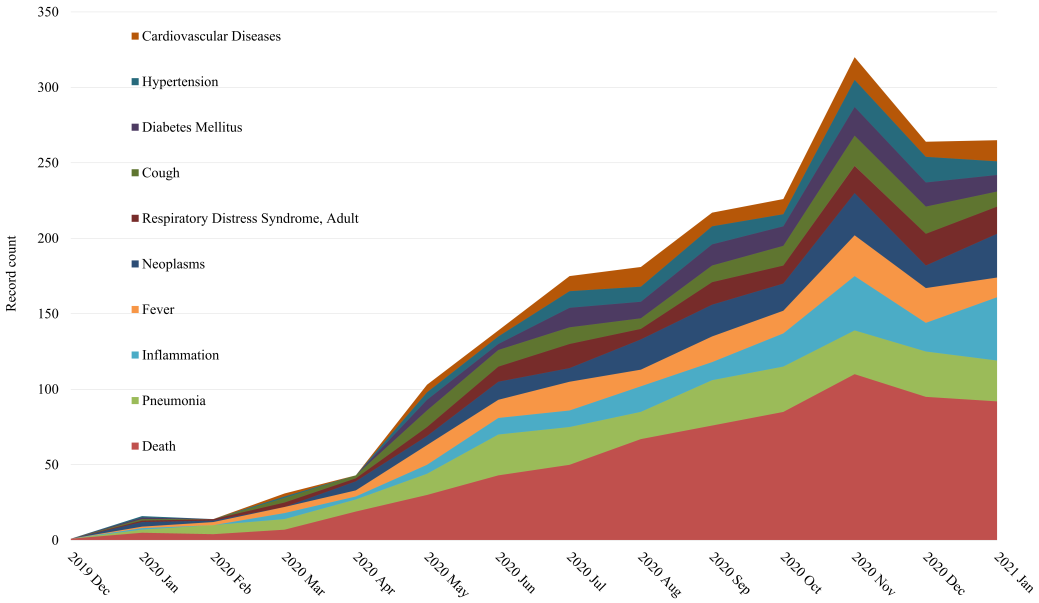


Figure 6.10 : Top ten disease (comorbidity) frequency changes

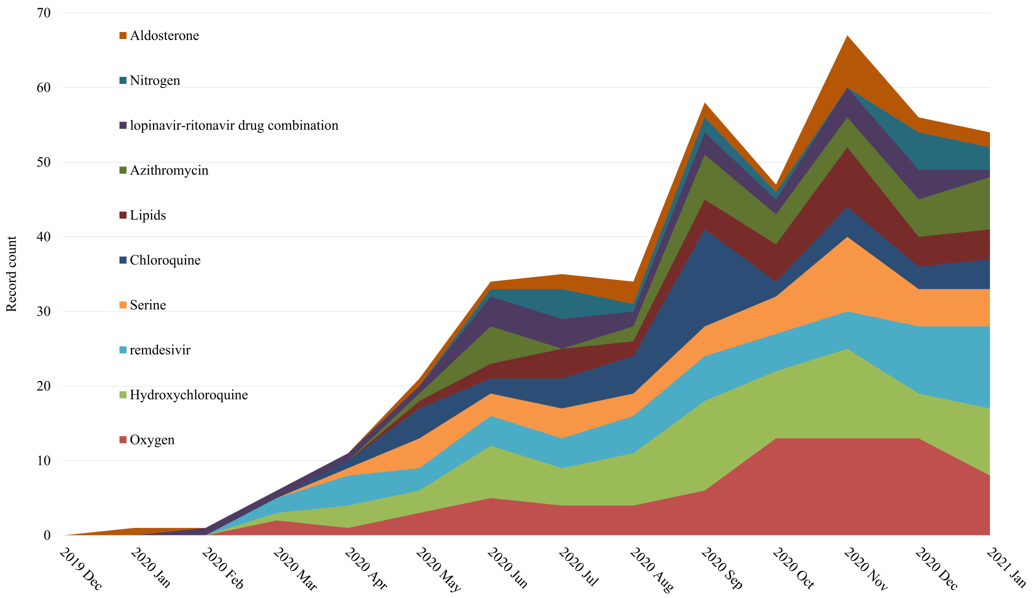


Figure 6.11 : Top ten chemical frequency changes

plasms, diabetes and hypertension. When it comes to chemicals, Figure 6.11 presents multiple prevalent drug treatments that were utilised and trialled for COVID-19, including hydroxychloroquine, azithromycin, remdesivir and the lopinavir-ritonavir drug combination, indicating that the pharmacogenomics of those drugs is also a key research interest in this field.

6.5.4 Emerging gene discovery

By applying centrality measurement, the non-dominated sorting centrality combination algorithm, and the intersection ratio calculation to the constructed network, we attributed every gene node two indicators: centrality combination and intersection ratio. We then normalised the two indicators as the intersection ratio on the X-axis and centrality combination on the Y-axis to locate all the genes in a coordinate system, as shown in Figure 6.12. According to our design, a high value of centrality combination indicates the importance/impact of a given gene to relatively broad domains of the target disease, while a high value of intersection ratio

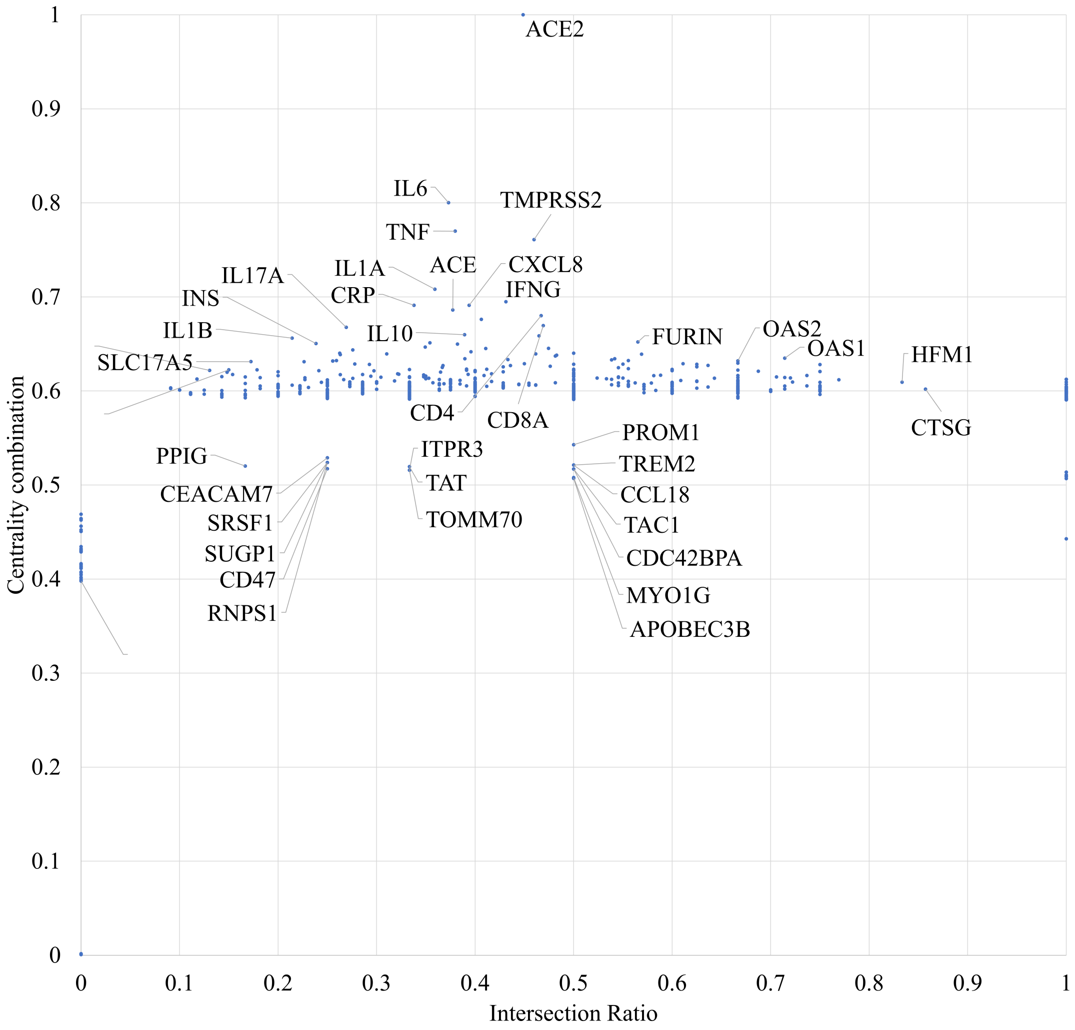


Figure 6.12 : Gene map derived from COVID-19 research

may represent the specialty of a given gene to the target disease.

The genes depicted are shown as a matrix based on their centrality combination, i.e., the strength of their contribution to the surveyed COVID-19 literature in biological terms, together with their intersection ratio, i.e., the strength of their relationship to COVID-19, as deduced from the published body of COVID-19 genetic research. For example, *ACE2* and *TMPRSS2* have a relatively strong presence in the literature, as well as a relatively crucial biological influence, and are also moderately strongly specific for COVID-19. From the centrality combination value perspective, we could identify *ACE2*, *IL6*, *TMPRSS2*, and *TNF* as frequently highlighted genes. *ACE2* is the primary functional receptor for the SARS-CoV-2 virus (Zheng et al., 2020). *TMPRSS2* is an enzyme that primes the spike S protein of the SARS-CoV-2 to promote virus entry (Hoffmann et al., 2020b). *IL6* and *TNF* are pro-inflammatory cytokines found generally elevated in severe COVID-19 patients (Cao, 2020).

Figure 6.13 zooms into some relatively recently mentioned genes in the literature that own both high intersection ratio and centrality combination in the first quadrant. The potential role of some of these ”emerging” genes and their products, together with their possible biological role in COVID-19, is discussed below:

FURIN: *FURIN* is an essential cleavage enzyme for the spike protein of SARS-CoV-2 in the virus infection process. From the biochemical perspective, Klimstra et al. (2020) identified the association between a putative furin cleavage signal generated by a novel insertion of the SARS-CoV-2 spike S glycoprotein and the expanded host range. Wrobel et al. (2020) discovered that the cleavage at the furin-cleavage site decreases the overall stability of SARS-CoV-2 S and facilitates the adoption of the open conformation required for the viral S (spike) protein to bind to the ACE2 receptor. From the treatment perspective, Hoffmann et al. (2020a) highlighted that obtaining an S1/S2 multibasic cleavage site was essential for COVID-19 infection and indicated furin as a potential target for therapeutic intervention. A similar finding was also presented by Sallenave and Guillot (2020), whose study identified a furin-like cleavage site in SARS-CoV-2 to facilitate the S protein priming. They also claimed that furin inhibitors could be targeted as potential drug therapies for SARS-CoV-2.

CXCL10: *CXCL10* is a frequently studied gene in multiple COVID-19 genetic studies (Bermejo-Martin et al., 2020; Chua et al., 2020; Han et al., 2021; Hou et al.,

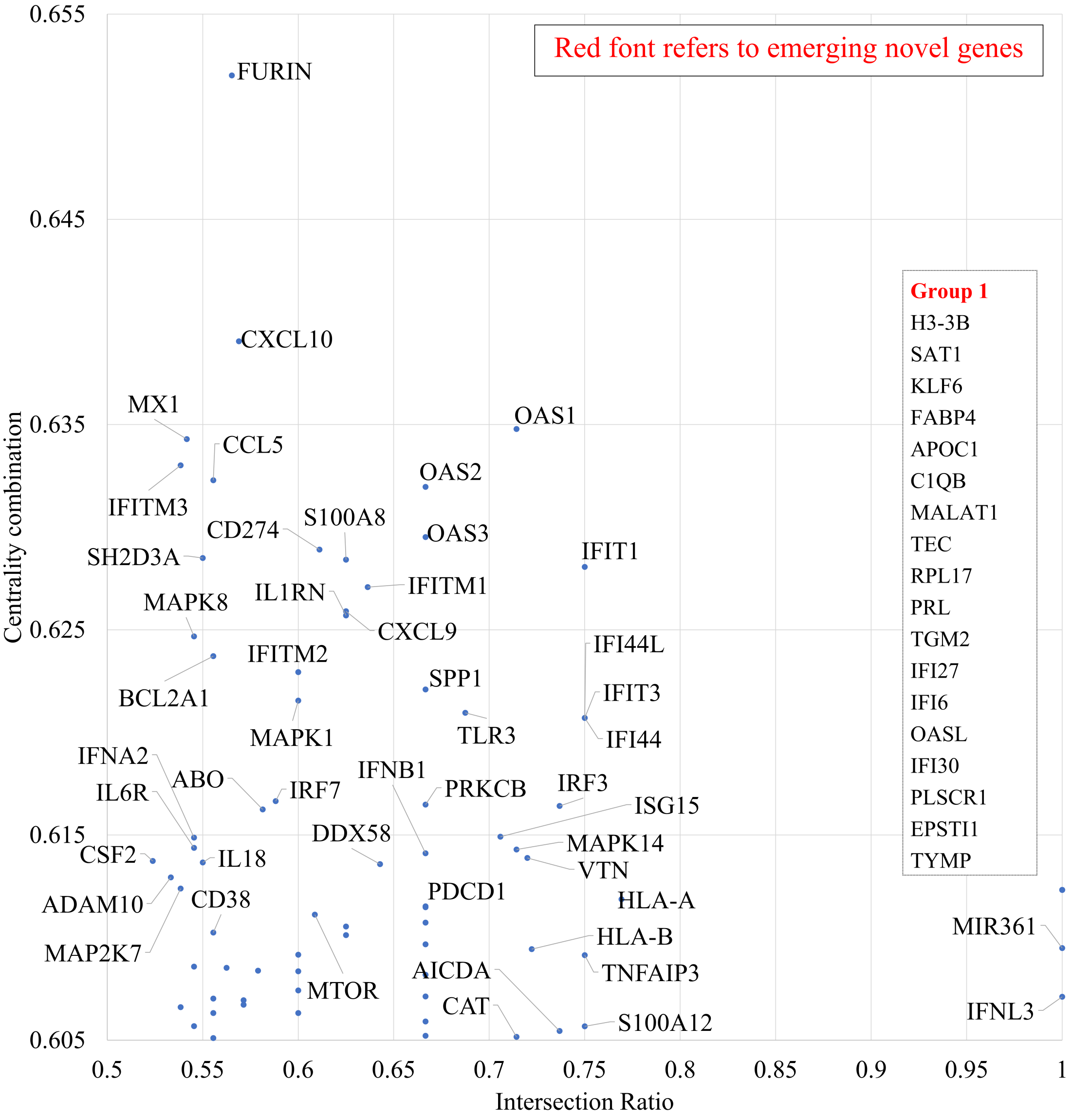


Figure 6.13 : Gene map from COVID-19 research – detailed partial view

2020; Parkinson et al., 2020; Tan et al., 2021; Xiong et al., 2020). Among those studies, a paper published in *Nature Biotechnology* identified that critical COVID-19 cases had shown stronger interactions between epithelial and immune cells which includes inflammatory macrophages expressing CXCL10 (Chua et al., 2020). Bermejo-Martin et al. (2020) reported that viral RNA load in plasma correlates with higher chemokines levels, including *CXCL10* and *CCL2*. Xiong et al. (2020) also indicated the association between COVID-19 pathogenesis and excessive cytokine release, including *CXCL10/IP-10*.

OAS1, *OAS2*, *OAS3*, *IFIT1*, *IFIT3*, *IFI44*, *IFI44L* and *IFITM1*: Current COVID-19 genetic studies incline to analyse those genes together. In a paper published on *Nature*, Pairo-Castineira et al. (2021) identified a significant genetic variant rs10735079 associated with critical illness of COVID-19 in the gene cluster encodes *OAS1*, *OAS2*, and *OAS3*. Interestingly, recent work on archaic human (Neandertal) DNA has identified an additional haplotype in Chromosome 12 containing *OAS1*, *OAS2*, and *OAS3* that protects against severe COVID-19 (Zeberg and Paabo, 2020). Klaassen et al. (2020) identified six genetic variants in innate immunity-related genes, including *OAS1* (p.Arg130His), which might have predictive value for COVID-19 infection. Besides, *IRF9*, *IFIT1*, *IFITM1*, *MX1*, *OAS2*, *OAS3*, *IFI44* and *IFI44L* were found to be upregulated in the COVID-19 infected normal human bronchial epithelial cells (Vishnubalaji et al., 2020). Similarly, Prasad et al. (2020) also found that some interferon-stimulated genes can be considered potential candidates for drug targets in COVID-19 treatment. Those genes include *IFIT1*, *IFITM1*, *IRF7*, *ISG*, *MX1*, and *OAS2*. Shi et al. (2021) showed that COVID-19 infections are generally restricted by *IFITM1*, *IFITM2* and *IFITM3* using gain- and loss-of-function approaches.

ISG15: The findings of *ISG15* are mostly related to the papain-like proteases (PLpro) encoded by the SARS-CoV-2 coronavirus. A paper published in *Nature* revealed a unique preference of SARS-CoV-2 coronavirus of cleaving ubiquitin-like interferon-stimulated gene 15 protein (*ISG15*), which is different from SARS-CoV (Shin et al., 2020). This study also indicated that SARS-CoV-2 papain-like protease contributes to the cleavage of ISG15 from interferon responsive factor 3 (*IRF3*) and attenuates type I interferon responses. Klemm et al. (2020) specified that the structure of the SARS-CoV-2 PLpro reveals that S1 ubiquitin-binding site is responsible for high *ISG15* activity, while the S2 binding site provides Lys48 chain specificity and cleavage efficiency. Freitas et al. (2020) evaluated the biochemical activity of SARS-CoV-2 PLpro and *ISG15* with its counterparts in MERS-CoV and SARS-CoV. They indicated that naphthalene-based PLpro inhibitors are shown to be effective at halting SARS-CoV-2 PLpro activity as well as SARS-CoV-2 replication.

TNFAIP3: Protein and protein interaction analysis from Islam et al. (2020c) indicated that *TNFAIP3* is one of the key hub genes that have good binding affinities with repurposed COVID-19 drug candidates, which includes dabrafenib, radicicol and AT-7519. Li et al. (2021) observed the bimodal gene expression of *TNFAIP3* in various immune cells from severely infected COVID-19 patients.

The overlapping genes in Group 1 are investigated in a single paper (Shaath et al., 2020). They identified neutrophils (*IFITM2*, *IFITM1*, *H3-H3B*, *SAT1* and *S100A8*) and macrophage cluster-1 (*CCL8*, *CCL3*, *CCL2*, *KLF6* and *SPP1*) as the main immune cell subsets associated with severe COVID-19 cases. They also found some upstream regulators (*IFNG*, *PRL*, *TLR7*, *PRL*, *TGM2*, *TLR9*, *IL1B*, *TNF*, *NFKB*, *IL1A*, *STAT3*, *CCL5*) were enriched in bronchoalveolar lavage cells in severe COVID-19 cases compared to the mild cases. Besides, common genes found in both mild and severe COVID-19 cases (*IFI27*, *IFITM3*, *IFI6*, *IFIT3*, *MX1*, *IFIT1*, *OASL*, *IFI30*, *OAS1*) and only in severe cases (*S100A8*, *IFI44*, *IFI44L*, *CXCL8*, *CCR1*, *PLSCR1*, *EPSTI1*, *FPR1*, *OAS2*, *OAS3*, *IL1RN*, *TYMP*, *BCL2A1*) are reported as well.

MIR361: miRNAs are essential regulators of viral pathogenesis, particularly among RNA viruses. Pierce et al. (2020) verified the biological plausibility of the predicted miRNA-target RNA interactions, in which miRNA361 binds to the SARS-CoV-2 *IFN*-α 3'-UTR. Li et al. (2021) showed that hsa-miR-361-3P is one of the top upregulated or downregulated genes in COVID-19 patients compared to the healthy controls.

IFNL3: Stevenson et al. (2021) claimed *IFNL3* as one of the predictive markers for severe symptoms of COVID-19 based on an analysis of serum chemokines and cytokines from COVID-19 patents, while another pharmacogenomic study did not find the potential of *IFNL3* in modifying treatments (Sugiyama et al., 2020). Instead, they identified *CYP2D6* and *CYP2C19* as likely best targets for treatment modification, especially for ondansetron, oxycodone, and clopidogrel.

6.5.5 Case summary

Due to the large number of studies done since the beginning of the COVID-19 pandemic, keeping up with the rapid changes is challenging for scientific researchers and policymakers. This empirical study comprehensively analyses COVID-19 genetic research papers published during the pandemic with traditional and intelligent bibliometric approaches. Our bioentity network analysis presents a gene coordinate system with every gene located by its network importance vs. specificity to COVID-19. From the node network importance perspective, we identify *ACE2*, *IL6*, *TMPRSS2*, and *TNF* to be frequently highlighted in the COVID-19 genetic studies. Combining the genes' network importance and specificity to COVID-19, we have used this method to identify candidate novel genes, such as *FURIN*, *CXCL10*, *OAS1*, *OAS2*, *OAS3*, *ISG15*, etc., as emerging genes that may require further research and attention in this field.

This study provides a suite of intelligent bibliometric tools for biomedical researchers to conduct medical knowledge discovery. For example, it could profile the research landscape of a given medical case with identified associations between genes and diseases. Compared with other bibliometrics conducted on COVID-19, this case study provides a systematic and adaptable research framework to profile the research landscape and exploit disease genetics-related knowledge from the literature. Additionally, this study specifically focused on COVID-19 genetic research. It targeted a set of frequently highlighted genes and emerging genes on COVID-19, which could be the clue for COVID-19 prevention and treatment. The results of this study could benefit 1) clinical researchers with longitudinal analyses on COVID-19 genetic research, and 2) policymakers with insights into recognising potential threats from COVID-19 and providing pre-emptive actions on national strategies, science policy, and public health and administration for gene-level prevention and treatments.

There are also some limitations to be addressed in our current study. From the methodology perspective, we designed a purely data-driven method to identify those primary genes and potentially emerging novel genes from literature data. However, we have not touched on some other valuable knowledge sources, such as clinical trials and curated medical knowledge databases, which also have the potential to facilitate novel knowledge discovery in this field. From the result validation perspective, we employ evidence from the literature to interpret and support our findings with assistance from our medical experts. Nonetheless, the validation is primarily qualitative without detailed measuring metrics. In future studies, we will put our efforts into 1) involving more data and information sources to improve the completeness and comprehensiveness of our method; and 2) designing a systematic validating method from both quantitative and qualitative perspectives.

The COVID-19 pandemic remains a worldwide threat to human health, as well as to the global economy and political landscape. Controlling the pandemic and improving prevention and treatment are top global priorities. Despite the welcome rollout of several different vaccines, there is still substantial knowledge about this virus waiting to be uncovered and explained, especially related to dealing with the ongoing mutations of this virus and optimising treatments for different patient groups. Our study provides a suite of novel bibliometrics-based tools for biomedical researchers to utilise to highlight the rapid changes in this field rapidly and may help accelerate our process of learning about this illness. Moreover, our research results could be of use to policymakers to help produce research priorities designed to mitigate future threats from SARS-CoV-2 and similar viruses, as well as to help plan post-COVID-19 health interventions.

6.6 COVID-19 Knowledge Hierarchy and Retrieval

Since the COVID-19 outbreak in 2020, scientists around the globe have published more than 200,000 research papers on the nature of this virus and ways to help mitigate its negative impacts. While beneficial, the sheer volume of information published has caused an information crisis (Chahrour et al., 2020; Xie et al., 2020). Apart from the problem of misinformation and rumours, the overwhelming influx of research papers has resulted in severe information overload, challenging scientists, healthcare professionals, and the general public to 1) keep up with the rapid accumulation of new knowledge, 2) accurately and comprehensively obtain knowledge on specific topics; and 3) understand the new knowledge emerging (Hossain, 2020; Wise et al., 2020; Yu et al., 2021). Although several open literature datasets and search tools are available online (Trewartha et al., 2020; Zhang et al., 2020a), a comprehensive framework incorporating effective analytical tools is still missing to help scientists meet these challenges. What is needed is a solution that can help researchers answer the following three case research questions:

- Case research question 1 (CRQ1): What are the key research topics in the emerging COVID-19 knowledge system?

- Case research question 2 (CRQ2): How can we retrieve established knowledge for specific COVID-19 research topics?
- Case research question 3 (CRQ3): How do we understand and utilise the retrieved knowledge?

Previous efforts to answer these questions mostly consist of COVID-19 topic analysis (Colavizza et al., 2021; Pourhatami et al., 2021; Tran et al., 2020; Zhang et al., 2020a), literature-based discovery studies (Wise et al., 2020; Zhang et al., 2021d; Yu et al., 2021), and literature search tools (Chen et al., 2021; Trewartha et al., 2020). A common approach in current topic studies is to apply co-word clustering (Pourhatami et al., 2021) or topic modelling (Colavizza et al., 2021; Tran et al., 2020) to the COVID-19 literature. Such studies have helped to track newly emerging knowledge but have often overlooked the relationships between new evidence and previously established coronavirus knowledge. For example, what are the similarities and differences between the diagnosing criteria, treatments, and social responses for SARS and SARS-CoV-2? In such cases, established knowledge can be a significant means of discovering and synthesising new knowledge (Haghani and Bliemer, 2020; Haghani and Varamini, 2021; Hu et al., 2021; Petrosillo et al., 2020). In addition, current literature-based discovery studies are conducted on macro levels and do not focus on specific knowledge domains to discover targeted knowledge for people pursuing different interests (Wise et al., 2020; Yu et al., 2021). For example, the interests of virologists and pathologists lie in tracing the spike protein mutations of SARS-CoV-2 (Pairo-Castineira et al., 2021; Starr et al., 2021), while clinical doctors are eager to follow the latest progress in diagnosis and treatment (Felsenstein et al., 2020; Merrill et al., 2020). For this reason, combining topic analysis and literature-based discovery approaches is a promising way to fill these two gaps. Further, few of the available COVID-19 knowledge search tools provide visualisations or other efficient ways to assist users in understanding the retrieved results (Trewartha et al., 2020; Zhang et al., 2020a). A concise and appropriate visualisation could save a huge amount of time in finding the right papers to follow or in narrowing down their search scope. Aiming to fill these research gaps, we developed a research framework that provides a systematic solution to answering the three cited research questions.

CRQ1 is answered via a strategy that involves two topic extraction methods, Principal Component Decomposition (PCD) (Watts and Porter, 1999; Watts et al., 1999) and hierarchical topic tree (HTT) analysis (Wu and Zhang, 2021). This approach identifies research topics from research papers and yields a bird's eye view of COVID-19's knowledge system. Compared to other topic extraction approaches like K-means text clustering (Wartena and Brussee, 2008) or topic modelling (Blei et al., 2003; Yau et al., 2014), PCD can generate robust document clustering results without introducing any randomisation processes. HTT, on the other hand, profiles the research topics in a hierarchical structure to highlight the differences and inner connections between topics. The two topic profiling approaches complement each other in generating both macro-level knowledge overviews and meso-level knowledge hierarchies. With the COVID-19 topics identified, we further developed a document retrieval approach based on a knowledge model that supports topic-specific document retrieval. The approach parses the entire PubMed database and links each identified topic with semantically similar pre-COVID literature in PubMed. In this way, new knowledge is linked to foundational knowledge. CRQ2 is answered by combining the topic analysis with the results of the knowledge model. Targeting CRQ3, the focus is on hierarchy, a specific dimension of knowledge composition, where the hierarchical structures of a topic's knowledge body are profiled and visualised. This helps researchers to efficiently understand the knowledge structures in the retrieved papers, further supporting knowledge discovery. All in all, this study blends multiple data-driven bibliometric approaches to reveal insights into COVID-19 knowledge deconstruction, effective retrieval, and understanding. It is in line with the direction what we called "intelligent bibliometrics" (Zhang et al., 2020b), targeting problems in science, technology, and innovation (ST&I) studies and highlighting the development of computational models incorporating artificial intelligence and data science techniques with bibliometric indicators. Despite a specific focus on COVID-19 knowledge in this chapter, the proposed framework is adaptable for knowledge deconstruction and retrieval in broad domains and scenarios.

To conduct our case study, we collected the abstracts of 127,971 research papers published in 2020 and 2021 from the PubMed database. Feeding those papers into the PCD analysis, we generated 35 PCD topics and revealed how the emphasis on different topics changed over different periods. Initially, the focus was on the epidemiological and clinical characteristics of the virus. However, over time the emphasis changed to the impacts of COVID-19 on different societies. The HTT results divided the explored knowledge into a clinical branch and a public health branch. The clinical branch focuses on discovering COVID-19-associated clinical factors and treatments. The public health branch addresses six particular public health concerns. Additionally, we constructed a knowledge model based on the most popular PCD topic of vaccination and ran a global search on PubMed for records published before 2020 to retrieve the knowledge foundations of this topic, resulting in 92,286 retrieved papers. Lastly, we used HTT to visualise the knowledge structures of the retrieved results. The HTT results highlighted multiple vaccination-related disciplines, including immunology, molecular biology, virology, etc. From the branches of those disciplines, we identified four future promising research directions: monoclonal antibody treatments, vaccination in diabetic patients, vaccination effectiveness in SARS-CoV-2 antigenic drift, and vaccination-related allergic sensitisation. We empirically evaluated the results by matching evidence identified from the lit-

Table 6.4 : Stepwise results of term clumping

Step	Description	# Terms
1	Raw terms retrieved with NLP	1,603,542
	Remove terms starting/ending with non-alphabetic characters	
	Remove common terms in scientific articles, e.g., “research framework”	
2	Remove meaningless terms, e.g., pronouns, prepositions, and conjunctions	1,367,374
	Consolidate synonyms based on expert knowledge, e.g., “COVID-19” and “COVID”	
	Consolidate terms with the same stem, e.g., “severe patient” and “severe patients”	
3	Filter terms with term frequency above 10	33,281

erature and identified research evidence in the latest studies. This empirical case not only demonstrates the reliability of our method but also derives insights to support potential COVID-related R&D and strategic management for funding agencies, individual researchers, and institutions.

6.6.1 Data collection and pre-processing

For topic analysis, the search process returned 127,971 relevant research papers from 1 January 2020 – 1 January 2022 as of early March 2022. Then we further applied the natural language processing function of VantagePoint to extract topic terms from titles and abstracts. The list of extracted terms was cleaned to remove stop words, consolidate similar terms, and eliminate all terms appearing in less than ten records (Zhang et al., 2014b). The term clumping process and stepwise results are given in Table 6.4.

6.6.2 Data overview

Trends in COVID-19 publications can help us glimpse the response patterns of the scientific community given catastrophic events. To capture such trends, we first applied a descriptive bibliographic analysis to profile the external characteristics of COVID-19 studies regarding the monthly growth, institution ranking, and

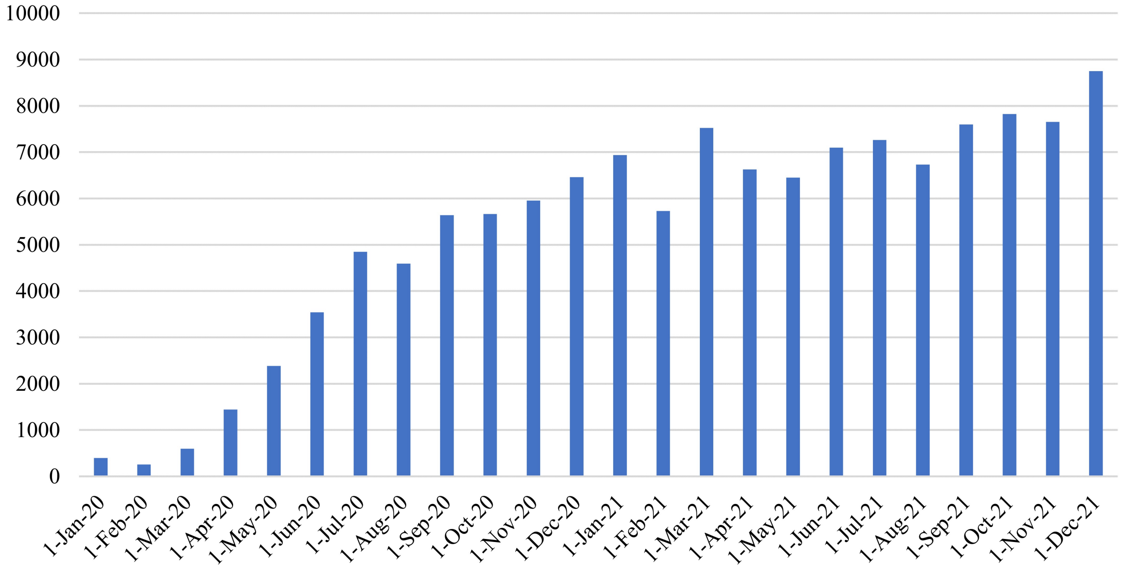


Figure 6.14 : Monthly increasing trend of COVID-19 research papers

geographical distribution.

Figure 6.14 illustrates the basic monthly numbers of COVID-19 research papers. Early in 2020, these numbers were rapidly increasing, but by 2021, they had become relatively steady. The burst of COVID-19 publications can easily be attributed to the disruptiveness and uncertainty that COVID-19 has brought to previously established knowledge systems (Zhang et al., 2020a), which attracts research attention from massive new researchers (Wagner et al., 2022). However, the slowing increase might be due to multiple reasons: Is it due to research capacity limitations (e.g., journals, review speed, funding, etc.)? Or does it indicate that newly discovered knowledge is converging to a new stage? Will there be a decay period following? These possibilities only trigger more research questions to be answered and examined in future studies.

Figures 6.15 and 6.16 respectively profile the global distribution and ranking changes of COVID-19 research papers among worldwide countries/regions. Figure 6.17 lists the top 20 productive institutions. In terms of the absolute number of papers published at the national level in Figure 6.15, the United States and China unsurprisingly hold leading positions, followed by Italy, India, Germany, Canada, etc. From a retrospective view, the ranking changes in Figure 6.16 intuitively indicate the association between productivity and local epidemic severity (Wagner et al., 2022). For example, China took first place in the initial few months because it was the first victim of COVID-19 and had first-hand access to massive numbers of clinical cases. However, the US soon overtook China and has held the first position since the middle of 2020. This may be because the US has solid research strength, but it could also be the result of how severely the COVID-19 pandemic hit the US (Burki, 2020). Italy maintained third place for a long time from March 2020 as it became the European COVID-19 epicentre, suffering high cases and mortality rates (Remuzzi and Remuzzi, 2020). Following a sharp decrease in March 2020, which could be a result of the 21-day nationwide lockdown at that time, India has remained high in the ranking list. The pandemic hit India severely, and multiple SARS-CoV-2 variants have emerged there (Bernal et al., 2021).

Diving into the institution level, we found that, compared to the earlier China-led trends in COVID-19 research (Fry et al., 2020), the momentum for US institutions to lead in this domain has continued to grow (Wu et al., 2021b). This indicates that even though China has published a substantial volume of papers, individual Chinese universities and research institutions have not demonstrated equal strength in competition with their global counterparts, particularly those from the States.

6.6.3 COVID-19 study overall research landscape

PCD is essentially a robust and reproducible variant of principal components analysis (PCA) that groups scientific documents according to their textual features (Watts et al., 1999; Watts and Porter, 1999). Compared to the original PCA, PCD automatically decides the number of factors (derived PCA groupings) by minimising

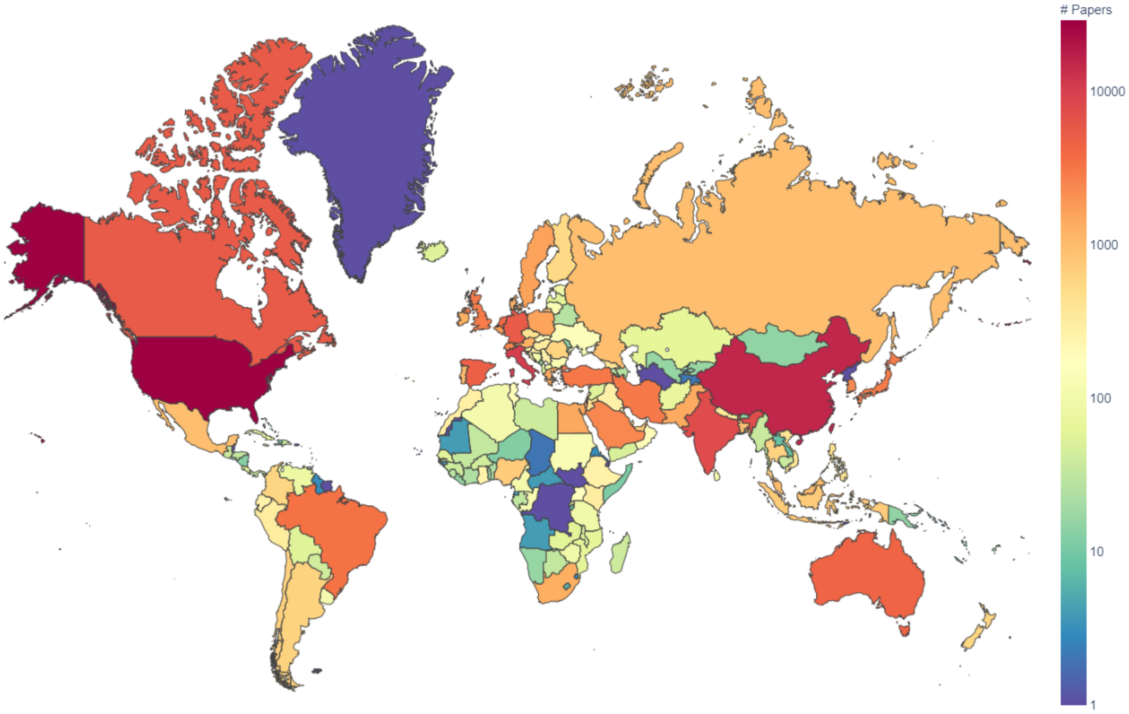


Figure 6.15 : The geographical distribution of COVID-19 papers

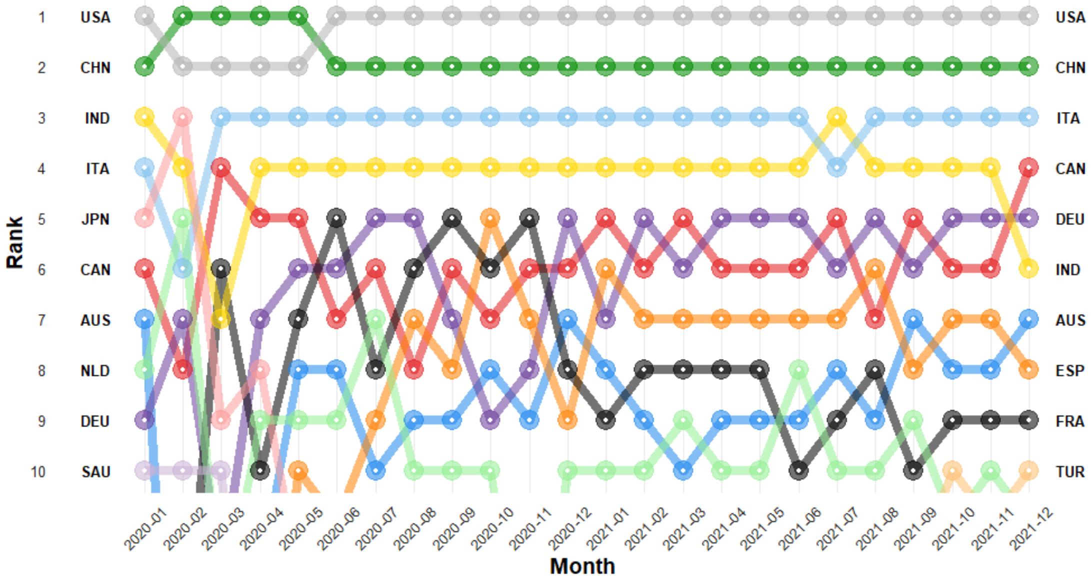


Figure 6.16 : The ranking changes of countries

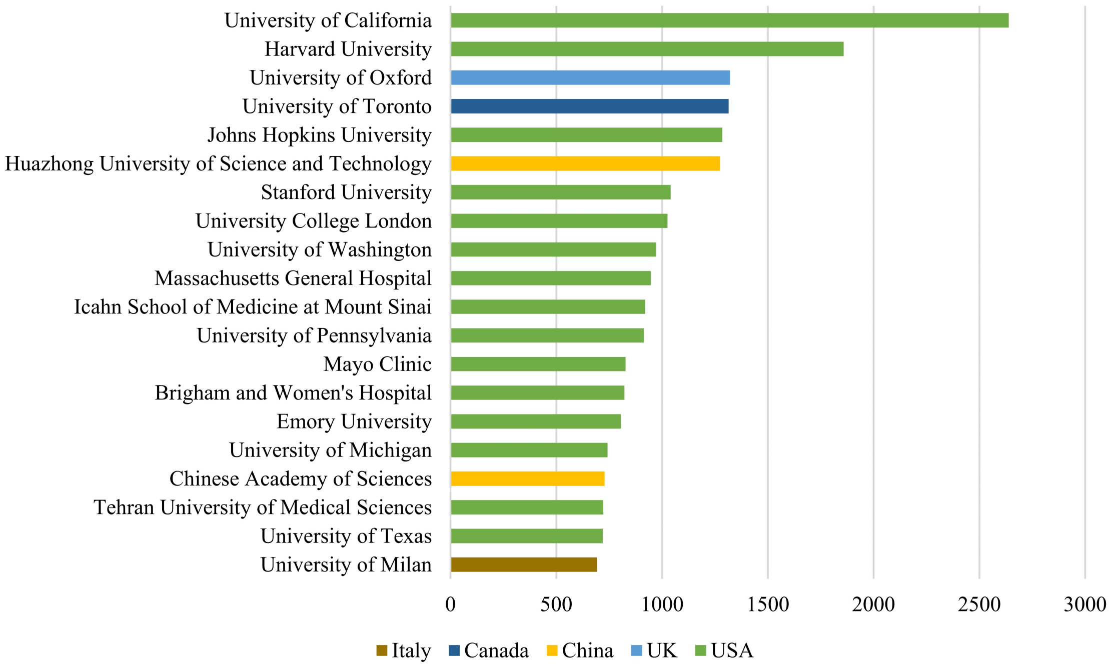


Figure 6.17 : Top twenty prolific research institutions

the entropy and maximising the cohesiveness of the derived factor groups. In our case, we exploited processed terms extracted from the titles and abstracts as document feature vectors. We then ran PCD on the document-term matrix to decide the factors automatically. The retained factors were deemed to be PCD topics.

Feeding the extracted topic terms into the PCD analysis, we distilled 35 research topics. Further, we plotted a topic correlation map in Figure 6.18, with each bubble representing a PCD research topic and the size denoting its associated paper count, and the links denoting a cosine correlation above 0.5 (Salton and McGill, 1986). The correlation map of the 35 topics highlights a core topic cluster in the middle, representing a set of clinical manifestations and hospitalisation factors of COVID-19. The other scattered topics cover a broad range, including public health, education, economics, etc. More details are provided on those topics in the following analysis.

The monthly ranking changes of the top ten topics are given in Figure 6.19,

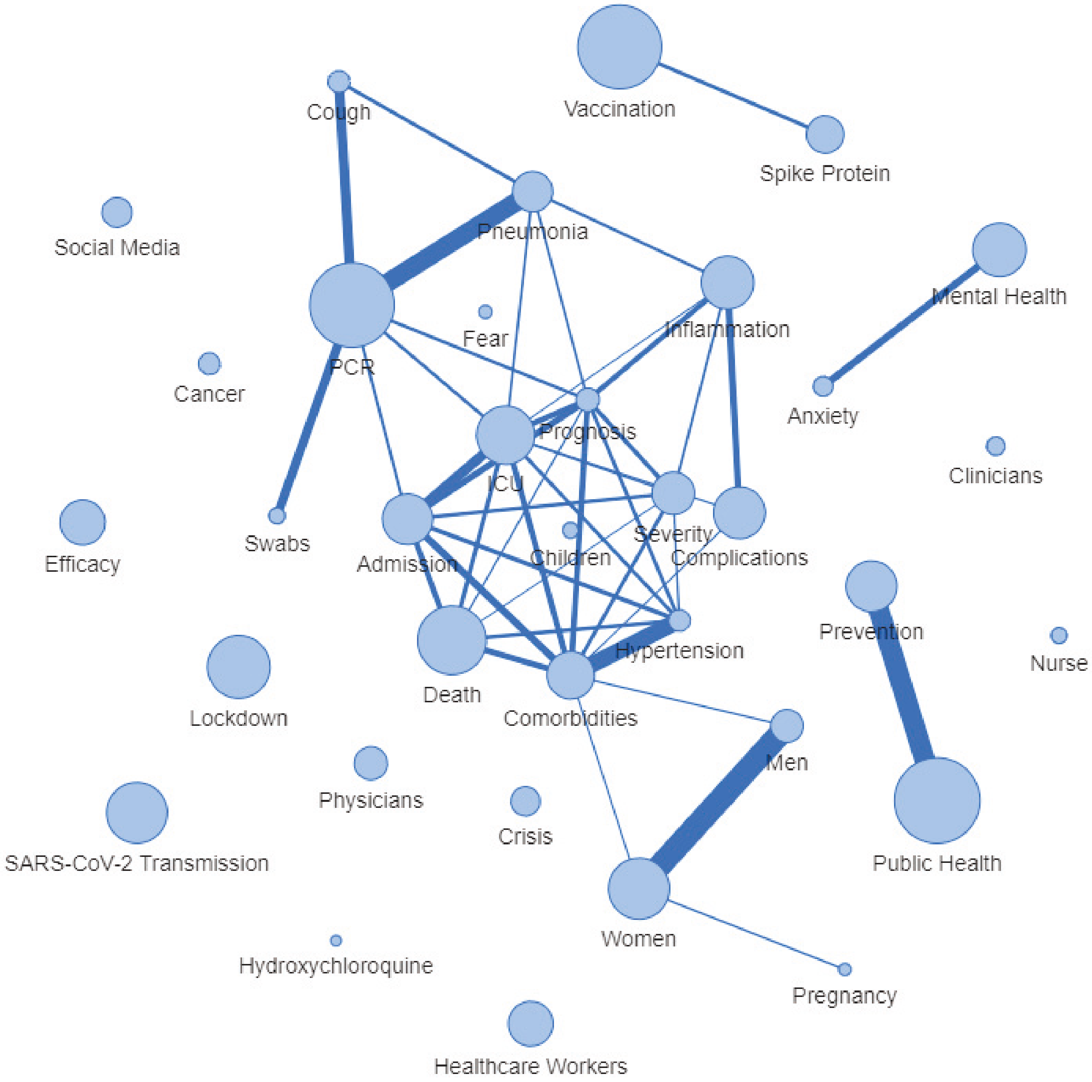


Figure 6.18 : The distribution and cross-correlation of PCD topics

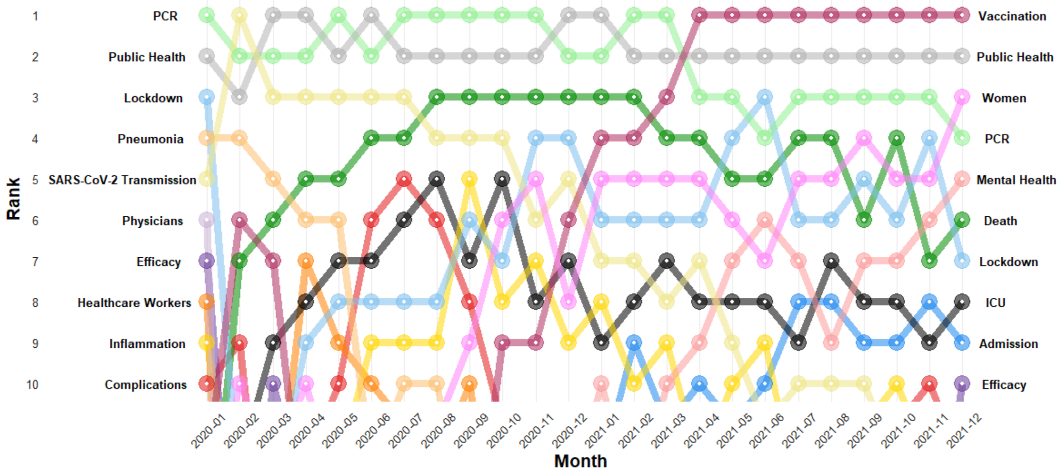


Figure 6.19 : Monthly increasing trend of PCD topics

indicating different stages of COVID-19 research. Among these topics, the rankings of PCR and Public Health maintain the top, while other topics show significant fluctuating trends.

At the beginning of the COVID-19 breakout in Wuhan, the PCD topics Pneumonia and SARS-CoV-2 Transmission attracted massive attention, as first-hand clinical and epidemiological investigations were urgently needed to improve COVID treatments and control its transmission (Huang et al., 2020a; Li et al., 2020b; Lu et al., 2020; Wang et al., 2020b). In such studies and following clinical trials, the gender difference is an essential analysing factor as indicated by the continuing ranking rise of PCD topics in Women and Men. Additional attention was put on the female group due to studies on the vulnerabilities of pregnant women or women at lactating ages (Chen et al., 2020). As COVID-19 turned from regional transmissions into a global pandemic, scientists started to look into the social impacts of COVID-19 as illustrated by the rise of topics Lockdown (Ruktanonchai et al., 2020; Shepherd et al., 2021) and Mental Health. The former broadly covers the social impacts of lockdown measures on healthcare services (Shepherd et al., 2021), economy (Bonaccorsi et al., 2020), education (Engzell et al., 2021), and environment (Venter et al., 2020), etc.; The latter topic discusses mental health issues among the general public (Brülhart et al., 2021; Shi et al., 2020) and healthcare workers (Lai et al., 2020). As the COVID pandemic progressed, the rankings of *Death* and *ICU* topics decreased relatively steadily.

Notably, the change in Vaccination-related papers illustrates two waves of vaccine studies. The first wave appeared at the beginning of the COVID-19 breakout and peaked in February 2020. These early-stage papers mainly focus on reviewing past coronavirus vaccines, calling for rapid vaccine development procedures, and proposing possible vaccine development approaches (Ahmed et al., 2020; Ahn et al., 2020; Pang et al., 2020; Prompetchara et al., 2020). With the advent of multiple available vaccines, the next wave emerged in the third quarter of 2020 and continued to rise. In addition to the massive numbers of basic medicine and clinical trial studies around the safety and efficacy of those vaccines (Polack et al., 2020; Thomas et al., 2021; Xia et al., 2020), the rollout of vaccines also triggers researchers' concerns about the social implications, including the vaccine hesitancy phenomena (Biswas et al., 2021; Dror et al., 2020), vaccine allocation strategies (Duch et al., 2021) and vaccination incentives (Campos-Mercade et al., 2021; Dai et al., 2021). As vaccination offers one of the most effective measures in preventing COVID-19, we will demonstrate how we used our knowledge model to retrieve historical knowledge of vaccination studies in the next section.

The PCD results yield a flat view of the COVID-19 research landscape. To dive further into the hierarchy of COVID-19 knowledge and obtain more details about research topics, we ran the HTT algorithm and constructed a co-term network using terms with a frequency above ten. The characteristics of our input network are given in Table 6.5. The generated HTT map is shown in Figure 6.18, with the tree trimmed to only show nontrivial branches. The node size indicates the prevalence

Table 6.5 : The characteristics of COVID-19 term co-occurrence network

	Number	**Weight**			
		Max	**Min**	**Avg.**	**Std.**
Node	3,281	14,914	10	45.448	237.11
Edge	7,504,641	3,618	1	1.568	6.246
Average degree			450.98		

of the topic, and the edge thickness denotes the co-occurring strength of the two linked topics.

The hierarchical topic tree shows more detail on every individual topic. The HTT map covers most PCD topics and arranges them hierarchically according to their topological importance in the co-occurrence network. This empirical evidence, discovered through PCD and HTT, coincides with knowledge manually identified from the literature, which can endorse the method's practical effectiveness. Mortality and Public health are two HTT topics that hold the top positions in this hierarchy and represent the two major research branches: clinical and public health studies.

The clinical branch spans efforts to uncover the associated clinical factors of COVID-19 and find effective therapies. As illustrated in Figure 6.20, such explored clinical factors include gender – women, men (Jin et al., 2020), complications – inflammation, cytokine storm (Jose and Manuel, 2020), thrombosis (Levi et al., 2020), age – elderly (Liu et al., 2020a), and comorbidities – diabetes (Muniyappa and Gubbi, 2020), hypertension (Fang et al., 2020). The treatments studied in clinical case reports and clinical trials consist of mechanical ventilation, hydroxychloroquine (Gautret et al., 2020), remdesivir (Beigel et al., 2020), and bamlanivimab (Gottlieb et al., 2021), etc. In summary, this branch contains various clinical case reports

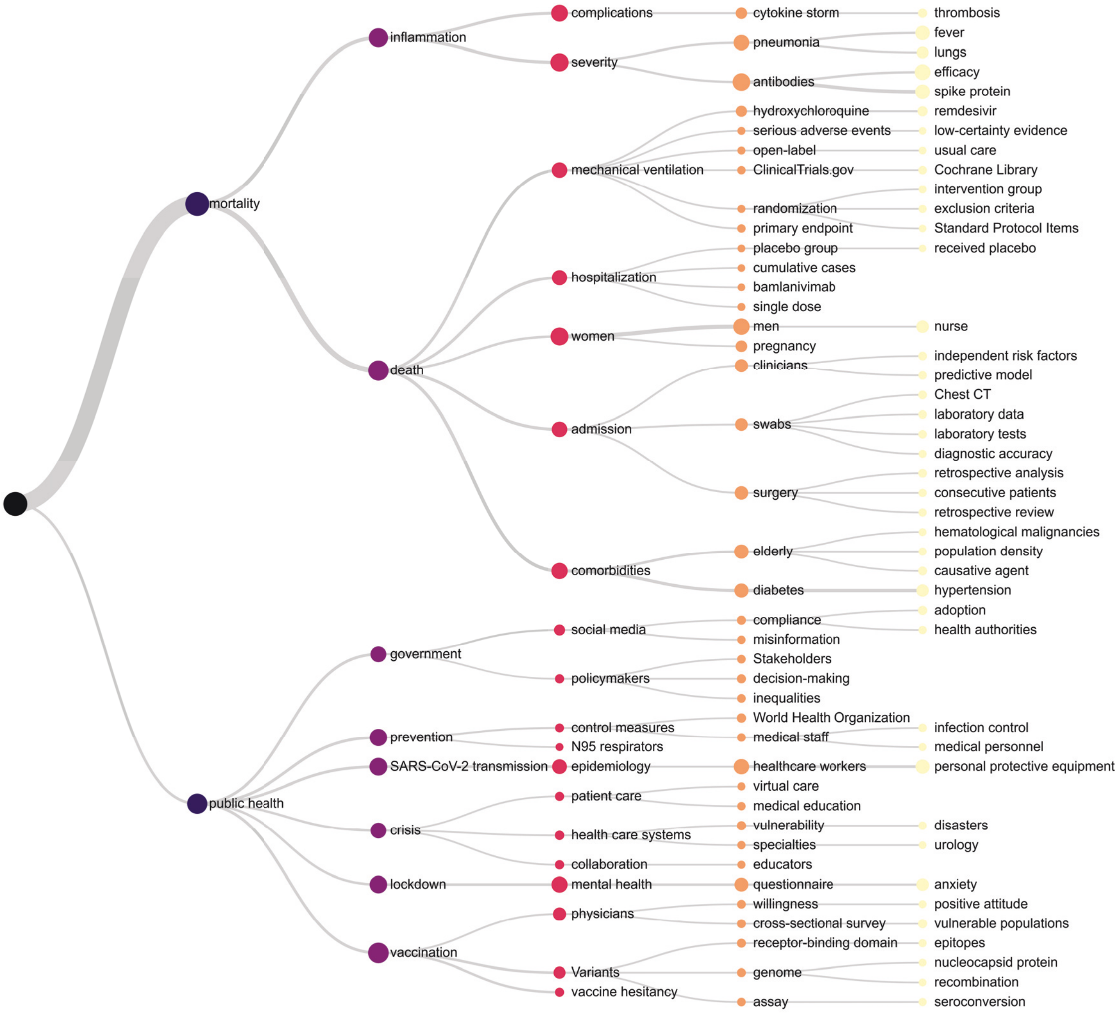


Figure 6.20 : The hierarchical knowledge landscape of COVID-19 literature

and clinical trial studies devoted to revealing the associated factors of COVID-19 severity/mortality/prognosis and finding effective treatments.

For the public health branch, six subtopics are highlighted as follows.

i. Government: This branch discusses the role of government in fighting COVID-19. One of its subordinate branches points to policymakers, and, within this, handling inequalities in different groups of people has become a notable concern in the policy-making process (Chu et al., 2020; Garcia et al., 2021). The other subordinate branch of social media indicates the role of social media as a double-edged sword for governments regarding information dissemination and evaluation (Islam et al., 2020b; Li et al., 2020a; Tsao et al., 2021), given the presence of misinformation.

ii. Prevention: This set of HTT topics reflects some of the major explorations of COVID -19 prevention: Face mask production and use issues (Brooks et al., 2021; Long et al., 2020; Wu et al., 2020); identifying effective control measures (Nussbaumer-Streit et al., 2020; Wang et al., 2020c); and how to protect frontline healthcare workers (Ding et al., 2020; Islam et al., 2020a).

iii. SARS-CoV-2 transmission: This set of topics explores the epidemiological characteristics of COVID-19, among which the transmission between healthcare workers (Bergwerk et al., 2021; Sikkema et al., 2020) and the use of personal protective equipment (Mick and Murphy, 2020) have attracted substantial research attention.

iv. Crisis: This topic set discusses the implications of COVID-19 on healthcare systems (Liu et al., 2020b; Spinelli and Pellino, 2020) and medical education (Hall et al., 2020).

v. Lockdown: As one of the strictest restrictions, lockdown measures were frequently explored for their associations with mental health issues in the general public and medical staff (Wang et al., 2020a; Williams et al., 2020).

vi. Vaccination: Apart from one branch highlighting the basic biomedical studies for vaccine development (Polack et al., 2020; Xia et al., 2020), the other two branches respectively address attention to vaccination in healthcare workers (Bergwerk et al., 2021) and the vaccination hesitancy issue (Dai et al., 2021; Machingaidze and Wiysonge, 2021).

6.6.4 The HTT result of COVID-19 vaccination studies

This section demonstrates the utility of our knowledge model and HTT approaches in retrieving historical knowledge from the entire PubMed database, using the most prominent PCD research topic, *Vaccination*, as an example. We extracted 15,967 papers related to this topic and calculated the TF-IDF values of all the extracted terms of those papers. Then, a knowledge model was built up with its top 50 and bottom 50 term stems[9]. Further, we ran the knowledge model-based document retrieval method (Cassidy, 2020) over the entire PubMed database and retrieved 92,286 historical records out of the COVID dataset. We removed records containing the /vaccin/ stem and empirically set the cosine similarity retrieving threshold to 0.25. The next section demonstrates how to deconstruct the retrieved results and mine the knowledge structures.

We further mapped the 92,286 records to Open Academic Graph (OAG) and retrieved 89,951 records with the field of study (FoS) information. The FoS is essentially constituted by Wikipedia entities assigned to scholarly papers via a Naïve Bayes-based tagging process (Shen et al., 2018). OAG originates from Microsoft academic graph (MAG) and currently covers more than 240 million publications. Compared to scientific terms extracted from titles and abstracts the FoS system adopts Wikipedia entity entries as the topics of each paper, which we consider more suitable for representing established knowledge foundations. To efficiently understand and visualise the knowledge in the search results, we constructed the FoS co-occurrence network of the 89,951 records and ran our HTT algorithm over it. The detail of the constructed network is given in Table 6.6.

We trimmed this HTT to retain the main body of knowledge. This is presented in Figure 6.21, yielding a hierarchical overview of the search results. Immunology is the

[9] The knowledge modelling process is introduced in (Wu et al., 2023)

Table 6.6 : The characteristics of the FoS network

	Number	**Weight**			
		Max	**Min**	**Avg.**	**Std.**
Node	27,596	39,542	1	3.459	44.336
Edge	922,252	18,737	1	28,1358	427.105
Average degree			66.840		

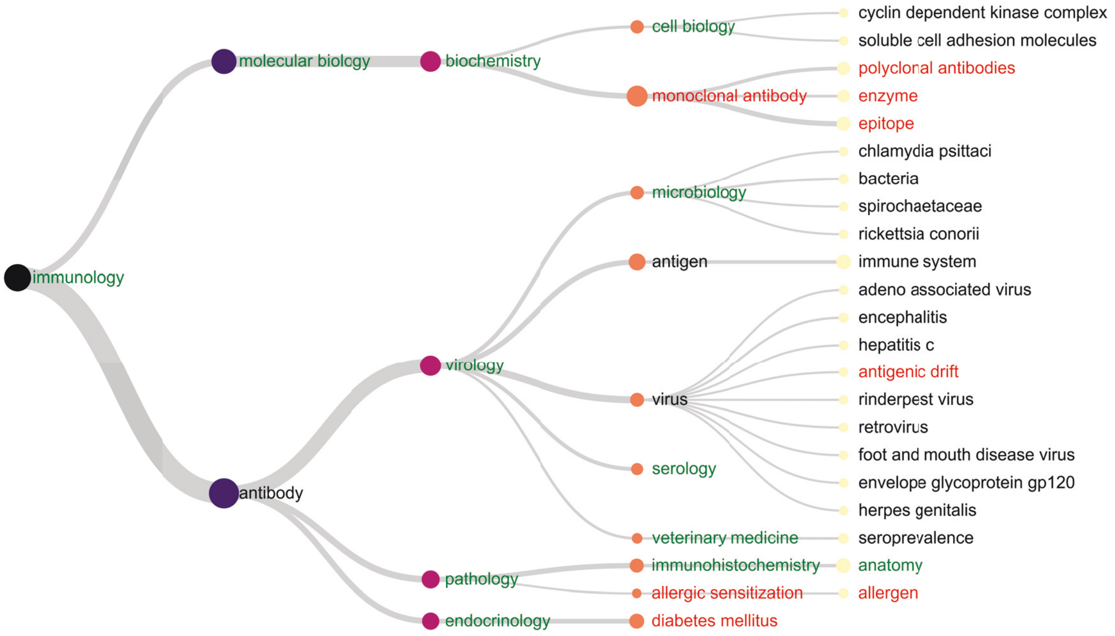


Figure 6.21 : The hierarchical knowledge landscape of retrieved results

root topic of this HTT, indicating that vaccination studies are mainly constructed based on immunology knowledge. The presented topics are primarily highlighted as discipline-level topics (green font) and entity-level topics (red font). By comparing and contrasting the historical records (regarded as the knowledge foundations) with the latest research evidence, we drew insights on four essential topics: Monoclonal antibodies, Antigenic drift, Diabetes, and Allergic sensitisation.

Then we validated our empirical results with literature-based evidence and dived into the historical papers and the newest COVID-19 research articles related to the four topics. The knowledge connections we identified from the papers are presented as follows:

- Monoclonal antibodies: This topic is positioned in the branch of molecular biology – biochemistry. Diving into this topic, we can trace many historical studies on developing monoclonal antibodies as a treatment for existing human and animal coronaviruses, including severe acute respiratory syndrome (SARS) coronavirus (Traggiai et al., 2004; Zhu et al., 2007) and bovine coronaviruses (Deregt and Babiuk, 1987; Mockett et al., 1984). Such studies can provide instructive research clues for developing novel monoclonal antibody treatments for COVID-19. With the approval of multiple monoclonal antibody treatments for COVID-19, more efforts will predictably be put into finding efficient methods of extracting and producing such monoclonal antibodies (Taylor et al., 2021).

- Antigenic drift: This topic exists in the virus branch, describing a natural phenomenon of antigen genetic mutations that also happens in the SARS-CoV-2 virus (Yuan et al., 2021). Medical experts can trace historical studies of influenza viruses (Pica et al., 2012; Yu et al., 2008) and other possibly related viruses (Coulson et al., 1985) in search results to infer and analyse the impacts of antigenic drift on vaccination implementations. The effectiveness and immune durability of various SARS-CoV-2 variants (including the Omicron subtype that is currently circulating) may need deeper exploration (Coulson et al., 1985; Koyama et al., 2020).

- Diabetes: Located in the endocrinology branch, this topic consists of historical papers clarifying the autoimmune-mediated beta-cell damage mechanisms (Van Belle et al., 2011), significant autoantigens (Wenzlau et al., 2007), and different subtypes of type 1 diabetes (Imagawa et al., 2000; Stenstrom et al., 2005). Recent studies report that two types of diabetes are associated with higher odds of COVID-19 hospital deaths (Barron et al., 2020; Holman et al., 2020), and SARS-CoV-2 infection possibly induces adverse effects on beta-cells (Apicella et al., 2020; Bornstein et al., 2020; Lim et al., 2021; Marchand et al., 2020). Consequently, vaccination in diabetic patients has become a trending topic among vaccination studies. On the one hand, many researchers have called for prioritising vaccination in diabetic patients as they are more vulnerable to COVID-19 (Pal et al., 2021; Powers et al., 2021). On the other hand, associating the knowledge from our search results with COVID vaccinations (especially for Type 1 diabetes) is worth deeper exploration because the current evidence is still limited (Boddu et al., 2020; Marchand et al., 2020).

- Allergic sensitisation: Historical studies on this topic comprehensively discuss the reactivity of immunoglobulin E in allergic reactions (Aalberse et al., 2001; Eibensteiner et al., 2000; Jenmalm et al., 2001), which can provide instructive insights for COVID-19 vaccination allergy studies (Cabanillas et al., 2020; Kounis et al., 2021).

6.6.5 Case summary

COVID-19 has brought about a global public health pandemic and an overwhelming flood of research knowledge. Aiming to efficiently discover and use the knowledge contained in the massive body of COVID-19 scientific studies, we devised a research framework that: 1) profiles the COVID-19 knowledge landscape and research topics at both the flat and hierarchical levels; 2) retrieves the foundations of knowledge related to specific topics; and 3) visualises the retrieved knowledge to support knowledge understanding and discovery. We anticipate that this research methodology and our key findings will support a) scientific researchers to quickly absorb new knowledge and identify their future study directions and b) research policymakers to make informed decisions about research funding allocations.

We exploited PCD and HTT analysis to profile the COVID-19 research landscape. The PCD results highlight 35 research hotspots and multiple research emphases over different periods. The changing trends in PCD topic rankings indicate that early COVID-19 studies focus on uncovering the clinical and epidemiology characteristics of COVID-19, while the subsequent studies throw more light on the societal impacts of the pandemic. Intriguingly, the change in PCD topic vaccination papers reflects two waves of vaccination studies – the first appearing at the start of the COVID outbreak and the second after the rollout of multiple available vaccines. The HTT results consistently reveal clinical and public health studies as two significant branches of research in this domain. Complementarily, the HTT results generate more detailed insights on 1) the clinically investigated factors associated with COVID-19 mortality/severity and effective treatments; and 2) six segments of public health concern: government, prevention, SARS-CoV-2 transmission, crisis, lockdown, and vaccination.

We ran our HTT algorithm over the search results from the knowledge model to reveal the hierarchies of topics. At the top levels of the HTT, we identified multiple significant medical disciplines, including immunology, molecular biology, virology, and so on. In addition to these disciplines, we uncovered four directions worthy of more attention in future vaccination-related studies. These are 1) monoclonal antibody treatments, 2) vaccination priority and immune responses in diabetic patients, 3) the effectiveness and durability of vaccines on various SARS-CoV-2 mutations, and 4) vaccination allergies.

There are three methodological contributions of this case study worth highlighting. Initially, incorporating PCD topic analysis and knowledge model searches provides an effective topic-based mode of knowledge retrieval. This approach first clusters research papers into research topics. Then it searches the entire PubMed dataset for the foundational knowledge on the target topic, generating a more narrowed, focused, and accurate search scope in knowledge retrieval. Additionally, our HTT results allow researchers to visualise and understand thousands of papers efficiently. The HTT can help researchers quickly clarify complex knowledge structures and identify intriguing topics by highlighting the topologically significant terms in the co-occurrence network. Last but not least, our research framework provides a paradigm for research profiling and knowledge retrieval. This methodology is adaptable to various cases and can be transferred with little cost.

From the practical perspective, this chapter profiles the knowledge landscape of COVID-19 studies both in flat (PCD) and hierarchical (HTT) manners, yielding hotspots for researchers to follow. Furthermore, the insights offered in the case study identify four intriguing vaccination-related research directions. Such insights can: 1) inspire medical researchers to conduct future studies with enriched knowledge foundations and 2) assist scientific policymakers in making informed decisions about research funding allocations.

Chapter 7

Conclusion and Further Study

This chapter concludes the thesis and presents further directions for relevant research trajectories.

7.1 Conclusion

The main contributions of this thesis are as follows:

i. **It develops a heterogeneous bioentity analysis methodology (to achieve Objective 1) for knowledge association analysis and prediction in Chapter 3.**

   The heterogeneous bioentity analysis methodology (HBAM) develops a systematic work pipeline for processing biomedical literature data, sorting heterogeneous bioentities and predicting bioentity associations. It incorporates a heterogeneous entity network construction procedure, a non-dominated sorting genetic algorithm-based scoring scheme, a bioentity2vec training model and a semantics-enhanced link prediction method to rank bioentity importance/specificity and predict unobserved emerging entity associations. A case study of atrial fibrillation reveals critical disease, gene and genetic variation bioentities associated with this disease; Further predictions uncover potential genes and genetic factors that are worth further investigation with empirical evidence. The main contributions of this work include 1) a cohesive methodology based on a combination of centrality and intersection ratio measurements for identifying diseases, chemicals, genes, and genetic variants core to disease; 2) a semantic similarity-enhanced link prediction algorithm for generating more accurate predictions of the possible associations between bioentities; 3) an adaptable and transferable framework for general use in genetic factor analysis and prediction.

ii. **It develops an adaptive hierarchical topic tree model - HTT I (to achieve Objective 2) for knowledge hierarchy extraction in Chapter 4.**

Hierarchy is an innate characteristic of scientific knowledge. The automatic construction of knowledge hierarchy from scientific textual data can support domain professionals and newcomers in developing an in-depth understanding of fine-grained knowledge components and benefit many downstream applications, including document recommendation and text classification. With the input of a term co-occurrence network from scientific literature data, the HTT-I model provides a feasible and handy approach for topic hierarchies extraction. It exploits the idea of density peak clustering to identify term nodes with high density and relatively long distances from other high-density nodes as community centroids. Then an overlapping community allocation algorithm applies to partition the rest nodes to affiliate the identified centroids. The process iterates until no density peak nodes can be found. The values of three evaluation indicators on the HTT-I model demonstrate that it can generate consistent topics and solid parent-child topic associations with reasonable information loss. The empirical studies on computer science topics profiling, AI ethic issues mining and digital transformation conceptualisation validate the model's practical effectiveness and generate fruitful research insights into the three research case themes.

iii. **It develops a non-parametric hierarchical topic extraction model -**

HTT II (to achieve Objective 2) and provides a more adaptive way to fit different real-world inputs in Chapter 5.

In Chapter 5, we proposed a refined version of the HTT-II model to improve adaptability and fit a broader range of network inputs with different degrees of clustering tendency. Still using the term co-occurrence network as the input, the HTT-II model adopts k-shell decomposition and the Louvain algorithm to partition parent and child layers of terms and terms belonging to different topics. Compared with the HTT-I model, the HTT-II model is parameter-free and embraces a different design to partition terms into parent and child topics. This design can better retain coupling knowledge and differentiate terms in parent and child topics. The results from the comparison experiment demonstrate that the HTT-II model can generate consistent topics, solid parent-child topic associations and exclusive sibling topics. The empirical study on information sciences topics profiling validates the new model's practical effectiveness and highlights five significant research directions in this discipline.

iv. **It constructs an intelligent bibliometric system and GUI (to achieve Objective 3) to integrate the proposed methods and empower future case studies in Chapter 6.**

Despite the proposed methodologies, it remains a challenge for non-technical background users to access the functions and leverage the value of scientific literature data. In Chapter 6, we developed a Python-based GUI that enables users to approach the proposed functions and perform systematic data analysis of scientific literature data. Moreover, we presented two case studies on COVID-19 literature datasets to 1) highlight core bioentities investigated in COVID-19 research, 2) profile COVID-19 research hot spots and segmented research directions, and 3) uncover the knowledge foundation for COVID-19 vaccination studies.

7.2 Social and Industry Implications

Notwithstanding the technical contributions, this thesis encompasses practical implications for both society and pertinent industry sectors. Firstly, the developed HBAM presents an adaptive methodological framework that enables researchers to mine entity associations and predict latent gene-disease links. This functionality can be of immense assistance to: 1) biomedical researchers in prioritising candidate genes for diseases, particularly rare diseases and those with unclear genetic mechanisms; and 2) clinicians in devising targeted diagnostic and therapeutic approaches.

Secondly, the developed Hierarchical Topic Tree (HTT) models, namely HTT I and HTT II, offer flexible topic hierarchy profiling methods from various perspectives. These models can be advantageous for the academic community by quantitatively capturing the research landscape and identifying cutting-edge areas. Moreover, they can facilitate librarians in rapidly comprehending the hierarchical topic profiles of document collections, thereby facilitating document classification.

Lastly, the entire thesis wraps up with the integration of an accessible software called BiblioEngine, which consolidates all the developed functions. The implementation of this system will significantly aid users without a technical background in accessing these functions and leveraging them within their respective industry sectors. Consequently, it will furnish knowledge associations and hierarchical intelligence, enabling them to gain a better understanding of their domains and attain competitive advantages.

7.3 Further Study

This thesis also has several limitations for investigation in future studies.

Regarding the HBAM, there are three directions we are heading to improve. The first comes from a technical standpoint. We emphasised the need to identify strong associations by adopting co-occurrence analysis. However, this process inevitably retrieves negative associations along with the positives because it does not recognise the causalities between entities. Embedding SAO triples could be a challenging but significant task in this area. We have conducted certain pilot studies on this trail (Zhang et al., 2021c) by incorporating word embedding techniques with SAO triples. It is foreseeable to enhance its capabilities in identifying the causalities of entities. Second, due to the lack of available rules for entity extraction, we dealt with disease entities at the granularity defined in MeSH and did not further classify disease entities into other subtypes. Yet, in the context of biomedical entity inference, the molecular mechanisms that underpin the different types of atrial fibrillation are pretty diverse. A rule-based filter could overcome this problem, which is also on our agenda. From a biomedical standpoint, there is also an issue with selecting the core entities using inference. Guilt-by-association is a prevalent hypothesis for establishing genetic associations for diseases. Thus, we instinctively filtered out the neighbours of the core entities to narrow down the model input, but it did not contribute significantly to our recall performance. Despite this limitation, however, we believe that emphasising the core genetic factors for link prediction is still a promising approach for improving performance.

We also plan a few improvements for future HTT models. First, HTT analysis is a method that reflects the knowledge component of a field. Yet, it does not generate the developing trending of topics along with time, which might be more significant and intriguing to scientists. Topic composition and hierarchies are constantly changing; Hence, we intend to build a variant of HTT that considers the temporal relationship between topics and how those research topics evolve. Second, HTT only focuses on textual data, exploiting the semantic relationships in the research literature. However, this may overlook more available data patterns, such as author collaborations, venue associations, co-citations, and citation coupling. These heterogeneous data sources have the potential to help HTT yield more accurate and explainable topic hierarchy results. In subsequent studies, we anticipate embedding more external information like author collaborations, publication types, and geographical locations to build a more sophisticated hierarchical topic model that incorporates external features. Last, the hard partition of communities will inevitably result in coupled terms in different research topics. To generate more comprehendible and informative hierarchical topic results, we plan to equip the HTT-II model with overlapping community detection approaches and enable it with the capability of revealing topic overlaps.

Bibliography

Aalberse, R. C., Akkerdaas, J. & Van Ree, R., 2001, ‘Cross-reactivity of ige antibodies to allergens’, *Allergy*, vol. 56, no. 6, pp. 478–490.

Abad-García, M., González-Teruel, A. & González-Llinares, J., 2018, ‘Effectiveness of openaire, base, recolecta, and google scholar at finding spanish articles in repositories’, *Journal of the Association for Information Science and Technology*, vol. 69, no. 4, pp. 619–622.

Abbasi, A., Altmann, J. & Hossain, L., 2011, ‘Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures’, *Journal of Informetrics*, vol. 5, no. 4, pp. 594–607.

Abdelfattah, R., Kamran, H., Lazar, J. & Kassotis, J., 2018, ‘Does caffeine consumption increase the risk of new-onset atrial fibrillation?’, *Cardiology*, vol. 140, pp. 106–114.

Abdelhamid, A. S., Brown, T. J., Brainard, J. S., Biswas, P., Thorpe, G. C., Moore, H. J., Deane, K. H., AlAbdulghafoor, F. K., Summerbell, C. D., Worthington, H. V. et al., 2018, ‘Omega-3 fatty acids for the primary and secondary prevention of cardiovascular disease’, *Cochrane Database of Systematic Reviews*.

Abdi, A., Shamsuddin, S. M. & Aliguliyev, R. M., 2018, ‘Qmos: Query-based multi-documents opinion-oriented summarization’, *Information Processing & Management*, vol. 54, no. 2, pp. 318–338.

Adamic, L. A., Wilkinson, D., Huberman, B. A. & Adar, E., 2002, 'A literature based method for identifying gene-disease connections', *Proceedings. IEEE Computer Society Bioinformatics Conference*, IEEE, pp. 109–117.

Afshar Ali, M., Alam, K. & Taylor, B., 2020, 'Do social exclusion and remoteness explain the digital divide in australia? evidence from a panel data estimation approach', *Economics of Innovation and New Technology*, vol. 29, no. 6, pp. 643–659.

Agarwal, R., Gao, G., DesRoches, C. & Jha, A. K., 2010, 'Research commentary—the digital transformation of healthcare: Current status and the road ahead', *Information Systems Research*, vol. 21, no. 4, pp. 796–809.

Ahmad, F., 2018, 'Knowledge sharing in a non-native language context: Challenges and strategies', *Journal of Information Science*, vol. 44, no. 2, pp. 248–264.

Ahmed, S. F., Quadeer, A. A. & McKay, M. R., 2020, 'Preliminary identification of potential vaccine targets for the covid-19 coronavirus (sars-cov-2) based on sars-cov immunological studies', *Viruses*, vol. 12, no. 3, p. 254.

Ahn, D.-G., Shin, H.-J., Kim, M.-H., Lee, S., Kim, H.-S., Myoung, J., Kim, B.-T. & Kim, S.-J., 2020, 'Current status of epidemiology, diagnosis, therapeutics, and vaccines for novel coronavirus disease 2019 (covid-19)', *Journal of Microbiology and Biotechnology*, vol. 30, no. 3, pp. 313–324, <http://jmb.or.kr/journal/view.html?doi=10.4014/jmb.2003.03011>.

Al-Aamri, A., Taha, K., Al-Hammadi, Y., Maalouf, M. & Homouz, D., 2019, 'Analyzing a co-occurrence gene-interaction network to identify disease-gene association', *BMC bioinformatics*, vol. 20, no. 1, p. 70.

Alaimo, C. & Kallinikos, J., 2017, 'Computing the everyday: Social media as data platforms', *The Information Society*, vol. 33, no. 4, pp. 175–191.

Alhassani, S., Deif, B., Conacher, S., Cunningham, K. S. & Roberts, J. D., 2018, ‘A large familial pathogenic plakophilin-2 gene (pkp2) deletion manifesting with sudden cardiac death and lone atrial fibrillation: Evidence for alternating atrial and ventricular phenotypes’, *HeartRhythm Case Reports*, vol. 4, no. 10, pp. 486–489.

Almeida, M. B., 2013, ‘Revisiting ontologies: A necessary clarification’, *Journal of the American Society for Information Science and Technology*, vol. 64, no. 8, pp. 1682–1693.

Anastasiadou, E., Jacob, L. S. & Slack, F. J., 2018, ‘Non-coding rna networks in cancer’, *Nature Reviews Cancer*, vol. 18, no. 1, p. 5.

Annosi, M. C., Brunetta, F., Monti, A. & Nati, F., 2019, ‘Is the trend your friend? an analysis of technology 4.0 investment decisions in agricultural smes’, *Computers in Industry*, vol. 109, pp. 59–71.

Antonucci, Y. L., Fortune, A. & Kirchmer, M., 2020, ‘An examination of associations between business process management capabilities and the benefits of digitalization: All capabilities are not equal’, *Business Process Management Journal*.

Aouadi, H., Torjmen-Khemakhem, M. & Jemaa, M. B., 2012, ‘Combination of document structure and links for multimedia object retrieval’, *Journal of Information Science*, vol. 38, no. 5, pp. 442–458.

Apicella, M., Campopiano, M. C., Mantuano, M., Mazoni, L., Coppelli, A. & Del Prato, S., 2020, ‘Covid-19 in people with diabetes: Understanding the reasons for worse outcomes’, *The Lancet Diabetes & Endocrinology*, vol. 8, no. 9, pp. 782–792.

Ardolino, M., Rapaccini, M., Saccani, N., Gaiardelli, P., Crespi, G. & Ruggeri, C., 2018, 'The role of digital technologies for the service transformation of industrial companies', *International Journal of Production Research*, vol. 56, no. 6, pp. 2116–2132.

Aria, M. & Cuccurullo, C., 2017, 'bibliometrix: An r-tool for comprehensive science mapping analysis', *Journal of informetrics*, vol. 11, no. 4, pp. 959–975.

Arias, T. D., Jorge, L. & Barrantes, R., 1991, 'Uses and misuses of definitions of genetic polymorphism. a perspective from population pharmacogenetics', *British Journal of Clinical Pharmacology*, vol. 31, no. 1, p. 117.

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*, vol. 119, no. 3, pp. 1455–1486.

Bai, X., Yang, P. & Shi, X., 2017, 'An overlapping community detection algorithm based on density peaks', *Neurocomputing*, vol. 226, pp. 7–15.

Baker, N. C. & Hemminger, B. M., 2010, 'Mining connections between chemicals, proteins, and diseases extracted from medline annotations', *Journal of biomedical informatics*, vol. 43, no. 4, pp. 510–519.

Barabási, A.-L., Gulbahce, N. & Loscalzo, J., 2011, 'Network medicine: a network-based approach to human disease', *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68.

Barron, E., Bakhai, C., Kar, P., Weaver, A., Bradley, D., Ismail, H., Knighton, P., Holman, N., Khunti, K. & Sattar, N., 2020, 'Associations of type 1 and type 2 diabetes with covid-19-related mortality in england: A whole-population study', *The Lancet Diabetes & Endocrinology*, vol. 8, no. 10, pp. 813–822.

Bayer, S., Gimpel, H. & Rau, D., 2020, 'Iot-commerce-opportunities for customers through an affordance lens', *Electronic Markets*, pp. 1–24.

Begelman, G., Keller, P., Smadja, F. et al., 2006, 'Automated tag clustering: Improving search and exploration in the tag space', *collaborative web tagging workshop at WWW2006, Edinburgh, Scotland*, pp. 15–33.

Beigel, J. H., Tomashek, K. M., Dodd, L. E., Mehta, A. K., Zingman, B. S., Kalil, A. C., Hohmann, E., Chu, H. Y., Luetkemeyer, A. & Kline, S., 2020, 'Remdesivir for the treatment of covid-19', *New England Journal of Medicine*, vol. 383, no. 19, pp. 1813–1826.

Bentzen, B. H., Bomholtz, S. H., Simó-Vicens, R., Folkersen, L., Abildgaard, L., Speerschneider, T., Muthukumarasamy, K. M., Edvardsson, N., Sørensen, U. S. & Grunnet, M., 2020, 'Mechanisms of action of the kca2-negative modulator ap30663, a novel compound in development for treatment of atrial fibrillation in man', *Frontiers in Pharmacology*, vol. 11, p. 610.

Bergwerk, M., Gonen, T., Lustig, Y., Amit, S., Lipsitch, M., Cohen, C., Mandelboim, M., Levin, E. G., Rubin, C. & Indenbaum, V., 2021, 'Covid-19 breakthrough infections in vaccinated health care workers', *New England Journal of Medicine*, vol. 385, no. 16, pp. 1474–1484.

Bermejo-Martin, J. F., González-Rivera, M., Almansa, R., Micheloud, D., Tedim, A. P., Domínguez-Gil, M., Resino, S., Martín-Fernández, M., Murua, P. R. & Pérez-García, F., 2020, 'Viral rna load in plasma is associated with critical illness and a dysregulated host response in covid-19', *Critical Care*, vol. 24, no. 1, pp. 1–13.

Bernal, J. L., Andrews, N., Gower, C., Gallagher, E., Simmons, R., Thelwall, S., Stowe, J., Tessier, E., Groves, N. & Dabrera, G., 2021, 'Effectiveness of covid-19 vaccines against the b. 1.617. 2 (delta) variant', *New England Journal of Medicine*.

Bernstein, B., 2000, *Pedagogy, symbolic control, and identity: Theory, research, critique*, vol. 5, Rowman & Littlefield.

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A. & Venkatraman, N., 2013, 'Digital business strategy: Toward a next generation of insights', *MIS Quarterly*, pp. 471–482.

Bharadwaj, A. S., 2000, 'A resource-based perspective on information technology capability and firm performance: An empirical investigation', *MIS Quarterly*, pp. 169–196.

Bienhaus, F. & Haddud, A., 2018, 'Procurement 4.0: Factors influencing the digitisation of procurement and supply chains', *Business Process Management Journal*.

Biswas, N., Mustapha, T., Khubchandani, J. & Price, J. H., 2021, 'The nature and extent of covid-19 vaccination hesitancy in healthcare workers', *Journal of Community Health*, vol. 46, no. 6, pp. 1244–1251.

Björkdahl, J., 2020, 'Strategies for digitalization in manufacturing firms', *California Management Review*, p. 0008125620920349.

Black, J. S. & van Esch, P., 2020, 'Ai-enabled recruiting: What is it and how should a manager use it?', *Business Horizons*, vol. 63, no. 2, pp. 215–226.

Blei, D. M., 2012, 'Probabilistic topic models', *Communications of the ACM*, vol. 55, no. 4, pp. 77–84.

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)*, vol. 57, no. 2, pp. 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*, vol. 16, no. 16, pp. 17–24.

Blei, D. M., Ng, A. Y. & Jordan, M. I., 2003, 'Latent dirichlet allocation', *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E., 2008, 'Fast unfolding of communities in large networks', *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008.

Boddu, S. K., Aurangabadkar, G. & Kuchay, M. S., 2020, 'New onset diabetes, type 1 diabetes and covid-19', *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 6, pp. 2211–2217.

Bonaccorsi, G., Pierri, F., Cinelli, M., Flori, A., Galeazzi, A., Porcelli, F., Schmidt, A. L., Valensise, C. M., Scala, A. & Quattrociocchi, W., 2020, 'Economic and social consequences of human mobility restrictions under covid-19', *Proceedings of the National Academy of Sciences*, vol. 117, no. 27, pp. 15530–15535.

Borgatti, S. P., Mehra, A., Brass, D. J. & Labianca, G., 2009, 'Network analysis in the social sciences', *Science*, vol. 323, no. 5916, pp. 892–895.

Borner, K., 2015, *Atlas of knowledge: Anyone can map*, MIT Press.

Bornstein, S. R., Rubino, F., Khunti, K., Mingrone, G., Hopkins, D., Birkenfeld, A. L., Boehm, B., Amiel, S., Holt, R. I. & Skyler, J. S., 2020, 'Practical recommendations for the management of diabetes in patients with covid-19', *The Lancet Diabetes & Endocrinology*, vol. 8, no. 6, pp. 546–550.

Bossmann, J., 2016, 'Top 9 ethical issues in artificial intelligence', *World Economic Forum*, , vol. 21p. 1.

Bourfiss, M., Te Riele, A. S., Mast, T. P., Cramer, M. J., Van Der Heijden, J. F., Van Veen, T. A., Loh, P., Dooijes, D., Hauer, R. N. & Velthuis, B. K., 2016, 'Influence of genotype on structural atrial abnormalities and atrial fibrillation or flutter in arrhythmogenic right ventricular dysplasia/cardiomyopathy', *Journal of Cardiovascular Electrophysiology*, vol. 27, no. 12, pp. 1420–1428.

Brock, J. K.-U. & Von Wangenheim, F., 2019, 'Demystifying ai: What digital transformation leaders can teach you about realistic artificial intelligence', *California Management Review*, vol. 61, no. 4, pp. 110–134.

Brooks, J. T., Beezhold, D. H., Noti, J. D., Coyle, J. P., Derk, R. C., Blachere, F. M. & Lindsley, W. G., 2021, 'Maximizing fit for cloth and medical procedure masks to improve performance and reduce sars-cov-2 transmission and exposure, 2021', *Morbidity and Mortality Weekly Report*, vol. 70, no. 7, p. 254.

Browne, O., O'Reilly, P., Hutchinson, M. & Krdzavac, N. B., 2019, 'Distributed data and ontologies: An integrated semantic web architecture enabling more efficient data management', *Journal of the Association for Information Science and Technology*, vol. 70, no. 6, pp. 575–586.

Bruza, P. & Weeber, M., 2008, *Literature-based discovery*, Springer Science & Business Media.

Brülhart, M., Klotzbücher, V., Lalive, R. & Reich, S. K., 2021, 'Mental health concerns during the covid-19 pandemic as revealed by helpline calls', *Nature*, vol. 600, no. 7887, pp. 121–126.

Buchanan, S., Jardine, C. & Ruthven, I., 2019, 'Information behaviors in disadvantaged and dependent circumstances and the role of information intermediaries', *Journal of the Association for Information Science and Technology*, vol. 70, no. 2, pp. 117–129.

Bukowska, A., Schild, L., Keilhoff, G., Hirte, D., Neumann, M., Gardemann, A., Neumann, K. H., Röhl, F.-W., Huth, C. & Goette, A., 2008, 'Mitochondrial dysfunction and redox signaling in atrial tachyarrhythmia', *Experimental Biology and Medicine*, vol. 233, no. 5, pp. 558–574.

Burki, T., 2020, 'China's successful control of covid-19', *The Lancet Infectious Diseases*, vol. 20, no. 11, pp. 1240–1241.

Bush, W. S. & Moore, J. H., 2012, 'Genome-wide association studies', *PLoS Computational Biology*, vol. 8, no. 12.

Butler, J. S., Garg, R. & Stephens, B., 2020, 'Social networks, funding, and regional advantages in technology entrepreneurship: An empirical analysis', *Information Systems Research*, vol. 31, no. 1, pp. 198–216.

Börner, K., Scrivner, O., Cross, L. E., Gallant, M., Ma, S., Martin, A. S., Record, L., Yang, H. & Dilger, J. M., 2020, 'Mapping the co-evolution of artificial intelligence, robotics, and the internet of things over 20 years (1998-2017)', *PLoS One*, vol. 15, no. 12, p. e0242984.

Cabanillas, B., Akdis, C. & Novak, N., 2020, 'Allergic reactions to the first covid-19 vaccine: A potential role of polyethylene glycol', *Allergy*, vol. 76, no. 6, pp. 1617–1618.

Cai, X., Fry, C. V. & Wagner, C. S., 2021, 'International collaboration during the covid-19 crisis: Autumn 2020 developments', *Scientometrics*, vol. 126, no. 4, pp. 3683–3692.

Cameron, D., Bodenreider, O., Yalamanchili, H., Danh, T., Vallabhaneni, S., Thirunarayan, K., Sheth, A. P. & Rindflesch, T. C., 2013, 'A graph-based recovery and decomposition of swanson's hypothesis using semantic predications', *Journal of Biomedical Informatics*, vol. 46, no. 2, pp. 238–251.

Campos-Mercade, P., Meier, A. N., Schneider, F. H., Meier, S., Pope, D. & Wengström, E., 2021, 'Monetary incentives increase covid-19 vaccinations', *Science*, vol. 374, no. 6569, pp. 879–882.

Candi, M. & Beltagui, A., 2019, 'Effective use of 3d printing in the innovation process', *Technovation*, vol. 80, pp. 63–73.

Cao, X., 2020, 'Covid-19: immunopathology and its implications for therapy', *Nature Reviews Immunology*, vol. 20, no. 5, pp. 269–270.

Cariaso, M. & Lennon, G., 2012, 'Snpedia: A wiki supporting personal genome annotation, interpretation and analysis', *Nucleic Acids Research*, vol. 40, no. D1, pp. D1308–D1312.

Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y. & Shir, E., 2007, 'A model of internet topology using k-shell decomposition', *Proceedings of the National Academy of Sciences*, vol. 104, no. 27, pp. 11150–11154.

Cassidy, C., 2020, 'Parameter tuning naïve bayes for automatic patent classification', *World Patent Information*, vol. 61, p. 101968.

Castelo-Branco, I., Cruz-Jesus, F. & Oliveira, T., 2019, 'Assessing industry 4.0 readiness in manufacturing: Evidence for the european union', *Computers in Industry*, vol. 107, pp. 22–32.

Cechinel, C., Sicilia, M.-Á., Sánchez-Alonso, S. & García-Barriocanal, E., 2013, 'Evaluating collaborative filtering recommendations inside large learning object repositories', *Information Processing & Management*, vol. 49, no. 1, pp. 34–50.

Cetindamar, D., Kitto, K., Wu, M., Zhang, Y., Abedin, B. & Knight, S., 2022, 'Explicating ai literacy of employees at digital workplaces', *IEEE Transactions on Engineering Management*.

Cetindamar, D., Phaal, R. & Probert, D., 2016, *Technology management: Activities and tools*, Macmillan International Higher Education.

Chae, B. K., 2019, 'A general framework for studying the evolution of the digital innovation ecosystem: The case of big data', *International Journal of Information Management*, vol. 45, pp. 83–94.

Chae, H.-C., Koh, C. E. & Prybutok, V. R., 2014, 'Information technology capability and firm performance: Contradictory findings and their possible causes', *MIS Quarterly*, vol. 38, no. 1, pp. 305–326.

Chahrour, M., Assi, S., Bejjani, M., Nasrallah, A. A., Salhab, H., Fares, M. & Khachfe, H. H., 2020, 'A bibliometric analysis of covid-19 research activity: A call for increased output', *Cureus*, vol. 12, no. 3.

Chen, C., 2006, 'Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature', *Journal of the American Society for information Science and Technology*, vol. 57, no. 3, pp. 359–377.

Chen, H., Guo, J., Wang, C., Luo, F., Yu, X., Zhang, W., Li, J., Zhao, D., Xu, D. & Gong, Q., 2020, 'Clinical characteristics and intrauterine vertical transmission potential of covid-19 infection in nine pregnant women: A retrospective review of medical records', *The Lancet*, vol. 395, no. 10226, pp. 809–815.

Chen, Q., Allot, A. & Lu, Z., 2021, 'Litcovid: An open database of covid-19 literature', *Nucleic Acids Research*, vol. 49, no. D1, pp. D1534–D1540.

Chen, W.-J., Chang, S.-H., Chan, Y.-H., Lee, J.-L., Lai, Y.-J., Chang, G.-J., Tsai, F.-C. & Yeh, Y.-H., 2019, 'Tachycardia-induced cd44/nox4 signaling is involved in the development of atrial remodeling', *Journal of Molecular and Cellular Cardiology*, vol. 135, pp. 67–78.

Christophersen, I. E., Rienstra, M., Roselli, C., Yin, X., Geelhoed, B., Barnard, J., Lin, H., Arking, D. E., Smith, A. V. & Albert, C. M., 2017, 'Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation', *Nature Genetics*, vol. 49, no. 6, pp. 946–952.

Chu, I. Y.-H., Alam, P., Larson, H. J. & Lin, L., 2020, 'Social consequences of mass quarantine during epidemics: A systematic review with implications for the covid-19 response', *Journal of Travel Medicine*, vol. 27, no. 7, p. taaa192.

Chua, R. L., Lukassen, S., Trump, S., Hennig, B. P., Wendisch, D., Pott, F., Debnath, O., Thürmann, L., Kurth, F. & Völker, M. T., 2020, 'Covid-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis', *Nature Biotechnology*, vol. 38, no. 8, pp. 970–979.

Clauset, A., Moore, C. & Newman, M. E., 2008, 'Hierarchical structure and the prediction of missing links in networks', *Nature*, vol. 453, no. 7191, pp. 98–101.

Clemente-Casares, X., Blanco, J., Ambalavanan, P., Yamanouchi, J., Singha, S., Fandos, C., Tsai, S., Wang, J., Garabatos, N. & Izquierdo, C., 2016, 'Expanding antigen-specific regulatory networks to treat autoimmunity', *Nature*, vol. 530, no. 7591, pp. 434–440.

Cohen, A. M., Hersh, W. R., Dubay, C. & Spackman, K., 2005, 'Using co-occurrence network structure to extract synonymous gene and protein names from medline abstracts', *BMC Bioinformatics*, vol. 6, no. 1, p. 103.

Colavizza, G., Costas, R., Traag, V. A., Van Eck, N. J., Van Leeuwen, T. & Waltman, L., 2021, 'A scientometric overview of cord-19', *PloS One*, vol. 16, no. 1, p. e0244839.

Colavizza, G. & Franceschet, M., 2016, ‘Clustering citation histories in the physical review’, *Journal of Informetrics*, vol. 10, no. 4, pp. 1037–1051.

Collins, F. S. & Varmus, H., 2015, ‘A new initiative on precision medicine’, *New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M., 2009, ‘Mapping complex disease traits with global gene expression’, *Nature Reviews Genetics*, vol. 10, no. 3, pp. 184–194.

Cordasco, G. & Gargano, L., 2010, ‘Community detection via semi-synchronous label propagation algorithms’, *2010 IEEE international workshop on: Business applications of social network analysis (BASNA)*, IEEE, pp. 1–8.

Correani, A., De Massis, A., Frattini, F., Petruzzelli, A. M. & Natalicchio, A., 2020, ‘Implementing a digital strategy: Learning from the experience of three digital transformation projects’, *California Management Review*, vol. 62, no. 4, pp. 37–56.

Coulet, A., Shah, N. H., Garten, Y., Musen, M. & Altman, R. B., 2010, ‘Using text to build semantic networks for pharmacogenomics’, *Journal of Biomedical Informatics*, vol. 43, no. 6, pp. 1009–1019.

Coulson, B. S., Fowler, K., Bishop, R. & Cotton, R., 1985, ‘Neutralizing monoclonal antibodies to human rotavirus and indications of antigenic drift among strains from neonates’, *Journal of Virology*, vol. 54, no. 1, pp. 14–20.

Crichton, G., Baker, S., Guo, Y. & Korhonen, A., 2020, ‘Neural networks for open and closed literature-based discovery’, *PloS One*, vol. 15, no. 5, p. e0232891.

Crichton, G., Guo, Y., Pyysalo, S. & Korhonen, A., 2018, ‘Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches’, *BMC bioinformatics*, vol. 19, no. 1, p. 176.

Culot, G., Orzes, G., Sartor, M. & Nassimbeni, G., 2020, ‘The future of manufacturing: A delphi-based scenario analysis on industry 4.0’, *Technological Forecasting and Social Change*, vol. 157, p. 120092.

Curatolo, P. W. & Robertson, D., 1983, ‘The health consequences of caffeine’, *Annals of Internal Medicine*, vol. 98, pp. 641–653.

Dai, H., Saccardo, S., Han, M. A., Roh, L., Raja, N., Vangala, S., Modi, H., Pandya, S., Sloyan, M. & Croymans, D. M., 2021, ‘Behavioural nudges increase covid-19 vaccinations’, *Nature*, vol. 597, no. 7876, pp. 404–409.

Dan, Y. & Chieh, H. C., 2008, ‘A reflective review of disruptive innovation theory’, *PICMET’08-2008 Portland International Conference on Management of Engineering & Technology*, IEEE, pp. 402–414.

Danneels, E., 2004, ‘Disruptive technology reconsidered: A critique and research agenda’, *Journal of Product Innovation Management*, vol. 21, no. 4, pp. 246–258.

Day, G. S. & Schoemaker, P. J., 2016, ‘Adapting to fast-changing markets and technologies’, *California Management Review*, vol. 58, no. 4, pp. 59–77.

Demeter, K., Losonci, D. & Nagy, J., 2020, ‘Road to digital manufacturing–a longitudinal case-based analysis’, *Journal of Manufacturing Technology Management*.

Deregt, D. & Babiuk, L. A., 1987, ‘Monoclonal antibodies to bovine coronavirus: Characteristics and topographical mapping of neutralizing epitopes on the e2 and e3 glycoproteins’, *Virology*, vol. 161, no. 2, pp. 410–420.

Ding, J., Fu, H., Liu, Y., Gao, J., Li, Z., Zhao, X., Zheng, J., Sun, W., Ni, H. & Ma, X., 2020, ‘Prevention and control measures in radiology department for covid-19’, *European Radiology*, vol. 30, no. 7, pp. 3603–3608.

Ding, Y., 2011, ‘Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks’, *Journal of Informetrics*, vol. 5, no. 1, pp. 187–203.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L. & Chambers, T., 2013, ‘Entitymetrics: Measuring the impact of entities’, *PloS One*, vol. 8, no. 8, p. e71416.

Ding, Y., Yan, E., Frazho, A. & Caverlee, J., 2009, ‘Pagerank for ranking authors in co-citation networks’, *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2229–2243.

Dinneen, J. D., Asadi, B., Frissen, I., Shu, F. & Julien, C.-A., 2018, ‘Improving exploration of topic hierarchies: Comparative testing of simplified library of congress subject heading structures’, *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pp. 102–109.

Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F., 2006, ‘K-core organization of complex networks’, *Physical Review Letters*, vol. 96, no. 4, p. 040601.

Dougherty, M. & Meyer, E. T., 2014, ‘Community, tools, and practices in web archiving: The state-of-the-art in relation to social science and humanities research needs’, *Journal of the Association for Information Science and Technology*, vol. 65, no. 11, pp. 2195–2209.

Doulamis, I. P., Samanidis, G., Tzani, A., Antoranz, A., Gkogkos, A., Konstantopoulos, P., Pliaka, V., Minia, A., Alexopoulos, L. G. & Perrea, D. N., 2019, ‘Proteomic profile of patients with atrial fibrillation undergoing cardiac surgery’, *Interactive CardioVascular and Thoracic Surgery*, vol. 28, no. 1, pp. 94–101.

Dremel, C., Wulf, J., Herterich, M. M., Waizmann, J.-C. & Brenner, W., 2017, 'How audi ag established big data analytics in its digital transformation', *MIS Quarterly Executive*, vol. 16, no. 2.

Dror, A. A., Eisenbach, N., Taiber, S., Morozov, N. G., Mizrachi, M., Zigron, A., Srouji, S. & Sela, E., 2020, 'Vaccine hesitancy: The next challenge in the fight against covid-19', *European Journal of Epidemiology*, vol. 35, no. 8, pp. 775–779.

Du, M., Ding, S. & Jia, H., 2016, 'Study on density peaks clustering based on k-nearest neighbors and principal component analysis', *Knowledge-Based Systems*, vol. 99, pp. 135–145.

Duch, R., Roope, L. S., Violato, M., Becerra, M. F., Robinson, T. S., Bonnefon, J.-F., Friedman, J., Loewen, P. J., Mamidi, P. & Melegaro, A., 2021, 'Citizens from 13 countries share similar preferences for covid-19 vaccine allocation priorities', *Proceedings of the National Academy of Sciences*, vol. 118, no. 38.

Düzen, I. V., Yavuz, F., Vuruskan, E., Saracoglu, E., Poyraz, F., Göksülük, H., Candemir, B. & Demiryürek, S., 2017, 'Leukocyte trp channel gene expressions in patients with non-valvular atrial fibrillation', *Scientific Reports*, vol. 7, no. 1, pp. 1–7.

Ebadi, A., Xi, P., Tremblay, S., Spencer, B., Pall, R. & Wong, A., 2021, 'Understanding the temporal evolution of covid-19 research through machine learning and natural language processing', *Scientometrics*, vol. 126, no. 1, pp. 725–739.

Eibensteiner, P., Spitzauer, S., Steinberger, P., Kraft, D. & Valenta, R., 2000, 'Immunoglobulin e antibodies of atopic individuals exhibit a broad usage of vh-gene families', *Immunology*, vol. 101, no. 1, pp. 112–119.

Eisenhardt, K. M. & Martin, J. A., 2000, 'Dynamic capabilities: What are they?', *Strategic Management Journal*, vol. 21, no. 10-11, pp. 1105–1121.

El Sawy, O. A., Kræmmergaard, P., Amsinck, H. & Vinther, A. L., 2016, 'How lego built the foundations and enterprise capabilities for digital leadership', *MIS Quarterly Executive*, vol. 15, no. 2.

Ellinor, P. T., Lunetta, K. L., Glazer, N. L., Pfeufer, A., Alonso, A., Chung, M. K., Sinner, M. F., De Bakker, P. I., Mueller, M. & Lubitz, S. A., 2010, 'Common variants in kcnn3 are associated with lone atrial fibrillation', *Nature genetics*, vol. 42, no. 3, pp. 240–244.

Endrédi, A., Senánszky, V., Libralato, S. & Jordán, F., 2018, 'Food web dynamics in trophic hierarchies', *Ecological Modelling*, vol. 368, pp. 94–103.

Engzell, P., Frey, A. & Verhagen, M. D., 2021, 'Learning loss due to school closures during the covid-19 pandemic', *Proceedings of the National Academy of Sciences*, vol. 118, no. 17.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P. & Zalányi, L., 2013, 'Prediction of emerging technologies based on analysis of the us patent citation network', *Scientometrics*, vol. 95, no. 1, pp. 225–242.

Eryd, S. A., Smith, J. G., Melander, O., Hedblad, B. & Engström, G., 2011, 'Inflammation-sensitive proteins and risk of atrial fibrillation: A population-based cohort study', *European Journal of Epidemiology*, vol. 26, no. 6, p. 449.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A. & Pappas, G., 2008, 'Comparison of pubmed, scopus, web of science, and google scholar: Strengths and weaknesses', *The FASEB Journal*, vol. 22, no. 2, pp. 338–342.

Fang, L., Karakiulakis, G. & Roth, M., 2020, ‘Are patients with hypertension and diabetes mellitus at increased risk for covid-19 infection?’, *The Lancet Respiratory Medicine*, vol. 8, no. 4, p. e21.

Fang, Y., Cheng, R., Li, X., Luo, S. & Hu, J., 2017, ‘Effective community search over large spatial graphs’, *Proceedings of the VLDB Endowment*, vol. 10, no. 6, pp. 709–720.

Fang, Y., Zhang, H., Ye, Y. & Li, X., 2014, ‘Detecting hot topics from twitter: A multiview approach’, *Journal of Information Science*, vol. 40, no. 5, pp. 578–593.

Feghaly, J., Zakka, P., London, B., MacRae, C. A. & Refaat, M. M., 2018, ‘Genetics of atrial fibrillation’, *Journal of the American Heart Association*, vol. 7, no. 20, p. e009884.

Felsenstein, S., Herbert, J. A., McNamara, P. S. & Hedrich, C. M., 2020, ‘Covid-19: Immunology and treatment options’, *Clinical Immunology*, vol. 215, p. 108448.

Fernandes, C., Ferreira, J. J., Raposo, M. L., Estevão, C., Peris-Ortiz, M. & Rueda-Armengot, C., 2017, ‘The dynamic capabilities perspective of strategic management: A co-citation analysis’, *Scientometrics*, vol. 112, no. 1, pp. 529–555.

Fernández-Reyes, F. C., Hermosillo-Valadez, J. & Montes-y Gómez, M., 2018, ‘A prospect-guided global query expansion strategy using word embeddings’, *Information Processing & Management*, vol. 54, no. 1, pp. 1–13.

Ferreira, J. J., Fernandes, C. I. & Ferreira, F. A., 2019, ‘To be or not to be digital, that is the question: Firm innovation and performance’, *Journal of Business Research*, vol. 101, pp. 583–590.

Fiscarelli, A. M., Brust, M. R., Danoy, G. & Bouvry, P., 2019, ‘Local memory boosts label propagation for community detection', *Applied Network Science*, vol. 4, no. 1, pp. 1–17.

Frank, A. G., Dalenogare, L. S. & Ayala, N. F., 2019, 'Industry 4.0 technologies: Implementation patterns in manufacturing companies', *International Journal of Production Economics*, vol. 210, pp. 15–26.

Freeman, L. C., Roeder, D. & Mulholland, R. R., 1979, 'Centrality in social networks: Ii. experimental results', *Social Networks*, vol. 2, no. 2, pp. 119–141.

Freitas, B. T., Durie, I. A., Murray, J., Longo, J. E., Miller, H. C., Crich, D., Hogan, R. J., Tripp, R. A. & Pegan, S. D., 2020, 'Characterization and noncovalent inhibition of the deubiquitinase and deisgylase activity of sars-cov-2 papain-like protease', *ACS Infectious Diseases*, vol. 6, no. 8, pp. 2099–2109.

Frishammar, J., Richtnér, A., Brattström, A., Magnusson, M. & Björk, J., 2019, 'Opportunities and challenges in the new innovation landscape: Implications for innovation auditing and innovation management', *European Management Journal*, vol. 37, no. 2, pp. 151–164.

Fry, C. V., Cai, X., Zhang, Y. & Wagner, C. S., 2020, 'Consolidation in a crisis: Patterns of international collaboration in early covid-19 research', *PloS One*, vol. 15, no. 7, p. e0236307.

Ganegoda, G. U., Wang, J., Wu, F.-X. & Li, M., 2014, 'Prediction of disease genes using tissue-specified gene-gene network', *BMC Systems Biology*, vol. 8, no. S3, p. S3.

Gao, J., Wang, X., Wang, Y. & Xie, X., 2019, 'Explainable recommendation through attentive multi-view learning', *Proceedings of the AAAI Conference on Artificial Intelligence*, , vol. 33pp. 3622–3629.

Garay-Rondero, C. L., Martinez-Flores, J. L., Smith, N. R., Morales, S. O. C. & Aldrette-Malacara, A., 2019, 'Digital supply chain model in industry 4.0', *Journal of Manufacturing Technology Management*.

Garcia, M. A., Homan, P. A., García, C. & Brown, T. H., 2021, 'The color of covid-19: Structural racism and the disproportionate impact of the pandemic on older black and latinx adults', *The Journals of Gerontology: Series B*, vol. 76, no. 3, pp. e75–e80.

Garten, Y., Tatonetti, N. P. & Altman, R. B., 2010, 'Improving the prediction of pharmacogenes using text-derived drug-gene relationships', *Biocomputing 2010*, World Scientific, pp. 305–314.

Gautret, P., Lagier, J.-C., Parola, P., Meddeb, L., Mailhe, M., Doudier, B., Courjon, J., Giordanengo, V., Vieira, V. E. & Dupont, H. T., 2020, 'Hydroxychloroquine and azithromycin as a treatment of covid-19: results of an open-label non-randomized clinical trial', *International Journal of Antimicrobial Agents*, vol. 56, no. 1, p. 105949.

Ghadge, A., Kara, M. E., Moradlou, H. & Goswami, M., 2020, 'The impact of industry 4.0 implementation on supply chains', *Journal of Manufacturing Technology Management*.

Ghazawneh, A. & Henfridsson, O., 2015, 'A paradigmatic analysis of digital application marketplaces', *Journal of Information Technology*, vol. 30, no. 3, pp. 198–208.

Ghobakhloo, M. & Fathi, M., 2019, 'Corporate survival in industry 4.0 era: The enabling role of lean-digitized manufacturing', *Journal of Manufacturing Technology Management*.

Goldstein, D. B., 2009, ‘Common genetic variation and human traits’, *New England Journal of Medicine*, vol. 360, no. 17, p. 1696.

Gordon, M., Lindsay, R. K. & Fan, W., 2002, ‘Literature-based discovery on the world wide web’, *ACM Transactions on Internet Technology (TOIT)*, vol. 2, no. 4, pp. 261–275.

Gottlieb, R. L., Nirula, A., Chen, P., Boscia, J., Heller, B., Morris, J., Huhn, G., Cardona, J., Mocherla, B. & Stosor, V., 2021, ‘Effect of bamlanivimab as monotherapy or in combination with etesevimab on viral load in patients with mild to moderate covid-19: A randomized clinical trial’, *JAMA*, vol. 325, no. 7, pp. 632–644.

Goyal, N., Bron, M., Lalmas, M., Haines, A. & Cramer, H., 2018, ‘Designing for mobile experience beyond the native ad click: Exploring landing page presentation style and media usage’, *Journal of the Association for Information Science and Technology*, vol. 69, no. 7, pp. 913–923.

Gu, M. S. & Hwang, J., 2015, ‘Geosemantic information retrieval and its performance evaluation’, *Journal of Information Science*, vol. 41, no. 5, pp. 705–719.

Guo, J., Wang, X., Li, Q. & Zhu, D., 2016, ‘Subject–action–object-based morphology analysis for determining the direction of technological change’, *Technological Forecasting and Social Change*, vol. 105, pp. 27–40.

Guo, L., Sun, Z. & Hu, W., 2019, ‘Learning to exploit long-term relational dependencies in knowledge graphs’, *International Conference on Machine Learning*, PMLR, pp. 2505–2514.

Gurbaxani, V. & Dunkle, D., 2019, ‘Gearing up for successful digital transformation’, *MIS Quarterly Executive*, vol. 18, no. 3.

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L. & Leser, U., 2017, 'Deep learning with word embeddings improves biomedical named entity recognition', *Bioinformatics*, vol. 33, no. 14, pp. i37–i48.

Hacherouf, M., Bahloul, S. N. & Cruz, C., 2015, 'Transforming xml documents to owl ontologies: A survey', *Journal of Information Science*, vol. 41, no. 2, pp. 242–259.

Haghani, M. & Bliemer, M. C., 2020, 'Covid-19 pandemic and the unprecedented mobilisation of scholarly efforts prompted by a health crisis: Scientometric comparisons across sars, mers and 2019-ncov literature', *Scientometrics*, vol. 125, no. 3, pp. 2695–2726.

Haghani, M. & Varamini, P., 2021, 'Temporal evolution, most influential studies and sleeping beauties of the coronavirus literature', *Scientometrics*, vol. 126, no. 8, pp. 7005–7050.

Hall, A. K., Nousiainen, M. T., Campisi, P., Dagnone, J. D., Frank, J. R., Kroeker, K. I., Brzezina, S., Purdy, E. & Oswald, A., 2020, 'Training disrupted: Practical tips for supporting competency-based medical education during the covid-19 pandemic', *Medical Teacher*, vol. 42, no. 7, pp. 756–761.

Han, K., Blair, R. V., Iwanaga, N., Liu, F., Russell-Lodrigue, K. E., Qin, Z., Midkiff, C. C., Golden, N. A., Doyle-Meyers, L. A. & Kabir, M. E., 2021, 'Lung expression of human angiotensin-converting enzyme 2 sensitizes the mouse to sars-cov-2 infection', *American Journal of Respiratory Cell and Molecular Biology*, vol. 64, no. 1, p. 79.

Hao, J., Chen, M., Yu, W., Sun, Y. & Wang, W., 2019, 'Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts', *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1709–1719.

Harper, R., 2006, *Inside the smart home*, Springer Science & Business Media.

Hartley, J. L. & Sawaya, W. J., 2019, 'Tortoise, not the hare: Digital transformation of supply chain business processes', *Business Horizons*, vol. 62, no. 6, pp. 707–715.

Hein, A., Schreieck, M., Riasanow, T., Setzke, D. S., Wiesche, M., Böhm, M. & Krcmar, H., 2019, 'Digital platform ecosystems', *Electronic Markets*, pp. 1–12.

Helfat, C. E., Finkelstein, S., Mitchell, W., Peteraf, M., Singh, H., Teece, D. & Winter, S. G., 2009, *Dynamic capabilities: Understanding strategic change in organizations*, John Wiley & Sons.

Helfat, C. E. & Raubitschek, R. S., 2018, 'Dynamic and integrative capabilities for profiting from innovation in digital platform-based ecosystems', *Research Policy*, vol. 47, no. 8, pp. 1391–1399.

Henfridsson, O., Mathiassen, L. & Svahn, F., 2014, 'Managing technological change in the digital age: The role of architectural frames', *Journal of Information Technology*, vol. 29, no. 1, pp. 27–43.

Heo, G. E., Xie, Q., Song, M. & Lee, J.-H., 2019, 'Combining entity co-occurrence with specialized word embeddings to measure entity relation in alzheimer's disease', *BMC Medical Informatics and Decision Making*, vol. 19, no. 5, p. 240.

Hidalgo, A., Gabaly, S., Morales-Alonso, G. & Urueña, A., 2020, 'The digital divide in light of sustainable development: An approach through advanced machine learning techniques', *Technological Forecasting and Social Change*, vol. 150, p. 119754.

Hoffmann, M., Kleine-Weber, H. & Pöhlmann, S., 2020a, ‘A multibasic cleavage site in the spike protein of sars-cov-2 is essential for infection of human lung cells’, *Molecular Cell*, vol. 78, no. 4, pp. 779–784. e5.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T. S., Herrler, G., Wu, N.-H. & Nitsche, A., 2020b, ‘Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor’, *Cell*, vol. 181, no. 2, pp. 271–280. e8.

Holeman, I. & Kane, D., 2020, ‘Human-centered design for global health equity’, *Information Technology for Development*, vol. 26, no. 3, pp. 477–505.

Holman, N., Knighton, P., Kar, P., O’Keefe, J., Curley, M., Weaver, A., Barron, E., Bakhai, C., Khunti, K. & Wareham, N. J., 2020, ‘Risk factors for covid-19-related mortality in people with type 1 and type 2 diabetes in england: A population-based cohort study’, *The Lancet Diabetes & Endocrinology*, vol. 8, no. 10, pp. 823–833.

Horváth, D. & Szabó, R. Z., 2019, ‘Driving forces and barriers of industry 4.0: Do multinational and small and medium-sized companies have equal opportunities?’, *Technological Forecasting and Social Change*, vol. 146, pp. 119–132.

Hossain, M. M., 2020, ‘Current status of global research on novel coronavirus disease (covid-19): A bibliometric analysis and knowledge mapping’, *SSRN*.

Hossain, M. S., Muhammad, G. & Guizani, N., 2020, ‘Explainable ai and mass surveillance system-based healthcare framework to combat covid-19 like pandemics’, *IEEE Network*, vol. 34, no. 4, pp. 126–132.

Hou, J., Yang, X. & Chen, C., 2018, ‘Emerging trends and new developments in information science: A document co-citation analysis (2009–2016)’, *Scientometrics*, vol. 115, no. 2, pp. 869–892.

Hou, X., Zhang, X., Wu, X., Lu, M., Wang, D., Xu, M., Wang, H., Liang, T., Dai, J. & Duan, H., 2020, ‘Serum protein profiling reveals a landscape of inflammation and immune signaling in early-stage covid-19 infection’, *Molecular & Cellular Proteomics*, vol. 19, no. 11, pp. 1749–1759.

Hristovski, D., Peterlin, B., Mitchell, J. A. & Humphrey, S. M., 2005, ‘Using literature-based discovery to identify disease candidate genes’, *International Journal of Medical Informatics*, vol. 74, no. 2-4, pp. 289–298.

Hu, B., Guo, H., Zhou, P. & Shi, Z.-L., 2021, ‘Characteristics of sars-cov-2 and covid-19’, *Nature Reviews Microbiology*, vol. 19, no. 3, pp. 141–154.

Hu, Y.-H., Chen, Y.-L. & Chou, H.-L., 2017, ‘Opinion mining from online hotel reviews–a text summarization approach’, *Information Processing & Management*, vol. 53, no. 2, pp. 436–449.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J. & Gu, X., 2020a, ‘Clinical features of patients infected with 2019 novel coronavirus in wuhan, china’, *The Lancet*, vol. 395, no. 10223, pp. 497–506.

Huang, H., Wang, J. & Chen, H., 2017, ‘Implicit opinion analysis: Extraction and polarity labelling’, *Journal of the Association for Information Science and Technology*, vol. 68, no. 9, pp. 2076–2087.

Huang, L., Jia, X., Zhang, Y., Zhou, X. & Zhu, Y., 2018a, ‘Detecting hotspots in interdisciplinary research based on overlapping community detection’, *2018 Portland International Conference on Management of Engineering and Technology (PICMET)*, IEEE, pp. 1–6.

Huang, L., Liu, F. & Zhang, Y., 2020b, ‘Overlapping community discovery for identifying key research themes’, *IEEE Transactions on Engineering anagement*, vol. 68, no. 5, pp. 1321–1333.

Huang, L., Zhang, Y., Guo, Y., Zhu, D. & Porter, A. L., 2014, 'Four dimensional science and technology planning: A new approach based on bibliometrics and technology roadmapping', *Technological Forecasting and Social Change*, vol. 81, pp. 39–48.

Huang, L., Zhu, Y., Zhang, Y., Zhou, X. & Jia, X., 2018b, 'A link prediction-based method for identifying potential cooperation partners: A case study on four journals of informetrics', *2018 Portland international conference on management of engineering and technology (PICMET)*, IEEE, pp. 1–6.

Huang, T., Lo, L., Yamada, S., Chou, Y., Lin, W., Chang, S., Lin, Y., Liu, S., Cheng, W. & Tsai, T., 2019, 'Gastroesophageal reflux disease and atrial fibrillation: Insight from autonomic cardiogastric neural interaction', *Journal of Cardiovascular Electrophysiology*, vol. 30, no. 11, pp. 2262–2270.

Huq, A. Z., 2020, 'A right to a human decision', *Va. L. Rev.*, vol. 106, p. 611.

Imagawa, A., Hanafusa, T., Miyagawa, J.-i. & Matsuzawa, Y., 2000, 'A novel subtype of type 1 diabetes mellitus characterized by a rapid onset and an absence of diabetes-related antibodies', *New England Journal of Medicine*, vol. 342, no. 5, pp. 301–307.

Islam, M. S., Rahman, K. M., Sun, Y., Qureshi, M. O., Abdi, I., Chughtai, A. A. & Seale, H., 2020a, 'Current knowledge of covid-19 and infection prevention and control strategies in healthcare settings: A global analysis', *Infection Control & Hospital Epidemiology*, vol. 41, no. 10, pp. 1196–1206.

Islam, M. S., Sarkar, T., Khan, S. H., Kamal, A.-H. M., Hasan, S. M., Kabir, A., Yeasmin, D., Islam, M. A., Chowdhury, K. I. A. & Anwar, K. S., 2020b, 'Covid-19–related infodemic and its impact on public health: A global social media analysis', *The American Journal of Tropical Medicine and Hygiene*, vol. 103, no. 4, p. 1621.

Islam, T., Rahman, M. R., Aydin, B., Beklen, H., Arga, K. Y. & Shahjaman, M., 2020c, 'Integrative transcriptomics analysis of lung epithelial cells and identification of repurposable drug candidates for covid-19', *European Journal of Pharmacology*, vol. 887, p. 173594.

Ittipanuvat, V., Fujita, K., Sakata, I. & Kajikawa, Y., 2014, 'Finding linkage between technology and social issue: A literature based discovery approach', *Journal of Engineering and Technology Management*, vol. 32, pp. 160–184.

Ivanov, D., Dolgui, A. & Sokolov, B., 2019, 'The impact of digital technology and industry 4.0 on the ripple effect and supply chain risk analytics', *International Journal of Production Research*, vol. 57, no. 3, pp. 829–846.

Jantunen, A., Tarkiainen, A., Chari, S. & Oghazi, P., 2018, 'Dynamic capabilities, operational changes, and performance outcomes in the media industry', *Journal of Business Research*, vol. 89, pp. 251–257.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. & Zhao, L., 2019, 'Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey', *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211.

Jenmalm, M., Van Snick, J., Cormont, F. & Salman, B., 2001, 'Allergen-induced th1 and th2 cytokine secretion in relation to specific allergen sensitization and atopic symptoms in children', *Clinical & Experimental Allergy*, vol. 31, no. 10, pp. 1528–1535.

Jenssen, T.-K., Lægreid, A., Komorowski, J. & Hovig, E., 2001, 'A literature network of human genes for high-throughput analysis of gene expression', *Nature Genetics*, vol. 28, no. 1, pp. 21–28.

Jeong, Y., Jang, H. & Yoon, B., 2021, 'Developing a risk-adaptive technology roadmap using a bayesian network and topic modeling under deep uncertainty', *Scientometrics*, vol. 126, no. 5, pp. 3697–3722.

Jin, J.-M., Bai, P., He, W., Wu, F., Liu, X.-F., Han, D.-M., Liu, S. & Yang, J.-K., 2020, 'Gender differences in patients with covid-19: Focus on severity and mortality', *Frontiers in Public Health*, p. 152.

Jose, R. J. & Manuel, A., 2020, 'Covid-19 cytokine storm: The interplay between inflammation and coagulation', *The Lancet Respiratory Medicine*, vol. 8, no. 6, pp. e46–e47.

Josephs, N., Peng, S. & Crawford, F. W., 2022, 'Communication network dynamics in a large organizational hierarchy', *arXiv preprint arXiv:2208.01208*.

Kajikawa, Y., Mejia, C., Wu, M. & Zhang, Y., 2022, 'Academic landscape of technological forecasting and social change through citation network and topic analyses', *Technological Forecasting and Social Change*, vol. 182, p. 121877.

Kalluri, P., 2021, 'Don't ask if artificial intelligence is good or fair, ask how it shifts power', .

Kanehisa, M. & Goto, S., 2000, 'Kegg: kyoto encyclopedia of genes and genomes', *Nucleic acids research*, vol. 28, no. 1, pp. 27–30.

Kaplan, S. & Vakili, K., 2015, 'The double-edged sword of recombination in breakthrough innovation', *Strategic Management Journal*, vol. 36, no. 10, pp. 1435–1457.

Karimi, J. & Walter, Z., 2015, 'The role of dynamic capabilities in responding to digital disruption: A factor-based study of the newspaper industry', *Journal of Management Information Systems*, vol. 32, no. 1, pp. 39–81.

Kastrin, A., Rindflesch, T. C. & Hristovski, D., 2016, 'Link prediction on a network of co-occurring mesh terms: Towards literature-based discovery', *Methods of Information in Medicine*, vol. 55, no. 04, pp. 340–346.

Kay, L., Newman, N., Youtie, J., Porter, A. L. & Rafols, I., 2014, 'Patent overlay mapping: Visualizing technological distance', *Journal of the Association for Information Science and Technology*, vol. 65, no. 12, pp. 2432–2443.

Kholghi, M., De Vine, L., Sitbon, L., Zuccon, G. & Nguyen, A., 2017, 'Clinical information extraction using small data: An active learning approach based on sequence representations and word embeddings', *Journal of the Association for Information Science and Technology*, vol. 68, no. 11, pp. 2543–2556.

Kim, J., Kim, J.-j. & Lee, H., 2017, 'An analysis of disease-gene relationship from medline abstracts by digsee', *Scientific Reports*, vol. 7, no. 1, pp. 1–13.

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y. & Collier, N., 2004, 'Introduction to the bio-entity recognition task at jnlpba', *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, Citeseer, pp. 70–75.

King, T. C., Aggarwal, N., Taddeo, M. & Floridi, L., 2020, 'Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions', *Science and Engineering Ethics*, vol. 26, no. 1, pp. 89–120.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E. & Makse, H. A., 2010, 'Identification of influential spreaders in complex networks', *Nature Physics*, vol. 6, no. 11, pp. 888–893.

Klaassen, K., Stankovic, B., Zukic, B., Kotur, N., Gasic, V., Pavlovic, S. & Stojiljkovic, M., 2020, 'Functional prediction and comparative population analysis of variants in genes for proteases and innate immunity related to sars-cov-2 infection', *Infection, Genetics and Evolution*, vol. 84, p. 104498.

Klemm, T., Ebert, G., Calleja, D. J., Allison, C. C., Richardson, L. W., Bernardini, J. P., Lu, B. G., Kuchel, N. W., Grohmann, C. & Shibata, Y., 2020, 'Mechanism and inhibition of the papain-like protease, plpro, of sars-cov-2', *The EMBO Journal*, vol. 39, no. 18, p. e106275.

Klimstra, W. B., Tilston-Lunel, N. L., Nambulli, S., Boslett, J., McMillen, C. M., Gilliland, T., Dunn, M. D., Sun, C., Wheeler, S. E. & Wells, A., 2020, 'Sars-cov-2 growth, furin-cleavage-site adaptation and neutralization using serum from acutely infected hospitalized covid-19 patients', *Journal of General Virology*, vol. 101, no. 11, pp. 1156–1169.

Kohtamäki, M., Parida, V., Patel, P. C. & Gebauer, H., 2020, 'The relationship between digitalization and servitization: The role of servitization in capturing the financial potential of digitalization', *Technological Forecasting and Social Change*, vol. 151, p. 119804.

Kostoff, R. N. & Briggs, M. B., 2008, 'Literature-related discovery (lrd): potential treatments for parkinson's disease', *Technological Forecasting and Social Change*, vol. 75, no. 2, pp. 226–238.

Kostoff, R. N., Solka, J. L., Rushenberg, R. L. & Wyatt, J. A., 2008, 'Literature-related discovery (lrd): water purification', *Technological Forecasting and Social Change*, vol. 75, no. 2, pp. 256–275.

Kounis, N. G., Koniari, I., de Gregorio, C., Velissaris, D., Petalas, K., Brinia, A., Assimakopoulos, S. F., Gogos, C., Kouni, S. N. & Kounis, G. N., 2021, 'Allergic reactions to current available covid-19 vaccinations: Pathophysiology, causality, and therapeutic considerations', *Vaccines*, vol. 9, no. 3, p. 221.

Kousha, K. & Thelwall, M., 2020, ‘Covid-19 publications: Database coverage, citations, readers, tweets, news, facebook walls, reddit posts’, *Quantitative Science Studies*, vol. 1, no. 3, pp. 1068–1091.

Kousha, K., Thelwall, M. & Abdoli, M., 2018, ‘Can microsoft academic assess the early citation impact of in-press articles? a multi-discipline exploratory analysis’, *Journal of Informetrics*, vol. 12, no. 1, pp. 287–298.

Koyama, T., Weeraratne, D., Snowdon, J. L. & Parida, L., 2020, ‘Emergence of drift variants that may affect covid-19 vaccine development and antibody treatment’, *Pathogens*, vol. 9, no. 5, p. 324.

Kuken, B., Yang, Y., Liu, Z., He, P. & Wulasihan, M., 2020, ‘Relationship between m235t and t174m polymorphisms in angiotensin gene and atrial fibrillation in uyghur and han populations of xinjiang, china’, *International Journal of Clinical and Experimental Pathology*, vol. 13, no. 8, p. 2065.

La Bella, A., Fronzetti Colladon, A., Battistoni, E., Castellan, S. & Francucci, M., 2018, ‘Assessing perceived organizational leadership styles through twitter text mining’, *Journal of the Association for Information Science and Technology*, vol. 69, no. 1, pp. 21–31.

Lai, J., Ma, S., Wang, Y., Cai, Z., Hu, J., Wei, N., Wu, J., Du, H., Chen, T. & Li, R., 2020, ‘Factors associated with mental health outcomes among health care workers exposed to coronavirus disease 2019’, *JAMA Network Open*, vol. 3, no. 3, pp. e203976–e203976.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D. & Hoover, J., 2016, ‘Clinvar: public archive of interpretations of clinically relevant variants’, *Nucleic Acids Research*, vol. 44, no. D1, pp. D862–D868.

Lawson, C. E., Wu, S., Bhattacharjee, A. S., Hamilton, J. J., McMahon, K. D., Goel, R. & Noguera, D. R., 2017, 'Metabolic network analysis reveals microbial community interactions in anammox granules', *Nature Communications*, vol. 8, no. 1, pp. 1–12.

Lei, C. & Ruan, J., 2013, 'A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity', *Bioinformatics*, vol. 29, no. 3, pp. 355–364.

Levallet, N. & Chan, Y. E., 2018, 'Role of digital capabilities in unleashing the power of managerial improvisation', *MIS Quarterly Executive*, vol. 17, no. 1.

Lever, J., Gakkhar, S., Gottlieb, M., Rashnavadi, T., Lin, S., Siu, C., Smith, M., Jones, M. R., Krzywinski, M. & Jones, S. J., 2018, 'A collaborative filtering-based approach to biomedical knowledge discovery', *Bioinformatics*, vol. 34, no. 4, pp. 652–659.

Levi, M., Thachil, J., Iba, T. & Levy, J. H., 2020, 'Coagulation abnormalities and thrombosis in patients with covid-19', *The Lancet Haematology*, vol. 7, no. 6, pp. e438–e440.

Leydesdorff, L., Bornmann, L. & Mingers, J., 2019, 'Statistical significance and effect sizes of differences among research universities at the level of nations and worldwide based on the leiden rankings', *Journal of the Association for Information Science and Technology*, vol. 70, no. 5, pp. 509–525.

Leydesdorff, L. & Rafols, I., 2009, 'A global map of science based on the isi subject categories', *Journal of the American Society for Information Science and Technology*, vol. 60, no. 2, pp. 348–362.

Leydesdorff, L. & Rafols, I., 2011, 'Indicators of the interdisciplinarity of journals:

Diversity, centrality, and citations', *Journal of Informetrics*, vol. 5, no. 1, pp. 87–100.

Li, E. Y., Liao, C. H. & Yen, H. R., 2013, 'Co-authorship networks and research impact: A social capital perspective', *Research Policy*, vol. 42, no. 9, pp. 1515–1530.

Li, F., 2020, 'The digital transformation of business models in the creative industries: A holistic framework and emerging trends', *Technovation*, vol. 92, p. 102012.

Li, H. O.-Y., Bailey, A., Huynh, D. & Chan, J., 2020a, 'Youtube as a source of information on covid-19: A pandemic of misinformation?', *BMJ Global Health*, vol. 5, no. 5, p. e002604.

Li, L., Su, F., Zhang, W. & Mao, J., 2018, 'Digital transformation by sme entrepreneurs: A capability perspective', *Information Systems Journal*, vol. 28, no. 6, pp. 1129–1157.

Li, M., Porter, A. L. & Wang, Z. L., 2017, 'Evolutionary trend analysis of nanogenerator research based on a novel perspective of phased bibliographic coupling', *Nano Energy*, vol. 34, pp. 93–102.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H. & Wong, J. Y., 2020b, 'Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia', *New England Journal of Medicine*.

Li, T. C. & Chan, Y. E., 2019, 'Dynamic information technology capability: Concept definition and framework development', *The Journal of Strategic Information Systems*, vol. 28, no. 4, p. 101575.

Li, Y., Duche, A., Sayer, M. R., Roosan, D., Khalafalla, F. G., Ostrom, R. S., Totonchy, J. & Roosan, M. R., 2021, 'Sars-cov-2 early infection signature identified potential key infection mechanisms and drug targets', *BMC Genomics*, vol. 22, no. 1, pp. 1–13.

Li, Y.-y., Zhou, C.-w., Xu, J., Qian, Y. & Wang, B., 2012, 'Cyp11b2 t-344c gene polymorphism and atrial fibrillation: a meta-analysis of 2,758 subjects', *PLoS One*, vol. 7, no. 11, p. e50910.

Liben-Nowell, D. & Kleinberg, J., 2007, 'The link-prediction problem for social networks', *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031.

Lim, S., Bae, J. H., Kwon, H.-S. & Nauck, M. A., 2021, 'Covid-19 and diabetes mellitus: From pathophysiology to clinical management', *Nature Reviews Endocrinology*, vol. 17, no. 1, pp. 11–30.

Lin, H.-F., Su, J.-Q. & Higgins, A., 2016, 'How dynamic capabilities affect adoption of management innovations', *Journal of Business Research*, vol. 69, no. 2, pp. 862–876.

Lin, Z., Zhang, F., Lin, X., Zhang, W. & Tian, Z., 2021, 'Hierarchical core maintenance on large dynamic graphs', *Proceedings of the VLDB Endowment*, vol. 14, no. 5, pp. 757–770.

Lindsay, R. K. & Gordon, M. D., 1999, 'Literature-based discovery by lexical statistics', *Journal of the American Society for Information Science*, vol. 50, no. 7, pp. 574–587.

Liu, K., Chen, Y., Lin, R. & Han, K., 2020a, 'Clinical features of covid-19 in elderly patients: A comparison with young and middle-aged patients', *Journal of Infection*, vol. 80, no. 6, pp. e14–e18.

Liu, Q., Luo, D., Haase, J. E., Guo, Q., Wang, X. Q., Liu, S., Xia, L., Liu, Z., Yang, J. & Yang, B. X., 2020b, 'The experiences of health-care providers during the covid-19 crisis in china: A qualitative study', *The Lancet Global Health*, vol. 8, no. 6, pp. e790–e798.

Liu, X., Bollen, J., Nelson, M. L. & Van de Sompel, H., 2005, 'Co-authorship networks in the digital library research community', *Information Processing & Management*, vol. 41, no. 6, pp. 1462–1480.

Liu, Y., Tang, M., Zhou, T. & Do, Y., 2015, 'Improving the accuracy of the k-shell method by removing redundant links: From a perspective of spreading dynamics', *Scientific Reports*, vol. 5, no. 1, pp. 1–11.

Lobo, S. & Whyte, J., 2017, 'Aligning and reconciling: Building project capabilities for digital delivery', *Research Policy*, vol. 46, no. 1, pp. 93–107.

Loebbecke, C. & Picot, A., 2015, 'Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda', *The Journal of Strategic Information Systems*, vol. 24, no. 3, pp. 149–157.

Long, Y., Hu, T., Liu, L., Chen, R., Guo, Q., Yang, L., Cheng, Y., Huang, J. & Du, L., 2020, 'Effectiveness of n95 respirators versus surgical masks against influenza: A systematic review and meta-analysis', *Journal of Evidence-Based Medicine*, vol. 13, no. 2, pp. 93–101.

Lovász, L., 1993, 'Random walks on graphs: A survey', *Combinatorics, Paul Erdos is Eighty*, vol. 2, no. 1, pp. 1–46.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B. & Zhu, N., 2020, 'Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding', *The Lancet*, vol. 395, no. 10224, pp. 565–574.

Lubani, M., Noah, S. A. M. & Mahmud, R., 2019, ‘Ontology population: Approaches and design aspects’, *Journal of Information Science*, vol. 45, no. 4, pp. 502–515.

Lumsden, J., Hall, H. & Cruickshank, P., 2011, ‘Ontology definition and construction, and epistemological adequacy for systems interoperability: A practitioner analysis’, *Journal of Information Science*, vol. 37, no. 3, pp. 246–253.

Lyzinski, V., Tang, M., Athreya, A., Park, Y. & Priebe, C. E., 2016, ‘Community detection and classification in hierarchical stochastic blockmodels’, *IEEE Transactions on Network Science and Engineering*, vol. 4, no. 1, pp. 13–26.

Lü, L., Jin, C.-H. & Zhou, T., 2009, ‘Similarity index based on local paths for link prediction of complex networks’, *Physical Review E*, vol. 80, no. 4, p. 046122.

Lü, L. & Zhou, T., 2010, ‘Link prediction in weighted networks: The role of weak ties’, *EPL (Europhysics Letters)*, vol. 89, no. 1, p. 18001.

Machingaidze, S. & Wiysonge, C. S., 2021, ‘Understanding covid-19 vaccine hesitancy’, *Nature Medicine*, vol. 27, no. 8, pp. 1338–1339.

Malhotra, D. & Chug, A., 2021, ‘A modified label propagation algorithm for community detection in attributed networks’, *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100030.

Mallory, E. K., Zhang, C., Ré, C. & Altman, R. B., 2016, ‘Large-scale extraction of gene interactions from full-text literature using deepdive’, *Bioinformatics*, vol. 32, no. 1, pp. 106–113.

Mao, J. & Cui, H., 2018, ‘Identifying bacterial biotope entities using sequence labeling: performance and feature analysis’, *Journal of the Association for Information Science and Technology*, vol. 69, no. 9, pp. 1134–1147.

Marchand, L., Pecquet, M. & Luyton, C., 2020, 'Type 1 diabetes onset triggered by covid-19', *Acta Diabetologica*, vol. 57, no. 10, pp. 1265–1266.

Marsi, E., Ozturk, P., Aamot, E., Sizov, G. & Ardelan, M. V., 2014, 'Towards text mining in climate science: Extraction of quantitative variables and their relations', Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. & Piperidis, S. (eds.) *LREC 2014 - NINTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION*, Holmes Semant Solut; European Media Lab GmBH; EML; VoiceBox Technologies; KDICTIONARIES, p. 1, 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik, ICELAND, MAY 26-31, 2014.

Mejia, C., Wu, M., Zhang, Y. & Kajikawa, Y., 2021, 'Exploring topics in bibliometric research through citation networks and semantic analysis', *Frontiers in Research Metrics and Analytics*, vol. 6.

Melo, P. F., Dalip, D. H., Junior, M. M., Gonçalves, M. A. & Benevenuto, F., 2019, '10sent: A stable sentiment analysis method based on the combination of off-the-shelf approaches', *Journal of the Association for Information Science and Technology*, vol. 70, no. 3, pp. 242–255.

Merrill, J. T., Erkan, D., Winakur, J. & James, J. A., 2020, 'Emerging evidence of a covid-19 thrombotic syndrome has treatment implications', *Nature Reviews Rheumatology*, vol. 16, no. 10, pp. 581–589.

Mick, P. & Murphy, R., 2020, 'Aerosol-generating otolaryngology procedures and the need for enhanced ppe during the covid-19 pandemic: A literature review', *Journal of Otolaryngology-Head & Neck Surgery*, vol. 49, no. 1, pp. 1–10.

Mikolov, T., Chen, K., Corrado, G. & Dean, J., 2013, 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781*.

Mir, T. A. & Ausloos, M., 2018, 'Benford's law: A "sleeping beauty" sleeping in the dirty pages of logarithmic tables', *Journal of the Association for Information Science and Technology*, vol. 69, no. 3, pp. 349–358.

Mockett, A. A., Cavanagh, D. & Brown, T. D. K., 1984, 'Monoclonal antibodies to the s1 spike and membrane proteins of avian infectious bronchitis coronavirus strain massachusetts m41', *Journal of General Virology*, vol. 65, no. 12, pp. 2281–2286.

Müller, V. C., 2020, 'Ethics of artificial intelligence and robotics', https://plato.stanford.edu/entries/ethics-ai/, accessed: 2020-09-30.

Muninger, M.-I., Hammedi, W. & Mahr, D., 2019, 'The value of social media for innovation: A capability perspective', *Journal of Business Research*, vol. 95, pp. 116–127.

Muniyappa, R. & Gubbi, S., 2020, 'Covid-19 pandemic, coronaviruses, and diabetes mellitus', *American Journal of Physiology-Endocrinology and Metabolism*.

Nasab, F. R. & Rahim, F., 2020, 'Bibliometric analysis of global scientific research on sars-cov-2 (covid-19)', *MedRxiv*.

Neirotti, P. & Pesce, D., 2019, 'Ict-based innovation and its competitive outcome: The role of information intensity', *European Journal of Innovation Management*.

Newburger, D. E. & Bulyk, M. L., 2009, 'Uniprobe: an online database of protein binding microarray data on protein–dna interactions', *Nucleic acids research*, vol. 37, no. suppl_1, pp. D77–D82.

Nicolescu, R., Huth, M., Radanliev, P. & De Roure, D., 2018, ‘Mapping the values of iot’, *Journal of Information Technology*, vol. 33, no. 4, pp. 345–360.

North, K., Aramburu, N. & Lorenzo, O. J., 2019, ‘Promoting digitally enabled growth in smes: A framework proposal’, *Journal of Enterprise Information Management*.

Nussbaumer-Streit, B., Mayr, V., Dobrescu, A. I., Chapman, A., Persad, E., Klerings, I., Wagner, G., Siebert, U., Ledinger, D. & Zachariah, C., 2020, ‘Quarantine alone or in combination with other public health measures to control covid-19: A rapid review’, *Cochrane Database of Systematic Reviews*, vol. 9.

Nylén, D. & Holmström, J., 2015, ‘Digital innovation strategy: A framework for diagnosing and improving digital product and service innovation’, *Business Horizons*, vol. 58, no. 1, pp. 57–67.

Olesen, M. S., Bentzen, B. H., Nielsen, J. B., Steffensen, A. B., David, J.-P., Jabbari, J., Jensen, H. K., Haunsø, S., Svendsen, J. H. & Schmitt, N., 2012, ‘Mutations in the potassium channel subunit kcne1 are associated with early-onset familial atrial fibrillation’, *BMC Medical Genetics*, vol. 13, no. 1, pp. 1–9.

Onan, A. & Korukoğlu, S., 2017, ‘A feature selection model based on genetic rank aggregation for text sentiment classification’, *Journal of Information Science*, vol. 43, no. 1, pp. 25–38.

Opap, K. & Mulder, N., 2017, ‘Recent advances in predicting gene-disease associations’, *F1000Research*, vol. 6, pp. 578–578, <https://pubmed.ncbi.nlm.nih.gov/28529714https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5414807/>.

Ott, J., 1999, *Analysis of human genetic linkage*, JHU Press.

Pagoropoulos, A., Maier, A. & McAloone, T. C., 2017, ‘Assessing transformational change from institutionalising digital capabilities on implementation and development of product-service systems: Learnings from the maritime industry’, *Journal of Cleaner Production*, vol. 166, pp. 369–380.

Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A. D., Rawlik, K., Pasko, D., Walker, S., Parkinson, N., Fourman, M. H. & Russell, C. D., 2021, ‘Genetic mechanisms of critical illness in covid-19’, *Nature*, vol. 591, no. 7848, pp. 92–98.

Pal, R., Bhadada, S. K. & Misra, A., 2021, ‘Covid-19 vaccination in patients with diabetes mellitus: Current concepts, uncertainties and challenges’, *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 15, no. 2, pp. 505–508.

Palla, G., Derényi, I., Farkas, I. & Vicsek, T., 2005, ‘Uncovering the overlapping community structure of complex networks in nature and society’, *Nature*, vol. 435, no. 7043, pp. 814–818.

Palla, G., Tibély, G., Mones, E., Pollner, P. & Vicsek, T., 2015, ‘Hierarchical networks of scientific journals’, *Palgrave Communications*, vol. 1, no. 1, pp. 1–9.

Pan, Y., Wang, Y. & Wang, Y., 2020, ‘Investigation of causal effect of atrial fibrillation on alzheimer disease: A mendelian randomization study’, *Journal of the American Heart Association*, vol. 9, no. 2, p. e014889.

Pandey, A. C., Rajpoot, D. S. & Saraswat, M., 2017, ‘Twitter sentiment analysis using hybrid cuckoo search method’, *Information Processing & Management*, vol. 53, no. 4, pp. 764–779.

Pang, J., Wang, M. X., Ang, I. Y. H., Tan, S. H. X., Lewis, R. F., Chen, J. I.-P., Gutierrez, R. A., Gwee, S. X. W., Chua, P. E. Y. & Yang, Q., 2020, ‘Potential rapid diagnostics, vaccine and therapeutics for 2019 novel coronavirus

(2019-ncov): A systematic review', *Journal of Clinical Medicine*, vol. 9, no. 3, p. 623.

Park, I. & Yoon, B., 2018, 'Technological opportunity discovery for technological convergence based on the prediction of technology knowledge flow in a citation network', *Journal of Informetrics*, vol. 12, no. 4, pp. 1199–1222.

Park, Y. & Mithas, S., 2020, 'Organized complexity of digital business strategy: A configurational perspective', *MIS Quarterly*, vol. 44, no. 1.

Parkinson, N., Rodgers, N., Fourman, M. H., Wang, B., Zechner, M., Swets, M. C., Millar, J. E., Law, A., Russell, C. D. & Baillie, J. K., 2020, 'Dynamic data-driven meta-analysis for prioritisation of host genes implicated in covid-19', *Scientific Reports*, vol. 10, no. 1, pp. 1–12.

Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K. & Fortunato, S., 2015, 'Attention decay in science', *Journal of Informetrics*, vol. 9, no. 4, pp. 734–745.

Peixoto, T. P., 2014, 'Hierarchical block structures and high-resolution model selection in large networks', *Physical Review X*, vol. 4, no. 1, p. 011047.

Pera, M. S. & Ng, Y., 2018, 'Recommending books to be exchanged online in the absence of wish lists', *Journal of the Association for Information Science and Technology*, vol. 69, no. 4, pp. 541–552.

Pershina, R., Soppe, B. & Thune, T. M., 2019, 'Bridging analog and digital expertise: Cross-domain collaboration and boundary-spanning tools in the creation of digital innovation', *Research Policy*, vol. 48, no. 9, p. 103819.

Pessot, E., Zangiacomi, A., Battistella, C., Rocchi, V., Sala, A. & Sacco, M., 2020, 'What matters in implementing the factory of the future', *Journal of Manufacturing Technology Management*.

Petriĕ, I., Urbanĕiĕ, T., Cestnik, B. & Macedoni-Lukšiĕ, M., 2009, ‘Literature mining method rajolink for uncovering relations between biomedical concepts’, *Journal of Biomedical Informatics*, vol. 42, no. 2, pp. 219–227.

Petrosillo, N., Viceconte, G., Ergonul, O., Ippolito, G. & Petersen, E., 2020, ‘Covid-19, sars and mers: Are they closely related?’, *Clinical Microbiology and Infection*, vol. 26, no. 6, pp. 729–734.

Pica, N., Hai, R., Krammer, F., Wang, T. T., Maamary, J., Eggink, D., Tan, G. S., Krause, J. C., Moran, T. & Stein, C. R., 2012, ‘Hemagglutinin stalk antibodies elicited by the 2009 pandemic influenza virus as a mechanism for the extinction of seasonal h1n1 viruses’, *Proceedings of the National Academy of Sciences*, vol. 109, no. 7, pp. 2573–2578.

Pierce, J. B., Simion, V., Icli, B., Pérez-Cremades, D., Cheng, H. S. & Feinberg, M. W., 2020, ‘Computational analysis of targeting sars-cov-2, viral entry proteins ace2 and tmprss2, and interferon genes by host micrornas’, *Genes*, vol. 11, no. 11, p. 1354.

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F. & Furlong, L. I., 2016, ‘Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants’, *Nucleic acids research*, p. gkw943.

Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X. & Jensen, L. J., 2015, ‘Diseases: Text mining and data integration of disease–gene associations’, *Methods*, vol. 74, pp. 83–89.

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Marc, G. P., Moreira, E. D. & Zerbini, C., 2020, ‘Safety and efficacy of the bnt162b2 mrna covid-19 vaccine’, *New England Journal of Medicine.*

Ponzetto, S. P. & Strube, M., 2007, ‘Deriving a large scale taxonomy from wikipedia’, *AAAI*, , vol. 7pp. 1440–1445.

Porter, A. L., Zhang, Y., Huang, Y. & Wu, M., 2020, ‘Tracking and mining the covid-19 research literature’, *Frontiers in Research Metrics and Analytics*, vol. 5, p. 12.

Pourhatami, A., Kaviyani-Charati, M., Kargar, B., Baziyad, H., Kargar, M. & Olmeda-Gómez, C., 2021, ‘Mapping the intellectual structure of the coronavirus field (2000–2020): A co-word analysis’, *Scientometrics*, vol. 126, no. 8, pp. 6625–6657.

Powers, A. C., Aronoff, D. M. & Eckel, R. H., 2021, ‘Covid-19 vaccine prioritisation for type 1 and type 2 diabetes’, *The Lancet Diabetes & Endocrinology*, vol. 9, no. 3, pp. 140–141.

Prasad, K., Khatoon, F., Rashid, S., Ali, N., AlAsmari, A. F., Ahmed, M. Z., Alqahtani, A. S., Alqahtani, M. S. & Kumar, V., 2020, ‘Targeting hub genes and pathways of innate immune response in covid-19: A network biology perspective’, *International Journal of Biological Macromolecules*, vol. 163, pp. 1–8.

Preiss, J., Stevenson, M. & Gaizauskas, R., 2015, ‘Exploring relation types for literature-based discovery’, *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 987–992.

Price, D. J., 1986, *Little science, big science... and beyond*, vol. 480, Columbia University Press New York.

Price, W. N., Gerke, S. & Cohen, I. G., 2019, ‘Potential liability for physicians using artificial intelligence’, *JAMA*, vol. 322, no. 18, pp. 1765–1766.

Pritchard, A., 1969, ‘Statistical bibliography or bibliometrics’, *Journal of Documentation*, vol. 25, p. 348.

Prompetchara, E., Ketloy, C. & Palaga, T., 2020, ‘Immune responses in covid-19 and potential vaccines: Lessons learned from sars and mers epidemic’, *Asian Pacific Journal of Allergy and Immunology*, vol. 38, no. 1, pp. 1–9.

Pyysalo, S., Baker, S., Ali, I., Haselwimmer, S., Shah, T., Young, A., Guo, Y., Högberg, J., Stenius, U., Narita, M. et al., 2019, ‘Lion lbd: a literature-based discovery system for cancer biology’, *Bioinformatics*, vol. 35, no. 9, pp. 1553–1561.

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*, vol. 14, no. 3, p. 101047.

Raghavan, U. N., Albert, R. & Kumara, S., 2007, ‘Near linear time algorithm to detect community structures in large-scale networks’, *Physical Review E*, vol. 76, no. 3, p. 036106.

Ravikumar, S., Agrahari, A. & Singh, S. N., 2015, ‘Mapping the intellectual structure of scientometrics: A co-word analysis of the journal scientometrics (2005–2010)’, *Scientometrics*, vol. 102, no. 1, pp. 929–955.

Ravindra, N., Sehanobish, A., Pappalardo, J. L., Hafler, D. A. & van Dijk, D., 2020, ‘Disease state prediction from single-cell data using graph attention networks’, *Proceedings of the ACM conference on health, inference, and learning*, pp. 121–130.

Reimers, N. & Gurevych, I., 2019, ‘Sentence-bert: Sentence embeddings using siamese bert-networks’, *arXiv preprint arXiv:1908.10084*.

Reis, J., Amorim, M., Melão, N. & Matos, P., 2018, 'Digital transformation: A literature review and guidelines for future research', *World Conference on Information Systems and Technologies*, Springer, pp. 411–421.

Remuzzi, A. & Remuzzi, G., 2020, 'Covid-19 and italy: What next?', *The Lancet*, vol. 395, no. 10231, pp. 1225–1228.

Ren, H. & Zhao, Y., 2021, 'Technology opportunity discovery based on constructing, evaluating, and searching knowledge networks', *Technovation*, vol. 101, p. 102196.

Rodriguez, A. & Laio, A., 2014, 'Clustering by fast search and find of density peaks', *Science*, vol. 344, no. 6191, pp. 1492–1496.

Rodríguez-García, M. Á., Valencia-García, R., Colomo-Palacios, R. & Gómez-Berbís, J. M., 2019, 'Blinddate recommender: A context-aware ontology-based dating recommendation platform', *Journal of Information Science*, vol. 45, no. 5, pp. 573–591.

Rong, X., 2014, 'word2vec parameter learning explained', *arXiv preprint arXiv:1411.2738*.

Roselli, C., Rienstra, M. & Ellinor, P. T., 2020, 'Genetics of atrial fibrillation in 2020: Gwas, genome sequencing, polygenic risk, and beyond', *Circulation Research*, vol. 127, no. 1, pp. 21–33.

Rosvall, M. & Bergstrom, C. T., 2011, 'Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems', *PloS One*, vol. 6, no. 4, p. e18209.

Ruiz-Castillo, J. & Costas, R., 2014, 'The skewness of scientific productivity', *Journal of Informetrics*, vol. 8, no. 4, pp. 917–934.

Ruktanonchai, N. W., Floyd, J., Lai, S., Ruktanonchai, C. W., Sadilek, A., Rente-Lourenco, P., Ben, X., Carioli, A., Gwinn, J. & Steele, J., 2020, 'Assessing the impact of coordinated covid-19 exit strategies across europe', *Science*, vol. 369, no. 6510, pp. 1465–1470.

Ruthven, I., Buchanan, S. & Jardine, C., 2018, 'Relationships, environment, health and development: The information needs expressed online by young first-time mothers', *Journal of the Association for Information Science and Technology*, vol. 69, no. 8, pp. 985–995.

Saldanha, T., Mithas, S. & Krishnan, M. S., 2017, 'Leveraging customer involvement for fueling innovation: The role of relational and analytical information processing capabilities', *MIS Quarterly*, vol. 41, no. 1, pp. 267–286.

Sallenave, J.-M. & Guillot, L., 2020, 'Innate immune signaling and proteolytic pathways in the resolution or exacerbation of sars-cov-2 in covid-19: key therapeutic targets?', *Frontiers in Immunology*, vol. 11.

Salton, G. & McGill, M. J., 1986, *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc.

Sato, Y., Honda, Y. & Jun, I., 2010, 'Long-term oral anticoagulation therapy and the risk of hip fracture in patients with previous hemispheric infarction and nonrheumatic atrial fibrillation', *Cerebrovascular Diseases*, vol. 29, no. 1, pp. 73–78.

Savransky, S. D., 2000, *Engineering of creativity: Introduction to TRIZ methodology of inventive problem solving*, CRC press.

Schaub, M. T. & Peel, L., 2020, 'Hierarchical community structure in networks', *arXiv preprint arXiv:2009.07196*.

Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. Ø., 2010, ‘Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions’, *BMC bioinformatics*, vol. 11, pp. 1–10.

Schlicker, A., Lengauer, T. & Albrecht, M., 2010, ‘Improving disease gene prioritization using the semantic similarity of gene ontology terms’, *Bioinformatics*, vol. 26, no. 18, pp. i561–i567.

Schwab, K., 2017, *The fourth industrial revolution*, Currency.

Seitner, J., Bizer, C., Eckert, K., Faralli, S., Meusel, R., Paulheim, H. & Ponzetto, S. P., 2016, ‘A large database of hypernymy relations extracted from the web’, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pp. 360–367.

Setia, P., Setia, P., Venkatesh, V. & Joglekar, S., 2013, ‘Leveraging digital technologies: How information quality leads to localized capabilities and customer service performance’, *MIS Quarterly*, pp. 565–590.

Shaath, H., Vishnubalaji, R., Elkord, E. & Alajez, N. M., 2020, ‘Single-cell transcriptome analysis highlights a role for neutrophils and inflammatory macrophages in the pathogenesis of severe covid-19’, *Cells*, vol. 9, no. 11, p. 2374.

Shams, M. & Baraani-Dastjerdi, A., 2017, ‘Enriched lda (elda): Combination of latent dirichlet allocation with word co-occurrence analysis for aspect extraction’, *Expert Systems with Applications*, vol. 80, pp. 136–146.

Shang, J., Liu, J., Jiang, M., Ren, X., Voss, C. R. & Han, J., 2018, ‘Automated phrase mining from massive text corpora’, *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1825–1837.

Shang, J., Zhang, X., Liu, L., Li, S. & Han, J., 2020, ‘Nettaxo: Automated topic taxonomy construction from text-rich network', *Proceedings of the Web Conference 2020*, pp. 1908–1919.

Shang, N., Xu, H., Rindflesch, T. C. & Cohen, T., 2014, 'Identifying plausible adverse drug reactions using knowledge extracted from the literature', *Journal of Biomedical Informatics*, vol. 52, pp. 293–310.

Sheikh, O., Vande Hei, A. G., Battisha, A., Hammad, T., Pham, S. & Chilton, R., 2019, 'Cardiovascular, electrophysiologic, and hematologic effects of omega-3 fatty acids beyond reducing hypertriglyceridemia: as it pertains to the recently published reduce-it trial', *Cardiovascular Diabetology*, vol. 18, no. 1, p. 84, <https://doi.org/10.1186/s12933-019-0887-0>.

Shen, X., Li, Y., Sun, Y., Chen, J. & Wang, F., 2019, 'Knowledge withholding in online knowledge spaces: Social deviance behavior and secondary control perspective', *Journal of the Association for Information Science and Technology*, vol. 70, no. 4, pp. 385–401.

Shen, Z., Ma, H. & Wang, K., 2018, 'A web-scale system for scientific knowledge exploration', *arXiv preprint arXiv:1805.12216*.

Shepherd, J. P., Moore, S. C., Long, A., Kollar, L. M. M. & Sumner, S. A., 2021, 'Association between covid-19 lockdown measures and emergency department visits for violence-related injuries in cardiff, wales', *JAMA*, vol. 325, no. 9, pp. 885–887.

Shi, G., Kenney, A. D., Kudryashova, E., Zani, A., Zhang, L., Lai, K. K., Hall-Stoodley, L., Robinson, R. T., Kudryashov, D. S. & Compton, A. A., 2021, 'Opposing activities of ifitm proteins in sars-cov-2 infection', *The EMBO Journal*, vol. 40, no. 3, p. e106501.

Shi, L., Lu, Z.-A., Que, J.-Y., Huang, X.-L., Liu, L., Ran, M.-S., Gong, Y.-M., Yuan, K., Yan, W. & Sun, Y.-K., 2020, ‘Prevalence of and risk factors associated with mental health symptoms among the general population in china during the coronavirus disease 2019 pandemic’, *JAMA Network Open*, vol. 3, no. 7, pp. e2014053–e2014053.

Shiau, W.-L., Dwivedi, Y. K. & Yang, H. S., 2017, ‘Co-citation and cluster analyses of extant literature on social networks’, *International Journal of Information Management*, vol. 37, no. 5, pp. 390–399.

Shin, D., 2019, ‘How do users experience the interaction with an immersive screen?’, *Computers in Human Behavior*, vol. 98, pp. 302–310.

Shin, D., Mukherjee, R., Grewe, D., Bojkova, D., Baek, K., Bhattacharya, A., Schulz, L., Widera, M., Mehdipour, A. R. & Tascher, G., 2020, ‘Papain-like protease regulates sars-cov-2 viral spread and innate immunity’, *Nature*, vol. 587, no. 7835, pp. 657–662.

Sikkema, R. S., Pas, S. D., Nieuwenhuijse, D. F., O’Toole, Á., Verweij, J., van der Linden, A., Chestakova, I., Schapendonk, C., Pronk, M., Lexmond, P. et al., 2020, ‘Covid-19 in health-care workers in three hospitals in the south of the netherlands: a cross-sectional study’, *The Lancet Infectious Diseases*, vol. 20, no. 11, pp. 1273–1280.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. & Wang, K., 2015, ‘An overview of microsoft academic service (mas) and applications’, *Proceedings of the 24th international conference on world wide web*, pp. 243–246.

Sinner, M. F., Pfeufer, A., Akyol, M., Beckmann, B.-M., Hinterseer, M., Wacker, A., Perz, S., Sauter, W., Illig, T. & Näbauer, M., 2008, ‘The non-synonymous coding ikr-channel variant kcnh2-k897t is associated with atrial fibrillation:

Results from a systematic candidate gene-based analysis of kcnh2 (herg)', *European Heart Journal*, vol. 29, no. 7, pp. 907–914.

Sisodia, D. S., Verma, S. & Vyas, O. P., 2017, 'Augmented intuitive dissimilarity metric for clustering of web user sessions', *Journal of Information Science*, vol. 43, no. 4, pp. 480–491.

Small, H., 2010, 'Maps of science as interdisciplinary discourse: Co-citation contexts and the role of analogy', *Scientometrics*, vol. 83, no. 3, pp. 835–849.

Small, H., Boyack, K. W. & Klavans, R., 2014, 'Identifying emerging topics in science and technology', *Research Policy*, vol. 43, no. 8, pp. 1450–1467.

Song, J., Huang, Y., Qi, X., Li, Y., Li, F., Fu, K. & Huang, T., 2016, 'Discovering hierarchical topic evolution in time-stamped documents', *Journal of the Association for Information Science and Technology*, vol. 67, no. 4, pp. 915–927.

Song, L., Tso, G. & Fu, Y., 2019, 'Click behavior and link prioritization: Multiple demand theory application for web improvement', *Journal of the Association for Information Science and Technology*, vol. 70, no. 8, pp. 805–816.

Song, M., Kim, W. C., Lee, D., Heo, G. E. & Kang, K. Y., 2015, 'Pkde4j: Entity and relation extraction for public knowledge discovery', *Journal of Biomedical Informatics*, vol. 57, pp. 320–332.

Spinelli, A. & Pellino, G., 2020, 'Covid-19 pandemic: Perspectives on an unfolding crisis', *Journal of British Surgery*, vol. 107, no. 7, pp. 785–787.

Srivastava, S. C. & Shainesh, G., 2015, 'Bridging the service divide through digitally enabled service innovations: Evidence from indian healthcare service providers', *MIS Quarterly*, vol. 39, no. 1, pp. 245–267.

Stapley, B. J. & Benoit, G., 1999, ‘Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts’, *Biocomputing 2000*, World Scientific, pp. 529–540.

Starr, T. N., Greaney, A. J., Addetia, A., Hannon, W. W., Choudhary, M. C., Dingens, A. S., Li, J. Z. & Bloom, J. D., 2021, ‘Prospective mapping of viral mutations that escape antibodies used to treat covid-19’, *Science*, vol. 371, no. 6531, pp. 850–854.

Stenstrom, G., Gottsater, A., Bakhtadze, E., Berger, B. & Sundkvist, G., 2005, ‘Latent autoimmune diabetes in adults: definition, prevalence, β-cell function, and treatment’, *Diabetes*, vol. 54, no. suppl_2, pp. S68–S72.

Stevenson, J. M., Alexander, G. C., Palamuttam, N. & Mehta, H. B., 2021, ‘Projected utility of pharmacogenomic testing among individuals hospitalized with covid-19: A retrospective multicenter study in the united states’, *Clinical and Translational Science*, vol. 14, no. 1, pp. 153–162.

Stoeckli, E., Dremel, C. & Uebernickel, F., 2018, ‘Exploring characteristics and transformational capabilities of insurtech innovations to understand insurance value creation in a digital world’, *Electronic Markets*, vol. 28, no. 3, pp. 287–305.

Sugiyama, M., Kinoshita, N., Ide, S., Nomoto, H., Nakamoto, T., Saito, S., Ishikane, M., Kutsuna, S., Hayakawa, K. & Hashimoto, M., 2020, ‘Serum ccl17 level becomes a predictive marker to distinguish between mild/moderate and severe/critical disease in patients with covid-19’, *Gene*, vol. 766, p. 145145.

Sun, P. G., Miao, Q. & Staab, S., 2021, ‘Community-based k-shell decomposition for identifying influential spreaders’, *Pattern Recognition*, vol. 120, p. 108130.

Sun, Q., Wang, C., Zhou, Y., Zuo, L. & Tang, J., 2020a, ‘Dominant platform capability, symbiotic strategy and the construction of “internet+ weee collection" business ecosystem: A comparative study of two typical cases in china', *Journal of Cleaner Production*, vol. 254, p. 120074.

Sun, Z., Chen, M., Hu, W., Wang, C., Dai, J. & Zhang, W., 2020b, 'Knowledge association with hyperbolic knowledge graph embeddings', *arXiv preprint arXiv:2010.02162*.

Sun, Z., Wang, C., Hu, W., Chen, M., Dai, J., Zhang, W. & Qu, Y., 2020c, 'Knowledge graph alignment network with gated multi-hop neighborhood aggregation', *Proceedings of the AAAI Conference on Artificial Intelligence*, , vol. 34pp. 222–229.

Suominen, A. & Toivanen, H., 2016, 'Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification', *Journal of the Association for Information Science and Technology*, vol. 67, no. 10, pp. 2464–2476.

Swanson, D. R., 1986, 'Fish oil, raynaud's syndrome, and undiscovered public knowledge', *Perspectives in Biology and Medicine*, vol. 30, no. 1, pp. 7–18.

Szalavetz, A., 2019, 'Industry 4.0 and capability development in manufacturing subsidiaries', *Technological Forecasting and Social Change*, vol. 145, pp. 384–395.

Tan, S., Chen, W., Xiang, H., Kong, G., Zou, L. & Wei, L., 2021, 'Screening druggable targets and predicting therapeutic drugs for covid-19 via integrated bioinformatics analysis', *Genes & Genomics*, vol. 43, no. 1, pp. 55–67.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. & Su, Z., 2008, 'Arnetminer: extraction and mining of academic social networks', *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 990–998.

Taylor, P. C., Adams, A. C., Hufford, M. M., De La Torre, I., Winthrop, K. & Gottlieb, R. L., 2021, 'Neutralizing monoclonal antibodies for treatment of covid-19', *Nature Reviews Immunology*, vol. 21, no. 6, pp. 382–393.

Teece, D. J., 2007, 'Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance', *Strategic Management Journal*, vol. 28, no. 13, pp. 1319–1350.

Teece, D. J., Pisano, G. & Shuen, A., 1997, 'Dynamic capabilities and strategic management', *Strategic Management Journal*, vol. 18, no. 7, pp. 509–533.

Thelwall, M., 2018a, 'Dimensions: A competitor to scopus and the web of science?', *Journal of Informetrics*, vol. 12, no. 2, pp. 430–435.

Thelwall, M., 2018b, 'Do females create higher impact research? scopus citations and mendeley readers for articles from five countries', *Journal of Informetrics*, vol. 12, no. 4, pp. 1031–1041.

Thomas, S. J., Moreira Jr, E. D., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Polack, F. P. & Zerbini, C., 2021, 'Safety and efficacy of the bnt162b2 mrna covid-19 vaccine through 6 months', *New England Journal of Medicine*, vol. 385, no. 19, pp. 1761–1773.

Tong, H., Faloutsos, C. & Pan, J.-Y., 2008, 'Random walk with restart: Fast solutions and applications', *Knowledge and Information Systems*, vol. 14, no. 3, pp. 327–346.

Tortorella, G. L., Vergara, A. M. C., Garza-Reyes, J. A. & Sawhney, R., 2020, 'Organizational learning paths based upon industry 4.0 adoption: An empirical study with brazilian manufacturers', *International Journal of Production Economics*, vol. 219, pp. 284–294.

Trabucchi, D. & Buganza, T., 2019, ‘Data-driven innovation: Switching the perspective on big data’, *European Journal of Innovation Management*.

Traggiai, E., Becker, S., Subbarao, K., Kolesnikova, L., Uematsu, Y., Gismondo, M. R., Murphy, B. R., Rappuoli, R. & Lanzavecchia, A., 2004, ‘An efficient method to make human monoclonal antibodies from memory b cells: Potent neutralization of sars coronavirus’, *Nature Medicine*, vol. 10, no. 8, pp. 871–875.

Tran, B. X., Ha, G. H., Nguyen, L. H., Vu, G. T., Hoang, M. T., Le, H. T., Latkin, C. A., Ho, C. S. & Ho, R., 2020, ‘Studies of novel coronavirus disease 19 (covid-19) pandemic: A global analysis of literature’, *International Journal of Environmental Research and Public Health*, vol. 17, no. 11, p. 4095.

Trantopoulos, K., von Krogh, G., Wallin, M. W. & Woerter, M., 2017, ‘External knowledge and information technology: Implications for process innovation performance’, *MIS Quarterly*, vol. 41, no. 1, pp. 287–300.

Trewartha, A., Dagdelen, J., Huo, H., Cruse, K., Wang, Z., He, T., Subramanian, A., Fei, Y., Justus, B. & Persson, K., 2020, ‘Covidscholar: An automated covid-19 research aggregation and analysis platform’, *arXiv preprint arXiv:2012.03891*.

Tsao, S.-F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L. & Butt, Z. A., 2021, ‘What social media told us in the time of covid-19: A scoping review’, *The Lancet Digital Health*, vol. 3, no. 3, pp. e175–e194.

Tsourikov, V. M., Batchilo, L. S. & Sovpel, I. V., 2000, ‘Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (sao) structures’, US Patent 6,167,370.

Tumelero, C., Sbragia, R. & Evans, S., 2019, ‘Cooperation in r & d and eco-innovations: The role in companies' socioeconomic performance', *Journal of Cleaner Production*, vol. 207, pp. 1138–1149.

Ukko, J., Nasiri, M., Saunila, M. & Rantala, T., 2019, 'Sustainability strategy as a moderator in the relationship between digital business strategy and financial performance', *Journal of Cleaner Production*, vol. 236, p. 117626.

Uçar, E., Uzun, E. & Tüfekci, P., 2017, 'A novel algorithm for extracting the user reviews from web pages', *Journal of Information Science*, vol. 43, no. 5, pp. 696–712.

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., Cau, P., Remy, E. & Baudot, A., 2019, 'Random walk with restart on multiplex and heterogeneous biological networks', *Bioinformatics*, vol. 35, no. 3, pp. 497–505.

Van Belle, T. L., Coppieters, K. T. & Von Herrath, M. G., 2011, 'Type 1 diabetes: Etiology, immunology, and therapeutic strategies', *Physiological Reviews*, vol. 91, no. 1, pp. 79–118.

Van Dam, S., Vosa, U., van der Graaf, A., Franke, L. & de Magalhaes, J. P., 2018, 'Gene co-expression analysis for functional classification and gene–disease predictions', *Briefings in Bioinformatics*, vol. 19, no. 4, pp. 575–592.

Van Eck, N. J. & Waltman, L., 2010, 'Software survey: Vosviewer, a computer program for bibliometric mapping', *Scientometrics*, vol. 84, no. 2, pp. 523–538.

Velden, T., Boyack, K. W., Gläser, J., Koopman, R., Scharnhorst, A. & Wang, S., 2017, 'Comparison of topic extraction approaches and their results', *Scientometrics*, vol. 111, no. 2, pp. 1169–1221.

Venter, Z. S., Aunan, K., Chowdhury, S. & Lelieveld, J., 2020, 'Covid-19 lockdowns cause global air pollution declines', *Proceedings of the National Academy of Sciences*, vol. 117, no. 32, pp. 18984–18990.

Vial, G., 2019, ‘Understanding digital transformation: A review and a research agenda’, *The Journal of Strategic Information Systems*, vol. 28, no. 2, pp. 118–144.

Vicente-Gomila, J. M., 2014, ‘The contribution of syntactic–semantic approach to the search for complementary literatures for scientific or technical discovery’, *Scientometrics*, vol. 100, no. 3, pp. 659–673.

Vishnubalaji, R., Shaath, H. & Alajez, N. M., 2020, ‘Protein coding and long noncoding rna (lncrna) transcriptional landscape in sars-cov-2 infected bronchial epithelial cells highlight a role for interferon and inflammatory response’, *Genes*, vol. 11, no. 7, p. 760.

Wagner, C. S., Cai, X., Zhang, Y. & Fry, C. V., 2022, ‘One-year in: Covid-19 research at the international level in cord-19 data’, *Plos One*, vol. 17, no. 5, p. e0261624.

Walter, A., Finger, R., Huber, R. & Buchmann, N., 2017, ‘Smart farming is key to developing sustainable agriculture’, *Proceedings of the National Academy of Sciences*, vol. 114, no. 24, pp. 6148–6150.

Waltman, L. & Van Eck, N. J., 2013, ‘A smart local moving algorithm for large-scale modularity-based community detection’, *The European Physical Journal B*, vol. 86, no. 11, p. 471.

Wang, C., Danilevsky, M., Desai, N., Zhang, Y., Nguyen, P., Taula, T. & Han, J., 2013, ‘A phrase mining framework for recursive construction of a topical hierarchy’, *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 437–445.

Wang, C., Liu, J., Desai, N., Danilevsky, M. & Han, J., 2015a, ‘Constructing topical hierarchies in heterogeneous information networks', *Knowledge and Information Systems*, vol. 44, no. 3, pp. 529–558.

Wang, C., Pan, R., Wan, X., Tan, Y., Xu, L., McIntyre, R. S., Choo, F. N., Tran, B., Ho, R. & Sharma, V. K., 2020a, 'A longitudinal study on the mental health of general population during the covid-19 epidemic in china', *Brain, Behavior, and Immunity*, vol. 87, pp. 40–48.

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z. & Xiong, Y., 2020b, 'Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china', *JAMA*, vol. 323, no. 11, pp. 1061–1069.

Wang, J., Liu, Q., Yuan, S., Xie, W., Liu, Y., Xiang, Y., Wu, N., Wu, L., Ma, X. & Cai, T., 2017, 'Genetic predisposition to lung cancer: comprehensive literature integration, meta-analysis, and multiple evidence assessment of candidate-gene association studies', *Scientific Reports*, vol. 7, no. 1, pp. 1–13.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Eide, D., Dong, Y., Qian, J., Kanakia, A., Chen, A. & Rogahn, R., 2019a, 'A review of microsoft academic services for science of science studies', *Frontiers in Big Data*, vol. 2, p. 45.

Wang, R. D. & Miller, C. D., 2020, 'Complementors' engagement in an ecosystem: A study of publishers'e-book offerings on amazon kindle', *Strategic Management Journal*, vol. 41, no. 1, pp. 3–26.

Wang, W., Mahmood, A., Sismeiro, C. & Vulkan, N., 2019b, 'The evolution of equity crowdfunding: Insights from co-investments of angels and the crowd', *Research Policy*, vol. 48, no. 8, p. 103727.

Wang, X., Nie, Y., Ning, S., Shi, Y., Zhao, Y., Niu, S., Guo, C., Meng, X. & Yuan, Y., 2018, 'Rs17042171 at chromosome 4q25 is associated with atrial fibrillation in the chinese han population from the central plains', *Journal of Central South University. Medical Sciences*, vol. 43, no. 6, p. 594.

Wang, Y., Wang, Y., Chen, Y. & Qin, Q., 2020c, 'Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (covid-19) implicate special control measures', *Journal of Medical Virology*, vol. 92, no. 6, pp. 568–576.

Wang, Z., Zhao, H. & Wang, Y., 2015b, 'Social networks in marketing research 2001–2014: A co-word analysis', *Scientometrics*, vol. 105, no. 1, pp. 65–82.

Warner, K. S. & Wäger, M., 2019, 'Building dynamic capabilities for digital transformation: An ongoing process of strategic renewal', *Long Range Planning*, vol. 52, no. 3, pp. 326–349.

Wartena, C. & Brussee, R., 2008, 'Topic detection by clustering keywords', *2008 19th International Workshop on Database and Expert Systems Applications*, IEEE, pp. 54–58.

Wartena, C., Brussee, R. & Slakhorst, W., 2010, 'Keyword extraction using word co-occurrence', *2010 workshops on database and expert systems applications*, IEEE, pp. 54–58.

Watts, R. J. & Porter, A. L., 1999, 'Mining foreign language information resources', *PICMET'99: Portland International Conference on Management of Engineering and Technology. Proceedings Vol-1: Book of Summaries (IEEE Cat. No. 99CH36310)*, IEEE, pp. 198–206.

Watts, R. J., Porter, A. L. & Courseault, C., 1999, 'Functional analysis: Deriving systems knowledge from bibliographic information resources', *Information Knowledge Systems Management*, vol. 1, no. 1, pp. 45–61.

Webber, B. L., Raghu, S. & Edwards, O. R., 2015, ‘Is crispr-based gene drive a biocontrol silver bullet or global conservation threat?’, *Proceedings of the National Academy of Sciences*, vol. 112, no. 34, pp. 10565–10567.

Wei, C.-H., Allot, A., Leaman, R. & Lu, Z., 2019, ‘Pubtator central: automated concept annotation for biomedical full text articles’, *Nucleic Acids Research*, vol. 47, no. W1, pp. W587–W593.

Wei, C.-H., Kao, H.-Y. & Lu, Z., 2013, ‘Pubtator: a web-based text mining tool for assisting biocuration’, *Nucleic Acids Research*, vol. 41, no. W1, pp. W518–W522.

Wenzlau, J. M., Juhl, K., Yu, L., Moua, O., Sarkar, S. A., Gottlieb, P., Rewers, M., Eisenbarth, G. S., Jensen, J. & Davidson, H. W., 2007, ‘The cation efflux transporter znt8 (slc30a8) is a major autoantigen in human type 1 diabetes’, *Proceedings of the National Academy of Sciences*, vol. 104, no. 43, pp. 17040–17045.

Williams, S. N., Armitage, C. J., Tampe, T. & Dienes, K., 2020, ‘Public perceptions and experiences of social distancing and social isolation during the covid-19 pandemic: A uk-based focus group study’, *BMJ Open*, vol. 10, no. 7, p. e039334.

Wise, C., Ioannidis, V. N., Calvo, M. R., Song, X., Price, G., Kulkarni, N., Brand, R., Bhatia, P. & Karypis, G., 2020, ‘Covid-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature’, *arXiv preprint arXiv:2007.12731*.

Wong, W., Liu, W. & Bennamoun, M., 2012, ‘Ontology learning from text: A look back and into the future’, *ACM Computing Surveys (CSUR)*, vol. 44, no. 4, pp. 1–36.

Wrobel, A. G., Benton, D. J., Xu, P., Roustan, C., Martin, S. R., Rosenthal, P. B., Skehel, J. J. & Gamblin, S. J., 2020, 'Sars-cov-2 and bat ratg13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects', *Nature Structural & Molecular Biology*, vol. 27, no. 8, pp. 763–767.

Wu, H.-l., Huang, J., Zhang, C. J., He, Z. & Ming, W.-K., 2020, 'Facemask shortage and the novel coronavirus disease (covid-19) outbreak: Reflections on public health measures', *EClinicalMedicine*, vol. 21, p. 100329.

Wu, M., Kozanoglu, D. C., Min, C. & Zhang, Y., 2021a, 'Unraveling the capabilities that enable digital transformation: A data-driven methodology and the case of artificial intelligence', *Advanced Engineering Informatics*, vol. 50, p. 101368.

Wu, M. & Zhang, Y., 2021, 'Hierarchical topic tree: A hybrid model comprising network analysis and density peak search', *18th International Conference on Scientometrics and Informetrics (ISSI)*, INT SOC SCIENTOMETRICS & INFORMETRICS-ISSI, pp. 1241–1252.

Wu, M., Zhang, Y., Grosser, M., Tipper, S., Venter, D., Lin, H. & Lu, J., 2021b, 'Profiling covid-19 genetic research: A data-driven study utilizing intelligent bibliometrics', *Frontiers in Research Metrics and Analytics*, vol. 6, p. 30.

Wu, M., Zhang, Y., Markley, M., Cassidy, C., Newman, N. & Porter, A., 2023, 'Covid-19 knowledge deconstruction and retrieval: Solutions of intelligent bibliometrics', *Scientometrics*.

Wu, M., Zhang, Y., Zhang, G. & Lu, J., 2021c, 'Exploring the genetic basis of diseases through a heterogeneous bibliometric network: A methodology and case study', *Technological Forecasting and Social Change*, vol. 164, p. 120513.

Xia, S., Duan, K., Zhang, Y., Zhao, D., Zhang, H., Xie, Z., Li, X., Peng, C., Zhang, Y. & Zhang, W., 2020, 'Effect of an inactivated vaccine against sars-cov-2 on safety and immunogenicity outcomes: Interim analysis of 2 randomized clinical trials', *JAMA*, vol. 324, no. 10, pp. 951–960.

Xiao, L., Chen, G., Sun, J., Han, S. & Zhang, C., 2016, 'Exploring the topic hierarchy of digital library research in china using keyword networks: a k-core decomposition approach', *Scientometrics*, vol. 108, no. 3, pp. 1085–1101.

Xie, B., He, D., Mercer, T., Wang, Y., Wu, D., Fleischmann, K. R., Zhang, Y., Yoder, L. H., Stephens, K. K. & Mackert, M., 2020, 'Global health crises are also information crises: A call to action', *Journal of the Association for Information Science and Technology*, vol. 71, no. 12, pp. 1419–1423.

Xie, W.-H., Chang, C., Xu, Y.-J., Li, R.-G., Qu, X.-K., Fang, W.-Y., Liu, X. & Yang, Y.-Q., 2013, 'Prevalence and spectrum of nkx2. 5 mutations associated with idiopathic atrial fibrillation', *Clinics*, vol. 68, no. 6, pp. 777–784.

Xie, Z., Ouyang, Z., Li, J., Dong, E. & Yi, D., 2018, 'Modelling transition phenomena of scientific coauthorship networks', *Journal of the Association for Information Science and Technology*, vol. 69, no. 2, pp. 305–317.

Xiong, Y., Liu, Y., Cao, L., Wang, D., Guo, M., Jiang, A., Guo, D., Hu, W., Yang, J. & Tang, Z., 2020, 'Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in covid-19 patients', *Emerging Microbes & Infections*, vol. 9, no. 1, pp. 761–770.

Xu, J., Kim, S., Song, M., Jeong, M., Kim, D., Kang, J., Rousseau, J. F., Li, X., Xu, W. & Torvik, V. I., 2020, 'Building a pubmed knowledge graph', *Scientific Data*, vol. 7, no. 1, pp. 1–15.

Xu, Y., Yin, J., Huang, J. & Yin, Y., 2018, 'Hierarchical topic modeling with automatic knowledge mining', *Expert Systems with Applications*, vol. 103, pp. 106–117.

Yablonsky, S., 2020, 'A multidimensional platform ecosystem framework', *Kybernetes*.

Yamagishi, S.-i., 2019, 'Concerns about clinical efficacy and safety of warfarin in diabetic patients with atrial fibrillation', *Cardiovascular Diabetology*, vol. 18, no. 1, p. 12.

Yan, E. & Ding, Y., 2009, 'Applying centrality measures to impact analysis: A coauthorship network analysis', *Journal of the American Society for Information Science and Technology*, vol. 60, no. 10, pp. 2107–2118.

Yan, E. & Ding, Y., 2012, 'Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other', *Journal of the American Society for Information Science and Technology*, vol. 63, no. 7, pp. 1313–1326.

Yan, E. & Guns, R., 2014, 'Predicting and recommending collaborations: An author-, institution-, and country-level analysis', *Journal of Informetrics*, vol. 8, no. 2, pp. 295–309.

Yang, C., Park, H. & Heo, J., 2010, 'A network analysis of interdisciplinary research relationships: the korean government's r&d grant program', *Scientometrics*, vol. 83, no. 1, pp. 77–92.

Yang, S., Zou, L., Wang, Z., Yan, J. & Wen, J.-R., 2017, 'Efficiently answering technical questions—a knowledge graph approach', *Thirty-First AAAI Conference on Artificial Intelligence*, pp. 3111–3118.

Yang, X., An, N., Zhong, C., Guan, M., Jiang, Y., Li, X., Zhang, H., Wang, L., Ruan, Y. & Gao, Y., 2020a, 'Enhanced cardiomyocyte reactive oxygen species signaling promotes ibrutinib-induced atrial fibrillation', *Redox Biology*, vol. 30, p. 101432.

Yang, X., Cao, D., Chen, J., Xiao, Z. & Daowd, A., 2020b, 'Ai and iot-based collaborative business ecosystem: A case in chinese fish farming industry', *International Journal of Technology Management*, vol. 82, no. 2, pp. 151–171.

Yau, C.-K., Porter, A., Newman, N. & Suominen, A., 2014, 'Clustering scientific documents with topic modeling', *Scientometrics*, vol. 100, no. 3, pp. 767–786.

Yeh, J., Wu, C. & Chen, M., 2008, 'Ontology-based speech act identification in a bilingual dialog system using partial pattern trees', *Journal of the American Society for Information Science and Technology*, vol. 59, no. 5, pp. 684–694.

Yeung, C., Enriquez, A., Suarez-Fuster, L. & Baranchuk, A., 2019, 'Atrial fibrillation in patients with inherited cardiomyopathies', *Ep Europace*, vol. 21, no. 1, pp. 22–32.

Young, A., Selander, L. & Vaast, E., 2019, 'Digital organizing for social impact: Current insights and future research avenues on collective action, social movements, and digital technologies', *Information and Organization*, vol. 29, no. 3, p. 100257.

Yu, Q., Ding, Y., Song, M., Song, S., Liu, J. & Zhang, B., 2015, 'Tracing database usage: Detecting main paths in database link networks', *Journal of Informetrics*, vol. 9, no. 1, pp. 1–15.

Yu, Q., Wang, Q., Zhang, Y., Chen, C., Ryu, H., Park, N., Baek, J.-E., Li, K., Wu, Y. & Li, D., 2021, 'Analyzing knowledge entities about covid-19 using entitymetrics', *Scientometrics*, vol. 126, no. 5, pp. 4491–4509.

Yu, X., Tsibane, T., McGraw, P. A., House, F. S., Keefer, C. J., Hicar, M. D., Tumpey, T. M., Pappas, C., Perrone, L. A. & Martinez, O., 2008, 'Neutralizing antibodies derived from the b cells of 1918 influenza pandemic survivors', *Nature*, vol. 455, no. 7212, pp. 532–536.

Yu, Y., Li, Y., Shen, J., Feng, H., Sun, J. & Zhang, C., 2020, 'Steam: Self-supervised taxonomy expansion with mini-paths', *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1026–1035.

Yuan, M., Huang, D., Lee, C.-C. D., Wu, N. C., Jackson, A. M., Zhu, X., Liu, H., Peng, L., Van Gils, M. J. & Sanders, R. W., 2021, 'Structural and functional ramifications of antigenic drift in recent sars-cov-2 variants', *Science*, vol. 373, no. 6556, pp. 818–823.

Yuan, Y., Xu, H. & Wang, B., 2014, 'An improved nsga-iii procedure for evolutionary many-objective optimization', *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*, pp. 661–668.

Zeberg, H. & Paabo, S., 2020, 'A genetic variant protective for covid-19 is inherited from neandertals', *BioRxiv*.

Zhai, Y., Ding, Y. & Wang, F., 2018, 'Measuring the diffusion of an innovation: A citation analysis', *Journal of the Association for Information Science and Technology*, vol. 69, no. 3, pp. 368–379.

Zhang, C., Tao, F., Chen, X., Shen, J., Jiang, M., Sadler, B., Vanni, M. & Han, J., 2018a, 'Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering', *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2701–2709.

Zhang, E., Gupta, N., Nogueira, R., Cho, K. & Lin, J., 2020a, 'Rapidly deploying a neural search engine for the covid-19 open research dataset: Preliminary thoughts and lessons learned', *arXiv preprint arXiv:2004.05125*.

Zhang, G.-Q., Zhang, G.-Q., Yang, Q.-F., Cheng, S.-Q. & Zhou, T., 2008, 'Evolution of the internet and its cores', *New Journal of Physics*, vol. 10, no. 12, p. 123027.

Zhang, J., Xie, J., Hou, W., Tu, X., Xu, J., Song, F., Wang, Z. & Lu, Z., 2012, 'Mapping the knowledge structure of research on patient adherence: knowledge domain visualization based co-word analysis and social network analysis', *PloS One*, vol. 7, no. 4, p. e34497.

Zhang, Y., Ahmed, A., Josifovski, V. & Smola, A., 2014a, 'Taxonomy discovery for personalized recommendation', *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, pp. 243–252.

Zhang, Y., Cai, X., Fry, C. V., Wu, M. & Wagner, C. S., 2021a, 'Topic evolution, disruption and resilience in early covid-19 research', *Scientometrics*, vol. 126, no. 5, pp. 4225–4253.

Zhang, Y., Chen, H., Lu, J. & Zhang, G., 2017a, 'Detecting and predicting the topic change of knowledge-based systems: A topic-based bibliometric analysis from 1991 to 2016', *Knowledge-Based Systems*, vol. 133, pp. 255–268.

Zhang, Y., Guo, Y., Wang, X., Zhu, D. & Porter, A. L., 2013, 'A hybrid visualisation model for technology roadmapping: Bibliometrics, qualitative methodology and empirical study', *Technology Analysis & Strategic Management*, vol. 25, no. 6, pp. 707–724.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H. & Zhang, G., 2018b, 'Does deep learning help topic extraction? a kernel k-means clustering method with word embedding', *Journal of Informetrics*, vol. 12, no. 4, pp. 1099–1117.

Zhang, Y., Porter, A. L., Cunningham, S., Chiavetta, D. & Newman, N., 2020b, 'Parallel or intersecting lines? intelligent bibliometrics for investigating the involvement of data science in policy analysis', *IEEE Transactions on Engineering Management*, pp. 1–13.

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y. & Newman, N. C., 2014b, '"term clumping" for technical intelligence: A case study on dye-sensitized solar cells', *Technological Forecasting and Social Change*, vol. 85, pp. 26–39.

Zhang, Y., Qian, Y., Huang, Y., Guo, Y., Zhang, G. & Lu, J., 2017b, 'An entropy-based indicator system for measuring the potential of patents in technological innovation: rejecting moderation', *Scientometrics*, vol. 111, no. 3, pp. 1925–1946.

Zhang, Y., Wang, X., Zhang, G. & Lu, J., 2018c, 'Predicting the dynamics of scientific activities: A diffusion-based network analytic methodology', *Proceedings of the Association for Information Science and Technology*, vol. 55, no. 1, pp. 598–607, <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2018.14505501065>.

Zhang, Y., Wu, M., Hu, Z., Ward, R., Zhang, X. & Porter, A., 2021b, 'Profiling and predicting the problem-solving patterns in china's research systems: A methodology of intelligent bibliometrics and empirical insights', *Quantitative Science Studies*, vol. 2, no. 1, pp. 409–432.

Zhang, Y., Wu, M. & Lu, J., 2022, 'Stepping beyond your comfort zone: Diffusion-based network analytics for knowledge trajectory recommendation', *arXiv preprint arXiv:2205.15504*.

Zhang, Y., Wu, M., Miao, W., Huang, L. & Lu, J., 2021c, 'Bi-layer network analytics: A methodology for characterizing emerging general-purpose technologies', *Journal of Informetrics*, vol. 15, no. 4, p. 101202, <https://www.sciencedirect.com/science/article/pii/S1751157721000730>.

Zhang, Y., Wu, M., Tian, G. Y., Zhang, G. & Lu, J., 2021d, 'Ethics and privacy of artificial intelligence: Understandings from bibliometrics', *Knowledge-Based Systems*, vol. 222, p. 106994.

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D. & Lu, J., 2016, 'Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research', *Technological Forecasting and Social Change*, vol. 105, pp. 179–191, <http://www.sciencedirect.com/science/article/pii/S0040162516000160>.

Zhang, Y., Zhang, G., Zhu, D. & Lu, J., 2017c, 'Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics', *Journal of the Association for Information Science and Technology*, vol. 68, no. 8, pp. 1925–1939.

Zhao, L.-q., Wen, Z.-j., Wei, Y., Xu, J., Chen, Z., Qi, B.-z., Wang, Z.-m., Shi, Y.-y. & Liu, S.-w., 2015, 'Polymorphisms of renin-angiotensin-aldosterone system gene in chinese han patients with nonfamilial atrial fibrillation', *PloS One*, vol. 10, no. 2, p. e0117489.

Zheng, T., Ardolino, M., Bacchetti, A., Perona, M. & Zanardini, M., 2019, 'The impacts of industry 4.0: A descriptive survey in the italian manufacturing sector', *Journal of Manufacturing Technology Management*.

Zheng, Y.-Y., Ma, Y.-T., Zhang, J.-Y. & Xie, X., 2020, 'Covid-19 and the cardiovascular system', *Nature Reviews Cardiology*, vol. 17, no. 5, pp. 259–260.

Zhou, T., Lü, L. & Zhang, Y.-C., 2009, ‘Predicting missing links via local information’, *The European Physical Journal B*, vol. 71, no. 4, pp. 623–630.

Zhu, J., Han, L., Gou, Z. & Yuan, X., 2018, ‘A fuzzy clustering-based denoising model for evaluating uncertainty in collaborative filtering recommender systems’, *Journal of the Association for Information Science and Technology*, vol. 69, no. 9, pp. 1109–1121.

Zhu, L., He, Y. & Zhou, D., 2019, ‘Hierarchical viewpoint discovery from tweets using bayesian modelling’, *Expert Systems with Applications*, vol. 116, pp. 430–438.

Zhu, Z., Chakraborti, S., He, Y., Roberts, A., Sheahan, T., Xiao, X., Hensley, L. E., Prabakaran, P., Rockx, B. & Sidorov, I. A., 2007, ‘Potent cross-reactive neutralization of sars coronavirus isolates by human monoclonal antibodies’, *Proceedings of the National Academy of Sciences*, vol. 104, no. 29, pp. 12123–12128.

Zhukov, D., Khvatova, T., Lesko, S. & Zaltcman, A., 2018, ‘Managing social networks: Applying the percolation theory methodology to understand individuals’ attitudes and moods’, *Technological Forecasting and Social Change*, vol. 129, pp. 297–307.

Özgür, A., Vu, T., Erkan, G. & Radev, D. R., 2008, ‘Identifying gene-disease associations using centrality on a literature mined gene-interaction network’, *Bioinformatics*, vol. 24, no. 13, pp. i277–i285.

Øiestad, S. & Bugge, M. M., 2014, ‘Digitisation of publishing: Exploration based on existing business models’, *Technological Forecasting and Social Change*, vol. 83, pp. 54–65.