Multi-Recommender Framework to Aid Identifying and Addressing Research Themes Using Bibliographic Metadata and AI

a thesis submitted for obtaining
the scientific title of PhD in Engineering
from
Politehnica University Timișoara
in the field of
Computer and Information Technology
by

M.Sc. Christian-Daniel CURIAC

PhD Committee Chair:
PhD Supervisor: Prof. Dr. Eng. Mihai Victor MICEA
Scientific Reviewers:

Date of the PhD Thesis Defense:

Acknowledgement

This thesis has been developed during my tenure at the Faculty of Automation and Computer Science at Politehnica University Timișoara, Romania. While I proudly acknowledge this dissertation as the product of my own work, it would not have been possible without the guidance, help, support, and companionship of many people. To each of them, I want to express my most sincere gratitude and appreciation.

Firstly, I would like to express my sincere appreciation and deepest gratitude to my supervisor, Professor Mihai Micea, for his unwavering support, understanding and insightful guidance throughout my research endeavors.

I want to express my profound gratitude to Professor Alex Doboli from Stony Brook University, NY, USA for our long and fruitful collaboration on various research themes. I thank him for his kindness, patience, and the motivation and encouragement he provided to bring this thesis to fruition.

I also extend my heartfelt thanks to Professor Helmut Gräb, Professor Eckehard Steinbach, and Dr. Andreas Noll who guided my scientific research during my Master's and Bachelor's degree studies at Technical University Munich, Germany, laying a solid foundation for my research journey.

To my work colleagues from msg security advisors, thank you for your encouragement and friendship. I am so fortunate to work with people who have always offered me an incredible amount of support.

Last but certainly not least, I would like to express my deepest gratitude to my family and my closest friends for their love and endless support throughout my doctoral studies and research work. A special dedication goes to my beloved parents, Nona and Daniel. Thank you for believing in me, ensuring I received the best education and opportunities possible, and giving me the strength and courage to pursue my dreams.

Timişoara, June 2024 Christian-Daniel CURIAC

Curiac, Christian-Daniel

Multi-Recommender Framework to Aid Identifying and Addressing Research Themes Using Bibliographic Metadata and AI

Keywords:
recommender system, bibliographic metadata, AI techniques, research gap, research theme, research trend prediction, knowledge transfer, research team formation

Abstract:

Every now and then, researchers need to consider new scientific topics to work on for an assortment of reasons originating either in the way scientific knowledge recently evolved (e.g., the development of new technologies, theories, and methods) or in the existing research theme itself (e.g., the research problem was already solved; no remarkable outcomes are expected; the research question leads to a dead end because of a lack of new ideas, or a material, financial or human resource shortage). The discovery of new research themes has never been so problematic as today mainly due to the dynamism, complexity, and segmentation of the research scene and also due to the abundance of scientific publications that need to be surveyed. In these circumstances, the demand for automatic or semi-automatic tools to aid researchers in discovering and starting working on new promising research themes that meet both their expectations and expertise is particularly increasing. This thesis provides solutions to cover this gap by developing a content-based human-in-the-loop recommender system framework where adequately structured, classified, and ranked contextual information coming from a large body of scientific publications is been employed to derive hot and feasible research themes alongside identifying suitable research teams, cross-domain knowledge transfers, or scientific literature to start with. To evaluate the proposed multi-recommender framework and its associated techniques, methods, and implementations, a series of case studies and experiments are presented, yielding promising results.

Contents

Abbreviations

AI	Artificial Intelligence
AIC	Akaike Information Criterion
API	Application Programming Interface
ARIMA	Auto Regressive Integrated Moving Average
ARMA	Auto Regressive Moving Average
BERT	Bidirectional Encoder Representations from Transformers
CB	Content-Based recommender
CMOS	Complementary Metal–Oxide–Semiconductor
CSV	Comma-Separated Values
DDTF	Data-Driven Team Formation
DOI	Digital Object Identifier
EDA	Electronic Design Automation
FPGA	Field Programmable Gate Array
FS	Familiarity Score
GDPR	General Data Protection Regulation
GGO	Generalized Global Objective
GMO	Generalized Mean Objective
HL	Human-in-the-Loop
HLRS	Human-in-the-Loop Recommender System
ICISG	Interpersonal Collaborations Inside Specified Groups
IEEE	Institute of Electrical and Electronics Engineers
IF	Impact Factor
IoT	Internet of Things
LDA	Latent Dirichlet Allocation
LSTM	Long-Short Term Memory
MCOT	Mean Co-Occurrence per Term
MK	Mann-Kendall test
ML	Machine Learning
NERD	Named-Entity Recognition and Disambiguation
NEL	Named-Entity Linking

NLP	Natural Language Processing
NSGA-II	Non-dominated Sorting Genetic Algorithm II
nsaMK	n-steps-ahead Mann-Kendall
PKS	Personal Knowledge Score
RAE	Researcher's Areas of Expertise
RCA	Researcher's Collaboration Ability
RLEGA	Researcher's Level of Expertise in a Given Area
RGE	Researcher's General Expertise
RS	Recommender System
SDK	Software Development Kit
TIM	Technology and Innovation Management
UPT	Politehnica University of Timisoara

Terminology

Bibliographic database – an organized collection of bibliographic records;

Bibliographic record/metadata – a file containing information about a publication that encompasses fields like the title, keywords, abstract, author names and affiliation, publication data, etc.;

Compound abstract – a short text document obtained by concatenating the title, keywords and abstract of a given publication;

Context (of a term) – the terms lying in the same topic with a given term;

Emerging domain – a research domain that is currently on a highly rising trend within the research community;

Feasible research gap – a research gap that can be approached at a given time based on existing methods, techniques and materials already published or used in the past;

Hot theme – a theme that is popular and on an increasing trend at a specified moment in time;

Key term – the term representing a relevant concept for a scientific area or a research theme;

Processed abstract – a short text document comprising a list of relevant terms from a compound abstract obtained through an entity linking procedure;

Recommender system – a subtype of information filtering system able to recommend items that are most relevant to the user;

Research domain – a disciplinary branch of knowledge and research representing a specific field of expertise;

Research gap – an area of research which suffers from a substantial lack of information and knowledge, being significantly immature;

Research theme – a clearly defined problem currently having no or partial solutions which is worth investigating;

Term (or mention) – a word or a sequence of words relevant to a given domain extracted using entity linking methods;

Topic – a computationally derived cluster of relatively related terms;

Twin domain – a research domain having similar or close methods and materials with a given domain;

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Over the last decade, Artificial Intelligence (AI) and especially Machine Learning (ML) have proved their transformative potential, promising to have a revolutionary impact on the economy and society in general. They have a ubiquitous contribution to many domains ranging from image and speech recognition to medical diagnosis and self-driving vehicles [1, 2]. Being able to understand and extract correlations, causalities, and patterns from real-world processes, AI is increasingly likely to replace humans not only for predictive or repetitive activities but also in fulfilling cognitive or complex decision-making tasks.

Recommender systems are a special category of AI applications. They are automated or semi-automated systems intended to advise users during decision-making [3, 4]. Recommender systems have been devised for a variety of everyday life activities, like music streaming, video on demand, or online retail [5, 6, 7]. Faced with an abundance of information, users have an increasingly difficult task anytime they search for relevant information to base their judgments or to make decisions. In this context, AI techniques can deeply mine and rigorously analyze large volumes of data, and then present results in a readable, comprehensible, and user-friendly manner.

The overall goal of this thesis is to provide a package of recommender systems to aid researchers in discovering and start working on new promising research themes. We mainly focus on content-based human-in-the-loop recommender modules where adequately structured, classified, and ranked contextual information coming from a large body of scientific publications is being employed.

The discovery and framing of new promising research themes have always been a great challenge for academia [8] and industry [9]. This activity requires extensive human effort, expertise, and, not least, intuition. The process is usually grounded in a systematic and critical literature review backed by the need to identify patterns, changes, and trends within an extensive body of knowledge. It includes two important stages, namely research gap identification and problem framing, each of which includes time-consuming, meticulous, and sometimes tedious activities, which alternate with decision-making where human expertise plays a decisive role. Following the research theme framing, three other tasks may be undertaken: investigating possible knowledge transfers that may help solve the problem; identifying relevant scientific literature to help start the evaluation of the state of the art in the field; and, probably the most important, forming an appropriate and effective team to carry out the intended research work.

With the emergence and widespread adoption of AI techniques, developing effective tools to assist researchers in framing and addressing relevant research themes becomes a natural step forward, motivated by the following reasons:

- *An increasing body of knowledge must be surveyed.* Since the number of scientific publications has exponential growth, extracting useful information without automated tools becomes increasingly difficult. This task is further complicated by several other factors, e.g., publications may contain incomplete, biased, misleading, or even erroneous information; or, the access to some publications is restricted. Recommender systems can eliminate unreliable inputs and/or predict the missing pieces of information based on existing information from similar documents.

- *Every research has its own life cycle.* From time to time, when their themes become saturated or declining, researchers need to switch to more timely domains [10].

- *Research theme framing needs to be correlated with the related team formation which can be stated as a complex optimization process* that besides discovering research gaps must carefully assess, predict, and consider a set of constraints that includes the number of available research team members, their profile and expertise, time deadlines, and available technical needs and financial resources [4].

- *Necessity to correlate the framed research themes with current trends* and advancements in research and innovation that originate from emergent domains, which are proven to have influential effects upon the entire scientific community.

- *Limited view of the researcher regarding the overall body of knowledge.* Generally, the researchers are more likely to search for new themes inside their own domains of expertise. In this context, an automatic or semi-automatic methodology may help explore a broader research area, thus increasing the productivity and scalability of the process.

- *Interest in research themes can be driven by funding agencies* through grants or projects with a specified area of interest. In this case researchers must adjust their ongoing interests or themes, or must find new ones within the field defined by the call for proposals.

- *Need to reduce the degree of subjectivity* in selecting new research themes and research teams, as any manually driven procedures incorporate subjective decisions linked to inherent fears of novelty and uncertainty, concerns regarding the long time and effort needed for researcher's recalibration to a totally new theme, or worries that the projected results will not materialize. While recommender systems can suggest potentially rewarding topics and problems, they can be used to identify future research collaborators who can complement one's research expertise.

- *Research projects tend to be more complex* often requiring a multidisciplinary and highly collaborative approach.

1.2 Objectives

The primary objective of this thesis is to design a multi-recommender system framework to aid in identifying and addressing high-impact and timely research themes.

The proposed architecture implements a Human-in-the-Loop (HL) methodology where the human expert is needed to supervise the research themes framing and addressing process mainly because of two reasons. The first one is given by the lack of historical data to train AI or ML models since recommendations need to be highly customized for a researcher or a research team which rarely does this activity, while the second one is related to the need to improve the accuracy and relevance of the obtained recommendations considering the noisy and imprecise information that inherently characterizes the scientific activity.

Our multi-recommender system relies on paper metadata records collected from bibliographic/bibliometric databases (e.g., IEEE Xplore) as a valuable and reliable source of research-related information. In this regard, the existing scientific publications are seen not only as the direct result of research activities but also as a means to understand past, current and future research trends. Moreover, by investigating the scientific production we may help objectivize the research team formation process.

The secondary objectives are related to the design, implementation and validation of the main functional modules that constitute the multi-recommender framework (i.e., research theme recommender; cross-domain knowledge transfer recommender; and, research team recommender) considering that they are meant to also function as standalone units.

1.3 Major Contributions

This thesis describes a human-in-the-loop multi-recommender framework designed to aid in discovering and addressing high-impact research themes using bibliographic metadata by artificial intelligence means. Particularly, the topics that are covered include automatic bibliographic metadata acquisition and preprocessing, scientific domain and research theme modeling, research trend assessment, cross-domain knowledge transfer, and research team formation. To address the mentioned research topics, a variety of Natural Language Processing (NLP) methods have been employed, including entity linking, document similarity assessment, topic modeling, and term co-occurrence analysis. These NLP methods are supplemented by prediction and multi-objective optimization techniques.

Considering the objectives stated in the previous section, the major contributions provided by this thesis are:

- *A human-in-the-loop multi-recommender system architecture to help researchers discover and address hot and timely research themes based on bibliographic metadata;*

  Analyzing the knowledge development process for a scientific field, a semi-automatic framework encompassing four recommender modules (i.e., research theme recommender, cross-domain knowledge transfer recommender, scientific literature recommender, and research team recommender) is designed. The findings were reported in [11].

- *A method to evaluate research trends from journal paper metadata, considering the research publication latency;*

  To incorporate the unfavorable influence of the time lag between the research ending and its results' publication on research trend assessments, we propose a trend detection methodology combining auto-ARIMA prediction method with the Mann–Kendall test. This contribution was reported in our journal paper [12].

- *A method to identify hot research topics using topic modeling and multivariate prediction techniques;*

  By representing the research themes as collections of key terms we proposed an approach to discover impactful research topics from bibliographical records using Latent Dirichlet Allocation (LDA) topic modeling coupled with a multivariate version of the Mann-Kendall test. This contribution was detailed in two of our papers, namely [13] and [14].

- *A method to evaluate the feasibility of a research theme using a co-occurrence-based double thresholding method;*

  We developed an automated mechanism to identify the feasible research gaps to be covered by using a double-threshold procedure that filters out the themes that are either difficult to study using existing knowledge or have limited novelty prospects. The method was the subject of our journal paper [15].

- *A cross-domain knowledge transfer recommender based on the concept of twin scientific domain;*

The thesis offers a practical approach that employs paper metadata to identify the twin domains that are closely connected to a given scientific domain and from which knowledge transfer might be successful, as well as the information that should be transferred.

- *A publicly available dataset for bibliographic/bibliometric data-driven research team formation;*

  The dataset consists of de-identified information regarding the technical expertise and collaborative proficiency of scholars affiliated with Politehnica University of Timisoara extracted from IEEE Xplore paper metadata for the time interval 2010-2022. The dataset is available on the Mendeley Data public repository [16] and is detailed in our journal data paper [17].

- *A formalization of research team formation as a generalized multi-objective set cover optimization problem;*

  We mathematically formulate the research team formation process as a customizable multi-objective optimization by generalizing the classic set multicover problem. Our optimization model is especially suited for egalitarian team formation and completion but can also be used in covering non-managerial positions inside hierarchical teams.

- *A research team recommender using a genetic multi-objective optimization algorithm and extended bibliometric data.*

  We used an extended set of paper metadata fields to derive four synthetic indicators about the candidates' expertise and interpersonal skills and solve the combinatorial multi-objective team formation problem using the NSGA-II genetic algorithm to suggest a list of optimal teams. The recommender's design and validation details were presented in our article [18].

Besides the already mentioned major contributions, the thesis contains several minor contributions (e.g., methods tuning and calibration; selection and analysis of representative case studies; etc.) that will be detailed in the subsequent chapters. Moreover, the author of this thesis developed all the necessary software using Python programming language and related libraries and also designed and conducted the theoretical and experimental studies.

During the PhD studies, the author has contributed to ten research papers in peer-reviewed journals and conference proceedings, one book chapter, and one public dataset:

1. **C.-D. Curiac**, O. Banias, and M. Micea, "Evaluating research trends from journal paper metadata, considering the research publication latency", Mathematics, vol. 10(2), 233, MDPI, 2022. *[journal paper]*

2. **C.-D. Curiac**, A. Doboli, and D.-I. Curiac, "Co-occurrence-based double thresholding method for research topic identification", Mathematics, vol. 10(17), 3115, MDPI, 2022. *[journal paper]*

3. **C.-D. Curiac**, and A. Doboli, "Combining informetrics and trend analysis to understand past and current directions in electronic design automation", Scientometrics, vol. 127(10), pp. 5661-5689, Springer, 2022. *[journal paper]*

4. **C.-D. Curiac**, and M. Micea, "Evaluating research trends using key term occurrences and multivariate Mann-Kendall test", in Proceedings of the International Symposium on Electronics and Telecommunications (ISETC 2022), pp. 1–4, IEEE, 2022. *[conference paper]*

5. **C.-D. Curiac**, and M. Micea, "Identifying hot information security topics using LDA and multivariate Mann-Kendall test", IEEE Access, vol. 11, pp. 18374-18384, IEEE, 2023. *[journal paper]*

6. **C.-D. Curiac**, M. Micea, T.-R. Plosca, D.-I. Curiac, and A. Doboli "Dataset for bibliometric data-driven research team formation", Mendeley Data, version 1, doi: 10.17632r4vrvhb23h.1, 2023. *[public dataset]*

7. **C.-D. Curiac**, M. Micea, T.-R. Plosca, D.-I. Curiac, and A. Doboli "Dataset for bibliometric data-driven research team formation: case of Politehnica University of Timisoara scholars for the interval 2010-2022", Data in Brief, vol. 53, 110275, Elsevier, 2024. *[journal paper]*

8. **C.-D. Curiac**, M. Micea, T.-R. Plosca, D.-I. Curiac, and A. Doboli "Optimized interdisciplinary research team formation using a genetic algorithm and extended bibliometric data". *[journal paper - under review]*

9. **C.-D. Curiac**, M. Micea, T.-R. Plosca, D.-I. Curiac, S. Doboli and A. Doboli "Towards automating new research problem framing and exploration based on symbolic-numerical knowledge extracted from bibliometric data", in Bibliometrics - An Essential Methodological Tool for Research Projects. IntechOpen, London, UK, 2024. *[book chapter]*

10. T. Andreica, **C.-D. Curiac**, C. Jichici and B. Groza, "Android head units vs. in-vehicle ECUs: performance assessment for deploying in-vehicle intrusion detection systems for the CAN bus", IEEE Access, vol. 10, pp. 95161-95178, IEEE, 2022. *[journal paper]*

11. M.D. Baciu, E.A. Capota, C.S. Stângaciu, **C.-D. Curiac**, and M. Micea, "Multicore time-triggered OCBP-based scheduling for mixed criticality periodic task systems", in Proceedings of the International Symposium on Electronics and Tele-

communications (ISETC 2022), pp. 1–4, IEEE, 2022. *[conference paper]*

12. T.-R. Plosca, **C.-D. Curiac**, and D.-I. Curiac. "Investigating semantic differences in user-generated content by cross-domain sentiment analysis means", Applied Sciences, vol. 14(6), 2421, MDPI, 2024. *[journal paper]*

The first nine of these works represent the main pillars of the thesis, while the rest address machine learning and task scheduling topics.

1.4 Thesis Outline

The remainder of this thesis is structured as follows.

Chapter 2 provides a critical review of related literature in the field of recommendation systems for research aiding.

Chapter 3 outlines the proposed architecture of a multi-recommender system meant to aid researchers in identifying and exploring relevant research topics using information extracted from publication bibliographical records. This framework incorporates four recommender modules that, based on paper metadata, assist researchers in finding new research themes, identifying the knowledge that is suitable to be transferred from other scientific domains, finding a set of relevant publications to start the literature review with, and composing a suitable team of experts to address the theme. Our endeavor is grounded in the way new knowledge may arise in a given scientific domain, pointing out that significant research ideas may be derived from appropriate combinations of already existing in-domain knowledge, may come from closely related and emergent domains, or, less frequently, may result from sudden insights.

Chapter 4 describes the data acquisition and preprocessing procedure. We rely on Application Programming Interfaces (APIs) to automatically collect publication metadata records corresponding to top-tier scientific journals or annual conferences to effectively summarize the research in a given field. Using an appropriate entity linking technique (i.e., TagMe), the title, keywords and abstract metadata fields are transformed into lists of relevant key terms to represent the essence of the publication content.

Chapter 5 is devoted to the detailed presentation of the semi-automatic research theme recommender module able to suggest new and high-impact research topics. Our HL methodology starts with identifying the set of key terms that characterize the scientific domain and clusters these key terms in research themes. Subsequently, the research themes are investigated in terms of their opportunity and feasibility by employing trend and statistical analysis.

Chapter 6 describes the recommender module designed to identify possible cross-domain knowledge transfers from twin or emerging scientific domains. In this respect, document similarity and topic modeling techniques have been used.

Chapter 7 presents the research team recommender module. To formalize the team formation process, we propose a generalized multicriteria set cover optimization model that may cope with a large variety of team objectives and constraints. By employing an extended set of bibliographic and bibliometric data, we evaluate each candidate's technical expertise and collaborative skills based on four carefully designed descriptors and solve the resulting problem using the NSGA-II elitist evolutionary algorithm.

Chapter 8 offers a summary of our work, followed by conclusions and a brief discussion of future work.

Chapter 2

Review of Related Literature

This chapter presents a critical review of the state-of-the-art recommendation systems for research aiding, emphasizing the existing methods and frameworks to recommend scientific content, citations, new research themes, and appropriate team members for research projects.

Recommendation systems, also known as recommender systems, are a special category of information filtering systems that can provide pertinent suggestions for items or content to a particular user [19, 20]. Such suggestions mainly refer to diverse decision-making activities, including products or services to be purchased, online news to be read, or music to be listened. Recommendation systems are especially beneficial in situations when a user has to select from an overwhelming number of potential alternatives.

Depending on the information used in identifying and ranking the suggestions presented to the user, recommender systems may be classified into three main categories: (1) collaborative-filtering, (2) content-based, and (3) hybrid recommender systems [21, 4]. Collaborative filtering methods construct a behavioral model of the user based on previous interactions, and then make suggestions according to this model [4]. Content-based recommender systems identify the relevant features (e.g., concept attributes, keywords) from contextual information, and classify them according to the user's needs [4]. The two approaches can also be combined as hybrid recommender systems [4]. Content-based methods have been used for text or webpage recommendation, but are challenged by limited content analysis (e.g., only keywords are used, not the relations between them), over-specialization due to the frequent keywords that bias the search, and tackling new keywords [4]. Collaborative filtering has been used in systems like Amazon's [22] or PHOAKS [23] for creating personalized online shopping experiences, finding useful web information, and joke recommendations [4]. However, they are challenged by addressing new users, new keywords, and keyword sparsity as few keywords from the entire set are used to describe a document [4]. Hybrid methods combine the advantages of the two methods but are challenged by multi-criteria requirements, quality, and scalability [4].

2.1 Recommender Systems for Research Aiding: State of the Art

In many fields of activity, and the research field makes no exception, the amount of information rises exponentially, needing appropriate means to aid decision-making processes. Researchers must devote significant resources and effort to finding personalized academic/scientific material that relates to their work. Even though such research-related tasks can in principle be fully or partially automatized using artificial intelligence techniques, conclusive results are still pending. In this context, little research has been reported on specialized recommender systems as helpful tools to scholars when starting and conducting research [24]. More specifically, the existing research is limited and fragmented and was directed toward four objectives: proposing customized scientific content, suggesting citations to accompany a research theme, discovering research hotspots based on trend analysis, and assembling research teams.

Scientific content recommendation

In their endeavor to evaluate the state of knowledge in a scientific area, scholars are generally searching online libraries and bibliographic databases for adequate scientific materials using keywords or text phrases. The use of appropriately designed recommender systems to provide this service enables researchers to easily and quickly collect and make use of a variety of digital materials from all around the world, composed mainly of textual publications (e.g., scientific papers, books and patents). Academic search engines like Google Scholar provide lists of scientific publications based on the submitted user query [25]. They are helpful in the sense of systematicity, transparency or reproducibility, but they don't make a lot of difference when it comes to filtering or personalization, whereas the recommender systems are key to managing information overload [26].

In [27], the authors proposed a content-based filtering article recommender, named PURE, that performs a model-based clustering to provide the highly-rated articles containing a set of chosen keywords from PubMed database. Gipp et al. [28] developed the first hybrid scientific paper recommender, coined as Scienstein, as a powerful alternative to traditional academic search engines. It improves the usually employed keyword-based search by mixing it with various methods (e.g., citation, author or source analysis; implicit and explicit ratings; 'Distance Similarity Index' and the 'In-text Impact Factor' evaluation). The users are prompted to enter not only keywords but also entire documents when searching for additional scientific materials. An open-source Python library, named Science Concierge, that implements a content-based recommender system for scientific literature search is reported in [29]. The library processes the documents using a scalable vectorization of texts through Latent Semantic Analysis and employs a mixture of the Rocchio algorithm with an approximate nearest neighbor search to derive the recommendations. Guo et al. [30] were the first to include semantic representation, obtained using a Long-Short Term Memory (LSTM) neural network, by considering the relevance of the words from paper abstracts with respect to words in the paper title, while Haruna et al. [31] and Sakib et al. [32] developed hybrid recommenders that employ both the magnitude of the collaborative similitudes between papers and correlations between their contents.

Other representative examples of services that provide scientific content suggestions are the ones provided by university digital libraries that actively employ recommender systems to support learning, education, and research. The hybrid fuzzy linguistic recommender system described in [33] uses a combination of two approaches to rank items in a university digital library. While in the first step, the items are ranked by their content using distance similarity measures, in the second step the items are re-ranked based on their quality (i.e., the popularity of items among users). Serrano-Guerrero et al. [34] proposed another fuzzy linguistic recommender system that suggests useful scientific resources and potential collaborators for given research topics using information from university digital libraries. By using the Google Wave capabilities, their system disseminates information between several researchers interested in the same topic.

Citation recommendation

The task of specifying correct citations for a given text passage in a document is generally referred to as citation recommendation [35]. Given the abundance of published articles and the need to cite appropriate publications when writing scientific texts, the topic of citation recommendations has become increasingly important for researchers.

The seminal paper of He et al. [36] provides a probabilistic non-parametric model to assess the context-based relevance of a given citation context for a document. Their system recommends the bibliography corresponding to a manuscript and offers a ranked list of citations for a specified citation placeholder. Wang et al. [37] propose SentCite, a tool that identifies the sentences that need to be backed up with references by employing a convolutional recurrent neural network and recommends citations based on the similarity between citation sentences and target papers. Yang et al. [38] designed a citation recommender that uses the semantic similarity between citation context and scientific papers, and also information about authors to improve the accuracy. Jeong et al. [39] proposed a deep learning-based model for context-aware paper citation recommendation that employs graph convolutional network layers, a bidirectional encoder, and a pre-trained model of textual data.

Research topics recommendation based on trend evaluations

Existing academic publications are not only the direct result of scientific activities but they also influence the current and future research trends. By employing content or citation analysis, a large spectrum of valuable evidence may be revealed, including the impact and attractiveness of specific research topics.

In 2018, Lee and Kang [8] adopted a Latent Dirichlet Allocation (LDA) topic modeling approach to automatically uncover research topics in the Technology and Innovation

Management (TIM) field. Their method explores topic trends by investigating the fluctuations in topic rankings based on a citation analysis over diverse time periods and identifies hot and cold topics. The approach is specifically tailored for TIM by using bibliometric data only from representative domain-specific journals, while possible connections with other domains have not been considered.

Huang et al. [40] constructed a framework that uses a large-scale academic graph and is able to perform time and space analysis of research frontiers. The goal of this framework is to effectively encourage the implementation of data-driven knowledge discovery and was exemplified by a case study from medical and health sciences. Here, the Rapid Automatic Keyword Extraction algorithm was employed for keyword extraction and the Mann-Kendall (MK) test was utilized to identify trend changes in research topics. Even the framework can effectively achieve deep and dynamic evaluation of research frontiers within diverse fields, it does not identify the research gaps to base the research themes framing process.

A more general and also practice-oriented way to identify the research opportunities is ResGap [41], a tool that uses text mining procedures to extract topics and map topic trends from a body of publications, and an entity linking method for recognizing and disambiguating terms using unambiguous identifiers from Wikipedia [42, 43]. ResGap can be perceived as a means to discover promising research areas that can be later manually formalized into fruitful future research themes. The same basic idea to implement such research gap discovery systems was put forward by Wang et al. [44]. Their framework of a hybrid recommender system uses Hierarchical Latent Tree Analysis for topic modeling and Google Trends to evaluate the topics' trends. These two approaches are able to discover interesting research gaps but are neither meant to offer research themes nor to provide research gaps customized to the researchers' technical expertise and shared interests.

Research team recommendation

The success of any research project largely depends on the team assigned to perform the tasks. In this respect, besides members' technical expertise, there is a series of individual psychological, organizational, and teamwork-related factors that need to be considered when assembling high-performing teams. Such factors, including team coherence, interpersonal relationships, positive attitude, or conflict management potential, are hard to objectively quantify, making the research team formation process very difficult. However, some promising approaches have been reported.

In their pioneering work, Lappas et al. [45] developed a research team recommender that gathers together experts by considering three key pillars: the research theme, the pool of candidates with specialized and diverse skills, and a documented social network to assess the compatibility between individuals. For this, they employed candidate-related information extracted from the DBLP bibliographic database.

Srivastava et al. [46] proposed ULTRA - an AI-based approach for aiding the team

Table 2.1: Related work summary

Ref.	Academic Service	RS Model	RS Description
[33]	Recommendation in Digital Library	Hybrid System	fuzzy linguistic recommender system that provides personalized research resources that are relevant to the users and also have a quality that was previously certified by other users.
[34]	Recommendation in Digital Library	Hybrid System	fuzzy linguistic recommender system that suggests scientific resources and potential collaborators
[27]	Paper recommendation	CB	model-based clustering to provide the highly-rated articles containing a set of keywords
[28]	Paper recommendation	Hybrid System	keyword-based search accompanied by various methods (e.g., citation, author or source analysis; implicit and explicit ratings; 'Distance Similarity Index' and the 'In-text Impact Factor' evaluation).
[29]	Paper recommendation	CB	employed Latent Semantic Analysis to the content of scientific papers to derive recommendations
[30]	Paper recommendation	CB	includes semantic relationship for paper recommendation
[31]	Paper recommendation	Hybrid System	employs both the magnitude of the collaborative similitudes between papers and correlations between their contents
[32]	Paper recommendation	Hybrid System	public contextual metadata and paper-citation information are incorporated to enhance the recommendation accuracy
[36]	Citation recommendation	CB	proposes a non-parametric probabilistic model to evaluate the context-based relevance between a document to be cited and citation context.
[37]	Citation recommendation	CB	proposes SentCite to identify the sentences that need to be backed by references using a convolutional recurrent neural network and recommends citations based on the similarity between sentences
[38]	Citation recommendation	CB	uses the semantic similarity between citation context and scientific papers, and also information about authors to improve the accuracy of citation recommendations
[39]	Citation recommendation	CB	employs graph convolutional network layers, a bidirectional encoder and a pre-trained textual data model
[8]	Research topics recommendation	CB	method to identify research topics in the Technology and Innovation Management field using LDA
[40]	Research topics recommendation	CB	proposes a framework that uses a large-scale academic graph and is able to perform time and space analysis of research frontiers
[41]	Research topics recommendation	CB	proposes ResGap, a practice-oriented method to identify the research opportunities using text mining procedures to extract topics and to map topic trends from a body of publications
[44]	Research topics recommendation	Hybrid System	proposes a research gap discovery system that uses Hierarchical Latent Tree Analysis and Google Trends
[45]	Research Team recommendation	CB	provides expert team recommendations based on the research theme, a given pool of skilled candidates and their previous collaborations.
[46]	Research team recommendation	CB	describes an AI-based recommender system, named ULTRA, to assemble teams of experts to meet the requirements extracted using NLP techniques from calls for proposals.
[47]	Research team recommendation	CB	proposes a recommender based on LANT architecture that comprises unsupervised transfer learning and neural team recommendation.

formation process in response to calls for proposals from funding entities. This recommender system employs NLP techniques to extract the required technical skills from proposal calls and then identifies potential team members by investigating their matching calls using specialized techniques (e.g., the SPECTER representation learning method derived from BERT). Finally, the team suggestions are filtered based on business constraints. Another approach worth mentioning was reported by Kaw et al. [47] who recommend expert teams based on the LANT architecture. Their technique incorporates transfer learning and neural team recommendation based on Deep Graph Infomax for vector representations of skills on graph-structured data.

From the above analysis, the following conclusions are worth mentioning: (a) all the mentioned works employ text-mining approaches to investigate the scientific publication corpora, such NLP techniques showing promising results; (b) the research in the field is still in its infancy, failing to provide integrated recommender frameworks to adequately help scholars when starting and conducting their research; and, (c) the existing research is limited and fragmented, being directed toward only four objectives (i.e., proposing customized scientific content, suggesting citations to accompany a research theme, discovering research hotspots based on trend analysis, and assembling research teams) while neglecting important issues like identifying viable and timely research themes or cross-domain knowledge transfers.

Table 2.1 summarizes the main works on recommender systems for aiding research-related activities.

In our perspective, an integrated human-in-the-loop recommender system to help researchers discover and frame new customized research themes and aid them to start working on these topics may accelerate the research process and offer an increased level of objectivity. In our case, the need for a human expert to supervise the research theme framing and selection and its related activities (i.e., finding appropriate cross-domain knowledge transfers, identifying the initial bibliography to start with; and research team formation) is driven by two reasons. The first one is given by the lack of data to train machine learning or artificial intelligence models since recommendations need to be customized for a researcher or a research team which rarely does this activity, while the second one is related to the need to improve the precision and relevance of the recommendations.

Chapter 3

Overview of the Multi-Recommender System

This chapter presents a modular human-in-the-loop recommendation system architecture meant to guide the research activities toward discovering and exploring new and promising research themes. The resulting suggestions exploit the insights emerging from bibliographic information about scientific publications. The chapter encapsulates the multi-recommender framework presented in [11].

Finding a relevant, timely and feasible research theme is, especially for mature domains, a challenging task. In recent years, the quantity of published scientific information has significantly increased. This fact has made the relevant information extraction and identification of current trends within the domain even harder. Moreover, the large volume of information further complicates the discovery of the research gaps that can be exploited both within a field of study, or through cross-disciplinary research. Another issue to consider is the configuration of the team to fulfill the research theme based on individuals' technical expertise and collaborative traits, especially when the topic is complex and involves cross-disciplinary research.

In order to overcome these challenges, we propose a data-driven semi-automated methodology, implemented as a human-in-the-loop multi-recommender system, to aid hot research problem framing and addressing. This framework is based on a sequence of natural language processing techniques (e.g., topic modeling, document clustering, and keyword extraction) applied to relevant datasets of publication bibliographic metadata. To devise the methodology, we start by investigating the state-of-the-art recommender systems for assisting research-related activities, and then, by analyzing the knowledge development flow for a selected scientific domain, we provide a strong basis for the rationale behind the framework design.

3.1 Knowledge Development Process for a Research Domain

As a systematic procedure, research can be characterized as a complex process of exploring existing knowledge sources and materials to establish facts and reach new developments [48]. The newly obtained research results are rooted in either reusing, linking and combining parts of the existing body of knowledge or, more rarely, in purely creative processes (e.g., sudden insights), where new knowledge is framed (which apparently comes from nowhere).

In order to describe the way a knowledge repository for a particular research domain is established and developed in time, Figure 3.1 summarizes the entire process and its related knowledge flows.

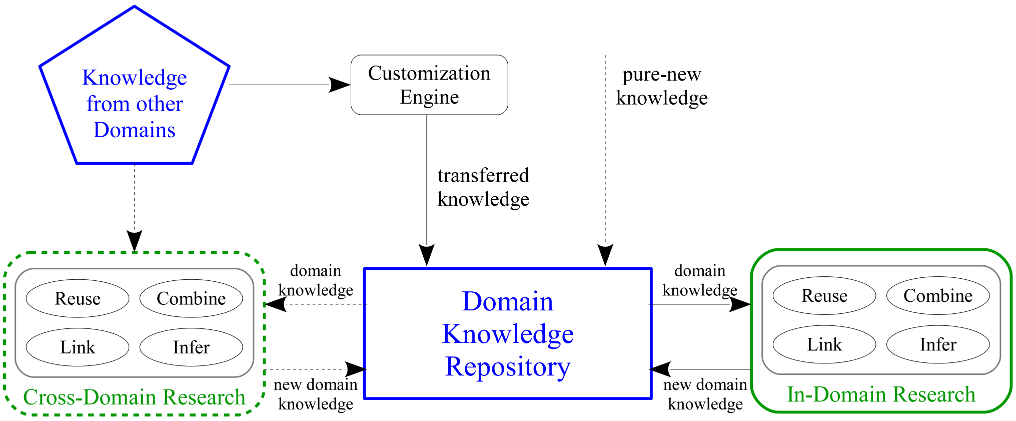


Figure 3.1: Knowledge development for a specific field [11]

Let us consider a domain that we are interested in. The central component, which collects all the results provided by the research and development process in this specific field, is the knowledge repository related to that field. It encompasses the full body of knowledge gained throughout time, including theories, methodologies, datasets, case studies, experiments, and scientific literature. There is a variety of ways in which, in our view, this domain repository may be enlarged [11]:

(a) in-domain research,

(b) knowledge transferred from other domains,

(c) inter-, trans- and multi-domain research, and

(d) pure-new knowledge coming from sudden insights.

In the case of in-domain research, we deal with incremental research, the needed knowledge being already available within the domain under investigation. This is why it focuses on combining, linking and reusing in a new fashion the already existing ideas in order to obtain novel research results.

Another possibility to augment the body of knowledge in the scientific area under investigation is to customize and transfer knowledge from other fields. Of all the existing fields, the fastest results for new research ideas can be arguably achieved by transferring knowledge from related fields, having conceptually similar topics, algorithms and methods. Important contributions may also come from areas that have had an accelerated rising trend in recent years or are anticipated by the scientific literature to be peak domains in the near future, i.e., emergent domains.

When we consider a collaboration between researchers from several fields, we talk about cross-disciplinary research, which includes multi-, inter-, or trans-disciplinary types of research, where parts of the results may enhance the knowledge repository of the domain under investigation [49]. The last way we consider enriching the repository is represented by innovative knowledge that apparently emerged from nowhere, i.e., sudden insight [50].

In the context of providing new research theme framings, in this thesis, we rely only on information that can be obtained either directly from the same field or by transferring knowledge from similar or emerging domains. These tracks are marked in Figure 3.1 with solid black lines.

The next section presents an overview of the proposed framework, developed to accelerate the research theme framing and addressing process.

3.2 Modular Recommender System Architecture

Our methodology aims towards automating the entire process that precedes the scientific research, and starts with framing new hot and feasible research themes, continues with recommending the relevant scientific literature and ends with team formation. The proposed approach is defined by a semi-supervised procedure, as in some places the human expert, i.e., the researcher, intervenes to channel the process according to her/his expectations and expertise.

The input of this complex recommending procedure is represented by a rich dataset containing scientific paper metadata (i.e., bibliographic records containing publication-related information including author names and affiliations, titles, keywords, and abstracts) for top-tier publications which may effectively summarize the research in the field and related or emerging domains, and may also track the publication profiles of researchers. In order to acquire the needed dataset we may extract records from influential bibliometric databases like Clarivate Web of Science, Scopus, PubMed or IEEE Xplore. For our case studies and experiments, since we direct our attention toward research themes from information technology and electric and electronic engineering fields, we selected IEEE Xplore as the bibliometric data source.

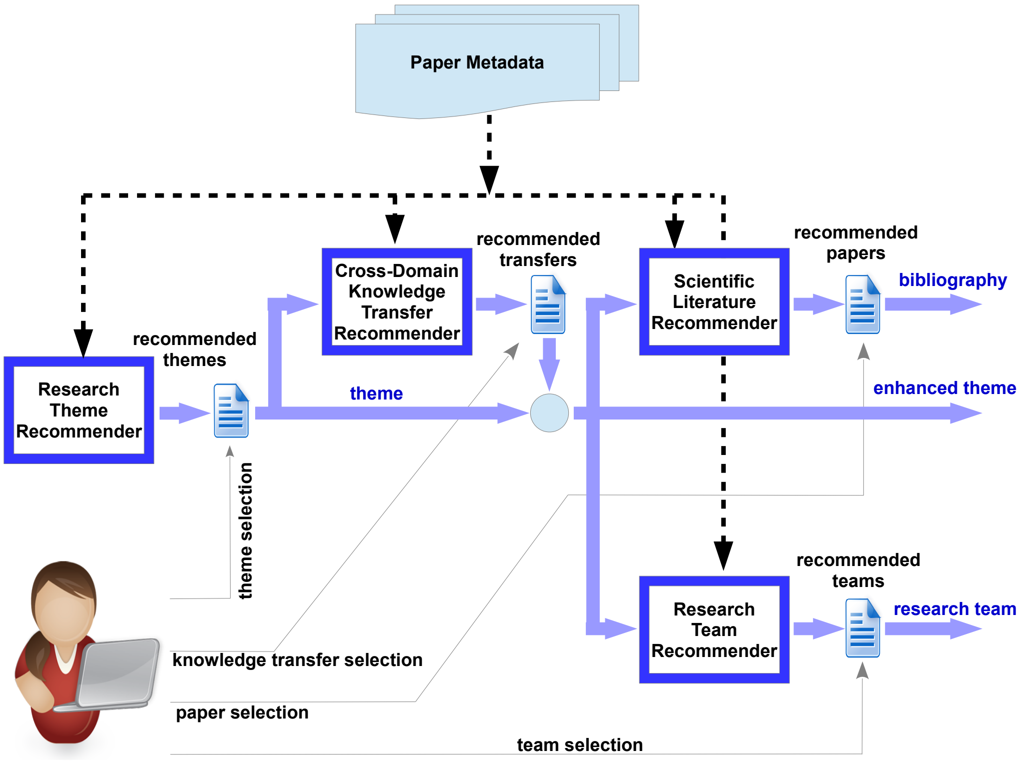


Figure 3.2: Multi-recommender system architecture [11]

Our proposed multi-recommender system framework is made up of four recommender modules that may act either interconnected as in Figure 3.2 or as standalone recommenders:

I. Research theme recommender,
II. Cross-domain knowledge transfer recommender,
III. Scientific literature recommender, and
IV. Research team recommender.

Research theme recommender. This human-in-the-loop recommender aims to provide a list of hot and feasible research themes by investigating the state of research in the domain and by judging the opportunity and viability of the themes by conducting trend and statistical analysis. In this context, a collection of relevant publication metadata is used to identify an extensive list of domain-specific terms, to discover the existing research gaps inside the domain, and also to assess the timeliness and achievability to investigate the scientific questions that lie behind these research gaps.

Cross-domain knowledge transfer recommender. This recommender is meant to identify possible sources for relevant knowledge transfers that may help solve the recommended research theme. Using document similarity assessment and topic modeling techniques, we explore the twin and emerging domains to find methods or materials, able to be related to the research theme, that have already proved their effectiveness.

Scientific literature recommender. In order to help the researchers establish a suitable starting point from where the literature review may begin, we direct our search in two directions: an in-domain exploration to find seminal works regarding the research theme; and, twin/emerging domain explorations to find relevant papers concerning the knowledge we intend to transfer.

Research team recommender. Analyzing the corpus of paper metadata to extract insights about the expertise and teamwork skills associated to each of the available researchers, a set of teams that may carry out the specified research theme is proposed by solving a complex and multi-objective team formation optimization problem.

At the end of our proposed recommendation process, a list of hot and feasible research themes, accompanied by knowledge transfer opportunities, scientific bibliography proposals, and research team recommendations, is provided.

In the following chapters, three of the recommender modules, namely research theme recommender, cross-domain knowledge transfer recommender, and research team recommender, are detailed and validated based on several case studies and experiments. As concerning the scientific literature recommender module, a series of such systems have already been proposed [51, 52], but none of them can appropriately cope with real-life situations that require particular customization to researchers' expertise, dynamism and granularity of the scientific domain, or, the existent/nonexistent highly influential researchers across the field under investigation. In this context, we consider that carefully tailored search queries to customize the bibliographical database search Application Programming Interfaces (APIs) are, at this stage, an adequate answer. While knowing that the development of a more effective scientific paper recommender module would be beneficial but difficult to perform, we leave it as a promising and important future work.

Chapter 4

Data Acquisition and Preprocessing

Bibliographic records are a valuable source of insights regarding the evolution and state of the art of scientific domains. This chapter presents two preliminary steps, namely automatic data acquisition and key term extraction, in the endeavor to obtain meaningful information from raw bibliographic metadata belonging to journal and/or conference scientific papers. Part of the results of this chapter was reported in [17] and offered as a public dataset on Mendeley Data [16].

Framing and addressing new research themes is arguably based on analyzing the state of the art and trends within an existing, evolving body of knowledge. In this respect, a comprehensive and reliable source of information is represented by top-tier scientific journals or conference proceedings that summarize the research status and reveal the hotspots and development trends. Bibliographic databases[1] comprise a plethora of scholarly publications along with corresponding impact indices (e.g., number of accessions, citation counts), being widely recognized as rich and valuable resources of both innovative ideas and insights for the scientific community. By analyzing their collections of publication-related metadata we may investigate different research facets including its impact and significance, or also to discover research gaps and decide research priorities. Since almost all bibliographic metadata records encapsulate three data fields that are intrinsically intended to summarize the corresponding research publication, specifically title, abstract and keywords, extracting relevant scientific insights occurs almost naturally. This, combined with the fact that such bibliographic information may generally be accessed through specially designed Application Programming Interfaces (APIs), sets the premises to develop and implement adequate automated procedures to aid in identifying and addressing research topics of interest.

[1] Since influential bibliographic databases are collections of bibliographic records that also include bibliometric fields, in this thesis the terms bibliographic and bibliometric are used interchangeably.

In the next sections, we will present the automatic procedures designed to collect and preprocess the bibliographic records and also the datasets that were employed in our experiments. If needed, more specific details regarding the datasets' preparation will be given in the preamble of each of the case studies described in the next three chapters.

4.1 Bibliographic Records Acquisition

Bibliographic databases are structured digital repositories holding bibliographic records that precisely represent and describe the publications. Such formatted records contain specific fields (i.e., entities) that are meant to help users identify and search for library resources and may also provide additional information to summarize their content (e.g., keywords and abstracts). In the case of influential databases, pure bibliographic information is enhanced with bibliometric fields to reveal the publication impact (e.g., citation or download counts). In this particular instance, the terms bibliographic and bibliometric are used interchangeably.

Automatic acquisition of such bibliographic records is generally done using the native API functions offered by all the prominent scientific databases including PubMed, Scopus, Clarivate Web of Science, Google Scholar, or IEEE Xplore, and consists of gathering the metadata corresponding to each publication. A journal paper metadata example, acquired using the IEEE Xplore API, is presented in Figure 4.1. In this respect, for each publication, a set of specific fields may be acquired, including title, keywords, abstract, publication date, and authors. The set of bibliographic fields that, in our perspective, are suitable to be used in configuring and fulfilling research-related activities are listed in Table 4.1.

Since the case studies and experiments that accompany this thesis are directed toward the electric, electronic, and computer engineering fields, we selected the IEEE Xplore scientific database as the source of bibliometric data. To automatically extract data from this database, the following three steps were pursued:

i) obtain an API access key from IEEE;
ii) install the Software Development Kit (SDK); and
iii) build a Python data acquisition program.

Thus, for every considered journal or conference proceedings, a Comma-Separated Values (CSV) file is provided. In these CSV files, for each scientific paper, the following corresponding attributes are retained: title, keywords, abstract, authors, affiliations, digital object identifier, publication title, publication year, citing paper count, and download count.

```
<article>
  <doi>10.1109/ACCESS.2023.3247588</doi>
  <title>Identifying Hot Information Security Topics Using LDA and
      Multivariate Mann-Kendall Test</title>
  <publisher>IEEE</publisher>
  <issn>2169-3536</issn>
  <rank>3</rank>
  <volume>11</volume>
  <authors>
   <author>
    <affiliation>Computer and Information Technology Department,
        Politehnica University of Timisoara, Timisoara, Romania</
        affiliation>
    <authorUrl>https://ieeexplore.ieee.org/author/37087077864</
        authorUrl>
    <id>37087077864</id>
    <full_name>Christian-Daniel Curiac</full_name>
    <author_order>1</author_order>
    <authorAffiliations>
     <authorAffiliation>Computer and Information Technology Department,
          Politehnica University of Timisoara, Timisoara, Romania</
         authorAffiliation>
    </authorAffiliations>
   </author>
   <author>
    <affiliation>Computer and Information Technology Department,
        Politehnica University of Timisoara, Timisoara, Romania</
        affiliation>
    <authorUrl>https://ieeexplore.ieee.org/author/37299909400</
        authorUrl>
    <id>37299909400</id>
    <full_name>Mihai V. Micea</full_name>
    <author_order>2</author_order>
    <authorAffiliations>
     <authorAffiliation>Computer and Information Technology Department,
          Politehnica University of Timisoara, Timisoara, Romania</
         authorAffiliation>
    </authorAffiliations>
   </author>
  </authors>
  <accessType>open-access</accessType>
  <content_type>Journals</content_type>
  <abstract>Discovering promising research themes in a scientific
      domain by evaluating semantic information extracted from
      bibliometric databases represents a challenging task for Natural
      Language Processing (NLP). While existing NLP methods generally
      characterize the research topics using unique key terms, we take
      a step further by more accurately modeling the research themes as
       finite sets of key terms. The proposed approach involves two
```

Figure 4.1: Paper metadata record from IEEE Xplore

```
        stages: identifying the research themes from paper metadata
        using LDA topic modeling; and, evaluation of research theme
        trends by employing a version of the Mann-Kendall test that is
        able to cope with multivariate time series of term occurrences.
        The results obtained by applying this general methodology to
        Information Security domain confirm its viability.
  </abstract>
  <article_number>10049568</article_number>
  <pdf_url>https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber
      =10049568</pdf_url>
  <html_url>
      https://ieeexplore.ieee.org/document/10049568/
  </html_url>
  <abstract_url>
      https://ieeexplore.ieee.org/document/10049568/
  </abstract_url>
  <publication_title>IEEE Access</publication_title>
  <publication_number>6287639</publication_number>
  <is_number>10005208</is_number>
  <publication_year>2023</publication_year>
  <publication_date>2023</publication_date>
  <start_page>18374</start_page>
  <end_page>18384</end_page>
  <citing_paper_count>0</citing_paper_count>
  <citing_patent_count>0</citing_patent_count>
  <download_count>117</download_count>
  <insert_date>20230222</insert_date>
  <index_terms>
   <ieee_terms>
    <term>Market research</term>
    <term>Bibliometrics</term>
    <term>Natural language processing</term>
    <term>Databases</term>
    <term>Data mining</term>
    <term>Time series analysis</term>
    <term>Information security</term>
    <term>Metadata</term>
   </ieee_terms>
   <author_terms>
    <terms>LDA topic modeling</terms>
    <terms>multivariate Mann-Kendall test</terms>
    <terms>natural language processing</terms>
    <terms>paper metadata</terms>
    <terms>research theme</terms>
    <terms>research trend</terms>
   </author_terms>
  </index_terms>
</article>
```

Figure 4.1: Paper metadata record from IEEE Xplore (continued)

Table 4.1: Metadata fields used in research-related activities [11]

Publication Metadata Field	**publication content categorization**	**scientific trend analysis**	**research gap identification**	**institution/researcher assessment**	**research topic framing**	**bibliography recommendation**
publication ID	–	–	–	✓	✓	✓
access (e.g., open, close, unavailable)	–	–	–	–	–	✓
publication type (e.g., article, review)	–	–	–	–	–	✓
title	✓	✓	✓	✓	✓	✓
keywords	✓	✓	✓	✓	✓	✓
abstract	✓	✓	✓	✓	✓	✓
author ID	–	–	–	✓	✓	✓
author name	–	–	–	✓	✓	✓
author affiliation	–	–	–	✓	–	✓
references	–	–	–	–	–	✓
year	✓	✓	✓	✓	✓	✓
citation count	–	✓	–	✓	✓	✓
citing patent count	–	✓	–	✓	✓	✓
download count	–	✓	–	✓	✓	✓

4.2 Bibliographic Metadata Preprocessing

To obtain a list of relevant key terms to describe the content of each publication, collected data fields that summarize the publication content (i.e., title, keywords, and, abstract) are preprocessed by following the methodology described in Figure 4.2. The procedure begins with a data cleansing stage aimed at filtering out all irrelevant records of the acquired bibliographic dataset. These entries can be easily removed by locating empty author, abstract, or keywords fields. They often relate to lists of authors, reviewers, or editors, tables of contents, errata, or indexes.

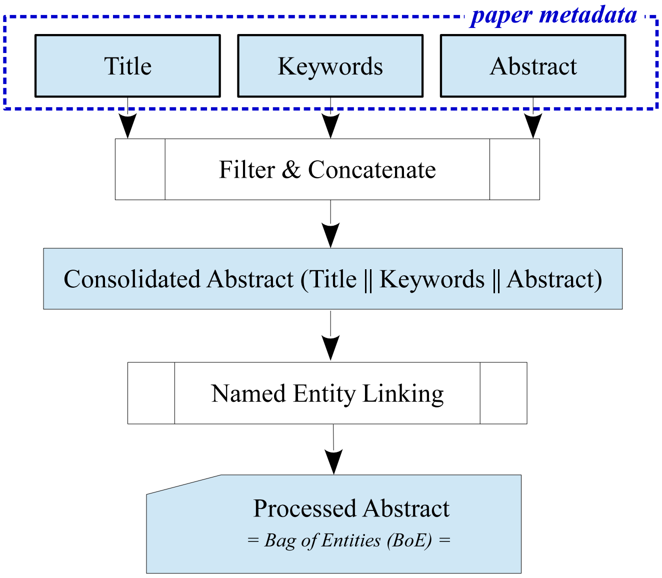


Figure 4.2: Data cleansing and key term extraction [14]

For each scientific publication, the associated bibliographic record generally includes three fields that are assumed and configured by the authors and that indubitably encapsulate the quintessence of their publication: title, keywords and abstract. As a direct consequence, employing these condensed forms of the publication in summarizing its content is, from our point of view, a justified decision. Thus, we formed a textual sequence, that we named consolidated abstract, by concatenating title, keywords and abstract fields.

In our view, there are two possibilities to further process the compound abstracts in order to extract relevant and meaningful terms to encapsulate the content of the paper: the classic bag-of-words model approach and the bag-of-entities model approach. These two approaches are briefly presented in the following.

The first option to process compound abstracts is the standard bag-of-words related procedure [53], which is based on the following four steps:

1. *Tokenization* – process of breaking text documents into simple units called tokens.
2. *Forming of n-grams* – building of tokens consisting of n consecutive words (e.g., "machine learning", "integrated circuit", "system on a chip"). n-grams have a specific meaning as a sequence of words rather than separate words. The major drawback here is given by the fact that in order to automatically build n-grams, the sequence of n consecutive words must occur at least twice in the text document. This was not the case in our scenario, because usually the same sequences were not used several times in a short text, such as our compound abstract.

3. *Data cleaning and stop words removal* – elimination of most common words in a language (e.g., "the", "a", "an", "in", etc.) that help in formation of sentences, but do not provide any significance in language processing. Here, we also need to include the list of insignificant words to the domain (e.g., "paper", "conclusion", "approach", etc.) provided by the user.
4. *Lemmatization* – it considers the morphological analysis of the tokens and returns meaningful words in a dictionary form.

The second possibility is to use a bag-of-entities model approach [54] that employs an entity linking method to extract meaningful representations, named mentions, from texts by relating them with entities from a knowledge base [55]. In this respect, one of the most popular entity linking software tool is TagMe [56], presented in detail in the following subsection, which provides the list of mentions from a given text that are also included as links in Wikipedia. Thus, a filtering of non-essential terms (e.g., removal of stop words) and non-specific terms to the domain is automatically performed. This process is conditioned only by the choice of a single parameter, namely the link-probability (lp) and in our case is able to provide a processed abstract in the form of a text document, where the mentions are separated by spaces and the words within n-gram mentions are separated by underscore characters.

We chose to process the compound abstracts using an entity linking procedure based on TagMe due to the following reasons: (a) the obtained list of relevant terms is more consistent with the investigated scientific domain in the case of TagMe; (b) the formation of n-gram terms is implicit and automatically done when using TagMe (i.e., compound words and sequences of terms that appear together in Wikipedia are automatically retrieved) compared to the difficulty of generating n-grams from relatively short texts in the case of bag-of-words based approaches; (c) the entire process can be conducted in a simple manner by choosing a single parameter, namely lp, in the case of TagMe, while for the bag-of-words approach the need for user intervention in configuring the stop words set for data cleaning is tedious and time-consuming.

Our data preprocessing procedure also allows the user to force terms (e.g., brand-new terms that are not yet covered in Wikipedia or terms that were previously dropped by choosing an inadequate value of the threshold lp for TagMe) to be evaluated and included. For this, each compound abstract is once again parsed to detect the user-enforced terms within it, and all identified enforced terms are then added at the end of the processed abstract.

4.2.1 TagMe Method and Its Parameter Selection

Finding relevant information within documents is a fundamental task in Natural Language Processing. In the past, information retrieval from text documents has been predominantly formalized as a problem of identifying the most relevant terms in the document [57]. In the last decade this trend is gradually switching towards understanding the documents through named-entities [58] provided by entity linking techniques. The Named-Entity Linking (NEL), also referred to as Named-Entity Recognition and Disambiguation (NERD), is a class of NLP methods that are aimed to discover the word sequences of interest (i.e., named-entities or mentions) by linking them to entities from a given knowledge base, e.g., Wikipedia.

NEL techniques are generally performed in two stages: (i) named-entity recognition phase, where the word sequences that might refer to an entity from the knowledge base are identified within the text documents; and, (ii) named-entity disambiguation phase, where each recognized word sequence is linked to a unique entity from the knowledge base.

Over the recent years, a series of NEL approaches have been proposed, including DBpedia Spotlight, AIDA, WikiMiner or Babelfy [59]. Among these, TagMe is considered one of the most influential ones, especially for on-the-fly annotating of short texts [56] and for its ease of integration within NLP frameworks using the provided RESTful API. TagMe is implemented as a three-step pipeline as follows [56, 60].

Parsing. After tokenization of the input document, TagMe performs mention recognition for all n-grams of the document ($n \leq 6$) using a dictionary built by collecting entities from anchor texts of Wikipedia articles, Wikipedia page titles, and title variants, and redirect pages. The obtained mentions are filtered by applying a lower threshold upon the link probability lp defined as:

$$lp(m) = \frac{link(m)}{freq(m)}, \tag{4.1}$$

where m represents the mention under investigation, $link(m)$ denotes the number of mention's occurrences as a link in Wikipedia, while $freq(m)$ represents the number of mention's occurrences in Wikipedia either as a link or not. It is worth mentioning that an n-gram which is already contained in a longer and also with higher link probability n-gram, is discarded. At the end of this step, a set of pairs made out of a mention and its corresponding candidate entity is obtained.

Disambiguation. Entity disambiguation process is performed based on a voting scheme, where the score $rel(m, e)$ for each mention-entity pair (m, e) is obtained by summing the votes $vote(m', e)$ provided by candidate entities of all the other m' mentions:

$$rel(m, e) = \sum_{m' \in M \setminus \{m\}} vote(m', e), \tag{4.2}$$

where M is the set of all identified mentions. The votes given by candidate entities can be computed using the following formula:

$$vote(m', e) = \frac{\sum_{e' \in E(m')} relatedness(e, e') \cdot commonness(m', e')}{|E(m')|}. \tag{4.3}$$

In equation (4.3), $E(m')$ is the set of mentions m' that are involved in the voting process.

The relatedness is a measure of the semantic relationship between two entities [61] and is given by:

$$relatedness(e, e') = \frac{log(max(|in(e)|, |in(e')|)) - log(|in(e) \cap in(e')|)}{log(|E|) - log(min(|in(e)|, |in(e')|))}, \tag{4.4}$$

where $in(e)$ denotes the set of entities linked to entity e and $|E|$ is the number of entities.

The *commonness* represents the probability of an entity e' being linked with of a given mention m' [62]:

$$commonness(m', e') = \frac{link(m', e')}{link(m')}. \tag{4.5}$$

In equation (4.5), $link(m', e')$ denotes the number of times entity e' is the link target for m' and $link(m')$ represents the total number of times the mention m' appears as a link.

Once the scores of all candidate entities are computed using equation (4.2), TagMe selects the best entity for each mention m in a two-phase procedure: (i) a list of candidate entities having a score equal or close to the best $rel(m, e)$ value is formed; (ii) the entity e having the highest $commonness(m, e)$ score within the list is selected as the winner entity. Thus, at the end of the entity disambiguation step, each mention from the input text is linked with the most pertinent unique entity.

Pruning. This step filters out all meaningless mentions based on a pruning score ρ obtained as an average between link probability *lp* provided by equation (4.1) and a *coherence* score defined as the average *relatedness* between the candidate entity and all other identified entities:

$$\rho(m) = \frac{1}{2}(lp(m) + coherence(m)) \tag{4.6}$$

with

$$coherence(m) = \frac{1}{|E(T)| - 1} \sum_{e' \in E(T) \setminus \{e\}} vote(m', e), \tag{4.7}$$

where $E(T)$ is the set of distinct mentions from the input text T, $|E(T)|$ represents the cardinality of this set, and e is the target entity for m. The final set of mentions includes only the mentions for which $\rho > \rho_{NA}$, where ρ_{NA} is a user-specified threshold that allows a balance between recall and precision and has a nominal value of 0.2.

To exemplify how TagMe works, we employ the TagMe RESTful API to obtain the list of mentions for two threshold values for *lp*, namely 0 (i.e., no threshold applied) and 0.1, for the following sentence extracted from Electronic Design Automation - Wikipedia page [63]:

"Electronic design automation is a category of software tools for designing electronic systems such as integrated circuits and printed circuit boards."

The list of obtained mentions in the two cases is presented below.

List of extracted mentions and their link probabilities lp (threshold $lp = 0$)

```
Electronic design automation [0,28] lp=1.0
category [34,42] lp=0.015567442402243614
software [46,54] lp=0.06506886333227158
software tools [46,60] lp=0.023204419761896133
designing [65,74] lp=0.0019603450782597065
electronic systems [75,93] lp=0.008771929889917374
integrated circuits [102,121] lp=0.20196352899074554
printed circuit boards [126,148] lp=0.16010499000549316
```

List of extracted mentions and their link probabilities lp (threshold $lp = 0.1$)

```
Electronic design automation [0,28] lp=1.0
integrated circuits [102,121] lp=0.20196352899074554
printed circuit boards [126,148] lp=0.16010499000549316
```

As we may notice, TagMe is able to extract relevant word sequences (i.e., mentions) from a short text, this being a viable starting point for other NLP methods that we applied, including document clustering or topic modeling. It also offers a better alternative to text preprocessing procedures based on the bag-of-words model not only by employing its inherent semantic analysis but also by its ease of use, the link probability lp threshold being the only parameter that needs to be selected.

In the following, our practical procedure for selecting a proper lp value is described. The link probability basically represents a threshold aimed at dropping the less relevant mentions within a given text. If the user wants to include in his analysis all the terms from the original text that are mentioned in Wikipedia, then $lp = 0$, while a higher value for lp may discard some relevant terms. In order to find a reasonable compromise, the user must choose a set of relevant terms from the scientific domain under investigation and evaluate each of them with TagMe to obtain their corresponding lp score. Finally, the user may select an lp value that is less than or equal to the lowest obtained lp. Table 4.2 shows an example of how to select a proper lp value in the case of the EDA domain.

Since the user's goal was not to exclude any of the essential terms of the domain from the processed texts, the value of lp in this case must be chosen so that $lp \leq lp('optimization') = 0.10528$.

Table 4.2: Selection of the lp threshold for EDA domain

Term	lp**-score**
Logic gates	0.14468
Integrated circuit	1.00000
Optimization	***0.10528***
FPGA	0.92440
Transistor	0.54157
CMOS	0.60634
System-on-Chip	0.34883

4.3 Datasets

4.3.1 Dataset Used for the Information Security Domain

To adequately describe the state of the Information Security domain during the 2010-2022 interval, we considered a prominent conference in the field, namely IEEE Symposium on Security and Privacy (A-ranked conference) held annually in the USA since 1980, and two top-tier journals: IEEE Transactions on Information Forensics and Security (6.8 IF in 2022) published since March 2006, and, IEEE Security & Privacy (1.9 IF in 2022) published since 2003.

The initial corpus collected in November 2022, contains 12374 bibliographic records as follows: 5264 from IEEE Symposium on Security and Privacy, 4166 from IEEE Transactions on Information Forensics and Security, and 2944 from IEEE Security & Privacy.

Following the paper metadata preprocessing described in Figure 4.2, the set of processed abstracts was obtained using $lp = 0.1$ for the TagMe entity linking procedure.

4.3.2 Dataset Used to Model IEEE Domains and Emerging Domains

To analyze and model the subdomains within the IEEE broad field, in November 2021 we collected a corpus of journal paper metadata for the interval 2010-2020, using the IEEE Xplore API. For this, we selected the representative journals for all IEEE societies and councils based on their impact factors extracted from Clarivate's Journal Citation Reports, and their first publication dates. These flagship journals are listed in Table 4.3. For example, to evaluate the state of the Electronic Design Automation (EDA) domain, we considered the flagship journal published by the IEEE Council on Electronic Design Automation in association with the IEEE Circuits and Systems Society, namely the IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD) (2.9 IF in 2022) published since 1982. It is worth mentioning that the IEEE Product Safety Engineering Society does not publish any journal of its own.

Table 4.3: Representative journals for IEEE societies and councils

No.	**IEEE Society / IEEECouncil**	**Flagship Journal**	**Journal Abbrev.**
1.	IEEE Aerospace and Electronic Systems	IEEE Transactions on Aerospace and Electronic Systems	TAES
2.	IEEE Antennas and Propagation	IEEE Transactions on Antennas and Propagation	TAP
3.	IEEE Broadcast Technology	IEEE Transactions on Broadcasting	TBC
4.	IEEE Circuits and Systems	IEEE Transactions on Circuits and Systems I: Regular Papers	TCSI
5.	IEEE Communications	IEEE Transactions on Communications	TCOMM
6.	IEEE Computational Intelligence	IEEE Transactions on Neural Networks and Learning Systems	TNNLS
7.	IEEE Computer	IEEE Transactions on Computers	TC
8.	IEEE Consumer Technology	IEEE Transactions on Consumer Electronic	TCE
9.	IEEE Control Systems	IEEE Transactions on Automatic Control	TAC
10.	IEEE Dielectrics and Electrical Insulation	IEEE Transactions on Dielectrics and Electrical Insulation	TDEI
11.	IEEE Education	IEEE Transactions on Education	TE
12.	IEEE Electromagnetic Compatibility	IEEE Transactions on Electromagnetic Compatibility	TEMC
13.	IEEE Electron Devices	IEEE Transactions on Electron Devices	TED
14.	IEEE Electronics Packaging	IEEE Transactions on Components, Packaging, and Manufacturing Technology	TCPMT
15.	IEEE Engineering in Medicine and Biology	IEEE Transactions on Biomedical Engineering	TBME
16.	IEEE Geoscience and Remote Sensing	IEEE Transactions on Geoscience and Remote Sensing	TGRS
17.	IEEE Industrial Electronics	IEEE Transactions on Industrial Electronics	TIE
18.	IEEE Industry Applications	IEEE Transactions on Industry Applications	TIA
19.	IEEE Information Theory	IEEE Transactions on Information Theory	TIT
20.	IEEE Instrumentation and Measurement	IEEE Transactions on Instrumentation and Measurement	TIM
21.	IEEE Intelligent Transportation Systems	IEEE Transactions on Intelligent Transportation Systems	TITS
22.	IEEE Magnetics	IEEE Transactions on Magnetics	TMAG
23.	IEEE Microwave Theory and Techniques	IEEE Transactions on Microwave Theory and Techniques	TMTT
24.	IEEE Nuclear and Plasma Sciences	IEEE Transactions on Nuclear Science	TNS
25.	IEEE Oceanic Engineering	IEEE Journal of Oceanic Engineering	JOE
26.	IEEE Photonics	IEEE Journal of Photovoltaics	JPHOTOV

Table 4.3: Representative journals for IEEE societies and councils (continued)

No.	IEEE Society / IEEECouncil	Flagship Journal	Journal Abbrev.
27.	IEEE Power Energy	IEEE Transactions on Power Systems	TPWRS
28.	IEEE Product Safety Engineering	-	-
29.	IEEE Professional Communication	IEEE Transactions on Professional Communication	TPC
30.	IEEE Reliability	IEEE Transactions on Reliability	TR
31.	IEEE Robotics and Automation	IEEE Transactions on Robotics	TRO
32.	IEEE Signal Processing	IEEE Transactions on Signal Processing	TSP
33.	IEEE Social Implications of Technology	IEEE Transactions on Technology and Society	TTS
34.	IEEE Solid-State Circuits	IEEE Journal of Solid-State Circuits	JSSC
35.	IEEE Systems, Man, and Cybernetics	IEEE Transactions on Systems, Man, and Cybernetics: Systems	TSMC
36.	IEEE Technology and Engineering Management	IEEE Transactions on Engineering Management	TEM
37.	IEEE Ultrasonics, Ferroelectrics, and Frequency Control	IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control	TUFFC
38.	IEEE Vehicular Technology	IEEE Transactions on Vehicular Technology	TVT
39.	IEEE Biometrics Council	IEEE Transactions on Biometrics, Behavior, and Identity Science	TBIOM
40.	IEEE Council on Electronic Design Automation	IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems	TCAD
41.	IEEE Council on RFID	IEEE Journal of Radio Frequency Identification	JRFID
42.	IEEE Council on Superconductivity	IEEE Transactions on Applied Superconductivity	TASC
43.	IEEE Nanotechnology Council	IEEE Transactions on Nanotechnology	TNANO
44.	IEEE Sensors Council	IEEE Sensors Journal	JSEN
45.	IEEE Systems Council	IEEE Systems Journal	JSYST

Aside from the information about the scientific papers belonging to the flagship journals, we additionally collected bibliographic records for articles corresponding to the IEEE Internet of Things Journal (JIOT) to characterize the emergent Internet of Things domain.

In the case of each journal, we created a CSV file where we stored a set of bibliographic record fields (i.e., title, abstract, keywords, authors, publication date, and digital object identifier) for each article, and also the processed abstracts obtained by applying the TagMe procedure using $lp = 0.1$ to the concatenated title, keywords, and abstract fields.

4.3.3 Dataset Used for Research Team Formation

The raw bibliographic metadata corpus, including the fields 'authors', 'title', 'keywords', 'abstract', 'content type', 'citing paper count', 'download count', 'citing patent count', 'index terms', 'doi', and 'publication year', was extracted from IEEE Xplore on July 4, 2023. The records correspond to publications from the interval 2010 to 2022, having at least one author affiliated with Politehnica University Timisoara (UPT) – Romania. The corpus is made of 1992 records belonging to 1179 authors, anonymized to meet the regulations regarding data protection and privacy. The time distribution of the number of publications is displayed in Figure 4.3, while the total number of unique authors each year is displayed in Figure 4.4. In order to extract the key terms from 'title', 'abstract' and 'keywords' metadata fields, we used the TagMe entity linking procedure with a link probability threshold $lp = 0.1$.

To assess the use of bibliographic metadata for research team formation we provided information regarding researchers' expertise in the Electrical and Electronic Engineering field, and also information focused on their teamwork skills. The resulting dataset is structured in four collections, each of them being characterized by the metadata fields employed in extracting their list of key terms:

- Case_1: 6493 key terms were extracted from 'title', 'keywords', and 'abstract' metadata fields;
- Case_2: 2651 key terms were extracted from 'title' and 'keywords' metadata fields;
- Case_3: 1844 key terms from 'title' metadata fields;
- Case_4: 1254 key terms were extracted from 'keywords' fields.

Each of the four collections includes nine CSV tables:

- Individual_Expertise_and_Collaborators.csv - for each anonymized author it contains the number of authored publications, number of citations, number of citations in patents, number of downloads, number of distinct co-authors having the same affiliation, and number of distinct co-authors having other affiliations;

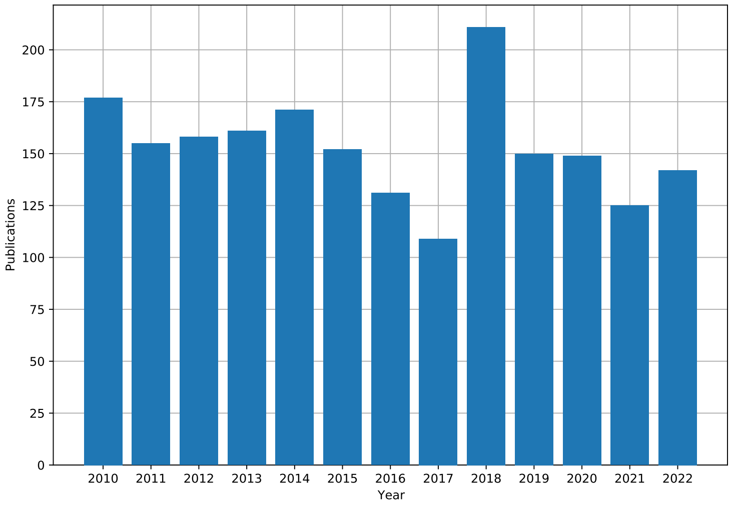


Figure 4.3: Publications per year [17]

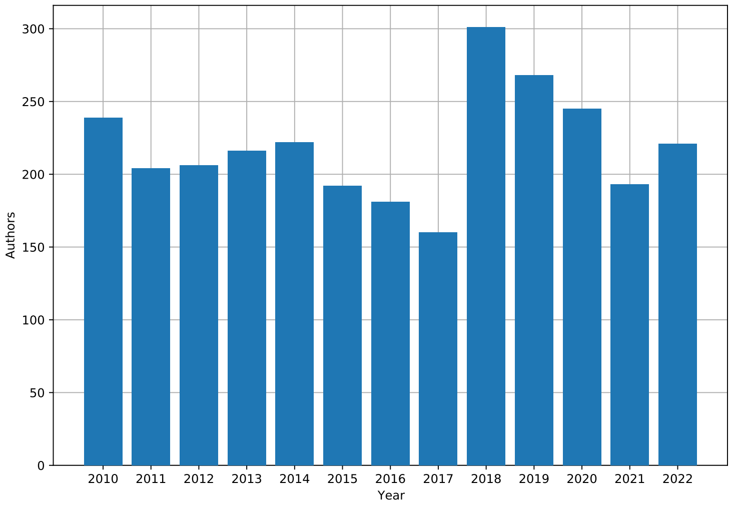


Figure 4.4: Unique authors per year [17]

- Collaborations_number.csv - a symmetric matrix that provides for each pair of researchers the corresponding number of co-authored papers.

- Collaborations_citations.csv - a symmetric matrix that provides for each pair of researchers the corresponding number of citations received by co-authored papers.

- Collaborations_citations_patent.csv - a symmetric matrix that provides for each pair of researchers the corresponding number of citations received by co-authored papers in patents.

- Collaborations_downloads.csv - a symmetric matrix that provides for each pair of researchers the corresponding number of downloads received by co-authored papers.

- KeyTerms_number.csv - offers the number of papers published by each author containing the identified key terms.

- KeyTerms_citations.csv - offers the number of citations received by author's papers that contain the identified key terms.

- KeyTerms_citations_patents.csv - offers the number of citations in patents received by author's papers that contain the identified key terms.

- KeyTerms_downloads.csv - offers the number of downloads received by author's papers that contain the identified key terms.

To provide a bird's eye view above the existing collaborations inside Politehica University, we constructed the collaborative graph using Collaborations_number.csv table and presented it in Figure 4.5. As we may notice the graph is extremely sparse, indicating that the research work is done in small and isolated teams.

The full dataset is hosted in the Mendeley Data repository [16], details being also presented in our journal data paper [17].

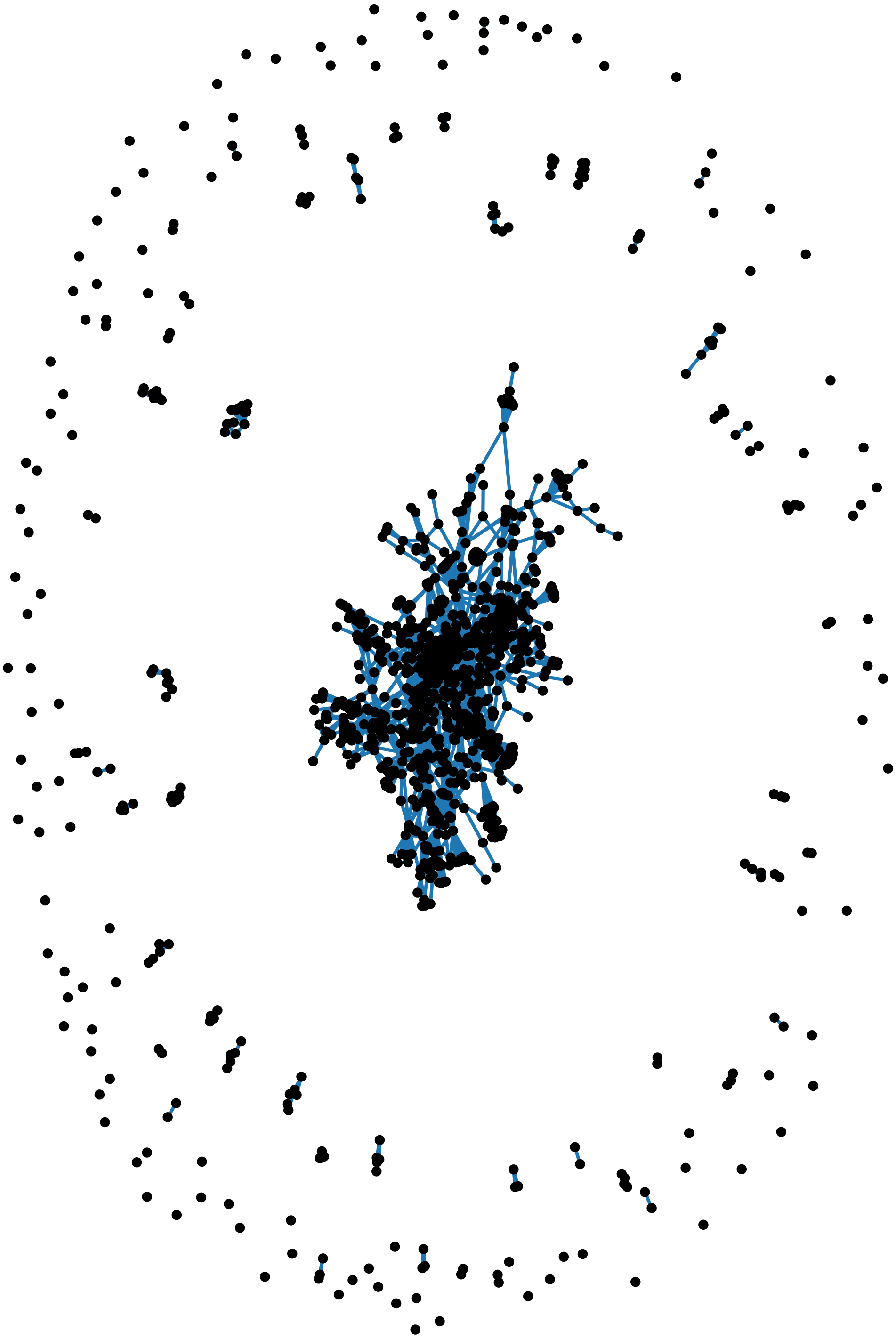


Figure 4.5: Collaborative graph for UPT scholars [17]

Chapter 5

Research Theme Recommender

Discovering realistic and suitable research themes to work on is a crucial activity for every researcher. This chapter provides a human-in-the-loop recommender module aimed to identify research themes from publication metadata and evaluate their hotness and feasibility. The chapter encapsulates the methods and results presented in our papers [12, 13, 14, 15].

5.1 Preliminaries and Related Work

Like all other types of projects, research themes have their own life cycle. Every now and then researchers must choose new projects to work on for a variety of reasons, which may be either related to the current research theme (e.g., the research question has already been answered; no outstanding outcomes are expected; investigation comes to an end because of a shortage of human, material, or financial resources, as well as a lack of novel ideas, etc.) or to the recent evolution of scientific knowledge (e.g., development of brand new methods, theories, or technologies). In this respect, framing new and feasible research themes is neither an easy nor quick-to-achieve endeavor since it has to be correlated with the latest research trends and recent developments in the field and moreover, to meet the researchers' expectations, interests and existing expertise.

Identifying novel research subjects within a specific scientific domain by analyzing the semantic information retrieved from bibliographic databases is a challenging Natural Language Processing (NLP) issue. While previous NLP approaches mainly model research subjects using unique key terms, we go one step further by appropriately characterizing research themes as sets of key terms. Our proposed approach consists of three steps: (i) employing LDA topic modeling to identify research themes from paper metadata; (ii) assessing research theme opportunity using a modified Mann-Kendall test that can handle multivariate time series of key term occurrences; and, (iii) assessing research theme viability using a statistical double-threshold technique. Based on the insights encapsulated in publication metadata, we propose a semi-automatic Human-in-the-Loop

Recommender System (HLRS) meant to assist researchers in framing promising and personalized research themes. The experimental results obtained when employing this recommender prove its effectiveness and feasibility.

Surveying the scientific literature to pinpoint research themes or topics and evaluate their perceived attractiveness and timeliness is an important objective in every scientific domain. While the conventional approach generally relies on expert knowledge derived from a top-down qualitative analysis of a large corpus of publications, a new NLP data-driven line of action arises. The latter generally uses the information acquired from bibliographic databases to reveal the research topics or trends based on the time analysis of key term occurrence, the temporal or spatial spreading of ideas, the context and content of citations, etc. Because of their monotonous, menial and repetitive nature, these quantitative bottom-up procedures are susceptible to be automatized by applying a two-step methodology [43] that starts with an automated research theme detection phase, followed by an automated research theme assessment in terms of suitability, opportunity and significance.

Even though a fully automatic research theme recommendation system remains an unfulfilled desideratum, some steps toward this goal have already been taken, the scientists proposing diverse methods to address the following challenges: identifying the research topics within scientific domains, research trend forecasting, and research hotspots detection.

Detecting research topics in scientific publications

A research topic may be defined either as a set of theories, concepts, phenomena, methods and technologies, or a broad problem area that is worth exploring to augment mankind's body of knowledge. To aid in identifying research topics from bibliographic metadata records, NLP offers a variety of methods which includes document classification [64], document clustering [65], or author co-citation analysis [66]. In this regard, the topic modeling-based technique is quickly becoming the norm, with Latent Dirichlet Allocation (LDA) [67] and its variations being used most frequently in discovering research topics inside scientific literature documents previously processed as bag-of-words [8, 68, 69] or bag-of-entities [15].

Two aspects must be carefully considered when trying to automate these types of topic modeling approaches [70]. The first one is related to the number of topics set to be extracted, this parameter decisively affecting the granularity of results [71]. The mean number of key terms in a topic will rise when fewer topics are sought, leading to the identification of wide study fields. Conversely, a sufficiently large number of topics will output research themes, which are often distinguished by a reduced number of key terms. While previously reported works were specifically designed to discover broad research areas within certain scientific fields, our technique will focus on extracting research themes, hence improving the topic granularity. The second aspect is related to the difficulty in interpreting the resulting topics, namely in transforming the set of key terms belonging to a topic into a plausible research theme. From our perspective, a parameter fine-tuning procedure has to be undertaken to devise pertinent research themes with adequate broadness.

Analyzing the burstiness and the trend of research topics

The last decade has witnessed a tremendous rise in the quest to understand the research dynamics based on information extracted from scholarly publications. Research trend assessments become extremely important anytime researchers are asking for new scientific niches to be investigated and are generally grounded in the methodical trend analysis within key terms or citation time series. As a result, diverse trend analysis techniques aimed at tracking topic evolution [72], unveiling research hotspots [43], or forecasting scientific trends [73] have been reported.

Relying the topic trend assessments on citation-based bibliometric indicators is detrimentally influenced by the time required for completing the citing research and the publication latency [12] and, for this reason, is inappropriate for rapidly-evolving fields such as Artificial Intelligence, Internet of Things, or Quantum Computing. This opens a wide window for content-based NLP techniques that employ time series analysis for the time evolution of key-term occurrences inside scientific publications. In this regard, the most popular trend analysis techniques include the Mann-Kendall trend test, often coupled with Sen's slope estimator [74, 43], and time-series regression procedures [69, 70], while for investigating the topic burstiness, Kleinberg's burst detection algorithm is frequently used [75, 76, 43, 77, 78]. All the mentioned techniques were designed to assess research trends based on univariate time series. Because in our perspective the research themes may be better characterized using sets of key terms, employing univariate time series is ineffective; instead, multivariate techniques are required. In our recent paper [15], we solved this issue by proposing a topic trend assessment mechanism based on a multivariate extension of the Mann-Kendall trend test.

To the best of our knowledge, our suggested technique for finding new research topics is the first to model the themes using a collection of key terms rather than a single or a pair of co-occurring key terms. This approach is justified by some notable advantages including the increased research theme modeling accuracy, the option to select the granularity of the research themes to be framed by choosing the number of key terms comprised by the theme model, and, the significantly improved specificity and interpretability of resulting research themes. Based on this, we aim to design a complex and modular recommender system able to aid both research theme framing and their evaluation in terms of hotness and feasibility, using a human-in-the-loop approach.

5.2 Methods

This section briefly presents the three methods (i.e., Latent Dirichlet Allocation, auto-ARIMA, and multivariate Mann-Kendall) representing the foundation upon which our research theme recommender is built.

5.2.1 Latent Dirichlet Allocation

In NLP, topic modeling is a procedure designed to identify and characterize topics occurring in a document corpus by employing a probabilistic model. It is commonly used as a text mining technique to discover semantic structures within texts. The most prominent such technique, namely Latent Dirichlet Allocation (LDA), was originally reported by Blei et al. [67] and uses an unsupervised generative probabilistic representation of the likelihood of term co-occurrences for extracting latent topics.

Let us consider a text corpus C containing N documents. LDA provides the set of K hidden topics associated to C by using a three-step generative process [67]:

1. Pick a symmetric multivariate beta distribution (i.e., Dirichlet distribution) prior, with β being its concentration parameter, over the φ_k multinomial term distribution for topic k ($k = 1, 2, ..., K$);

2. Pick a symmetric Dirichlet prior, with α being its concentration parameter, over the θ_i multinomial topic distribution for document d_i ($i = 1, 2, ..., N$);

3. Considering the pair (i, j) as identifying the position of each term $w_{i,j}$ in the document corpus (j refers to the location of the term inside the document i), with $i = 1, 2, ..., N$, $j = 1, 2, ..., N_{d_i}$ and N_{d_i} being the length of the document d_i:

   - From θ_i distribution, draw a topic assignment $z_{i,j}$;
   - From φ_k distribution, draw a term $w_{i,j}$.

The mentioned generative probabilistic process is characterized by the joint probability distribution of the observed and hidden variables:

$$p(w, z, \theta, \varphi \mid \alpha, \beta) = p(\varphi \mid \beta) \cdot p(\theta \mid \alpha) \cdot p(z \mid \theta) \cdot p(w \mid z, \varphi). \tag{5.1}$$

Considering that the draws of position terms $w_{i,j}$, topics k and documents d_i are independently done, the equation (5.1) becomes:

$$\begin{split} p(w, z, \theta, \varphi \mid \alpha, \beta) = \prod_{k=1}^{K} p(\varphi_k \mid \beta) \cdot \prod_{i=1}^{N} p(\theta_i \mid \alpha) \cdot \prod_{i=1}^{N} \prod_{j=1}^{N_{di}} p(z_{i,j} \mid \theta_i) \cdot \\ \cdot \prod_{i=1}^{N} \prod_{j=1}^{N_{di}} p(w_{i,j} \mid z_{i,j}, \varphi_i). \end{split} \tag{5.2}$$

In order to discover the latent topics in the document corpus based on the observed variables, the posterior distribution of hidden variables is calculated as follows:

$$p(z, \theta, \varphi \mid w, \alpha, \beta) = \frac{p(w, z, \theta, \varphi \mid \alpha, \beta)}{p(w \mid \alpha, \beta)} \tag{5.3}$$

After this, the three variables z, θ, and φ have to be recovered for the considered document corpus.

It is worth mentioning that the direct evaluation of the posterior distributions based on equation (5.3) is usually impracticable due to the normalization term $p(w|\alpha, \beta)$ which has the following formula:

$$p(w \mid \alpha, \beta) = \int_{\varphi} p(\varphi \mid \beta) \int_{\theta} p(\theta \mid \alpha) \int_{z} (p(z \mid \theta) p(w \mid z, \varphi)) \, dz \, d\theta \, d\varphi. \tag{5.4}$$

Because of the coupling of the φ and θ variables inside the most right integral (i.e., integral over the topic assignment variable z), the equation (5.4) is generally intractable. This problem may be circumvented by employing approximate or variational inference techniques such as collapsed Gibbs sampling [79], stochastic variational inference [80], or Bayesian variational inference [67].

5.2.2 Time-Series ARIMA Model Prediction. Auto-ARIMA Method

In many fields of activity, it is quite a common practice to gather observations over time. This type of information is usually represented in the form of discrete time series which are sequences of values X_t characterizing successive points in time t. The time series analysis is based on suitably chosen mathematical models and is mainly directed towards two objectives: (a) understanding the underlying factors and mechanisms that characterize the observed data sequence; or, (b) forecasting future observations based on past ones.

A commonly employed class of models, mainly for non-seasonal time series, is represented by the Autoregressive Integrated Moving Average (ARIMA) models originating in the work of Box and Jenkins [81] which extended the classic Autoregressive Moving Average (ARMA) model to improve the predictive performances in the case of nonstationary time series [82].

ARMA models [83], specified as $ARMA(p, q)$, characterize the stationary stochastic processes using two polynomials that respectively describe the autoregressive ($AR(p)$) and moving-average ($MA(q)$) components:

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-1} + \sum_{j=1}^{q} \theta_i \varepsilon_{t-1} + \varepsilon_t + c. \tag{5.5}$$

Here, p and q are the autoregressive and respectively the moving average orders, φ_i and θ_j are model parameters, ε_t is a white noise, while c is a constant.

The model depicted by (5.5) may be reshaped in a more compact form (5.6) by utilizing the lag operator L described by $LX_t = X_{t-1}$:

$$\varphi(L)X_t = \theta(L)\varepsilon_t + c. \tag{5.6}$$

In this case, the two polynomials in L have the following forms:

$$\varphi(L) = 1 - \varphi_1 L - \varphi_2 L^2 - \cdots - \varphi_p L^p \tag{5.7}$$

and

$$\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q. \tag{5.8}$$

In practice, ARMA models are utilized only for time-series with statistical properties that are not changing in time (i.e., stationary time-series) [82]. To cope with nonstationary time series, Box and Jenkins [81] employed a differencing technique to transform nonstationary into stationary time series. They basically replaced all the observations of a given time series X_t with their first difference resulting in a new time series Y_t described by:

$$Y_t = X_t - X_{t-1} = (1 - L)X_t. \tag{5.9}$$

By generalizing (5.9), we may formalize the d^{th} differences as:

$$Y_t = (1 - L)^d X_t. \tag{5.10}$$

Considering (5.10), we may generalize the ARMA model (5.6) into an AutoRegressive Integrated Moving Average (ARIMA) model, denoted as $ARIMA(p, d, q)$ and having the following form [82]:

$$\varphi(L)(1 - L)^d X_t = \theta(L)\varepsilon_t + c, \tag{5.11}$$

where the d parameter represents the differencing order applied for changing the original time series into a stationary one.

In the case that the triad (p, d, q) is known, the φ_i and θ_j coefficients may effectively be derived by employing a maximum likelihood parameter estimation technique [84].

The problem of obtaining the most fitted p, d, and q values can be solved by utilizing the auto-ARIMA method [85]. This simple method varies each of the p, d, and q ARIMA model orders inside given intervals and uses a goodness-of-fit test to select the most accurate $ARIMA(p, d, q)$ model. For our Python implementation, we chose as the goodness-of-fit indicator the Akaike Information Criterion (AIC):

$$AIC = 2k - 2log(\widehat{\mathfrak{L}}). \tag{5.12}$$

In this equation, $\widehat{\mathfrak{L}}$ represents the maximal value of the likelihood function of the ARIMA model, while k, being the number of estimated parameters, is either $k = p + q + 2$ for $c \neq 0$ or $k = p + q + 1$ for $c = 0$ [85]. The best trade-off between the goodness-of-fit and simplicity of the model is reached for a minimal AIC value.

5.2.3 Mutivariate Mann-Kendall Test

The Mann-Kendall (MK) [86, 87] test belongs to the non-parametric statistical techniques group for trend identification and has progressively become the standard methodology for assessing monotonic trends in NLP [13, 43]. Its appeal arises from its ability to handle censored and non-Gaussian data, as well as its ease of use. The next paragraphs provide a brief description of this method for both univariate and multivariate scenarios.

We consider X_i with $i = 1, 2, \dots, N$ being a time-stamped series of N observations. The MK test examines the shifts in signs that correspond to the differences between consecutive X_i data points by evaluating the S-statistic computed with:

$$S = \sum_{k=1}^{N-1} \sum_{j=k+1}^{N} sgn(X_j - X_k), \tag{5.13}$$

with $sgn(X)$ being the sign function. Analyzing (5.13) we may observe that a positive S value describes an ascending time series trend, while a negative S value characterizes a descending trend.

If a sufficiently large number N of observations (e.g., $N \geq 10$) is considered, the S-statistic has an approximately normal distribution with a zero mean (i.e., $E(S) = 0$) and a variance $\sigma^2(S)$ that can be obtained using the following equation:

$$\sigma^2(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{k=1}^{M} r_k(r_k - 1)(2r_k + 5)\right] \tag{5.14}$$

with M being the number of consecutive data points having the same values (i.e., tied group), while r_k being the rank corresponding to the tied group k.

Using equations (5.13) and (5.14), we are able to compute the MK Z-statistic, denoted by Z_{MK}, characterized by a zero mean and unit variance:

$$Z_{MK} = \begin{cases} \frac{S+1}{\sigma(S)} & for \quad S < 0 \\ 0 & for \quad S = 0 \\ \frac{S-1}{\sigma(S)} & for \quad S > 0 \end{cases}. \tag{5.15}$$

A positive value for Z_{MK} describes an increasing trend and a negative value is a characteristic of decreasing trends.

The univariate version of the Mann-Kendall test presented above was generalized to a multivariate MK test by Lettenmaier [88]. His proposed technique combines the trend information gained from each individual time series into an adapted S-statistics based on the covariance matrix [89], the trend being evaluated by performing the following steps [90, 91]:

1. The MK S-statistic is computed separately for each individual time series X using the equation (5.13).

2. Based on [89], a covariance matrix, denoted by Γ, is built by computing each of its Γ_{ij} elements with the following formula:

$$\Gamma_{XY} = \frac{1}{3}\left[K + 4\sum_{j=1}^{N} R_{jX}R_{jY} - N\left(N+1\right)^2\right], \tag{5.16}$$

   where N represents the number of time points in the multivariate time series, X and Y denote two univariate components of the multivariate time series, while coefficients K and R_jX are computed as follows:

$$K = \sum_{1\leq i<j\leq N} \operatorname{sgn}\left[\left(X_j - X_i\right)\left(Y_j - Y_i\right)\right] \tag{5.17}$$

$$R_{jX} = \frac{1}{2}\left[N + 1 + \sum_{i=1}^{N} \operatorname{sgn}\left(X_j - X_i\right)\right] \tag{5.18}$$

3. The Z-statistics of the multivariate time series is obtained using:

$$Z = \frac{\sum_{i=1}^{d} S_i}{\sqrt{\sum_{j=1}^{d}\sum_{i=1}^{d}\Gamma_{ij}}}, \tag{5.19}$$

   where d is the number of univariate components of the multivariate time series, S_i denotes the S-statistic for the i^{th} variate, and Γ_{ij} is an element of the covariance matrix previously computed using (5.16).

In our Python implementation, to evaluate the trends, we utilized the multivariate Mann-Kendall test in the form of *multivariate_test()* function from the *pyMannKendall* package [92].

5.3 Problem Formulation and Solving Strategy

Our proposed semi-automatic research themes recommender system is meant to solve the following problem:

Problem formulation. *Let us consider a given scientific domain D and a document corpus C made of processed bibliographic data (i.e., processed abstracts) that effectively characterize the domain. We aim to develop a NLP-based method to discover hot and feasible research themes within this domain.*

We consider that the given scientific domain D can be adequately described by a finite set of key terms T_q, with $q = 1, ..., Q$ and that each research theme RT_r, with $r = 1, ..., R$ can be modeled by a set of w key terms.

Solving strategy. After acquiring and preprocessing the bibliographic records corresponding to journals and conferences that are representative for D, we identify the set of domain-characteristic key terms T_q by analyzing the term frequency in the document corpus C, discover the domain's research themes RT_r by analyzing and clustering the semantic information from C and evaluate the trend and feasibility of these themes to offer valuable research topic recommendations.

Following this strategy, we designed and built a human-in-the-loop recommender system having the architecture presented in Figure 5.1, including the needed human interventions during execution.

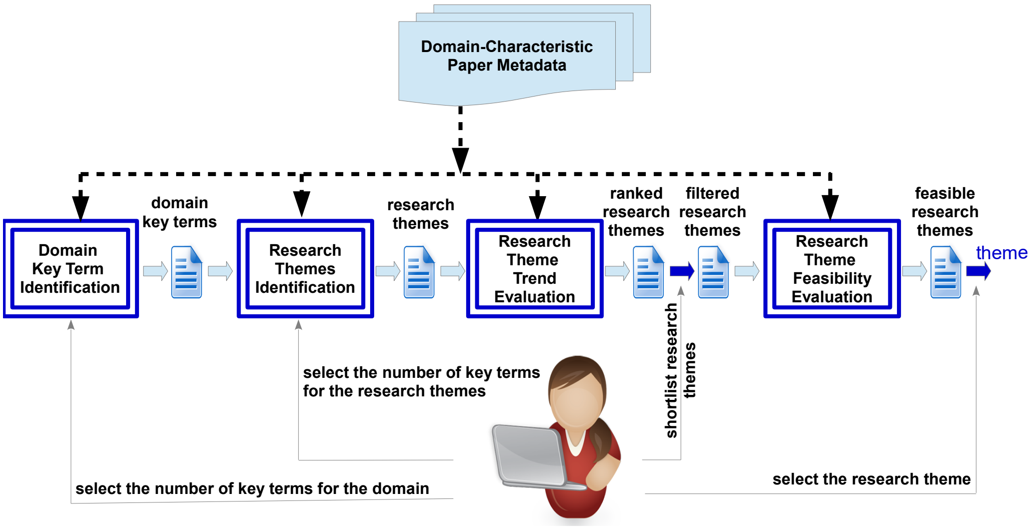

Domain-Characteristic Paper Metadata
domain key terms
research themes
ranked research themes
filtered research themes
feasible research themes
Domain Key Term Identification
Research Themes Identification
Research Theme Trend Evaluation
Research Theme Feasibility Evaluation
theme
shortlist research themes
select the number of key terms for the research themes
select the number of key terms for the domain
select the research theme

Figure 5.1: Research themes recommender system architecture

The recommender takes as input an adequately large corpus C of processed abstracts and provides a list of hot and also feasible research themes that will be subjected to the user's critical examination for selecting the best alternative to work on. The corpus C of processed paper metadata has to meet the following requirements: (i) to completely and, if possible, uniformly cover the entire domain D, leaving no sub-domain or scientific area within the domain left aside; (ii) to have a continuous time coverage of the domain for at least ten years to effectively identify the research trends; and, (iii) to include only peer-reviewed scientific materials to certify that the results of published research are original, logical, significant, and thorough. For this, the user needs to select the domain's flagship periodicals (e.g., renowned journals and yearly conferences) based on their reputation and recognition in the scientific community reflected by exceptional bibliometric indices (e.g., Clarivate's journal impact factor, Elsevier's CiteScore). For example, in the case of investigating the research topics inside the domain of Electronic Design Automation, we may consider journals like IEEE Transactions on Computer Aided Design of Integrated Circuits & Systems (the flagship journal of the IEEE Council on Electronic Design Automation) and ACM Transactions on Design Automation for Electronic Systems (flagship of ACM Special Interest Group on Design Automation) and prestigious yearly conferences like Design Automation (DAC); Design, Automation, Test in Europe (DATE); Asia and South Pacific Design Automation (ASPDAC); and, International Conference on Computer-Aided Design (ICCAD).

This recommender consists of four functional blocks that are sequentially excuted:

1. *Domain Key Term Identification:* derives a comprehensive and ranked list of key terms to accurately model the current state of the scientific domain D by evaluating the key term frequencies in the processed abstracts from the last one to three years;

2. *Research Themes Identification:* extracts the research topics that characterize the scientific domain D by performing topic modeling (i.e., LDA method) on the same document corpus used by the previous block;

3. *Research Theme Trend Evaluation:* investigates the "hotness" of each research theme by assessing its trend with a suitable multivariate variant of the classic Mann-Kendall trend test. In the case the publication latency corresponding to journal or conference papers cannot be ignored, to compute more accurate trends a novel method that combines auto-ARIMA and multivariate Mann-Kendall methods was designed;

4. *Research Theme Feasibility Evaluation*: examines each research theme in terms of its novelty and presumed success and categorizes it as feasible or not using a double-threshold method.

Each of these blocks is described in detail in the next four sections, while coping with the publication latency-related issue is presented in subsection 5.6.1.

The accuracy of a fully automatic version of our recommender system is drastically affected by a multitude of inherent factors: due to the sequential type of our proposed recommending process, the errors are propagating and accumulating; NLP procedures including the extraction of key terms from text documents, topic modeling and trend analysis are generally error-prone, needing expert assistance for tuning and performance optimization; bibliographic data are often biased or incomplete. In order to alleviate the accuracy degradation we opted for a human-in-the-loop approach, where the human expert is directly involved in selecting the appropriate parameters (i.e., the number of key terms to characterize the domain or research themes), shortlisting the intermediate research themes to be evaluated for feasibility according to her/his expectations and interests and finally in selecting the research theme to be undertaken.

5.4 Domain Key Term Identification

Our attempt to model a particular scientific domain derives from the plausible assumption that any scientific domain can be described by a finite set of relevant key terms and their complex interactions [14]. In our perspective, the set of domain-characteristic key terms can be obtained by employing a classic NLP procedure that ranks all terms from a given corpus C based on their occurrence frequency and retains a fair number of best-ranked terms. This procedure is detailed in the following paragraphs.

If we consider a given scientific domain D, we may derive its associated group of key terms T_q, $q = 1, ..., Q$, by computing the normalized document frequency ndf for all the entities (i.e., mentions) found inside the corpus of processed abstracts for the considered publications:

$$ndf(t, C) = \frac{df(t, C)}{N} \tag{5.20}$$

where N denotes the number of processed abstracts in the corpus C and $df(t, C)$ represents the document frequency of a mention t in that corpus.

Because we plan to evaluate the trends corresponding to current research themes, we must consider only the most recent descriptors for D. Consequently, we opt to compute the normalized document frequencies for articles released in the prior three years. After sorting the entities in decreasing order of ndf values, only the top Q entities will be kept in the set of key terms to characterize the scientific domain, Q being generally chosen based on the broadness of D and also on the percentage of outliers lying in the key term list. From this perspective, a suitable value may be selected inside the $[300, 1000]$ interval.

To illustrate how the domain key term identification procedure works, an example from the domain of Information Security [14] is presented next.

Example: Information Security domain key terms identification[2]

In the case of the Information Security domain, we took into account that the first-ranked $Q = 300$ key terms, when considering the publications from the last three years, are enough to model the current domain status [14]. To exemplify, Table 5.1 presents the first 34 key terms, together with corresponding ndf values. As expected, some inconsistent terms may be noticed, for example, the key term *leverage* (rank 28) and the key term pairs describing the same concept *internet_of_things* – *iot* (ranks 14 and 10), and *the_internet* – *internet* (ranks 25 and 16). Such inconsistencies are the direct result of duplicate entities included in Wikipedia and caught by TagMe. From this perspective, considering a number $Q' = 210$ of meaningful key terms (i.e., a percentage of $(1 - p) = 0.7$ from Q) seems a reasonable selection to cope with the issue.

[2] This example, uses only the processed abstracts belonging to the years 2020-2022 from the dataset described in subsection 4.3.1

Table 5.1: Top Information Security key terms [14]

Rank	Key Term	ndf	Rank	Key Term	ndf
1.	feature_extraction	0.103	18.	cryptographic	0.035
2.	algorithm	0.088	19.	optimization	0.034
3.	machine_learning	0.076	20.	semantics	0.034
4.	data_privacy	0.066	21.	encryption	0.033
5.	cryptography	0.066	22.	cybersecurity	0.032
6.	deep_learning	0.062	23.	differential_privacy	0.031
7.	task_analysis	0.056	24.	cloud_computing	0.031
8.	authentication	0.055	25.	the_internet	0.031
9.	computational_modeling	0.054	26.	face_recognition	0.030
10.	iot	0.045	27.	wireless_communication	0.028
11.	blockchain	0.043	28.	leverage	0.028
12.	computer_security	0.040	29.	perturbation_methods	0.026
13.	neural_networks	0.039	30.	android	0.024
14.	internet_of_things	0.039	31.	social_networking	0.024
15.	access_control	0.036	32.	computer_architecture	0.023
16.	internet	0.036	33.	security	0.022
17.	malware	0.036	34.	biometrics	0.021

For obtaining the set of Q key terms to model a domain D, we implemented a Python procedure based on $CountVectorizer()$ function from $sklearn.feature\_extraction$, and the $numpy$ and $pandas$ libraries.

5.5 Identifying Domain-Specific Research Themes Using LDA

To find the major research themes that characterize a given scientific domain D we investigate the semantic links between domain-specific keywords within the corpus of processed abstracts using the LDA topic modeling approach [14].

Topic modeling is an NLP approach that identifies latent clusters of linked terms from a collection of textual documents. Our methodology models the research themes as groups of key terms and extracts them as topics resulting from the standard Latent Dirichlet Allocation method. For this, selecting the number of topics k is of crucial importance for both the granularity and size of the obtained research themes. While employing existing automated techniques to extract the topic number k [93] fails to directly target the topic granularity, we derived this LDA parameter based on the following empirical equation:

$$k = \frac{(1-p)Q}{M} = \frac{Q'}{M} \tag{5.21}$$

In this formula, M represents the key term number assumed to model the research themes (it assures the needed level of granularity), p denotes the percentage of the inconsistent domain key terms selected for counteracting the imprecision in deriving the key terms from the processed abstracts, and Q' is the number of meaningful key terms describing the domain D (i.e., the number of domain's key terms excluding the outliers).

In this particular case, two essential insights are worth mentioning. First, considering the LDA's internal clustering mechanisms, the resulting topic number is less than or equal to k, therefore the research themes' granularity can not be raised beyond a certain limiting value. Second, the percentage of meaningless domain key terms, namely p is also contributing to the number of meaningful terms M':

$$M' = (1-p)M \tag{5.22}$$

As an illustrative example, let us consider a scientific domain modeled by a set of $Q = 400$ key terms, which includes both meaningful and irrelevant (i.e., outliers) key terms. If a quarter of these terms are meaningless ($p = 0.25$) and if the targeted research themes are each described by $M = 8$ key terms, equation (5.21) will provide $k \approx 38$ topics.

Our methodology to find the research themes that are specific to a given scientific domain includes the following three phases:

(a) apply the Latent Dirichlet Allocation algorithm to the collection of processed abstracts when considering the targeted value k for the number of clusters;

(b) for each of the obtained topics, keep only the terms included in the set of domain-relevant key terms T_q, $q = 1, ..., Q$;

(c) by targeting only the research themes that are modeled by at least M' relevant key terms, drop the topics that do not fulfill this requirement.

Example: Research themes identification in Information Security domain[3]

This example is a continuation of the example presented at the end of section 5.4. Besides the parameters already chosen, namely the number of key terms to describe the Information Security domain $Q = 300$ and the percentage of irrelevant key terms in this set $p = 0.30$, we carefully considered $M = 7$ (i.e., the number of key terms used to describe the broadness of the research themes that will be identified). Since we set $p = 0.30$, the number of meaningful key terms per research theme, according to equation

[3] This example, uses the processed abstracts belonging to the years 2020-2022 from the corpus obtained in section 4.3.1 and also the list of domain-specific key terms derived at the end of section 5.4

(5.22), will be $M' \approx 5$. Using equation (5.21) we may now compute the number of topics needed as a parameter for the LDA method: $k = 30$.

With the appropriate number of topics being set as $k = 30$, we used the LDA topic modeling approach to discover research themes in the Information Security area, by employing the $LdaModel()$ function included in the $gensim$ Python library. Since we intended to identify the most latest research topics in the field, we employed the processed abstracts belonging to the last three years. To exemplify the content of the obtained topics, Figure 5.2 presents the first eight of them in a wordcloud format.

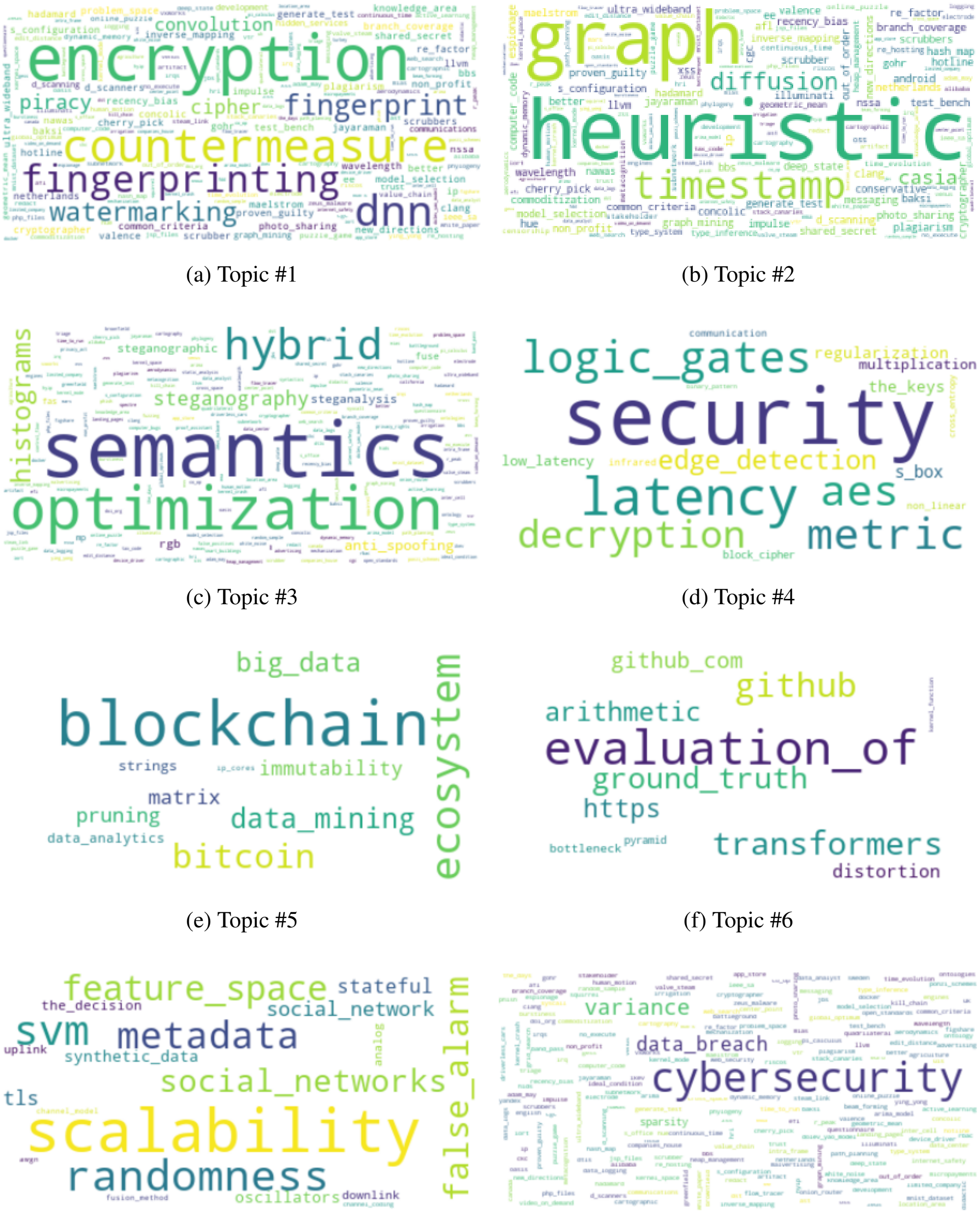
After that, we removed all the terms from the topics' content that did not belong to the set of domain-relevant key terms (this list contains $Q = 300$ terms and was obtained in the example presented at the end of section 5.4). The resulting topics, as they were ranked by the LDA procedure, together with their corresponding key terms, are listed in Table 5.2. It is worthwhile to notice that half of the topics do not meet the granularity-related constraint, namely to have a minimal number of terms greater than or equal to $M' = 5$, so they need to be filtered out. As examples of such topics, we may note Topic #11 which does not contain a single domain-relevant term, or Topic #26 which includes only an inconsistent term, namely 'upper_bound'.

5.6 Evaluating Research Theme Trends

This functional block belonging to the research theme recommender employs an effective and novel mechanism to investigate the research theme trends. It uses the multivariate Mann-Kendall test to process information regarding the interest among the scientific community that can be obtained from bibliographic metadata records. Our technique is particularly tailored for research themes represented as collections of key terms.

In the recent decade, NLP algorithms have improved rapidly, offering the needed means to quickly and also systematically investigate the bibliographic/bibliometric metadata records to reveal the research trends. In this particular context, the non-parametric trend assessment techniques are likely to be favored over parametric ones because of their lack of assumptions about data sample distribution [94] and homoscedasticity [95], and their proven reduced sensitivity to outliers [96]. Probably, the most utilized technique in this respect is the Mann–Kendall (MK) test which became an almost standard procedure for NLP applications [43, 97, 74, 98, 99, 100] because of its proven robustness in processing time series with missing values, censored data or non-Gaussian data [101].

The usual method for examining research trends is to track the time development of individual key term frequency in bibliographic records [15]. Traditionally, such approaches transform the title, abstract, and keywords metadata fields in either bag-of-entities or bag-of-words models and shape the key term occurrence counts during time into time series. Their basic purpose is to classify the key terms as "cold" or "hot". For example, Marrone [43] processed publication titles and abstracts in a bag-of-entities fash-



(a) Topic #1

(b) Topic #2

(c) Topic #3

(d) Topic #4

(e) Topic #5

(f) Topic #6

(g) Topic #7

(h) Topic #8

Figure 5.2: Wordclouds for representative topics [14]

Table 5.2: LDA topics and corresponding domain-relevant key terms [14]

Topic	Domain-Relevant Key Terms
1	encryption, countermeasure, dnn, fingerprinting, fingerprint, watermarking, convolution
2	heuristic, graph, android
3	semantics, optimization, hybrid, steganography, anti_spoofing, rgb, syntactics, spectre, static_analysis, fuzzing, computer_bugs, control_flow
4	security, latency, logic_gates, aes, metric, decryption, edge_detection, software
5	blockchain, ecosystem, bitcoin, data_mining, big_data, matrix, smart_contract, cryptocurrency
6	evaluation_of, github, transformers, ground_truth, arithmetic, https, github_com, distortion, backdoors
7	scalability, randomness, svm, metadata, feature_space, false_alarm, social_networks, tls, downlink
8	cybersecurity, variance
9	neural_networks, deep_learning, classifier, cnn, the_distance, biometric, modality, backdoors, filter
10	source_code, neural_network, redundancy, loss_function, open_source
11	—-
12	algorithm, gradient, convergence, android
13	malware, cloud_computing, forensics, ciphertext, ddos, attack_surface, phishing, helps, europe, android
14	smart_contracts, kernel, ethereum, logic, abstraction, wiretap, smart_contract, computer_bugs, cryptocurrency, compiler
15	cryptography, but_not
16	access_control, usability, password, biometrics, entropy, dns
17	deep_learning, the_power
18	data_privacy, internet, privacy, bandwidth, classifiers, vector, gdpr, android, europe
19	smartphones, smart_phones, data_security, smartphone, public_key, outsourced, pixel, android
20	forgery, jpeg
21	data_analysis, sampling
22	synchronization, data_set, physical_layer, randomization
23	task_analysis, nist, encoder
24	cryptographic, authentication, secret_key, fpga
25	iot, the_internet, google, perturbations, radio_frequency, mmwave, millimeter_wave, csi, smart_home, systematics
26	upper_bound
27	linear, eavesdropper, mimo
28	gaussian_noise, data_logging
29	leverage, computer_crime, to_show, risk_management, data_protection, secret_sharing, android, cyber_security, safety_critical, confidentiality
30	cache, defender, ip

ion and assessed the key term trends utilizing a combination of trend descriptors offered by Kleinberg burst detection, Mann–Kendall test, and Sen's slope estimation methods. Similarly, Marchini et al. [97] developed a method to evaluate urologic research trends associated with twelve previously selected key terms.

While in traditional approaches the research themes are described by unique key terms, our procedure is grounded in the belief that research subjects may be more exactly modeled by finite sets of meaningful M' key terms. For this, instead of using univariate time series trend detection techniques, we had to resort to analyzing the trends inside multivariate time series describing the time evolution of the occurrences of all M' key terms in the given corpus of processed abstracts. Following this line of thinking we have therefore employed a multivatiate Mann-Kendall test variant to evaluate and categorize the overall trends for the first M' key terms of each topic (i.e., research topics derived using the procedure presented in section 5.5) to be increasing, decreasing or no monotonic.

Example: Trend evaluation for Information Security research themes

We consider the research themes extracted using LDA from a corpus of processed abstracts from the Information Security domain and presented in Table 5.2. Following the rearrangement of the domain-relevant terms inside each topic according to their ndf relevancy score and the removal of any research theme described by less than M' key terms, we may now proceed to examine the trends. Accordingly, we developed and ran a Python script based on the $mk.multivariate\_test()$ function included in the $pymannkendall$ library to analyze the multivariate trends for best-ranked $M' = 5$ key terms inside each topic over the last ten years. The resulting topics (i.e, research themes), arranged in descending Z-statistic order, are listed in Table 5.3.

Considering our declared objective to identify hot and promising research subjects in the Information Security domain, we will analyze the interpretability and relevancy of the resulting themes. In order to determine the particular scientific focus corresponding to each research theme and to appropriately label them, we begin by investigating each topic's underlying key terms, presented in Table 5.3. Accordingly, we labeled the Research Themes (RTs) as follows:

- ○ RT1 (Topic#9): *Machine Learning Applications in Information Security domain*
- ○ RT2 (Topic#5): *Blockchain Technology and Its Applications*
- ○ RT3 (Topic#10): —-
- ○ RT4 (Topic#14): *Smart Contracts*
- ○ RT5 (Topic#18): *GDPR and Data Privacy*
- ○ RT6 (Topic#29): —-
- ○ RT7 (Topic#13): *Cloud Computing Security*

Table 5.3: Research themes ranked by their Z-statistic [14]

RT	**Topic**	**Domain-Relevant Key Terms (ranked by ndf)**	**Trend (Z-statistic)**
1	9	deep_learning, neural_networks, classifier, cnn, biometric	5.564
2	5	blockchain, ecosystem, data_mining, bitcoin, cryptocurrency	4.552
3	10	source_code, neural_network, open_source, redundancy, loss_function	2.664
4	14	smart_contracts, computer_bugs, kernel, ethereum, logic	2.553
5	18	data_privacy, internet, android, classifiers, gdpr	2.448
6	29	leverage, android, computer_crime, confidentiality, cyber_security	2.320
7	13	malware, cloud_computing, android, forensics, phishing	2.202
8	25	iot, the_internet, systematics, google, perturbation	2.159
9	3	optimization, semantics, computer_bugs, hybrid, fuzzing	2.122
10	6	evaluation_of, https, github, distortion, backdoors	2.049
11	1	encryption, fingerprinting, countermeasure, watermarking, dnn	1.911
12	7	scalability, randomness, svm, social_networks, tls	1.321
13	4	security, latency, metric, software, logic_gates	0.841
14	16	access_control, biometrics, usability, password, entropy	0.800
15	19	android, smartphones, smart_phones, data_security, pixel	0.240

- ○ RT8 (Topic#25): *Internet of Things Security*
- ○ RT9 (Topic#3): *Security Breaches Related To Computer Bugs*
- ○ RT10 (Topic#6): —
- ○ RT11 (Topic#1): *Watermarking and Fingerprinting*
- ○ RT12 (Topic#7): *Security and Scalability of the Transport Layer*
- ○ RT13 (Topic#4): *Software and Hardware Security-Related Implementations*
- ○ RT14 (Topic#16): *Access Control*
- ○ RT15 (Topic#19): *Smartphone Security*

We may observe that the research themes denoted by RT3, RT6, and RT10, by having no semantic meaning inside the Information Security domain (i.e., no domain-relevant terms characterize these themes), need to be dropped.

By further investigating the values for the Mann-Kendall Z-statistic presented in Table 5.3 we may classify the research themes into three categories: *(i) on a slowly-increasing trend*, described by a subunitary Z-statistic: RT13, RT14, RT15; *(ii) on an average-increasing trend*, described by a Z-statistic in the interval $[1, 3]$: RT4, RT5, ..., RT12; and, *(iii) on a rapidly-increasing trend*, described by a Z-statistic value higher than 4.5: RT1 and RT2.

Considering that the entire Information Security domain is on a significant upward trend, we may conclude that only the two best-ranked research subjects, namely Machine Learning Applications in the Information Security domain (RT1) and Blockchain Technology and Its Applications (RT2), may provide exceptionally promising prospects. This particular result is in line with current trends in the Information Security field since ML and blockchain are among the most popular and evolving research areas.

From our point of view, the analysis mentioned above might be strengthened by studying the time evolution of the scientific community's interest in the mentioned research themes. In this regard, in Figure 5.3 we show the evolution for the period 2001-2022 of the Z-statistic corresponding to seven RTs. We may observe that the best-ranked themes, namely RT1 and RT2, are characterized by general solid upward trends (i.e., Z-statistic has only positive values) in the entire time interval, while other research subjects like RT5, RT8, RT11, or RT14 have encountered both ascending and descending trends (i.e., Z-statistic changed its sign) during the same period.

5.6.1 Coping With Publication and Indexing Latency

When analyzing research trends using bibliographic information, an essential factor to consider is the time lag, which can be up to a year or more, between the completion of the

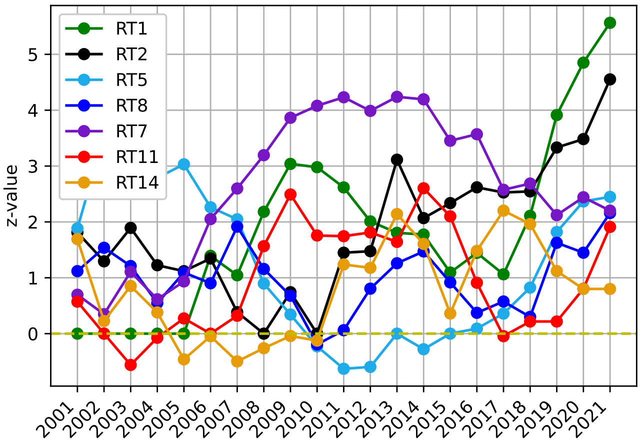


Figure 5.3: Trend evolution of relevant research themes during the last two decades [14]

research endeavor to the indexing of the related publication in the bibliographic database. Evidently, this delay has significant consequences, particularly in quickly evolving scientific fields [102, 103], where rapid theoretical and technological advances cause trends to shift suddenly. To fill the discovered gap, we suggest a new trend evaluation approach that combines the auto-ARIMA forecasting with an appropriate Mann-Kendall test variant in a single method, coined as n-steps-ahead Mann-Kendall (nsaMK).

Proposed methodology

Before delving into the details of the methodology for coping with publication and indexing latency, we define the following terms:

- *manuscript writing time* (δ_{WT})*:* the mean interval between research completion and the manuscript submission date. It involves certain tasks, such as: writing the first draft; editing, formatting and reviewing; and, identifying the most relevant and suitable publication for submitting the manuscript. Its duration strongly depends on the experience and expertise of the authors and also on the type of publication they are submitting their manuscript to and may range from a few days to even years.

- *publication latency* (δ_{PL})*:* the mean interval between the submission date of a manuscript and its initial publication date. This time lag is specific to the publication where the manuscript is submitted and represents the average time for reviewing, revising, and publishing, generally taking values between a month and two years.

✦ *research publication latency* (δ_{RPL})*:* the mean interval between research completion and the publication date of resultant publication. It may be computed using the following equation:

$$\delta_{RPL} = \delta_{WT} + \delta_{PL}, \tag{5.23}$$

✦ *indexing latency* (δ_{IL})*:* the average time interval between the publication date of a manuscript and the moment the publication is included in the bibliographic database.

✦ *bibliographic metadata latency* (δ)*:* the average time interval between the research completion and the moment the resulting publication is indexed in the database. This parameter may be computed using:

$$\delta = \delta_{RPL} + \delta_{IL} = \delta_{WT} + \delta_{PL} + \delta_{IL}. \tag{5.24}$$

As a direct consequence of the δ delay, the bibliographic metadata, instead of reflecting the actual state of research, describes a past state of the scientific research. To improve the precision when evaluating the current status of scientific knowledge and the underlying research trends, we need to mitigate the adverse effect of outdated bibliographic metadata by employing suitable forecasting methods.

Let us consider a research theme RT modeled by a set of key terms. To assess the overall RT trend in the case the bibliographic metadata latency cannot be disregarded, we propose the following three-stage methodology:

Stage I: select N (i.e., number of steps to be forecast) using the formula:

$$N = \left\lfloor \frac{\delta}{\Delta t} \right\rceil, \tag{5.25}$$

where $\lfloor . \rceil$ represents the nearest integer function (i.e., rounding function), while δ and the step function Δt use the same time measurement units (e.g., years).

Stage II: create the multivariate time series having a component (i.e., univariate time series) for each key term in RT. This is done by calculating the number of each key term's occurrences in bibliographic metadata during each time step Δt.

Stage III: employ the proposed multivariate nsaMK method described below, to evaluate the trend.

N-steps-ahead Mann–Kendall method [12]

Employing the multivariate time series that delineates the time evolution of each of the key term occurrences in the given corpus, and the number N of steps that need to be forecast due to bibliographic metadata latency as inputs, the nsaMK method evaluates the overall research trend as a Z-statistic based on a two-step procedure [13]:

1. each of the multivariate time series components x_i with $i = 1, 2, \dots, k$ are supplemented by N forecast values $x_{k+1}, x_{k+2}, \dots, x_{k+N}$ by employing the multivariate version of the auto-ARIMA technique, described in paragraph 5.2.2; if the predictions provided by auto-ARIMA have negative values, they will be automatically set to zero (the key term occurrences may take only non-negative values).

2. the multivariate Mann-Kendall test presented in paragraph 5.2.3, is applied to the time series obtained in the first step.

In the above procedure, that we termed as the n-steps-ahead Mann-Kendall (nsaMK) test, we utilized the auto-ARIMA prediction technique due to its effectiveness in forecasting various types of time series and its intuitive interpretability [81, 82, 85]. On the other side, auto-ARIMA comes with an inherent drawback that needs to be carefully mitigated: it enhances the existing serial correlations between the time series observations. To counteract this disadvantage, we may use two specially designed MK variants, proposed by Yue and Wang [104] or Hamed and Rao [105], to classify the trends using the Z-statistic score.

Example: Trend evaluation in EDA considering bibliographic metadata latency[4]
Our nsaMK method [12], proposed for research trend assessments when dealing with the bibliographic metadata latency, was evaluated against the standard MK test (variant reported by Yue and Wang [104]). We used the TCAD journal paper metadata for the interval 2010-2019, while the observations for the year 2020 were treated as ground truth. In the particular case of TCAD, by considering the influence of all the components in δ, we may approximate the value of N to one year according to equation (5.25).

It is important to mention that in this example, the most suitable ARIMA model was automatically chosen based on the Akaike information criterion, while the parameters for auto-ARIMA procedure were selected as follows: the order of autoregression $p \in \{1, 2, 3\}$; the order of the moving average $q \in \{0, 1, 2\}$; and, the degree of differencing $d \in \{0, 1, 2\}$. Moreover, because the order of autoregression p cannot be zero, employing the Yue and Wang MK test version is a viable option. All the experiments were conducted in Python 3.8 and used the *yue_wang_modification_test()* included in the *pyMannKendall 1.4.2* library [92], *CountVectorizer()* function from *scikit-learn 1.0.1* package, and, an ARIMA forecasting method derived from *tsa.statespace.SARIMAX()* from *statsmodels 0.13.0* library [106].

[4] This example, uses the processed abstracts belonging to TCAD journal for the years 2010-2020 from the corpus described in subsection 4.3.2

Table 5.4: Top 24 key terms in EDA domain [12]

Rank	Term	ndf	Rank	Term	ndf
1.	integrated circuit	0.211	13.	neural network	0.064
2.	optimization	0.163	14.	low power	0.059
3.	computer architecture	0.136	15.	hybrid	0.055
4.	algorithm	0.130	16.	system on chip	0.055
5.	logic gates	0.125	17.	mathematical model	0.053
6.	computational modeling	0.121	18.	power	0.044
7.	latency	0.094	19.	convolutional neural network	0.044
8.	fpga	0.090	20.	logic	0.044
9.	task analysis	0.084	21.	memory management	0.044
10.	energy efficiency	0.073	22.	real time systems	0.042
11.	machine learning	0.071	23.	cmos	0.042
12.	ram	0.067	24.	nonvolatile memory	0.041

The most important key terms for the EDA domain in 2020 have been discovered by employing the technique presented in section 5.4. The first 24 of them, together with their corresponding normalized document frequency score, are listed in Table 5.4. To quantitatively compare our nsaMK and classic MK methods, a set of evaluation metrics, composed of Z-statistic, Sen's slope, and p-value, was computed for all these 24 key terms in the following three cases:

MK2020: employs the MK test (Yue and Wang form) for the articles published in the interval 2011-2020. The obtained results are seen to be the ground truth.

nsaMK2020: employs the nsaMK test for the articles published in the interval 2010-2019 and the forecasts for 2020.

MK2019: employs the MK test (Yue and Wang form) for the articles published in the interval 2010-2019.

Thus, the results obtained using the traditional MK test (i.e., MK2019), are compared to the ones provided by our method (i.e., nsaMK2020), MK2020 being used as the ground truth.

Table 5.5 provides the comparative results for the mentioned set of 24 EDA key terms. In the last column, we labeled the key terms for which the proposed nsaMK method shows better performances (i.e., nsaMK2020 is closer to the ground truth than MK2019 when considering Sen's slope) with check marks. To summarize, in 75% of the cases nsaMK provides more reliable trend assessments and it displays a 48% less mean square error (i.e., $2.559 \cdot 10^{-6}$ versus $5.282 \cdot 10^{-6}$).

Table 5.5: Comparison results for nsaMK and MK methods [12]

	Term	nsaMK2020			MK2019			MK2020—Ground Truth			
		z	*p*-Value	Slope	z	*p*-Value	Slope	z	*p*-Value	Slope	
1	integrated circuit	−5.057	$4.2x10^{-7}$	-0.008	−2.715	0.00662	-0.005	−4.809	$1.5x10^{-6}$	-0.008	✓
2	optimization	8.494	0	0.005	5.964	$2.4x10^{-9}$	0.005	10.738	0	0.006	✓
3	computer architecture	4.856	$1.2x10^{-6}$	0.009	4.059	$4.9x10^{-5}$	0.009	6.828	$8.6x10^{-12}$	0.012	✓
4	algorithm	0	1	-0.000	−2.047	0.04060	-0.002	1.061	0.28868	0.001	✓
5	logic gates	0.597	0.54995	0.001	1.257	0.20871	0.002	0.300	0.76357	0.000	✓
6	computational modeling	0.590	0.55509	0.001	−0.522	0.60141	-0.000	1.714	0.08649	0.001	✓
7	latency	1.459	0.14438	0.001	2.027	0.04261	0.001	4.459	$8.2x10^{-6}$	0.004	✓
8	fpga	3.910	$9.2x10^{-5}$	0.006	1.842	0.06533	0.003	3.571	0.00035	0.008	✓
9	task analysis	1.862	0.06250	0	1.816	0.06932	0	2.031	0.04216	0.001	✓
10	energy efficiency	5.235	$1.6x10^{-7}$	0.007	4.005	$6.1x10^{-5}$	0.004	4.894	$9.8x10^{-7}$	0.007	✓
11	machine learning	5.856	$4.7x10^{-9}$	0.003	4.512	$6.3x10^{-6}$	0.003	5.025	$5.0x10^{-7}$	0.005	✓
12	ram	4.346	$1.3x10^{-5}$	0.002	4.232	$2.3x10^{-5}$	0.003	5.118	$3.0x10^{-7}$	0.004	
13	neural network	1.938	0.05250	0.002	0.907	0.36428	0	2.892	0.00382	0.004	✓
14	low power	0	1	0.000	1.712	0.08684	0.001	1.910	0.05607	0.001	
15	hybrid	3.298	0.00097	0.003	2.580	0.00987	0.002	4.213	$2.5x10^{-5}$	0.004	✓
16	system on chip	3.637	0.00027	0.003	6.111	$9.8x10^{-10}$	0.004	3.694	0.00022	0.003	✓
17	mathematical model	−5.016	$5.2x10^{-7}$	-0.005	−0.924	0.35519	-0.004	−3.686	0.00022	-0.004	
18	power	−4.734	$2.1x10^{-6}$	-0.002	−3.678	0.00023	-0.001	−2.532	0.01133	-0.001	
19	convolutional neural network	3.842	0.00012	0.001	3.275	0.00105	0.001	3.361	0.00077	0.002	✓
20	logic	0.497	0.61901	0.000	−0.685	0.49291	-0.001	2.419	0.01553	0.001	✓
21	memory management	5.262	$1.4x10^{-7}$	0.003	5.672	$1.4x10^{-8}$	0.003	9.513	0	0.004	✓
22	real time systems	0	1	0	2.307	0.02102	0.002	1.859	0.06289	0.003	
23	cmos	7.518	$5.5x10^{-14}$	0.003	7.646	$2.0x10^{-14}$	0.004	6.133	$8.6x10^{-10}$	0.002	✓
24	nonvolatile memory	2.058	0.03958	0.001	2.701	0.00690	0.002	5.340	$9.2x10^{-8}$	0.003	

Two illustrative examples, one for the key term ’algorithm’ and one for ’logic gates’, are displayed in Figures 5.4 and 5.5, respectively. In these figures: (a) the observations for the years 2010-2019, which are marked in light-blue, are used by MK2019 to derive the slope of the black-solid line; (b) the pink-marked value that is forecast by auto-ARIMA predictor and the last nine observations (i.e., for 2011-2019) are employed by nsaMK2020 to derive the slope of the red-dotted line; and, (c) the real trend (i.e., the blue-dashed line) or the ground true, is derived by MK2020 by considering the observations from 2011-2020, where the dark-blue value is the observation for 2020.

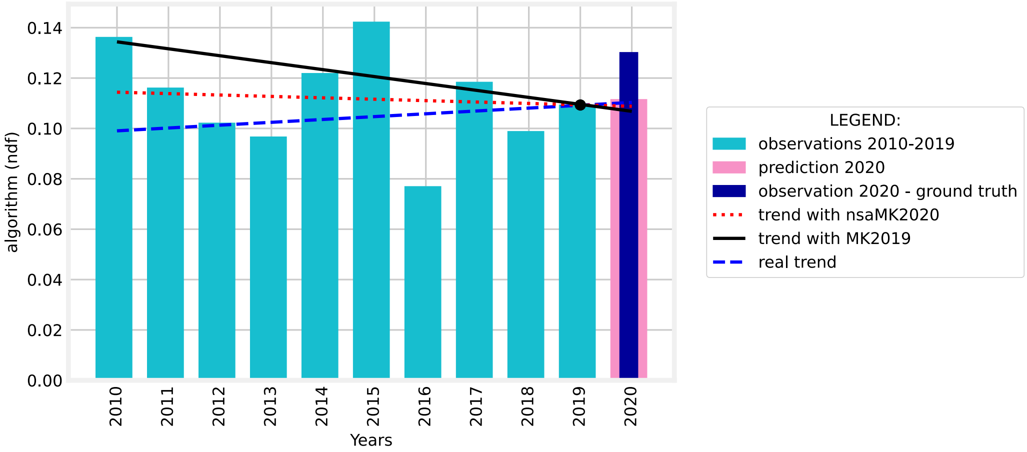


Figure 5.4: Trend comparison for the key term ’algorithm’[12]

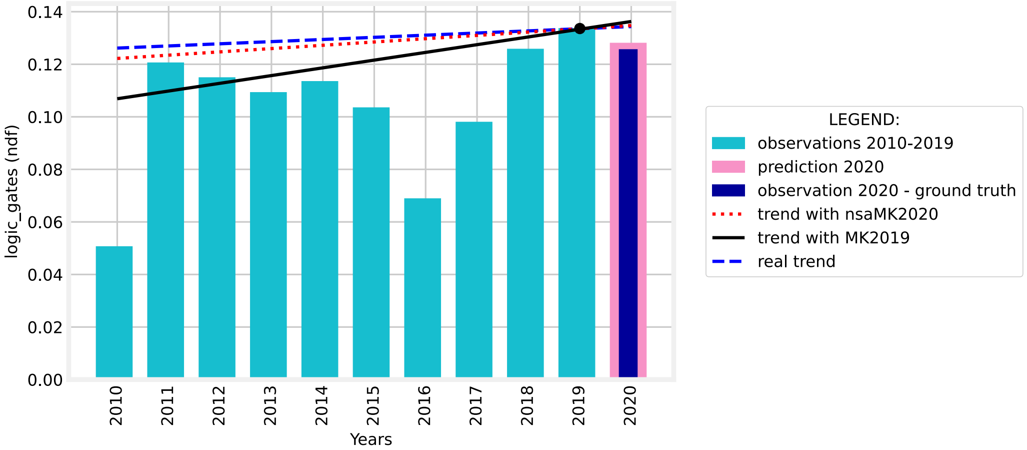


Figure 5.5: Trend comparison for the key term ’logic gates’ [12]

In the case of 'algorithm' key term (Figure 5.4), the best predicted value for 2020, characterized by an AIC score of -38.633, was provided when using $p = 1$, $q = 1$, and $d = 0$. Here, the trend is switching from descending, if the observations from 2010-2019 are considered, to ascending (i.e., ground truth), the proposed nsaMK test providing a more suitable trend evaluation than MK2019. In the second example, which is presented in Figure 5.5, the best auto-ARIMA forecast value for 2020 has an AIC score of -35.658 that was provided by $p = 1$, $q = 0$, and $d = 0$. It may be noticed that the slope provided by nsaMK is much closer to the real trend than MK2019.

It is noteworthy to point out that, like in every method that presumes a predictive component, our method's effectiveness drastically depends on the forecasting model. From this perspective, future work may employ switching from ARIMA models to exponential smoothing or neural networks.

5.7 Evaluating Research Theme Feasibility

This section presents our approach to the automated evaluation of the research theme feasibility using a carefully designed double-threshold procedure able to signal if the theme is currently difficult to study or offers little novelty. In the endeavor to derive such a method, we relied on the methodology that we presented in [15]. This recent work provided the first attempt to automatically identify feasible research gaps within a given domain by evaluating the correlations between pairs of key terms, followed by a double-threshold statistical approach to discard the research gaps that are either difficult to study with the existing knowledge or may provide insufficient originality. By switching the focus from pairs of key terms, characterized in terms of the number of co-occurrences, to groups of interconnected key terms described using a new overall co-occurrence-based metric, we are able to analyze the current state of knowledge and, based on this, to evaluate if a research theme is feasible or not.

In Natural Language Processing (NLP), the word co-occurrence is a widely used corpus-level statistic for modeling the connection between terms [107, 108]. It describes how frequently two terms appear together in a collection of documents [109]. The greater the co-occurrence value, the more significant the anticipated semantic link between terms. We consider $M(i, j)$ to be the number of documents in which the two key terms KT_i and KT_j simultaneously occur, normalized with the number of documents in the given corpus. Thus, we may interpret $M(i, j) \in [0; 1]$ as the mean frequency in which a pair of terms appears in a document corpus.

To develop a method for evaluating the feasibility of a research theme described by a set of s key terms using a NLP-based approach, we first analyzed the underlying information behind the term co-occurrence frequency $M(i, j)$ in a document corpus. We observed the following two aspects:

- A very low value for $M(i, j)$ not only indicates that the specified pair of key terms is hardly encountered in the documents but may also indicate that the present state of knowledge is not sufficiently advanced to link them or that the pair of key terms may even be incompatible. Accordingly, connecting two key terms KT_i and KT_j having $M(i, j)$ lower than a threshold α may likely be unfeasible, α playing in this case the "critical mass" role. We termed the α threshold as **success threshold**.

- A very high value of $M(i, j)$ generally indicates that the links between the terms are strong as the terms frequently appeared together in the same text. This circumstance might result from an intensively investigated term connection, anticipating that the research subject might likely be an implausible source of novelty. We coined β to be the **novelty threshold**, since $M(i, j) < \beta$ may provide an acceptable level of novelty.

It is worth noting that the potential novelty generated by a pair of key terms is diminishing when $M(i, j) \in [0, 1]$ gets larger, while the success ratio increases with $M(i, j)$. Based on this observation, our proposed double-thresholding approach, summarized by Table 5.6, considers that only the key terms having their co-occurrence frequency value $M(i, j) \in [\alpha, \beta]$ can be regarded to be feasible in terms of their novelty and success prospects. While $M(i, j) \in (\beta, 1]$ characterizes pairs of terms with reduced novelty expectations, $M(i, j) \in [0, \alpha]$ describes pairs that are unlikely to be tackled with existing scientific knowledge. Since the term co-occurrence values are corpus-dependent, picking the right success and novelty thresholds must emerge from an exploratory corpus examination.

Table 5.6: Double-thresholding approach for research theme feasibility assessment [15]

	$\boldsymbol{0 \leq M(i,j) < \alpha}$	$\boldsymbol{\alpha \leq M(i,j) \leq \beta}$	$\boldsymbol{\beta < M(i,j) \leq 1}$
Success ↗	low	high	high
Novelty ↘	high	high	low
Research Gap Type	Unfeasible	**Feasible**	None

Let us consider a scientific domain D characterized by a list of Q key terms and a research theme within D described by s key terms KT_k belonging to the domain-characteristic list. To evaluate the feasibility of the research themes we proposed the following three-stage methodology [15]:

- ***Stage I**: Determine the α and β thresholds*, using the following sequence of operations:

  (a) Calculate the term co-occurrence frequency values $M(i, j)$ for all domain-characteristic key term pairs (KT_i, KT_j) with $i, j = 1..Q$ and $i < j$;

  (b) Choose the α and β thresholds by appropriately trimming the left and right ends of the $M(i, j)$ distribution. The next paragraph provides the details about how this operation should be performed.

By investigating the $M(i,j)$ probability distribution inside a textual corpus of size $\mathcal{N}$, we developed a practical procedure to identify the success and novelty thresholds. We started with two key observations of the general $M(i,j)$ distribution: (i) based on the particular way the $M(i,j)$ values are computed, they may take only values that are multiples of $1/\mathcal{N}$; and, (ii) the term co-occurrence frequency values approximately follow an exponential distribution, being heavily skewed towards zero, with few values far from zero and no negative observations (Figure 5.6). From this perspective, we may consider α and β as the two thresholds that produce a two-sided trimmed exponential distribution. If we neglect the null $M(i,j)$ values, we may choose the α and β thresholds such that each of them filters out 15–30% of the $M(i,j)$ observations, a reasonable option considering $\alpha = Q1$ and $\beta = Q3$, where the first quartile (i.e., the lower quartile) is marked by $Q1$, while the third quartile (i.e., the upper quartile) by $Q3$.

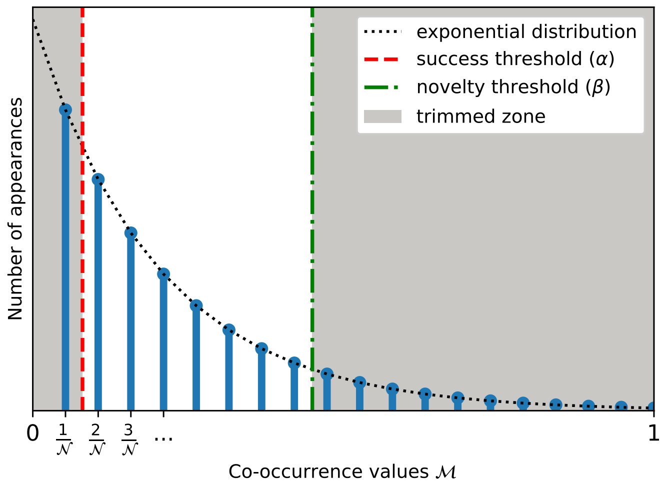


Figure 5.6: Distribution of $M(i,j)$ values in a corpus of size $\mathcal{N}$ [15]

- ***Stage II**: Drop all $M(i,j)$ that lay outside the considered feasibility interval* $[\alpha,\beta]$, retaining only $M'(i,j) \in [\alpha,\beta]$
- ***Stage III**: Evaluate the feasibility of the research theme* under investigation using a simple graph-based method: if we consider the research theme as a graph, where the vertices are the key terms and the edges correspond to $M'(i,j) \in [\alpha,\beta]$ values, the research theme is feasible only if the graph is connected (i.e., there exists a path between every pair of vertices).

  For feasible research themes having the same number of key terms s, we may use the mean co-occurrence value, to rank them. This indicator is computed using the following formula:

$$\mu = \frac{\sum_{i<j} M(i,j)}{0.5 \cdot s \cdot (s-1)}, \tag{5.26}$$

where the nominator is the summation of all co-occurrences between pairs of key terms, while the denominator is the number of all possible pairs. The larger the μ statistic value, the greater the expected success and the lesser the predicted novelty of the research subject modeled by the specified set of s key terms.

It is worth noting that the above double-threshold strategy may be adjusted to fit the researchers' risk profiles. Following the same line of thinking as in financial risk tolerance [110], we may describe research risk tolerance as the maximal value of uncertainty a researcher is ready to consider when framing research subjects. From this perspective, we may categorize the research themes as follows:

(a) conservative – characterized by thresholds α_c and β_c;

(b) moderate – characterized by thresholds α_m and β_m;

(c) aggressive – characterized by thresholds α_a and β_a;.

with the corresponding success thresholds holding $\alpha_a < \alpha_m < \alpha_c$, and the novelty thresholds holding $\beta_a \leq \beta_m \leq \beta_c$.

Example: Research theme feasibility assessment in EDA domain[5]

We exemplify the above methodology by trying to extract feasible research themes modeled by $s = 4$ key terms from an EDA subdomain described by a set of seven key terms denoted by: 'machine_earning' ($KT1$), 'energy_efficiency' ($KT2$), 'internet_of_things' ($KT3$), 'approximate_computing' ($KT4$), 'fault_tolerant' ($KT5$), 'biological_neural_networks' ($KT6$), and 'optimization_problem' ($KT7$) [15].

The following steps have been performed:

- The symmetric co-occurrence frequency matrix M, containing all $M(i,j)$ values was computed using the processed abstracts from 2019 and 2020. M is presented in the form of a heatmap in Figure 5.7, and the graph describing the subdomain described by $KT1 - KT9$ is presented as a chord diagram in Figure 5.8 (thinner edges correspond to lower $M(i,j)$ values).

- To derive the values of the two thresholds, namely the success and novelty thresholds, we employed the TCAD-related document corpus for the interval 2010-2020 and used the LDA algorithm [67] to obtain $r = 4$ topics from which we selected the most influential $p = 30$ key terms per topic. For these 120 key terms, all the $M(i,j)$ co-occurrences were computed, and the related histogram, together with the $\alpha = 0.000317$ and $\beta = 0.003$ selection, is displayed in Figure 5.9.

[5] This example, uses the processed abstracts belonging to TCAD journal for the years 2010-2020 from the corpus described in subsection 4.3.2

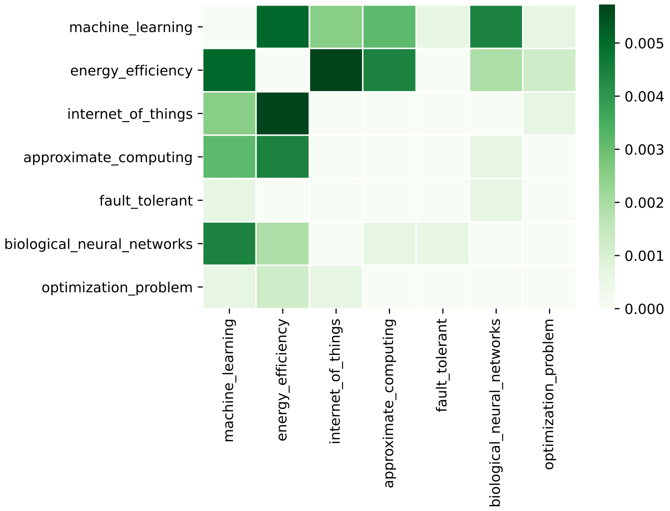


Figure 5.7: Heatmap visualization of the co-occurrence matrix M [15]

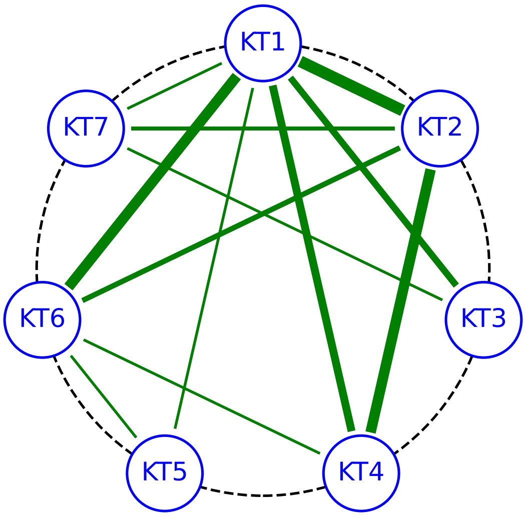


Figure 5.8: The original graph G of key terms [15]

- With $\alpha = 0.000317$ and $\beta = 0.003$ we filtered out all M components that do not lay inside the $[\alpha, \beta]$ interval, obtaining the double-thresholded version of M, denoted by M'. This new matrix is presented in Figure 5.10, while its related graph G' is displayed in Figure 5.11.

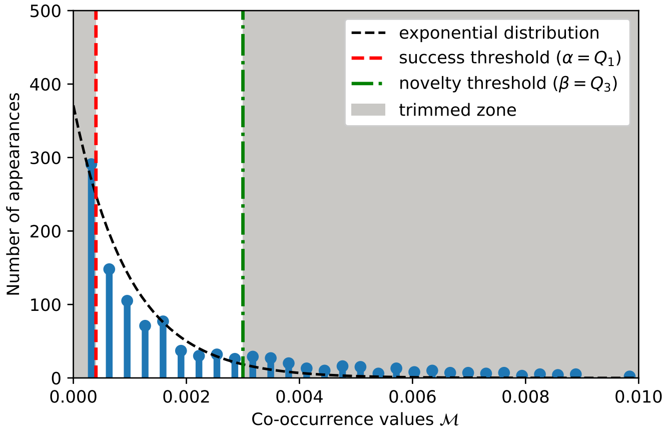


Figure 5.9: Histogram of the $\mathcal{M}$ co-occurrences and threshold values selection [15]

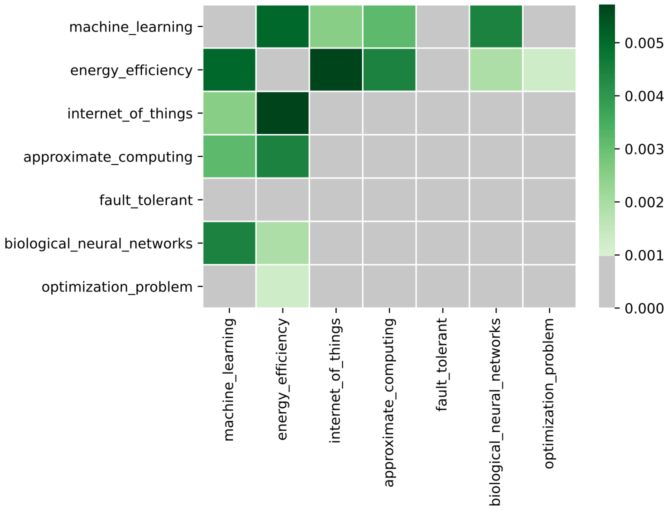


Figure 5.10: Visualization of feasible research gaps using heatmap [15]

- By extracting all the research themes made of four key terms (i.e., the connected graphs having four vertices) and ranking them according to their μ values, we obtained:

  (A) $KT1$, $KT2$, $KT3$, and $KT4$ ($\mu = 0.00174574$);

  (B) $KT1$, $KT2$, $KT3$, and $KT6$ ($\mu = 0.00164017$);

  (C) $KT1$, $KT2$, $KT4$, and $KT6$ ($\mu = 0.0013756$);

  (D) $KT1$, $KT2$, $KT3$, and $KT7$ ($\mu = 0.0012169$);

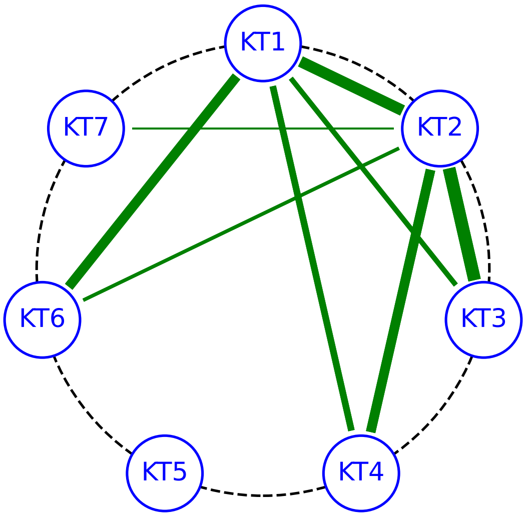


Figure 5.11: Visualization of feasible research gaps using chord diagram [15]

Table 5.7: Potential research subjects in EDA domain [15]

No.	Feasible Research Themes	Research Theme Interpretation	μ
1.	KT1: machine_learning KT3: internet_of_things KT4: approximate_computing KT6: biological_neura_networks	***Biological neural network*** *inspired algorithms for* ***approximate computing*** *in* ***ML*** *for* ***IoT*** *applications.*	0.00084652
2.	KT1: machine_learning KT2: energy_efficiency KT3: internet_of_things KT7: optimization_problem	*Design of integrated circuits for* ***IoT*** *applications* ***optimized*** *for* ***energy efficiency*** *by means of* ***ML.***	0.0012169

(E) $KT1$, $KT2$, $KT4$, and $KT7$ ($\mu = 0.0011639$);

(F) $KT2$, $KT3$, $KT4$, and $KT6$ ($\mu = 0.00100549$);

(G) $KT1$, $KT3$, $KT4$, and $KT6$ ($\mu = 0.00084652$);

(H) $KT2$, $KT3$, $KT6$, and $KT7$ ($\mu = 0.000740740$);

(I) $KT2$, $KT4$, $KT6$, and $KT7$ ($\mu = 0.0006348836$).

From these nine research recommendations, the two potential research topics that have practical interpretability are presented in Table 5.7, along with their textual interpretation.

If in the feasibility assessment process, we also consider the research risk tolerance by selecting the corresponding thresholds (e.g., aggressive framing: $\alpha_a = 0.0001$ and

Table 5.8: Examples of research subjects considering research risk [15]

No.	Feasible Research Themes	Research Theme Interpretation	Scenario
1.	KT1: machine_learning KT2: energy_efficiency KT3: internet_of_things	*Using **ML** for **energy efficient IoT**.*	conservative
2.	KT1: machine_learning KT3: internet_of_things KT4: approximate_computing KT6: biological_neural_networks	***Biological neural network** inspired algorithms for **approximate computing** in **ML** for **IoT** applications.*	moderate
3.	KT1: machine_learning KT2: energy_efficiency KT3: internet_of_things KT7: optimization_problem	*Design of integrated circuits for **IoT** applications **optimized** for **energy efficiency** by means of **ML**.*	moderate
4.	KT1: machine_learning KT4: approximate_computing KT5: fault_tolerant KT7: optimization_problem	***Approximate computing** for solving **optimization problems** in **fault tolerant ML**.*	aggressive
5.	KT1: machine_learning KT2: energy_efficiency KT5: fault_tolerant KT6: biological_neural networks	***Biological neural network** inspired methods for **fault tolerant** and **energy efficient ML**.*	aggressive

$\beta_a = 0.003$; moderate framing $\alpha_m = 0.0004$ and $\beta_m = 0.003$; and, conservative framing $\alpha_c = 0.001$ and $\beta_c = 0.003$) and the number of key terms to model possible research themes $s \in \{3, 4\}$, the recommendation samples listed in Table 5.8 are obtained.

Chapter 6

Cross-Domain Knowledge Transfer Recommender

Cross-domain knowledge transfer has a crucial role not only in solving concrete problems inside a given scientific domain but also in rejuvenating this target domain using fresh and already-proven information. This chapter provides a practical methodology where paper metadata are employed to discover the twin domains related to the target domain from where the transfer may be effective and also the pieces of knowledge from twin and emerging domains to be transferred and customized.

Cross-domain knowledge transfer is about learning from solutions that other scientific fields have discovered for themselves and transferring such solutions to their own topics. By transferring already proven knowledge we can minimize research-related risks and costs and also shorten research and development cycles. Moreover, the target domain may benefit from a perspective-changing that imported knowledge is able to induce. Besides the mentioned advantages, cross-domain knowledge transfers are always challenging. Due to a growing number of scientific papers being published, it is becoming increasingly difficult to keep track of what's happening in every research field and so many important logical linkages between the different sources of knowledge may go unrecognized.

The research on knowledge transfer between scientific domains is still in the early stage, the obtained results being sporadic and sparse. Analyzing the citation network corresponding to scientific papers from sustainability and aviation domains, Nakamura et al. [111] derived a recognized-unrecognized matrix that highlights the neglected problems and used this procedure to discover new knowledge for the development of an air and water transport system. By combining knowledge extracted from the fields of ammonia synthesis and fuel cells technologies, and employing keyword similarity and time-series analysis of the citation networks, Ogawa and Kajikawa [112] were able to suggest novel research ideas. A similar strategy, but this time attempting to uncover knowledge linkages between two completely unrelated subjects, namely gerontology and robotics, led to novel applications of assistive robots for elderly people [113]. These mentioned researches address isolated cases without intending to provide general knowledge transfer approaches.

Our endeavor to design a general methodology for cross-domain knowledge transfer purposes is built upon three pillars: (1) modeling the research domains or research themes as finite sets of key terms [15]; (2) employing publication metadata as a valuable source of information regarding scientific domains; and, (3) using appropriate NLP methods to extract and process meaningful information from textual data. Our approach considers that useful knowledge transfers are likely to be done from scientific domains that possess conceptually similar topics, methods and materials (i.e., twin domains), and also from emerging domains (e.g., Machine Learning, Internet of Things, or Blockchain Technology) which are capable to boost a large spectrum of other research domains.

6.1 Preliminaries and Problem Formulation

Since problems in one research field frequently have remedies in another, analogies have proved their importance in the development of science and technology throughout history, playing a key role in inspiring and fostering creativity across domains. However, as knowledge domains get more and more specialized, they interact less with one another, making analogical reasoning across research domains more difficult [114]. In this context, designing NLP-based solutions to facilitate cross-domain knowledge transfers becomes increasingly needed.

Problem formulation. Let us consider a target domain $\mathcal{D}$ defined by a group of Q key terms $\mathcal{KT}_i$ that we intend to enhance by transferring knowledge. We also consider a research theme RT within the target domain $\mathcal{D}$, specified by a group of J, $J \leq Q$ key terms KT_j. We aim to identify useful knowledge transfers from other scientific domains (i.e., source domains) $\mathcal{D}'_s$, with $s = 1, ..., S$ that may help enlarge the body of knowledge of the scientific theme under investigation RT.

In theory, every scientific domain can be a source of knowledge for a specified target domain. Since investigating all possible scientific domains is unrealistic, we have to focus on some categories of domains that may have a greater impact. In this respect, we consider that closely-related and emerging domains are most likely to make a difference.

Twin Domains. We define a twin domain as a scientific domain having similar or close research topics, methods or materials to a given scientific domain. Identifying such closely related domains offers various opportunities for knowledge or technology transfers, being an important component of our proposed methodology. In the attempt to find the twin domains from where to import relevant knowledge for a research theme RT, we intend to analyze the degree of similarity between corpora of paper metadata using a classic NLP procedure based on *tf-idf* scores and the cosine similarity metric. This procedure is presented in detail in the following section.

Emerging Domains. Emerging domains can be defined as "radically novel and relatively fast-growing" [115] fields that are frequently seen as having the power to directly and strongly influence the current state of science and technology. While often encompassing brand-new theories, methods, technologies and applications, such domains may also include older but reviving components. Such emerging domains can be simply identified by their explosive growth in the number of scientific publications or from their fast-growing interest within the scientific community. Three examples of such highly growing domains are Machine Learning (ML), Internet of Things (IoT), and Blockchain Technology.

A natural way to provide cross-domain recommendations is to take advantage of the existing overlapping entities, in our case key terms, that are shared between domains by using them as a bridge for knowledge transfers [116, 117]. Following this line of thinking, our approach is centered on the idea to identify the contexts (i.e., key terms that appear in the same cluster) in which the most key terms KT_j describing the research theme RT are found in twin and emerging domains and to recommend enhancing the investigated theme by including new key terms from these contexts.

6.2 Methods

To discover cross-domain knowledge transfer opportunities we will employ two classic NLP methods, namely Latent Dirichlet Allocation and document similarity evaluation using term frequency–inverse document frequency (*tf-idf*) vector space model and cosine similarity measure. While LDA was already presented in subsection 5.2.1, in the following paragraph we will provide a brief description of the second method.

6.2.1 Document Similarity Assessment

A widely used natural language processing procedure to perform document similarity evaluation method is based on the following two-step sequence: (i) text vectorization using term frequency – inverse term frequency (*tf-idf*) score; and, (ii) cosine similarity distance computation. These two steps are presented below.

Text vectorization using *tf-idf* is a technique that converts each text document from a given collection of documents, i.e., corpus, in a vector having the same length as the vocabulary (i.e., the set of unique terms in the corpus) and as elements the *tf-idf* scores for each term. The *tf-idf* score for a term t in the document d from the corpus C is calculated as follows:

$$tf\text{-}idf(t, d, C) = tf(t, d) \cdot idf(t, C). \tag{6.1}$$

Here, the term frequency $tf(t, d)$ is computed by counting the number of instances the term appears in the document, denoted by $freq(t, d)$, using the formula:

$$tf(t, d) = log(1 + freq(t, d)), \tag{6.2}$$

while the inverse term frequency $idf(t, C)$, which measures how common a term is in the entire corpus, is obtained by:

$$idf(t, C) = log\left(\frac{N}{count(d \in C, t \in d)}\right) \tag{6.3}$$

with N being the total number of documents in the corpus.

As a general rule, the higher the *tf-idf* score, the more significant that term is in a particular text document.

In the second step, the cosine similarity distance metric is computed for every vector pair (X, Y) using the following formula:

$$cos(X, Y) = \frac{\sum_{i=1}^{n} X_i \cdot Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \cdot \sqrt{\sum_{i=1}^{n} Y_i^2}}, \tag{6.4}$$

where X_i and Y_i, with $i = 1, 2, ..., n$ are the components of the two vectors having length n.

6.3 Proposed Approach

Our proposed cross-domain knowledge transfer recommender takes the processed abstracts of published papers corresponding to diverse scientific domains and the set of key terms that model the research theme as inputs and provides a set of knowledge transfers (i.e., key terms) from twin or emerging domains. Its architecture is presented in Figure 6.1 and contains four functional steps, described in the following paragraphs.

Step 1: Building a domain-characteristic text file for each scientific domain

Having the goal to represent as accurately as possible the body of knowledge of each scientific domain, we concatenate in a single text document the processed abstracts corresponding to all papers published within the domain during a given time interval. By this, all meaningful information extracted from publication titles, keywords and abstracts is summarized by a bag-of-entities (i.e., a list of key terms) that will be used by our method to identify the twin scientific domains of a given target domain.

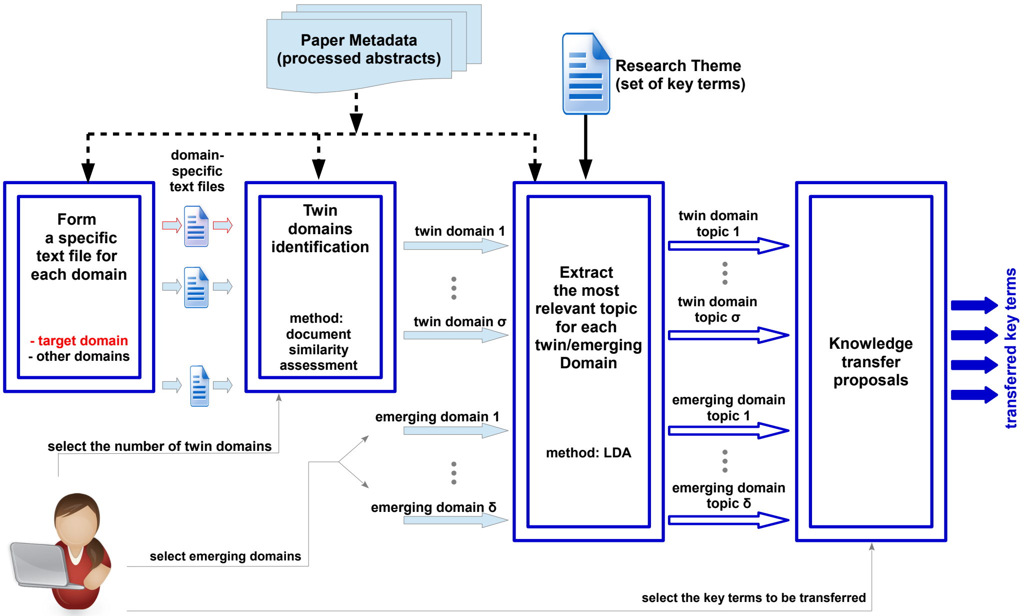

Paper Metadata (processed abstracts)
Research Theme (set of key terms)
domain-specific text files
Form a specific text file for each domain
- target domain
- other domains
Twin domains identification
method: document similarity assessment
twin domain 1
twin domain σ
emerging domain 1
emerging domain δ
Extract the most relevant topic for each twin/emerging Domain
method: LDA
twin domain topic 1
twin domain topic σ
emerging domain topic 1
emerging domain topic δ
Knowledge transfer proposals
transferred key terms
select the number of twin domains
select emerging domains
select the key terms to be transferred

Figure 6.1: Cross-domain knowledge transfer recommender system architecture

Step 2: Twin domains identification

To identify the source domains that due to their closeness in terms of topics, methods and materials to the target domain are plausible to export valuable knowledge, we examine the degree of similarity between pairs of target and source domain-characteristic text files. This step follows a classic document similarity assessment procedure that utilizes the term frequency–inverse document frequency (*tf-idf*) vector space model and cosine similarity measure [118]. This method includes two stages:

1. *tf-idf* vectorization of the document corpus [119] by computing the *tf-idf* weights for all encountered key terms (entities). The *tf-idf* score is meant to characterize the importance of terms for a document within a collection of documents.

2. Evaluate similarities between pairs of target and source domain-characteristic text files using the cosine similarity metric [120].

Finally, we will denote as twin domains the ones with the highest cosine similarity scores.

Step 3: Clustering the body of knowledge in each source domain using LDA topic modeling

To identify the context in which presumable knowledge related to the given scientific theme RT appears in twin or emerging domains from text files, we employ the Latent Dirichlet Allocation procedure, already described in subsection 5.2.1. Thus, for each of the identified twin or emerging domains, we apply LDA topic modeling to classify the twin/emerging domain's terms into 4-8 clusters and identify the topics that contain the largest number of key terms KT_j that describe RT. Afterward, we compute the term co-occurrences between defining terms KT_j of RT and other terms lying in the same topic and retain the terms from twin/emerging domains that have co-occurrence values above a chosen threshold to be analyzed for possible knowledge transfer.

It is worth mentioning that when analyzing the twin/emerging domains we are interested in identifying the areas where research is more advanced and can be a source of valuable knowledge transfers. Such advanced areas are characterized in the co-occurrence matrix $\mathcal{M}$ by a high score. In this way, we may transfer high-impact knowledge from related fields to RT.

Step 4: Knowledge transfer recommending

In this step, possible knowledge transfers are presented to the user in the form of sets of key terms that may accompany the existing set of key terms KT_j describing RT.

It is important to note that in order to alleviate the accuracy degradation we opted for a human-in-the-loop approach, where the human expert is directly involved in selecting the number of twin domains to be considered, in choosing the emerging domains and finally in selecting the appropriate knowledge transfers.

To exemplify how the proposed recommender works, in the following paragraph a detailed case study is presented.

Example: Knowledge transfer for a research theme in EDA domain[6]

Let us consider the following research theme in the Electronic Design Automation (EDA) domain:

*"Design of **integrated circuits** for **IoT** applications*
***optimized** for **energy efficiency** by means of **ML**".*

Analyzing the sentence, we may derive the following list of key terms KT that describe the research theme: 'integrated_circuit', 'IoT', 'optimization', 'energy_efficiency', and 'ML'.

[6] This example uses the processed abstracts belonging to all journals included in the corpus described in subsection 4.3.2, for the years 2010-2020.

Table 6.1: Cosine similarity between the flagship journals and TCAD

Journal	IEEE Society / Council	Cosine similarity to TCAD
TC	IEEE Computer	0.68692
TCSI	IEEE Circuits and Systems	0.47955
...	...	...
TR	IEEE Reliability	0.43722
...	...	...
JSSC	IEEE Solid-State Circuits	0.30843
TCE	IEEE Consumer Technology	0.29005
...	...	...
TE	IEEE Education	0.09875

Considering the specificity of the EDA domain and the research theme RT, we will focus our search for twin domains in the broad area of Electrical and Electronic Engineering. It is worth mentioning that the emerging domain that we choose to be appropriate for knowledge transfers (i.e., IoT) is encompassed by the same scientific area. Thus, we may use IEEE Xplore as a bibliographic database.

Step 1: Building a domain-characteristic text file for each scientific domain

For each of the 44 scientific domains extracted from Table 4.3 we prepared a text file, in the form of bag-of-entities by concatenating all processed abstracts from their representative journal.

Step 2: Twin domains identification

In the attempt to find suitable research domains from where to import relevant knowledge for our research theme, we analyzed the degree of similarity between the target domain represented by the IEEE Council on Electronic Design Automation and the other IEEE societies and councils (i.e., possible source domains). For this, we evaluated the cosine similarity between the text files already prepared in Step 1 corresponding to respective flagship journals. Some notable results are presented in Table 6.1, which suggest that the research domains covered by the IEEE Computer Society and IEEE Circuits and Systems Society are the most appropriate to export knowledge to our research theme under investigation.

In our case study, we also consider possible knowledge transfer from the emerging domain of Internet of Things, which encompasses an extremely wide range of innovative technologies able to link our physical world to the digital world, continuously proving its transformative power in reconfiguring the research strategies in many domains. In order to allow knowledge transfer from this emerging domain to RT as an EDA research topic, we considered paper-related metadata from IEEE Internet of Things Journal (JIOT).

Steps 3 and 4: Clustering the body of knowledge in each source domain and formulating the recommendations

In this stage of our methodology, we aim to detect relevant knowledge from closely related twin domains, namely the ones covered by the IEEE Computer and IEEE Circuits and Systems Societies, having TC and TCSI as flagship journals, and also from the IoT as an emerging domain described by bibliographic/bibliometric information acquired from the JIOT journal. The configuration of the relevant topics obtained using LDA on the three mentioned source domains (i.e., topics characterized by the highest number of encountered key terms KT_j) is presented in Table 6.2. In this table, the encountered key terms KT_j from RT are marked in bold.

Analyzing possible knowledge transfers from the two twin domains and the IoT emerging domain, by taking into account the paper metadata between 2010 and 2020, we identified the following strong links (i.e., high number of co-occurrences in the processed abstracts) which may signal knowledge transfer opportunities:

- ’fault_tolerant_systems’ strongly connected with ’optimization’ and ’energy_ efficiency’ in the TC journal papers;
- ’circuit_simulation’ strongly connected with ’integrated_circuit’ and ’optimization’ in the TCSI journal papers;
- the pair ’real-time_systems’ - ’smart _devices’ strongly connected with ’Internet_of_Things’, ’optimization’ and ’machine_learning’ in the JIOT journal papers;

These four key terms identified in the source domains which may possibly indicate knowledge transfers are presented in the last column of Table 6.2, while a graphic description of the considered research theme accompanied by the key terms imported from twin/emerging domains is presented in Figure 6.2.

Investigating Figure 6.2, the original research theme has to be augmented to consider both real-time and fault-tolerance analysis. Moreover, circuit simulation may be employed as a significant means to shorten the development cycle, while focusing the entire design process on possible applications in the field of smart devices may be beneficial.

Table 6.2: Source domains' topics and proposed knowledge transfer

Journal	Topic Terms	Proposed Transfer
TC	algorithm, **optimization**, heuristic, logic_gates, the_power, computer_architecture, scalability, vlsi, complexity_theory, topology, data_mining, voltage, capacitance, **energy_efficiency**, application_software, cpu, computer_science, redundancy, cryptography, approximation_algorithms, fault_tolerant_systems, leverage, error_correction	fault_tolerant_systems
TCSI	algorithm, **integrated_circuit**, linear, nonlinear, mathematical_model, circuit_simulation, convergence, computational_modeling, **optimization**, discret_time, neurons, matrix, semiconductor_device, mimo, time_domain, low_pass_filter, operational_amplifiers, neural_networks, vector, filter, domain_analysis, , adaptive_control	circuit_simulation
JIOT	**internet_of_things**, algorithm, the_internet, cloud_computing, task_analysis, latency, edge_computing, computational_modeling, **optimization**, computer_architecture, **machine_learning**, bandwidth, mobile_communication, blockchain, quality_of_service, reinforcement_learning, real_time_systems, hybrid, wireless_communication, authentication, smart_devices	real_time_systems smart_devices

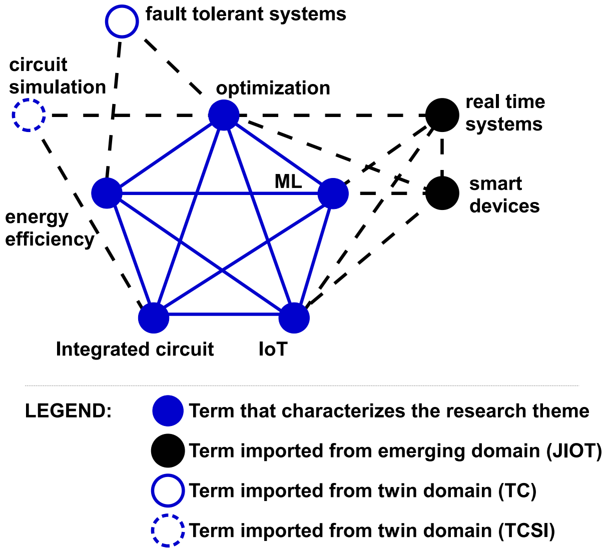

fault tolerant systems
circuit simulation
optimization
real time systems
ML
smart devices
energy efficiency
Integrated circuit
IoT
LEGEND:
● Term that characterizes the research theme
● Term imported from emerging domain (JIOT)
○ Term imported from twin domain (TC)
○ Term imported from twin domain (TCSI)

Figure 6.2: Research theme description after cross-domain knowledge transfers

Chapter 7

Research Team Recommender

The success of any research project substantially depends on the team that is assembled to accomplish it. In this chapter, we provide a practical methodology for data-driven team formation that uses bibliographic metadata to derive candidates' technical and teamwork skills and a carefully tailored multi-objective optimization model. The chapter is based on our paper [18].

7.1 Preliminaries

Any collaborative activity critically depends on selecting an appropriate team to be effectively fulfilled. In the case of research projects, due to their inherent innovative nature, the allocation of human resources becomes even more problematic. The almost standard approach builds the research team around an existing project leader in a top-down fashion, by considering not only the experience and expertise of team candidates but also by trying to build an appropriate interpersonal environment to unleash creativity. This manual allocation procedure [121] may provide beneficial results for small-sized projects where the pool of available experts taken into consideration is narrow and their interpersonal relationships are known. As we may notice, this type of team formation procedure has two evident drawbacks: it is subjective, being entirely dependent on project manager viewpoints and feelings; and, it has a limited scalability, being unsuitable for medium- to large-sized projects. This chapter aims to objectivize the research team formation process and to offer much-needed higher scalability by investigating the scientific production of the candidates with natural language processing means.

Our endeavor starts with a brief analysis of traditional team formation strategies and their possible enhancement using data-driven techniques, followed by a survey of related work in data-driven team formation. Afterward, the needed team member skills are investigated from both technical and teamwork perspectives to identify bibliographic record fields that may be used as metrics to quantify these traits. Our approach to automatic research team formation is built upon a generalized model of the team formation optimization problem, which is later customized and applied to derive the most appropriate team composition using bibliographic metadata. Finally, an illustrative case study is detailed.

7.1.1 Team Formation - From Traditional Strategies to Data-Driven Approaches

The success of any research project is undoubtedly determined by the quality and dedication of the team fulfilling it. Besides having a competitive pool of candidates to choose from, assembling a high-performing team is always a complex and challenging task that has no straightforward solution [122]. Bearing in mind that such a human resource allocation is done according to complex and sequential multicriterial decision-making actions, an appropriate strategy has to be selected. Such strategies are basically derived from the desired team organizational structure and their practical implementation is driven by available knowledge and expectations of involving actors.

Team formation strategies

Theoretically, three team selection strategies can be pursued, two presuming a hierarchical organization of the team (i.e., top-down and bottom-up approaches) and one assuming a flat organizational team structure (i.e., self-organized and self-managed team formation).

Top–down team formation strategy

A traditional and practical way to assemble a team is centered on the team leader's experience and competence in guiding this human resource allocation process. The top-down team selection approach is thus a sequential process initiated by the team leader, in which team members are individually chosen by their direct managers.

In order to select the appropriate candidates, any team (or sub-team) leader has to be aware of the entire picture of the team formation process. In this respect, deciding to pick a particular candidate is determined not only by his/her individual competencies and teamwork skills but also by the need to assemble a functional team covering all required technical and non-technical facets related to the team objectives.

The top-down approach is exemplified in Figure 7.1, on a three-level team hierarchy. In the first phase (specified by ①), the team leader selects the sub-team leaders, considering, on the one hand, the overall objectives of the team and, on the other hand, the sub-team leaders' organizational abilities and expertise. Afterward, in the second phase (specified by ②), the sub-team leaders will select the team members belonging to their own sub-teams.

The leader-controlled team formation is generally well-suited when the pool of candidates is relatively small and involved people already know the others. Initiated and controlled by team leaders and subordinate managers this type of strategy is profoundly

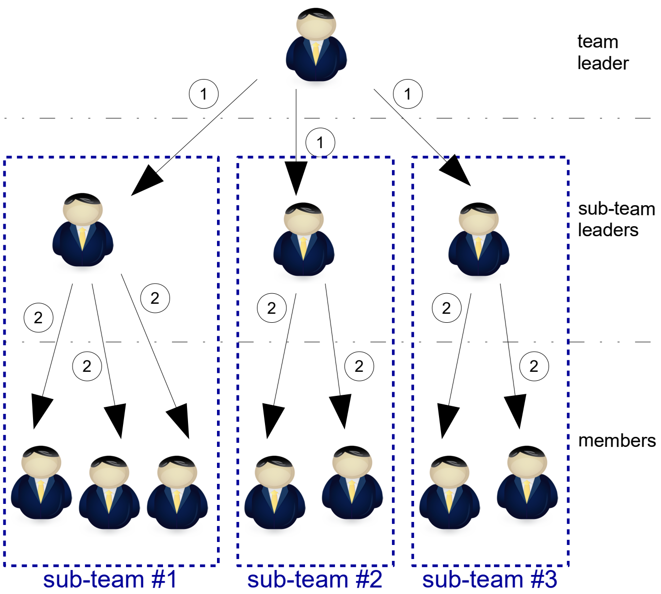


Figure 7.1: Top-down team formation for a three-level team hierarchy

biased since the team members are often selected based on feelings, opinions, and previous experiences. As a consequence, such teams are not always effective and successful [123].

Bottom–up team formation strategy

A bottom-up team formation is a team self-organizing approach in which the team members preferentially select their teammates and later choose the hierarchical leaders in a democratic-like manner. Being generated by the members themselves based on technical and personal affinities, such teams are usually characterized by stronger bonds resulting in improved team morale, a diminished amount of friction, and improved communication between members [124].

In Figure 7.2, a bottom-up team formation process for a three-level team hierarchy is presented. The process is initiated by team members who organize themselves into three separate groups (i.e., sub-teams #1, #2 and #3). Afterward, each of these sub-teams selects its own sub-team leader (phase ①), which is later involved in choosing the overall team leader (phase ②).

From the start of the team formation process, the team members have to be collectively aware of all the project objectives the team needs to achieve and also of the need to gather all the required competencies. From this perspective, assembling large teams

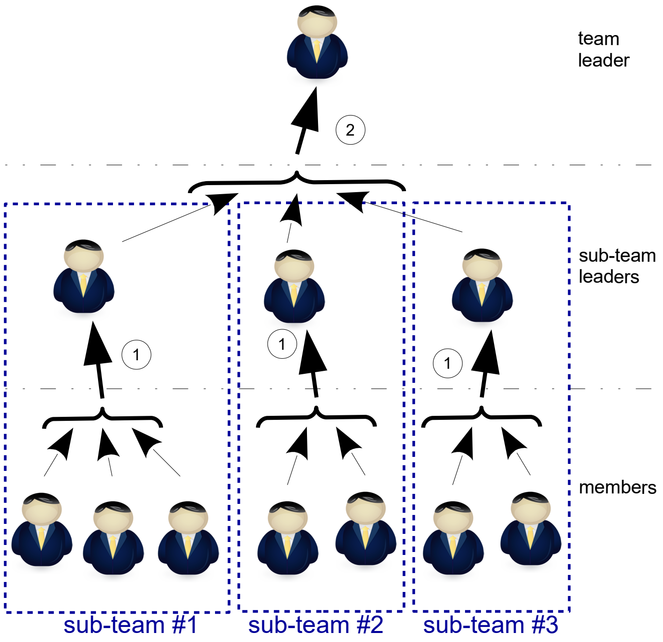


Figure 7.2: Bottom-up team formation for a three-level team hierarchy

using a self-organizing approach is strongly bounded by the number of already existing links between candidates. Moreover, due to the highly subjective nature of bottom-up team formation process, this type of teams may suffer from a potential lack of technical competencies in solving the project tasks [123].

Self-governing team formation strategy

Self-governing teams, also named egalitarian teams, are characterized by their complete and collectively exercising authority over their composition and resources, or in choosing and fulfilling their own goals. The team formation process is initiated by one or more team members who sequentially select their peers considering their abilities, skills, and interpersonal ties.

Figure 7.3 exemplifies such a self-organizing procedure where teammates are successively selected by their peers to meet the team's objectives.

Self-governing teams are usually encountered in domains characterized by creative thinking and problem-solving, and also by reduced formal interactions with the outside world, such as small collectives working in exploratory research areas, start-ups, special interest groups, think tanks or non-governmental organizations [125]. The freedom to jointly establish the rules to follow, to cooperatively allocate resources, or to organize the work obviously helps the team to reach its full potential. However, a team without a

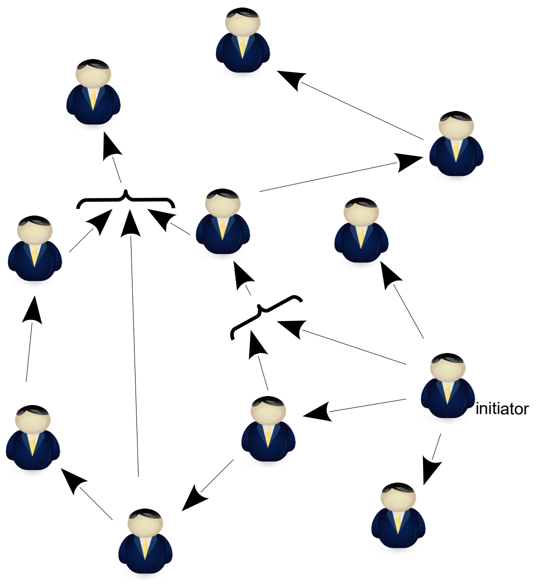


Figure 7.3: Self-managed team formation

leader (i.e., manager) may prove to be unrealistic in real-life circumstances [126], where in most cases a representative has to assume the decision-making responsibility in front of authorities or stakeholders (e.g., contracts or financial statements must be assumed by a representative and not by the entire team). Thus, even the team members perceive their team to be self-governing, in fact, in most cases, it is a bottom-up formed team. As a consequence, the self-managed team is more a theoretical concept than a practical team model.

The need for data-driven approaches for team formation

Traditional team formation strategies are in serious need of upgrading, mainly to improve their scalability and to increase their objectivity. The following reasons sustain this idea:

- screening of very large pools of candidates is tedious, time-consuming and almost unthinkable without automatic tools.
- every team formation problem is essentially a combinatorial multi-objective optimization that becomes progressively harder to solve as the number of candidates increases.
- evaluating candidates' competencies and teamwork skills and also their possible matching with other team members is biased by misconceptions, preconceptions or personal feelings.

- often, the preliminary filtering of candidates is done by outsiders (e.g., human resource departments), with superficial knowledge regarding the team's domain and its future objectives, further increasing subjectivity.

- when the pool of candidates contains both already known and previously unknown candidates, the latter are more prone to be excluded from newly formed teams.

- an appropriate level of diversity in team members' expertise, an important factor to be considered for highly-performing creative teams [127, 128], is hard to ensure given the subjectivity of the team formation process.

In this context, the artificial intelligence approaches, coupled with the abundance of available data regarding team candidates, may provide a solid basis for automatic or semi-automatic team formation procedures.

7.1.2 Data-Driven Team Formation

Considering the rapidly increasing data volume about the candidates' technical expertise and their non-technical skills (e.g., candidates' social media activity or profiles, publications, technical reports regarding projects, resumes, etc.), it becomes inappropriate to manually assess the suitability of a given candidate to assume a team member role. In this case, making use of the insights that the candidate-related data may provide represents a promising strategy.

Data-driven team formation concept

The Data-Driven Team Formation (DDTF) concept may be defined as the set of methods and methodologies meant to assist the decision-making process when choosing an optimal team using insights and information derived from data. Such a data-driven procedure to aid aggregating a team is triggered by a team initiator, who has the crucial role in defining the team's size, structure and objectives and who also makes the decisions during the team formation process. Depending on the future involvement in the teamwork, the initiator may be a team leader, a team member or even an outsider (e.g., managers or stakeholders who will not be a part of the team).

DDTF methodologies are centered on two main components: (a) a dataset comprising information on candidates' expertise and collaborative work capabilities; and, (b) a combinatorial optimization model to formalize both the objectives and constraints regarding the team selection process. By leveraging candidates-related data and managing it with the aid of algorithms and technology, we are able to generate relevant analytics to support humans in team formation decision-making. Such methodologies can naturally be described by the seven-phase sequence depicted in Figure 7.4 and briefly presented below.

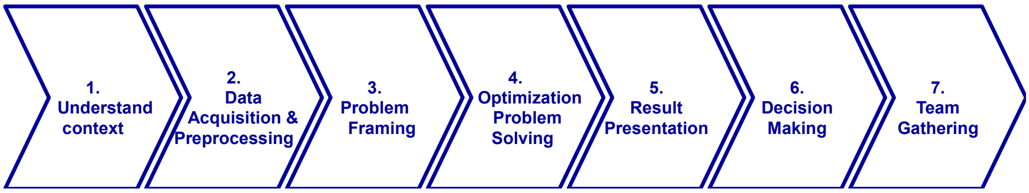


Figure 7.4: Phases of the data-driven team formation

✦ *Phase 1 – Understanding the context:* when triggering the team formation process, the team initiator must be keenly aware of the team-related issues including the team's mission, objectives, costs, deadlines, size, structure, etc.

✦ *Phase 2 – Data acquisition & preprocessing:* candidates-related information is extracted from diverse sources (e.g., documents, publications, social media streams) and processed to reveal candidates' characteristics such as technical or non-technical abilities or availability. Irrelevant candidates are filtered out to ease optimization problem-solving. It is noteworthy to mention that in order to increase the objectivity of the entire team formation process, unbiased data sources should be considered.

✦ *Phase 3 – Problem framing:* since team formation is deemed as a combinatorial multi-objective optimization problem, the team initiator needs to identify the criteria to be optimized and the constraints that need to be met, carefully considering the available data regarding candidates.

✦ *Phase 4 – Optimization problem solving:* a variety of discrete multicriterial optimization techniques and solvers may be employed to obtain qualified teams to complete the project.

✦ *Phase 5 – Results presentation:* the team optimization being multicriterial, a set of results (i.e., optimal team compositions), accompanied by their characteristics, are presented to the initiator.

✦ *Phase 6 – Decision making:* the team initiator selects the most qualified team to fulfill the objectives. If no team can be selected, the methodology will be restarted by considering adequate changes in Phases 2 and/or Phase 3.

✦ *Phase 7 – Team gathering:* the chosen team is mandated to fulfill the project.

Designing and implementing such a data-driven team formation decision-supporting system can yield significant advantages including:

- Increased objectivity by fostering data-based decisions over feeling-based ones;

- More transparent decisions, which are beneficial not only for selected team members by increasing their self-motivation and responsibility but also for rejected candidates to improve specific aspects;
- A higher level of proactivity in decisions since choosing a team from a list of recommended teams comes with known advantages and disadvantages that the team initiator is already aware of.
- Increased consistency of team formation decision making resulting from clear patterns and metrics extracted from unbiased data sources.
- More likely to be automatized by identifying team formation patterns that may lead to effective teams.
- Better control over certain aspects (e.g., functional diversity or shared history of the team members) of team performance that are otherwise left in the background.
- Much larger teams may be formed because it is not necessary for the initiator to have knowledge about candidates or their specific domains of expertise.
- Reducing the possibility of human error when selecting the team.

The DDTF methodology is feasible to be implemented in either an automatic or semi-automatic manner, the resulting system being classified as both a data-driven decision support system to help the initiator select a suitable team or a recommender system providing suitable team variants.

In the following sections of this chapter, we review the state of the art in the field and devise a general data-driven team formation procedure based on a novel formalization of the team formation optimization problem. This procedure is meant to help cope with all three team formation strategies (i.e., top-down, bottom-up, and self-governing team formation) and it will be particularized for employing bibliographic metadata in forming research teams.

In our endeavor, we rely on the reasonable assumption that a research team's success crucially depends on its members' expertise and teamwork skills, which can be inferred from recorded past performances.

Related works in data-driven team formation

The use of computational procedures to solve team formation problems became ubiquitous in various areas of our society, such as scientific research [122], manufacturing [129], sports [130, 131], or software development [126]. This trend stems from the necessity to consider a variety of criteria and constraints as well as to restore the team composition process tractability as the number of available candidates grows. Faced with the rising complexity of human resource allocation and management, organizations are increasingly relying on data-driven solutions, mostly to mitigate the risks related to uninformed and biased human judgments.

For the development of efficient data-driven team formation methodologies, two scientific disciplines must work closely together [132]: mathematics and computer science; as well as organizational psychology. While the former body of thinking provides techniques, methods, tools, and algorithms to formulate and resolve the team formation optimization problem, the latter comes with invaluable insights that should be taken into account while forming effective teams, especially regarding personal (e.g., conscientiousness, openness, extraversion, neuroticism, or agreeableness [133]) and interpersonal (e.g., leadership, conflict management, active listening, or clear communication) traits.

At first, the research regarding the team composition process in both computer science and organizational psychology disciplines has developed independently. Team composition research within organizational psychology was mainly centered on finding and experimentally analyzing the factors (i.e., personality traits of team members) correlated with successful and effective teams [132]. In contrast, researchers in the computer science field proposed team formation methodologies centered solely on applicants' technical ability while ignoring collaboration competencies. Askin and Huang [129] reported probably the first computational team composition technique back in 2001 where they developed an integer linear programming model to compose a cellular manufacturing team according to the workers' technical aptitudes. Their investigation was soon followed by several other related studies that are worth mentioning: de Korvin et al. [134] came with a fuzzy logic based approach to effectively match teammates to fulfil multiple stage projects by taking into consideration the required technical abilities and rather flexible cost allocation considerations; Hlaoittinun et al. [135] provided a method to compose multidisciplinary teams that clusters the potential members based on an incidence matrix, and draws the winning team by solving an integer programming optimization problem; Özceylan [131] developed a method to select a high-performing soccer team that first prioritizes the players using the Analytic Hierarchy Process and then uses a binary integer programming model to derive the best team; and, Shah et al. [136] offer a team formation method specifically suited for cybersecurity operations centers, that defines the team requirements and then picks individuals, using a collaborative score metric, to form collaborative teams that meet these requirements.

In the mid-2000s, there was a noticeable attempt to combine the expertise from the fields of organizational psychology and computer science to create high-performing teams, this tendency materializing in a consistent series of publications. Approaches that integrated the candidates' technical expertise with information about personal attributes enabling effective teamwork (i.e., soft skills) gleaned from interviews offered a first step in this direction. Two examples in this respect are presented by Fitzpatrick and Askin [137] who employed the Kolbe Conative Index for evaluating the candidates' temperament to complement their technical expertise, and, Chen and Lin [138] who utilized the Myers-Briggs personality test for deriving candidates' individual traits aimed to accompany the required technical competencies when forming multifunctional teams. While such methods enhanced team cohesion by taking into account specific member characteristics, they disregarded prior collaborations between the candidates. Lappas et al. [45] made a significant advancement when they utilized an expert social network to represent interpersonal collaboration. They proposed a greedy-based strategy that creates teams with the necessary expertise while minimizing communication costs (i.e., maximizing social compatibility) among members. A similar approach was reported by [139] which formulated the team member selection as a generalized densest k-subgraph problem, where the collaborative compatibility of a given team is represented by the edge density of the subgraph induced by vertices representing its members.

Over the past ten years, the problem of data-driven team composition has become increasingly complicated due to the addition of different kinds of constraints, goals, and criteria to be met. As a result, in addition to the already mentioned technical or non-technical abilities of candidates, a variety of concurrent factors were taken into account, including workload [140], members' geographical proximity [141, 142], and operating costs [143, 144]. Nevertheless, because of its strong correlation with specific scenarios, the formulation of the optimization problem describing the research team formation is still fragmented and incomplete. Besides some sporadic works (e.g. Selvarajah et al. [145] employed a set of four criteria, namely the candidate expertise, communication cost, collective trust, and, geographical proximity) a generalized optimization model of research team formation has not been proposed. According to D'Aniello et al. [123], a practical team formation strategy must take into account a set of characteristics that includes the specific competencies and skills of candidates; organizational restrictions such as time, resource, and budget constraints; and, also the candidates' desire to cooperate and help each other. In this context, we aim to devise a generic and consolidated formulation of the research team formation starting from a classic optimization model, namely the set cover problem.

Another problem that we identified when surveying the scientific literature on data-driven team formation methods is the notable lack of historical and unbiased public datasets about candidates [143, 145]. In these circumstances, team composition experiments often relied on artificially generated or simulated data. Following the line taken by Lappas et al. [45], which employed data extracted from the DBLP database when constructing research teams, we intend to use bibliographic sources to derive not only candidates' domains of expertise but also insights about their teamwork abilities.

Proposed methodology

A central role in our approach is played by a general data-driven team formation decision-making module, outlined in Figure 7.5. This module may be simply particularized to aid the team composition process, whatever team formation strategy we use.

A team formation procedure is triggered by a person we name *initiator*. Depending on the type of team he/she intends to construct, the initiator may be a team leader, team

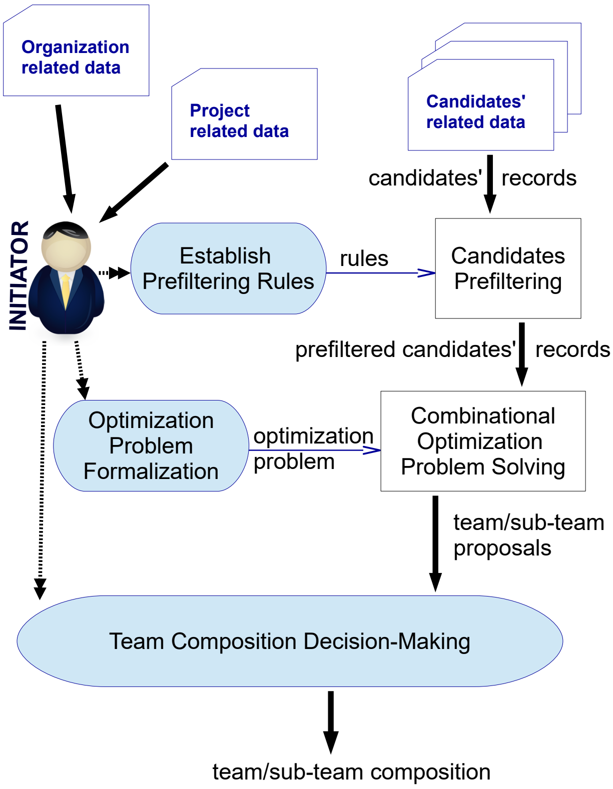

Organization related data
Project related data
Candidates' related data
candidates' records
INITIATOR
Establish Prefiltering Rules
rules
Candidates Prefiltering
prefiltered candidates' records
Optimization Problem Formalization
optimization problem
Combinational Optimization Problem Solving
team/sub-team proposals
Team Composition Decision-Making
team/sub-team composition

Figure 7.5: Data-driven team formation module

member, or outsider. Bearing in mind the research project (i.e., specifications, requirements, and objectives of the research) that the team has to fulfill and the organizational context (e.g., costs and availability of the candidates), the initiator is required to perform three actions:

(a) *formalize the optimization problem* that underlies team formation; this action demands a precise formulation of the objectives to be optimized and related constraints and, eventually, given that the problem is generally NP-hard to solve [146], a selection of the problem-solving approach.

(b) *establish the candidates' prefiltering rules;* these rules are closely related to the optimization problem previously formulated and are meant to provide a reasonably large pool of candidates by excluding the ones the initiator finds irrelevant.

(c) *decide the team to fulfill the research project;* since team formation is generally a non-trivial multicriterial optimization problem, there is no single solution to optimize all its objectives simultaneously. Thus, the team initiator needs to analyze the strengths and weaknesses of each team recommended by the semi-automated procedure and pick the best-suited one.

In the first phase, information regarding candidates' technical and non-technical competencies is acquired from reliable and, if possible, unbiased sources. Afterward, identified candidates are shortlisted using the set of prefiltering rules, tailored explicitly by the initiator in accordance with the project type and the formulated team formation optimization problem. Considering the shortlisted pool of candidates, the recommended teams are extracted by solving the multi-objective team formation optimization problem. In the final stage, the initiator selects the best team by making a trade-off between different criteria.

In the case of research teams, a suitable source of knowledge regarding possible team members is represented by bibliographic databases, which provide valuable information about competencies and past collaborations among researchers.

7.1.3 Researcher Skills and Their Reflection in Bibliographic Metadata Fields

A team can be defined as a group of two or more people accomplishing interdependent tasks toward attaining a common objective [147, 148]. When performing as a cohesive unit, a team is able to provide higher performances than the sum of all its members' performances when working alone [149]. Such a functioning social system is built on three pillars [150]:

- ***context*** – the circumstances and rationales related to team formation and its future dynamics. These conditionalities are closely related to the project type and objectives, to the estimated team cost and completion time, or to the needed institutional and technical support [151].
- ***identity*** – the need that the team members to recognize themselves and also act as a team; By this, members perceive their values to be aligned with the team characteristics and activities, allowing them to establish or maintain a healthy team relationship [152].
- ***teamwork*** – the members' capacity and desire to collaboratively work in order to achieve a shared goal. When performing a collective task, the coordinated effort needs to be backed up by a set of behaviors, skills and attitudes corresponding to each team member, including shared vision, mutual dependency, effective communication, conflict resolution, and trust [153].

By analyzing these three underlying components, we may conclude that high-functioning teams cannot be the result of coincidence but of a carefully designed team selection process relying on individual and team-related skills evaluation. Given the novelty-oriented nature of research teams, where members' complementary knowledge and skills are specifically geared toward boosting creativity, innovation, and problem-solving, team formation becomes even more challenging.

Our effort to develop data-driven research team formation methods started with a thorough analysis of how evaluations of candidate skills can be derived from publication-related data. Researchers have already confirmed the utility of bibliographic/bibliometric data when building research teams. They generally employed co-authorship graphs extracted from the DBLP database to identify candidates' domains of expertise or their previous work in teams [45, 154, 155]. In our perspective, besides the number and authorship of publications used to build such collaborative graphs, the bibliographic records may offer additional information worth exploiting when evaluating candidates, namely citation and accession counts or authors' affiliation.

Evaluating candidates' skills from bibliographic records

Assessing candidates to fill a member role inside a research team is generally done by considering their proven or potential individual and interpersonal skills. In this perspective, besides the candidate's technical knowledge, expertise, and experience, a set of teamwork-related abilities including communication, collaboration, listening, idea sharing or conflict resolution, have to be considered.

A. Technical skills

In the context of this work, we define technical skills related to a candidate as the specialized knowledge, expertise, and experience required to undertake research activities in a particular domain. These individual skills are widely recognized as critical elements of any collaborative research project [156, 157], but converting them to collective capabilities represents a crucial desideratum of any highly effective research team.

In the attempt to evaluate individual expertise and experience in a given scientific area, bibliographic records represent a practical and valuable resource. In our view, the following descriptive and quantitative information embedded in publication-related metadata may be employed to derive the researcher's areas of interest and corresponding expertise levels:

- publication's title, abstract, and keywords fields;
- publication's authors and their affiliation;
- publication date;
- number of citations received by the publication;
- downloading or viewing counts associated with the publication.

To identify if researchers have competencies in a given scientific domain, we may apply NLP procedures to identify occurrences of domain-characteristic key terms within titles, abstracts and keywords of their publications [158]. Additionally, the researchers' levels of expertise in a specified area can be derived from the number of publications, number of citations received and number of downloads/views of the publications written in that area.

Information extracted from publication-related metadata may be combined with the time distribution of the candidate's publications or with appropriate author/publication-level metrics (e.g., h-index, journal impact factor) to provide a much deeper interpretation.

B. Teamwork skills

Research team performance is crucially affected by the mixture of member personalities and attitudes which needs to functionally complement the ensemble of members' technical competencies. In this context, besides individual knowledge and expertise, a set of interpersonal competencies, that we refer to as teamwork skills, have to be thoroughly considered when trying to form a high-performing research team. Such personal traits, including collaboration and communication, conflict management, or maintaining a positive attitude, need to characterize relationships with other potential team members.

While the technical skills of a candidate can be adequately evaluated from recorded information (e.g., publications, reports), interpersonal skills are much harder to assess by outsiders only based on reported data [159]. This is both because interpersonal relations are primarily a matter of feeling and emotion, and because such historical information is generally not recorded during or after a research project is completed.

However, bibliographic metadata still offer useful information to rank the teamwork abilities of candidates. In this respect, we need to analyze the lists of authors corresponding to the candidate's publications. To build a high-performing team, we intend to favor candidates who have already worked effectively together (i.e., groups of candidates who were co-authors of some publications) and authors who have proven their teamwork skills in various collectives (e.g., candidates that have a higher total number of co-authors or a higher average number of co-authors per publication). The scientific literature on the subject reduces the complex cooperation inside the groups of co-authors to dyadic collaborations between pairs of co-authors by focusing the analysis on extracted co-authorship networks [45, 154, 155]. This approach obviously neglects the holistic nature of intra-group relationships. To solve this problem, we intend to facilitate the insertion of information regarding group relationships in the team formation optimization process.

7.2 Formalizing the Team Formation Optimization Problem

We aim to derive a general multi-objective combinatorial optimization model able to cope with research team formation specificities that besides the required task coverage also include aspects regarding team coherence, members' expertise, and redundancy-related issues. In this endeavor, we begin with a very simple optimization model and we successively generalize it to incorporate a variety of team formation facets.

7.2.1 A Set Cover Model of the Team Formation Problem

The team composition process can be mathematically modeled as a combinatorial optimization problem, that in our perspective may be suitably derived based on the standard set cover model [160] described as follows. Considering P to be a project encompassing n different tasks (i.e., $P = \{task_1, task_2, ..., task_n\}$) and a pool of m candidates, each characterized by a subset of competencies $S_i \subseteq P$ with $i = 1, ..., m$ in fulfilling the tasks included in the P universe, we aim to find a team having a minimal number of members that is able to cover all tasks from P. This optimization problem is coined as the set cover optimization and is described by the following model:

$$\text{minimize} \sum_{S_i \in S} x_{S_i} \tag{7.1}$$

$$\text{subject to} \sum_{S_i \in S} x_{S_i} \geq 1 \text{ for all } e \in P \tag{7.2}$$

with $x_{S_i} \in \{0, 1\}$ being binary flags, $x_{S_i} = 1$ denoting the inclusion of the candidate i in the optimal team. Considering X=<$x_{S_1}, ..., x_{S_m}$> an m-tuple containing all the x_{S_i} binary flags, we may reformulate the classic set cover problem, specified by (7.1) and (7.2), as the search for the minimum number of ones in X.

The set cover problem is one of the 21 standard NP-complete problems from the renowned Karp's list published in 1972 [160] and has been used in a large number of applications in the fields of computer science, combinatorics, operations research, or complexity theory.

In our attempt to generalize the set cover model, we first reshaped (7.1) by allocating weights w_{S_i} to each potential team member i and also included a new parameter (i.e., τ in (7.2)) to control the redundancy required in the team member expertise to fulfill the tasks $e \in P$. As a consequence, we obtained the weighted set multicover problem [161], described by the following model:

$$\text{minimize} \sum_{S_i \in S} w_{S_i} \cdot x_{S_i} \tag{7.3}$$

$$\text{subject to} \sum_{S_i \in S} x_{S_i} \geq \tau \text{ for all } e \in P \tag{7.4}$$

To further generalize the optimization model that characterizes the team formation process, we derived two generalized objective functions that besides the number or cost of team members are meant to effectively address issues regarding team coherence, team member expertise or team's organizational context:

(a) **Generalized Global Objective (GGO)** – is able to cope with per-team/global descriptors, like the overall team costs, or the team size; and,

(b) **Generalized Mean Objective (GMO)** – is able to cope with average value team descriptors, like the average number of past inter-member collaborations or the mean value of team expertise.

To obtain the GGO form of the objective functions, we expanded (7.3) by replacing candidates, each being specified by the corresponding x_{S_i} binary flag, with groups of candidates that can be indicated by products of individual binary flags. For this, let us consider a set $X^{(k)}$ of all sub-tuples drawn from the overall m-tuple X having k ($k \leq m$) ordered elements. In this case, the objective function may be written in the form:

$$(GGO): \qquad optimize \sum_{\substack{j=1\ldots\binom{m}{k} \\ S_i \in S}} W_j^{(k)} \cdot \sqcap_j^{(k)} \tag{7.5}$$

where $\sqcap_j^{(k)}$ are products of all elements x_{S_i} in the k-tuples $X^{(k)}$ of binary flags, while $W_j^{(k)}$ represent the corresponding weights.

As an illustrative example to help clarify the $\sqcap_j^{(k)}$ notation, let us consider a pool of five candidates ($m = 5$) and a group size $k = 3$. In this case, we will have $\binom{5}{3} = 10$ such products, denoted as follows: $\sqcap_1^{(3)} = x_{S_1} \cdot x_{S_2} \cdot x_{S_3}$; $\sqcap_2^{(3)} = x_{S_1} \cdot x_{S_2} \cdot x_{S_4}$; $\sqcap_3^{(3)} = x_{S_1} \cdot x_{S_2} \cdot x_{S_5}$; $\sqcap_4^{(3)} = x_{S_1} \cdot x_{S_3} \cdot x_{S_4}$; $\sqcap_5^{(3)} = x_{S_1} \cdot x_{S_3} \cdot x_{S_5}$; $\sqcap_6^{(3)} = x_{S_1} \cdot x_{S_4} \cdot x_{S_5}$; $\sqcap_7^{(3)} = x_{S_2} \cdot x_{S_3} \cdot x_{S_4}$; $\sqcap_8^{(3)} = x_{S_2} \cdot x_{S_3} \cdot x_{S_5}$; $\sqcap_9^{(3)} = x_{S_2} \cdot x_{S_4} \cdot x_{S_5}$; and, $\sqcap_{10}^{(3)} = x_{S_3} \cdot x_{S_4} \cdot x_{S_5}$.

It is worth mentioning that the GGO generic form (7.5) covers both maximization (e.g., objectives regarding the team expertise or its coherence) and minimization (e.g., objectives regarding the team size or team costs) objective functions.

In order to obtain the GMO type of objective function, we may simply divide the optimized quantity in (7.5) by the total number of k-sized groups existing inside the winning team:

$$(GMO): \qquad optimize \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{k}} \cdot \sum_{\substack{j=1\ldots\binom{m}{k} \\ S_i \in S}} W_j^{(k)} \cdot \sqcap_j^{(k)} \tag{7.6}$$

with $\binom{\sum_{S_i \in S} x_{S_i}}{k}$ being the total number of possible combinations of candidates in the winning team choose k.

To further generalize the considered set cover model, in (7.4) we may include fractions of coverage $\alpha_{S_i,e}$ (i.e., availability coefficients) in fulfilling the tasks $e \in P$ corresponding to each of the possible team members, and also a task-related value τ_e of the redundancy:

$$(CSTR): \qquad \sum_{S_i \in S} \alpha_{S_i,e} \cdot x_{S_i} \geq \tau_e \text{ for all } e \in P \tag{7.7}$$

7.2.2 A Generalized Optimization Model of the Team Formation Problem

Having the two generic optimization functions specified by (7.5) and (7.6) and the generalized version of the constraint (7.7), we may describe the team formation problem by the following complex multi-objective model:

$$\text{(GGOs): optimize} \qquad \sum_{\substack{j=1...\binom{m}{k_q} \\ S_i \in S}} W_j^{(k_q)} \cdot \sqcap_j^{(k_q)} \text{ with } q = 0, 1, ..., Q \text{ and } k_q \in \mathbb{N}^* \tag{7.8}$$

$$\text{(GMOs): optimize} \qquad \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{k_r}} \cdot \sum_{\substack{j=1...\binom{m}{k_r} \\ S_i \in S}} W_j^{(k_r)} \cdot \sqcap_j^{(k_r)} \text{ with } r = 0, 1, ..., R \text{ and } k_r \in \mathbb{N}^* \tag{7.9}$$

$$\text{(CSTR): subject to} \qquad \sum_{S_i \in S} \alpha_{S_i,e} \cdot x_{S_i} \geq \tau_e \text{ for all } e \in P \tag{7.10}$$

with Q representing the number of generalized global objectives (GGOs) and R specifying the number of generalized mean objectives (GMOs). Since our optimization model needs at least one objective function, Q and R need to satisfy the condition $Q + R > 0$.

The notations used in the (7.8)- (7.10) model are as follows:

	Notations:
P	project comprising a set of tasks
e	task from P ($e \in P$)
m	number of potential team members (candidates)
S_i	set of tasks that may be fulfilled by candidate i,
x_{S_i}	binary flag related to candidate i ($x_{S_i} = 1$ if the candidate i is a member of the team)
S	set of all candidates' abilities $S = \{S_i \vert i = 1, ..., m\}$ - is a set of sets
k	size of the considered group
$\sqcap_j^{(k)}$	j^{th} product of k different candidate flags x_{S_i}, with $j = 1...\binom{m}{k}$
$W_j^{(k)}$	weights corresponding to products $\sqcap_j^{(k)}$
Q	number of generalized global objectives (GGOs)
R	number of generalized mean objectives (GMOs)
τ_e	redundancy for covering the task $e \in P$

Regarding the (7.8)- (7.10) optimization model, the following observations are worth noting:

- if only a single objective function of GGO type is considered (i.e., $R = 0$ and $Q = 1$) and $k = 1$, the generic model (7.8)-(7.10) is reduced to the standard weighted set multicover model described by (7.3) and (7.4);

- the GGO and GMO objective functions may effectively be configured to cover either candidate-related characteristics if $k = 1$, or interpersonal aspects if $k \geq 2$;

- analyzing the sums in (7.8) or (7.9), we may notice that they only retain the relationships among the k team members belonging to the winning team (only in this case the products are non-zero);

- there may be more than a single objective associated with a given k_q (e.g., to form a research team, we may consider previous dyadic collaborations by maximizing not only the number of co-authored publications but also the number of co-authored patents or the research grants in which both the authors participated). This remark is also true in the case of k_r;

- if the optimization model (7.8)-(7.10) includes both minimization and maximization types of objectives, it may simply be reshaped into a multi-objective minimization model by switching the weights' signs in the maximization objective functions using either $\mathcal{W}_j^{(k_q)} = -W_j^{(k_q)}$ or $\mathcal{W}_j^{(k_r)} = -W_j^{(k_r)}$;

- if a given candidate i needs to be included in the winning team, we have to incorporate a supplementary constraint into the model (7.8)-(7.10), namely $x_{S_i} = 1$;

- if we intend compelling the mutual exclusion of two candidates i and j (i.e., the two candidates cannot be members of the same team), the following constraint needs to be incorporated in the (7.8)-(7.10) model: $x_{S_i} + x_{S_j} \leq 1$.

To illustrate how the GGO and GMO objective functions can be configured to catch certain team-related aspects, we provide the following set of examples:

a. *team size minimization:*

$$minimize \sum_{S_i \in S} x_{S_i} \tag{7.11}$$

b. *team cost minimization:*

$$minimize \sum_{S_i \in S} C_{S_i} \cdot x_{S_i} \tag{7.12}$$

where C_{S_i} represents the cost associated with the candidate i.

c. *optimization of the distance from the members' locations to the team workplace:*

$$minimize \sum_{S_i \in S} Dist_{S_i} \cdot x_{S_i} \tag{7.13}$$

with $Dist_{S_i}$ being the distance covered by the team member i when going to work.

d. *optimization of the mean expertise of team members:*

$$maximize \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{1}} \cdot \sum_{S_i \in S} Exp_{S_i} \cdot x_{S_i} \tag{7.14}$$

with Exp_{S_i} being an expertise-related parameter corresponding to the overall expertise of the candidate i.

e. *optimization of the mean past dyadic collaboration number:*

$$maximize \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{2}} \cdot \sum_{\substack{j=1 \ldots \binom{m}{2} \\ S_i \in S}} Dyad_j^{(2)} \cdot \sqcap_j^{(2)} \tag{7.15}$$

with $Dyad_j^{(2)}$ being the total number of existing collaborations involving the j pair of candidates.

f. optimization of the mean past triadic collaboration number:

$$maximize \ \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{3}} \cdot \sum_{\substack{j=1...\binom{m}{2} \\ S_i \in S}} Triad_j^{(3)} \cdot \sqcap_j^{(3)} \tag{7.16}$$

with $Triad_j^{(3)}$ being the total number of existing collaborations involving the j group made of three candidates.

Integrating the proposed generic team formation model into specific scenarios

The generic model, specified by (7.8)-(7.10), can be simply particularized to effectively and entirely formalize the egalitarian team formation procedure from scratch or the completion of an existent egalitarian team. Furthermore, the aforementioned optimization model may be used as a helpful instrument in both the top-down and bottom-up team formation paradigms to fill non-managerial positions.

A. Team formation

A team initiator, possessing all the needed knowledge (e.g., information regarding team objectives, size, or structure) about the team to be formed and being the one in charge of making all the strategic decisions throughout the team formation, will trigger this process. To deal with the anticipated position of the team initiator within the future research team (i.e., a member of the team or an outsider), we can reshape the optimization model (7.8)-(7.10) forcing the initiator's binary flag x_{S_ψ} on 1 if he/she intends to be a team member and 0 otherwise. As a result, the adjusted optimization model will have the following form:

$$\text{(GGOs): optimize} \quad \sum_{\substack{j=1...\binom{m}{k_q} \\ S_i \in S}} W_j^{(k_q)} \cdot \sqcap_j^{(k_q)} \text{ with } q = 0, 1, ..., Q \text{ and } k_q \in \mathbb{N}^* \tag{7.17}$$

$$\text{(GMOs): optimize} \quad \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{k_r}} \cdot \sum_{\substack{j=1...\binom{m}{k_r} \\ S_i \in S}} W_j^{(k_r)} \cdot \sqcap_j^{(k_r)} \text{ with } r = 0, 1, ..., R \text{ and } k_r \in \mathbb{N}^* \tag{7.18}$$

$$\text{(CSTR1): subject to} \quad \sum_{S_i \in S} \alpha_{S_i, e} \cdot x_{S_i} \geq \tau_e \text{ for all } e \in P \tag{7.19}$$

$$\text{(CSTR2):} \quad x_{S_\psi} = \begin{cases} 1 & \text{the initiator } \psi \text{ will be a team member} \\ 0 & \text{otherwise} \end{cases} \tag{7.20}$$

Here, (7.20) specifies the initiator's involvement in the resulting team, while Q and R need to satisfy the condition $Q + R > 0$ since at least an objective function is needed.

Even though the optimization model (7.17)-(7.20) was developed to formalize the egalitarian team formation process, it is also suitable to be utilized in the bottom-up or top-down approaches for filling non-managerial positions. For the top-down team formation procedures, the managers (i.e., team or sub-team leaders) may be selected by their superiors, while for the bottom-up approaches, a voting strategy may be employed to cover the managerial positions.

B. Completion of existing teams

A successful team may be perceived as a complex living organism that must constantly adapt to the environment where it operates. To strengthen its potential to generate added value, the team must be periodically reformed for an assortment of reasons like the necessity to eliminate lethargic, incompatible, inefficient, or unwilling members; the need to add new members able to cope with newly discovered challenges or to replace the ones who have left the team; etc. From this perspective, a team completion procedure needs to be immediately carried out every time a position becomes vacant. In such a scenario, the model (7.8) - (7.10) needs to be enhanced by including constraints that force the values of the binary flags x_{S_ϵ} associated with all already-occupied team member positions to be equal to one:

$$\text{(GGOs): optimize} \quad \sum_{\substack{j=1...\binom{m}{k_q} \\ S_i \in S}} W_j^{(k_q)} \cdot \sqcap_j^{(k_q)} \text{ with } q = 0, 1, ..., Q \text{ and } k_q \in \mathbb{N}^* \tag{7.21}$$

$$\text{(GMOs): optimize} \quad \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{k_r}} \cdot \sum_{\substack{j=1...\binom{m}{k_r} \\ S_i \in S}} W_j^{(k_r)} \cdot \sqcap_j^{(k_r)} \text{ with } r = 0, 1, ..., R \text{ and } k_r \in \mathbb{N}^* \tag{7.22}$$

$$\text{(CSTR1): subject to} \quad \sum_{S_i \in S} \alpha_{S_i,e} \cdot x_{S_i} \geq \tau_e \text{ for all } e \in P \tag{7.23}$$

$$\text{(CSTR2):} \quad x_{S_\epsilon} = 1 \text{ for all } \xi \in \Xi \tag{7.24}$$

with Ξ being the set of existing team members. Supplementary, to have at least one objective function, the $Q + R > 0$ condition must be satisfied.

Using the team completion problem described by the model (7.21)-(7.24) we are now able to effectively identify the optimal candidates to be added to the existing team.

Considerations on solving the generic team formation optimization problem

In the general case, the multi-objective optimization problems, since the objective functions normally conflict with each other, are not straightforward to solve. From this perspective, the problems (7.8)-(7.10), (7.17)-(7.20), or (7.21)-(7.24) make no exceptions. To overcome the challenges in tackling such complex problems, we may rely on the particularities of the research team formation. Firstly, we may observe that this process is usually not a very time-critical one (i.e., a research team is often built in days or even weeks). Secondly, a carefully designed candidate shortlisting process may be employed to reduce the complexity of the optimization problem. In this respect, we may for example retain only the candidates who have authored at least three scientific publications [45], drop the candidates with no authored publications in the last five years, or exclude the candidates having less than fifty citations. After that, we might employ the following strategies to address the issue:

- in case the shortlisted pool of candidates has a size of a couple of hundreds, a brute-force approach, that considers all possible combinations and afterward decides which one is the best, may be appropriate.

- in case the objective functions may be ranked according to their functional relevance, a lexicographic approach [162] may be employed.

- in case the size of the pool of candidates is very large and methods to decrease the number of objectives (e.g., ϵ - constraint method or linear scalarization [163]) are likely to be unsuitable, an evolutionary optimization method, like NSGA-II, may be chosen.

NSGA-II method

The general goal when solving a multi-objective optimization problem is to find the non-dominating set of solutions that cannot be improved simultaneously in all objectives (i.e., Pareto front). From the pool of available approaches for locating Pareto fronts, evolutionary methods have proved extremely successful by operating with a large population of solutions from which the Pareto dominance relationship may be constructed.

A popular evolutionary technique for solving combinatorial multi-objective optimization problems is NSGA-II [164]. It directly targets the non-dominated solutions by using: (i) an elitist concept that gives the opportunity for the population's elites to be replicated in the following generation; and, (ii) an explicit diversity-preserving mechanism based on crowding distance.

The NSGA-II algorithm is presented in Figure 7.6 and is briefly described as follows. The parent population $\mathcal{P}_t$ of size N and the standard genetic operators (i.e., binary tournament selection, crossover, and mutation) are initially used to construct the N-sized offspring population $\mathcal{Q}_t$ at any generation t. The two populations are then blended to create a new population $\mathcal{R}_t$ of size $2N$, which is classified into distinct non-domination fronts F_i. Only N slots will be retained for constructing the new population $\mathcal{P}_{t+1}$ by first dropping the populations not linked with a front and, if needed, the populations associated with the last front having the lowest crowding distances.

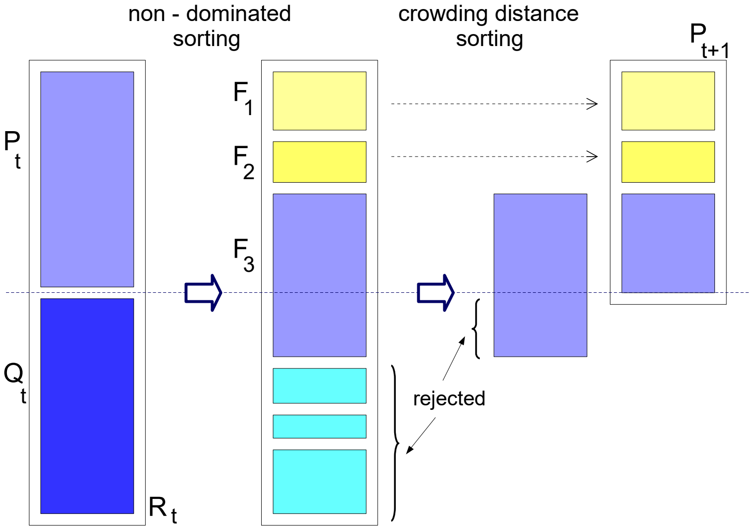


Figure 7.6: NSGA-II procedure

In our implementations, we used NSGA-II Python function from the *pymoo* library [165].

Example: Data-driven egalitarian team formation for a welding research project Let us consider a research team formation process having three objective functions, namely the minimization of the team size, maximization of the mean technical expertise of team members, and, maximization of the mean inter-member familiarity. We also consider the dataset depicted in [166] comprising information about the competence level, expressed as Personal Knowledge Scores (PKSs), of a pool of 45 researchers and their dyadic collaborative interactions, described as Familiarity Scores (FSs), in four scientific areas, namely arc welding, strip casting, water cooling, and magnesium alloy. The dataset contains insights extracted from 576 publications, including 158 patents, 197 project reports, and 221 articles, reported during the 2001–2006 time interval.

To derive the optimal teams to fulfill a project in the field described by the four scientific areas (i.e., P={'strip casting', 'magnesium alloy', 'arc welding', 'water cooling'}, we start with the particularization of the model (7.8)-(7.10). First, we selected the minimal redundancy for each of the four scientific areas to be $\tau = 2.2$. Since the candidate's expertise and their dyadic collaboration counts are naturally suited to maximization objectives, we had to switch from maximization objective functions to minimization ones, the resulting model having the following form:

$$\text{(f1):} \quad \text{minimize} \quad \sum_{S_i \in S} x_{S_i} \tag{7.25}$$

$$\text{(f2):} \quad \text{minimize} \quad \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{1}} \cdot \sum_{S_i \in S} InvExpertise_{S_i} \cdot x_{S_i} \tag{7.26}$$

$$\text{(f3):} \quad \text{minimize} \quad \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{2}} \cdot \sum_{\substack{j=1 \ldots \binom{m}{2} \\ S_i \in S}} InvCollaboration_j^{(2)} \cdot \sqcap_j^{(2)} \tag{7.27}$$

$$\text{subject to} \quad \tau_e - \sum_{S_i \in S} \alpha_{S_i,e} \cdot x_{S_i} \leq 0 \text{ for all } e \in P \tag{7.28}$$

The weights corresponding to (7.26)-(7.28) have been derived based on PKS and FS descriptors using the following relations:

$$InvExpertise_{S_i} = 1 - \frac{1}{4} \cdot \sum_{e} PKS_{S_i} \tag{7.29}$$

$$InvCollaboration_j^{(2)} = 1 - FS(j) \tag{7.30}$$

$$\alpha_{S_i,e} = PKS_{S_i,e} \tag{7.31}$$

Using the *pymoo* multi-objective optimization library [165], a Python program was created to solve the multi-objective combinatorial problem (7.25)-(7.28). We chose the *NSGA2()* function included in the *pymoo* package to implement the Non-dominated Sorting Genetic Algorithm II (NSGA-II) solver [164], defining the optimization model to be of *ElementwiseProblem* type (i.e., a single solution is evaluated at a time). This procedure is governed by the following set of parameters:

- number of objective functions: 3
- number of constraints: 4
- optimization type: *ElementwiseProblem* - it evaluates the candidate solutions one by one
- sampling mechanism: binary random sampling
- crossover mechanism: two-point crossover
- mutation mechanism: bit flip mutation
- size of population: 100
- number of generations: 50

- Python function: *NSGA2()* from *pymoo* library [165].

The obtained set of 28 non-dominant solutions is displayed in a parallel-coordinate representation (Figure 7.7) and also as a scattered plot (Figure 7.8). In these figures, we highlighted the four non-dominant solutions characterized by $f1 = 9$ (i.e., teams having a minimal size: nine members), their composition and objective function values being specified in Table 7.1.

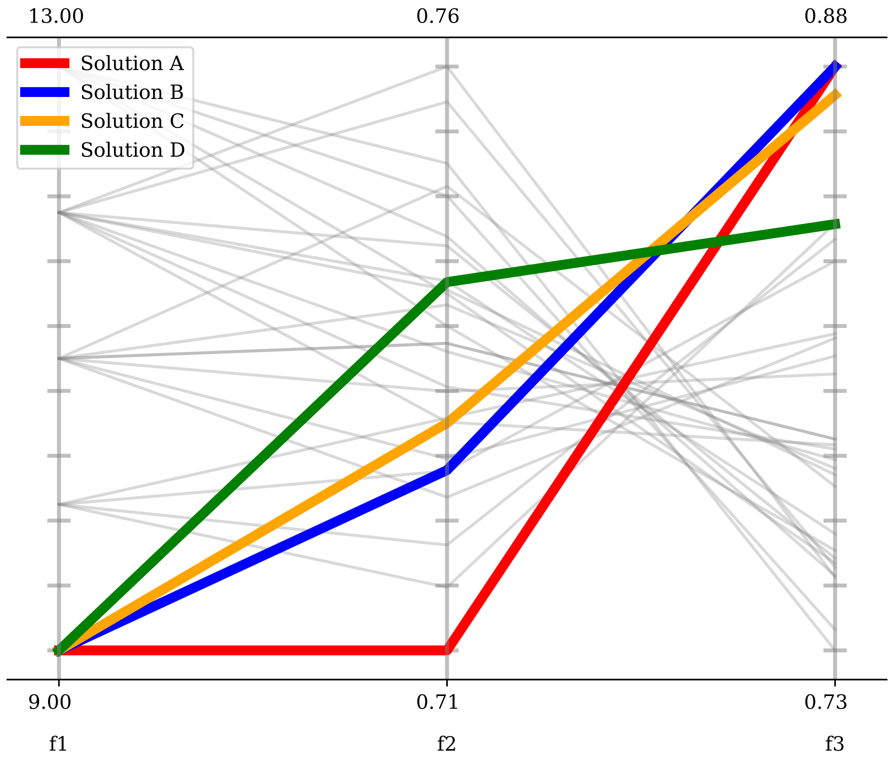


Figure 7.7: The set of non-dominant solutions as a parallel coordinate plot

We may notice that Solution A provides the best values for the first two objectives that control the team size and the average team expertise, namely $f1 = 9$ and $f2 = 0.7086$, while its collaborative prospects are low (i.e., $f3 = 0.8769$ is much higher than the minimum value $min(f3) = 0.7287$). The other three solutions having a minimal number of $f1 = 9$ members, namely B, C, and D, characterize teams with average values of expertise and also low collaborative expectations.

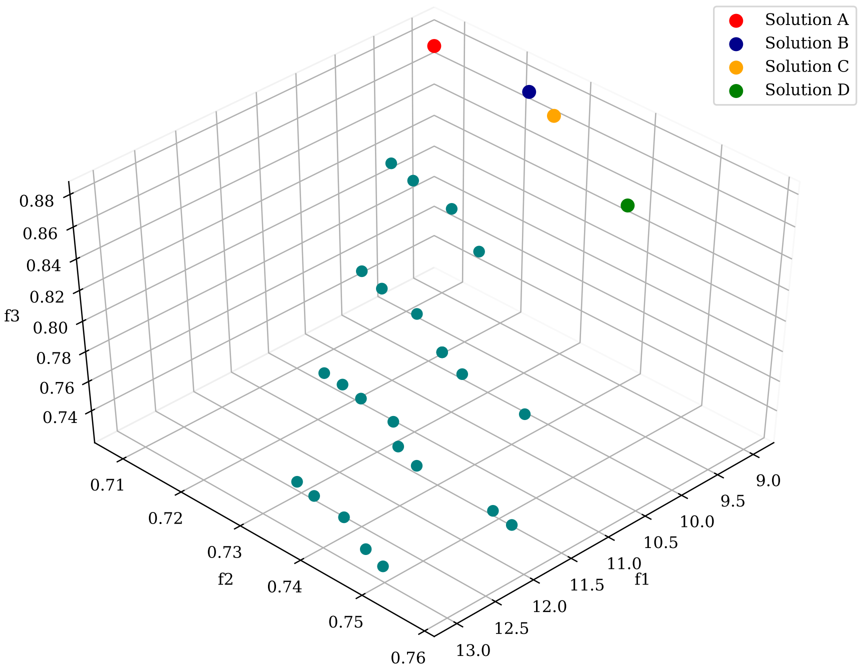


Figure 7.8: The set of non-dominant solutions as a scatter plot

7.3 Bibliographic Data-Driven Research Team Recommender

7.3.1 Formalization of Research Team Optimization Problem From Bibliographic Metadata

Formalizing a team formation optimization problem not only depends on the project the team has to fulfill or the potential pool of candidates but also on the type, size, quality, and reliability of available data regarding team candidates. From this perspective, bibliographic/bibliometric metadata provide meaningful insights about researchers' expertise and their personality and teamwork profiles.

The history of using bibliographic records for research team formation goes back to 2009, when Lappas et al. [45] interrogated the DBLP database to acquire raw data about candidates. They used paper title and authorship metadata fields to obtain researchers' areas of expertise, and paper authorship fields to infer the co-authorship graph. Afterward, the citation counts bibliometric field [167, 168] was considered for assessing the influence of an author or paper, while the keywords field was used to complement the title when discovering the researchers' areas of expertise [169]. Carefully analyzing the structure of paper metadata records, we consider that the list of mentioned bibliographic metadata fields that may be employed when evaluating researchers can be expanded by

Table 7.1: Optimal teams with nine members

Solution	f1	f2	f3	Member IDs
A	9	0.7086	0.8769	ID10;ID15;ID17;ID20;ID24;ID29;ID33;ID34;ID36
B	9	0.7238	0.8769	ID10;ID15;ID17;ID20;ID24;ID29;ID32;ID34;ID36
C	9	0.7278	0.8697	ID10;ID17;ID18;ID20;ID24;ID29;ID32;ID34;ID36
D	9	0.7397	0.8369	ID10;ID15;ID16;ID20;ID24;ID29;ID30;ID32;ID34

including the paper abstract alongside the title and keywords fields to find the candidates' areas of expertise, and download counts in correlation with citation counts for revealing the influence of an author or publication. Furthermore, we consider that interpersonal collaborations may be better understood by taking into account not only the past dyadic cooperations between candidates, but also existing candidates' collaborations inside groups larger than pairs (e.g., triadic, or tetradic inter-candidates collaborations). Based on these considerations, in this paragraph, we suggest a general research team formation model.

Our endeavor to develop a research team formation methodology driven by candidates-related information extracted from publication metadata begins with a thorough examination of bibliographic metadata from the perspective of candidates' assessment and continues with problem formalization and solving,

Employing bibliographic metadata fields in candidates' assessment [18]

In our view, the information provided by bibliographic records may be employed to:

- **identify the Researcher's Areas of Expertise (RAE)** using information from all the three metadata fields that are meant to effectively summarize the content of the publication (i.e., 'title', 'keywords', and 'abstract'). For this, we need to extract the pertinent key terms that best encapsulate the researcher's scientific production using appropriate NLP approaches and associate these terms with scientific areas.

- **compute four candidate-related indices** able to reflect the researcher's technical and teamwork abilities, as follows:

  ◇ **Researcher's General Expertise (RGE)** that can be derived based on the number of publications, number of citations, and number of downloads. Since the number of publications of a specified author can be determined by counting the number of her/his bibliographic records, the number of downloads or citations can be calculated by summing up the 'download counts' and 'citation counts' metadata fields, respectively. It is worth mentioning that the RGE is an overall indicator that considers the entire scientific output of a researcher, measuring the researcher's reputation. Alternatively, if already computed indexes are available (e.g., the h-index in Scopus or WoS databases), they may also be utilized.

- ◇ **Researcher's Level of Expertise in a Given Area (RLEGA)** that can be obtained from the number of scientific publications in that specific area and the related number of citations and downloads. The mechanism to derive these values is almost identical to the one employed when evaluating the RGE except it only considers the publications characterized by key terms belonging to the set of scientific area's relevant key terms.
- ◇ **Researcher's Collaboration Ability (RCA)** that can be assessed by using the total number of her/his co-authors and the number of co-authors having other affiliations. While the first one provides a general view of the researcher's collaborative prospects, the second one may be especially important in the case of projects involving multicultural, multilingual, and multinational research teams.
- ◇ **Interpersonal Collaborations Inside Specified Groups (ICISG)** that can be derived from the total number of already-existing collaborations inside that particular group. Since a group of researchers may have two, or more members, their fruitful collaboration can be described by a greater number of past collaborations (i.e., all individuals in that group are co-authors of the same publications.)

The utilization of bibliographic record fields in identifying and assessing the abilities of candidates or groups of candidates is displayed in Table 7.2.

The four indicators that can be computed using information extracted from a corpus of bibliographic metadata, namely RGE, RLEGA, RCA, and ICISG, represent the pillars of our methodology, their formalization being provided and discussed later in this section.

Proposed bibliographic data-driven egalitarian team formation methodology

Driven by the candidate-related insights extracted from the bibliographic metadata, we propose a general methodology for egalitarian research team formation. The flowchart of this methodology that implements a human-in-the-loop recommendation system is displayed in Figure 7.9. As we may notice, the recommender has the following set of inputs: (i) a carefully curated corpus of bibliographic metadata; (ii) the specifications of the research project to be fulfilled; and, (iii) details regarding the organizational context where the resulting research team will operate.

The first stage of this procedure is focused on publication metadata preprocessing and candidates' evaluation. It offers a core set of indicators regarding candidates' technical and non-technical abilities. By describing the overall scientific activity of the candidates, the RGE, RCA, and ICISG indices are not project-dependent. To obtain the RLEGA indicator, we have to focus only on the set of key terms that precisely characterize the

Table 7.2: Bibliographic metadata utilization in research team formation [18]

Publication Metadata Field	***Identify RAE***	***Evaluate RGE***	***Evaluate RLEGA***	***Evaluate RCA***	***Evaluate ICISG***
title	✓	–	–	–	–
abstract	✓	–	–	–	–
keywords	✓	–	–	–	–
author name	✓	✓	✓	✓	✓
author affiliation	–	–	–	✓	–
citing paper count	–	✓	✓	–	–
citing patent count	–	✓	✓	–	–
downloads count	–	✓	✓	–	–
paper ID	–	✓	✓	✓	✓

areas of expertise covered by the project. Thus, we must identify the set of project's relevant key terms from the overall list of terms output by the 'identify RAE' block. In the case one or more such key terms are not comprised in the overall key term list, the publication metadata corpus needs to be reprocessed by searching for these terms within the publications' 'title', 'keywords', and 'abstract' fields.

In the next stage of our methodology, the team formation problem formalization is performed considering the required candidates' features, available information about the research project, and organizational context (e.g., budget, interaction with other research projects, location, research infrastructure, etc.). As a result, a project-specific multi-objective combinatorial optimization problem is obtained, which may afterward be reshaped or even simplified to meet the requirements of a chosen problem solver method.

The list of suggested teams is provided to the initiator who may pick her/his favoured team composition. If the process outputs inadequate results (e.g., conflicting or inappropriate research teams), the initiator may restart the team formation sequence by making appropriate changes inside the preceding stages (e.g., trying to collect new and more extensive information, modifying the problem formulation by reshaping the objective functions or the constraints, or choosing another solver).

As it may be noticed from Figure 7.9, our proposed methodology is a human-assisted one, where the team initiator plays a decisive role not only in formulating the optimization problem, but also in choosing the solving method, or in picking the most appropriate team to fulfill the research subject.

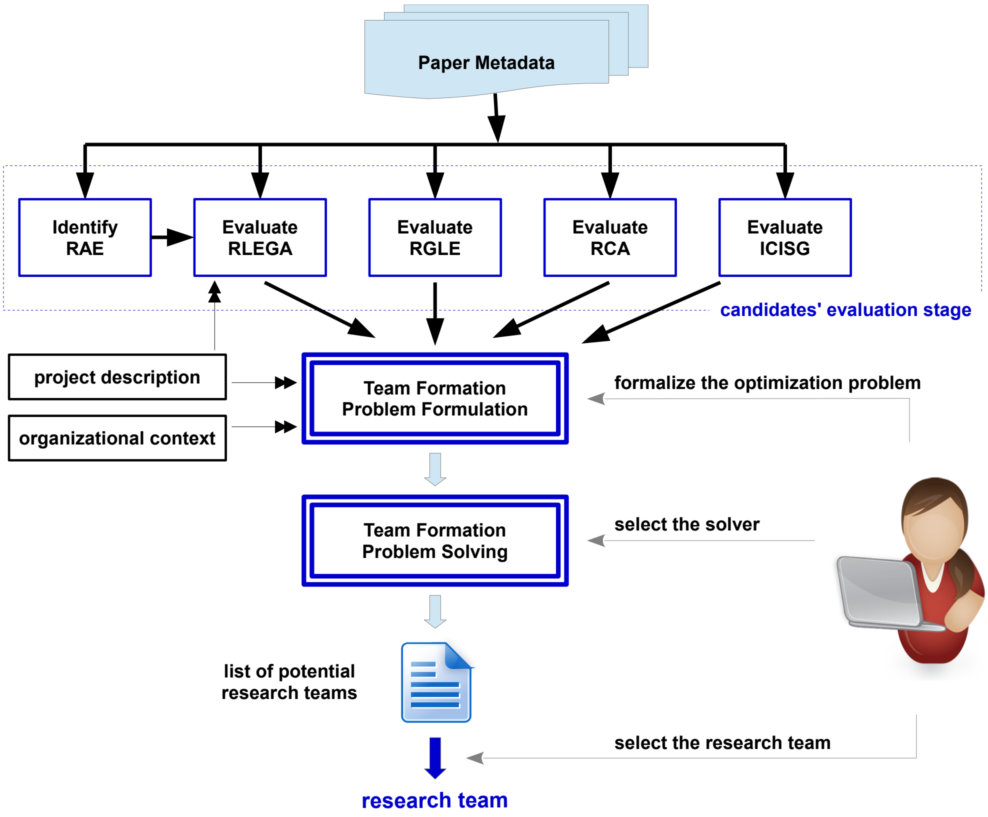


Figure 7.9: Bibliographic data-driven research team recommender framework [18]

General team formation problem with bibliographic metadata inputs

Driven by the four candidates-related indicators (i.e., RGE, RCA, ICISG, and RLEGA) extracted from bibliographic records, we suggest the following formalization for the egalitarian team formation optimization problem [18]:

$$\text{(F1):} \quad \text{minimize} \quad \sum_{S_i \in S} x_{S_i} \tag{7.32}$$

$$\text{(F2):} \quad \text{maximize} \quad \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{1}} \cdot \sum_{S_i \in S} RGE_{S_i} \cdot x_{S_i} \tag{7.33}$$

$$\text{(F3):} \quad \text{maximize} \quad \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{1}} \cdot \sum_{S_i \in S} RCA_{S_i} \cdot x_{S_i} \tag{7.34}$$

$$\text{(F4):} \quad \text{maximize} \quad \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{2}} \cdot \sum_{\substack{j=1...\binom{m}{2} \\ S_i \in S}} ICISG_j^{(2)} \cdot \sqcap_j^{(2)} \tag{7.35}$$

$$\text{subject to} \quad \sum_{S_i \in S} RLEGA_{S_i,e} \cdot x_{S_i} > \tau \text{ for all } e \in P \tag{7.36}$$

The objective functions (F1)-(F4) minimize the number of team members (F1), and maximize the average team member's general expertise (F2), the average team member collaboration prospects (F3), and the average collaboration between pairs of team members (F4). Moreover, the constraint guarantees that any scientific area, characterized by the e key term, of the research project, is adequately covered with expertise with a τ redundancy. The remainder of the notations are briefly described as follows:

	Notations:
P	project comprising a set of tasks
e	task from P ($e \in P$)
m	number of candidates
S_i	set of tasks from P that can be solved by candidate i, $i = 1, .., m$, $S_i \subseteq P$
x_{S_i}	binary flag corresponding to candidate i ($x_{S_i} = 1$ if candidate i is included in the team)
S	collection of all individual abilities $S = \{S_i \vert i = 1, ..., m\}$ - set of sets
k	size of the group taken into consideration
$\sqcap_j^{(k)}$	j^{th} product of k different flags x_{S_i}, $j = 1...\binom{m}{k}$
$\binom{\sum_{S_i \in S} x_{S_i}}{1}$	number of team members
$\binom{\sum_{S_i \in S} x_{S_i}}{2}$	number of pairs of team members

The multi-objective optimization problem (7.32)-(7.36) may be augmented by incorporating new project-specific criteria and constraints resulting from either the project specifications or organizational context (e.g., performance- or cost-related requirements). Additionally, information on previous collaborations inside groups of more than two team members, including triadic or tetradic collaborations, may serve as supplemental optimization goals as expressed by the following general objective function:

$$\text{(F5):} \qquad \text{maximize} \quad \frac{1}{\binom{\sum_{S_i \in S} x_{S_i}}{k}} \cdot \sum_{\substack{j=1\ldots\binom{m}{k} \\ S_i \in S}} ICISG_j^{(k)} \cdot \sqcap_j^{(k)} \tag{7.37}$$

where the considered group size is denoted by k.

Candidate-related indices formalization [18]

This paragraph presents the way the four candidate-related indices, namely RGE, RCA, ICISG, and RLEGA may be derived from a corpus of bibliographic metadata.

A. RGE formalization

In our view, when deriving the RGE index for a given candidate, two expertise-related facets are worth considering: (a) scientific output, characterized by the number of publications pn; and, (b) popularity among the scientific community, having two components i.e., the total amount of citations cn and the total amount of downloads dn that were received by all the candidate's publications.

$$\mathfrak{E}_i = \mu_1 \cdot pn_i + \mu_2 \cdot cn_i + \mu_3 \cdot dn_i \tag{7.38}$$

with $\mathfrak{E}_i$ denoting the candidate expertise, while $\mu_1, \mu_2, \mu_3 \in \mathbb{R}_+$ are carefully chosen coefficients that are meant to control the balance between the three expertise-related components. In the case $\mu_1 + \mu_2 + \mu_3 = 1$, the three weights correspond to the percentages in which each of the three components is considered in the overall expertise.

The RGE index for a candidate i may now be computed by normalizing the $\mathfrak{E}_i$ value using a classic normalization method, like min-max or z-score normalization, as follows:

$$RGE_i = \underset{i=1,\ldots,m}{normalize}(\mathfrak{E}_i) \tag{7.39}$$

If we decide to switch (F2) to a minimization objective type, RGE_i will be replaced by inv_RGE_i with

$$inv\_RGE_i = 1 - \underset{i=1,\ldots,m}{normalize}(\mathfrak{E}_i) \tag{7.40}$$

B. RCA formalization

We derive the researcher's collaboration ability $\mathfrak{C}_i$ from two values that can be extracted from authors' 'affiliation' fields, namely: (a) total number of candidate's co-authors from her/his institution (ci); and, (b) total number of candidate's co-authors from outside her/his institution (co). While the former reflects the natural collaborations appearing in an organization, the latter reveals the candidate's potential to work in more heterogeneous clusters that sometimes become multicultural, multilingual, or even multinational teams.

$$\mathfrak{C}_i = \delta_1 \cdot ci_i + \delta_2 \cdot co_i \tag{7.41}$$

with $\delta_1, \delta_2 \in \mathbb{R}_+$ being two chosen weights that control the balance between the two collaboration-related parts. If the selection of these coefficients satisfies the constraint $\delta_1 + \delta_2 = 1$, they will represent the percentages in which the two mentioned components are taken into account.

We may now obtain the RCA index for each candidate i by normalizing the corresponding $\mathfrak{C}_i$ value using a classic normalization method, like min-max or z-score normalization:

$$RCA_i = \underset{i=1,\dots,m}{normalize}(\mathfrak{C}_i) \tag{7.42}$$

If we decide to transform the (F3) objective function into a minimization objective, RCA_i will be replaced by inv_RCA_i:

$$inv\_RCA_i = 1 - \underset{i=1,\dots,m}{normalize}(\mathfrak{C}_i) \tag{7.43}$$

C. ICISG formalization

Interpersonal collaborations inside k-size groups quantifies the already-existing collaboration history of the specified k members. The ICISG value is computed based on the number of proven past collaborations of that particular group obtained by examining the bibliographic record fields related to publication authorship. For example, for a group composed of three members, ICISG is calculated using the total number of publications co-authored by all three members.

Since the relevance of even a single such cooperation is extremely high (a number of encountered collaborations higher than one only underlines the existing relationship), to compute the ICISG index we will use the hyperbolic tangent $tanh(x)$, a non-linear function that diminishes its slope if $x \in \mathbb{N}$ increases:

$$ICISG_j^{(k)} = tanh\left(\eta \cdot ig_j^{(k)}\right) \tag{7.44}$$

with η being a scaling factor and $ig_j^{(k)}$ denoting the total number of already-existing collaborations between the k members of the j^{th} group. It is worth noting that by employing the hyperbolic tangent function, the $ICISG_j^{(k)}$ values will already be normalized.

In team formation optimization problems the ICISG parameter is typically incorporated inside maximization objective functions. In case we decide to switch to a minimization objective, the $inv\_ICISG_j^{(k)}$ parameter may be used instead of $ICISG_j^{(k)}$, with

$$inv\_ICISG_j^{(k)} = 1 - tanh\left(\eta \cdot ig_j^{(k)}\right) \tag{7.45}$$

While already-existing one-to-one collaborations among candidates (i.e., $k = 2$) are almost ubiquitously employed when evaluating the collaboration abilities [45, 155, 145] during data-driven team formation procedures, we consider that past collaborations inside larger groups (i.e., $k > 2$) may provide new and deeper insights.

D. RLEGA formalization

To assess the researcher's level of expertise within a specified scientific area, we suggest using three types of information, namely the number of publications in the area (pna) along with the number of citations (cna) and downloads (dna) received by the researcher's publications in that particular area. For this, we may use a procedure similar to the one employed in the case of RGE, but which considers only the researcher's publications in that area (i.e., publications characterized by at least a key term that belongs to the scientific area's characteristic key term or key terms).

We may compute the candidate's expertise in a specified area e, denoted by $\mathfrak{A}_{i,e}$ as:

$$\mathfrak{A}_{i,e} = \gamma_1 \cdot pna_{i,e} + \gamma_2 \cdot cna_{i,e} + \gamma_3 \cdot dna_{i,e} \tag{7.46}$$

where $\gamma_1, \gamma_2, \gamma_3 \in \mathbb{R}_+$ are chosen coefficients that control the balance between the three expertise components. In the case these weights satisfy $\gamma_1 + \gamma_2 + \gamma_3 = 1$, they may be viewed as the percentages in which the three components are taken in the overall expertise value.

We may now obtain the expertise level for the candidate i in the scientific area denoted by e, using the following formula:

$$RLEGA_{i,e} = tanh(\rho \cdot \mathfrak{A}_{i,e}) \tag{7.47}$$

with ρ being a chosen scaling factor.

In our team formation methodology, the $RLEGA_{i,e}$ indices are employed to assess the probable contribution of the candidates i to the coverage with expertise of a given scientific area e (the equation (7.36) represents a typical constraint for a set multi-cover optimization problem, imposing the required τ_e redundancy). From this perspective, the τ_e values must not exceed the sum of all the $RLEGA$ indices for the candidates having expertise in the scientific area e:

$$\tau_e \leq \tau_{e,max} = \sum_{S_i \in S} RLEGA_{S_i,e} \tag{7.48}$$

Example: Research team formation with candidates from Politehnica University Timisoara[7]

To show how our proposed method works and to compare it against existing approaches, in this paragraph, we present an illustrative case study [18]. For this, let

[7] This example uses the dataset hosted in Mendeley Data repository [16], briefly described in subsection 4.3.3 and presented in detail in our journal data paper [17].

us consider the research team formation optimization problem formalized by (7.32)-(7.36). We intend to identify suitable teams of researchers from Politehnica University Timisoara – Romania, that are able to fulfill the scientific project P, modeled by the following set of key terms: 'hard_real_time', 'machine_learning', 'computer_vision', 'gesture_recognition', and 'image_processing'. Supplementary, each of the five scientific areas described by the mentioned key terms has to be covered by researchers' expertise with a specified redundancy τ.

Since the key terms in the considered dataset were extracted from the 'title', 'abstract' and 'keywords' fields using $lp = 0.1$, we first have to check if all the terms describing our research project are contained in the key term list. In our case, considering the lp values corresponding to each key term that models our research project presented in Table 7.3, the considered threshold value, namely $lp = 0.1$, needs to fulfill the condition $lp < 0.332503$, which is obviously true.

Table 7.3: Link probability scores for the key terms describing the research project [18]

Term	**lp-score**
hard_real_time	0.42307
machine_learning	0.75125
computer_vision	0.77634
gesture_recognition	0.76470
image_processing	***0.33250***

To build the research team we employed the NSGA-II evolutionary multi-objective optimization problem solver [164] and the following set of parameters:

1. Parameters used to compute the candidate-related indices:
   - RGE-related weights: $\mu_1 = 0.9$, $\mu_2 = 0.0999$, $\mu_3 = 0.0001$
   - RCA-related weights: $\delta_1 = 0.3$, $\delta_2 = 0.7$
   - ICISG-related weights: $\eta = 1$
   - RLEGA-related weights: $\gamma_1 = 0.9$, $\gamma_2 = 0.0999$, $\gamma_3 = 0.0001$, $\rho = 1$
2. Parameters to shortlist the pool of candidates
   - minimal number of publication relevant to the research project: 1
   - minimal number of citations received for publications relevant to the research project: 1
3. Parameters of the NSGA-II solver

- number of objective functions: 4
- number of constraints: 5
- coverage redundancy: $\tau = [4, 4, 4, 4, 4]$
- optimization type: *ElementwiseProblem* - it evaluates the candidate solutions one by one
- sampling mechanism: binary random sampling
- crossover mechanism: two-point crossover
- mutation mechanism: bit flip mutation
- size of population: 100
- number of generations: 100
- Python function: *NSGA2()* from *pymoo* library [165].

After the candidate prefiltering process, only 84 relevant candidates have been retained. This shortlisted pool of candidates provides the following set of maximal redundancy values (each of them correspond to one of the five key terms describing the project), computed using (7.48):

$$\tau_{max} = [6.51057, 29.837, 30.1965, 12.3179, 27.5543] \tag{7.49}$$

Since each component of the τ vector is less than its correspondent in τ_{max}, at least one optimal solution for the team formation problem exists.

The proposed method was implemented in Python 3.8 and is based on the *pymoo* multi-objective optimization package [165]. The set of 68 non-dominant solutions that have been obtained are displayed as a parallel coordinate plot in Figure 7.10. If we consider a lexicographic approach to rank the solutions (i.e., the objective functions are ranked in their descending relevance order, namely F1, F2, F3, F4) the best 9 solutions, are listed in Table 7.4. It is worth mentioning that the given research theme may be fulfilled by a minimal team consisting of twelve experts (Solutions A-H), but the other objectives (i.e., F2, F3, and F4) have greater values compared to the ones of other recommended teams.

Figure 7.11 describes how the expertise-related requirements are covered by the researchers composing the winning team (i.e., Solution A). We may observe, that this optimal solution provides a good balance of the number of researchers who have the necessary expertise to tackle each of the project's subdomains.

Comparative analysis with other methods

Existing research on bibliographic metadata utilization when building research teams often relies on the use of the DBLP database [45, 170, 171, 172, 173, 174], which covers a narrow scientific field, namely Computer Science, and is characterized by an extremely

Table 7.4: Optimal teams with fewer team members [18]

Solution	F1	F2	F3	F4	Team Member IDs
A	12	0.79941	0.78742	0.85796	ID20,ID440,ID757,ID759, ID799,ID802,ID803,ID900, ID942,ID944,ID984,ID1049
B	12	0.80300	0.78598	0.86410	ID20,ID440,ID757,ID759, ID799,ID803,ID804,ID900, ID942,ID944,ID984,ID1049
C	12	0.81633	0.85138	0.84642	ID20,ID440,ID757,ID759, ID793,ID799,ID802,ID803, ID900,ID942,ID984,ID1049
D	12	0.81993	0.84995	0.85256	ID20,ID440,ID757,ID759, ID793,ID799,ID803,ID804, ID900,ID942,ID984,ID1049
E	12	0.82319	0.84469	0.85058	ID440,ID732,ID757,ID759, ID773,ID799,ID802,ID803, ID900,ID942,ID984,ID1049
F	12	0.83003	0.84243	0.82750	ID440,ID757,ID759,ID773, ID799,ID802,ID803,ID804, ID900,ID942,ID984,ID1049
G	12	0.83152	0.83978	0.84949	ID20,ID440,ID757,ID759, ID773,ID799,ID803,ID804, ID900,ID942,ID984,ID1049
H	12	0.84752	0.86358	0.82695	ID440,ID757,ID759,ID799, ID802,ID803,ID804,ID888, ID900,ID942,ID984,ID1049
I	13	0.7752	0.77607	0.89424	ID20,ID138,ID732,ID757, ID759,ID793,ID799,ID803, ID848,ID942,ID944,ID984, ID1127

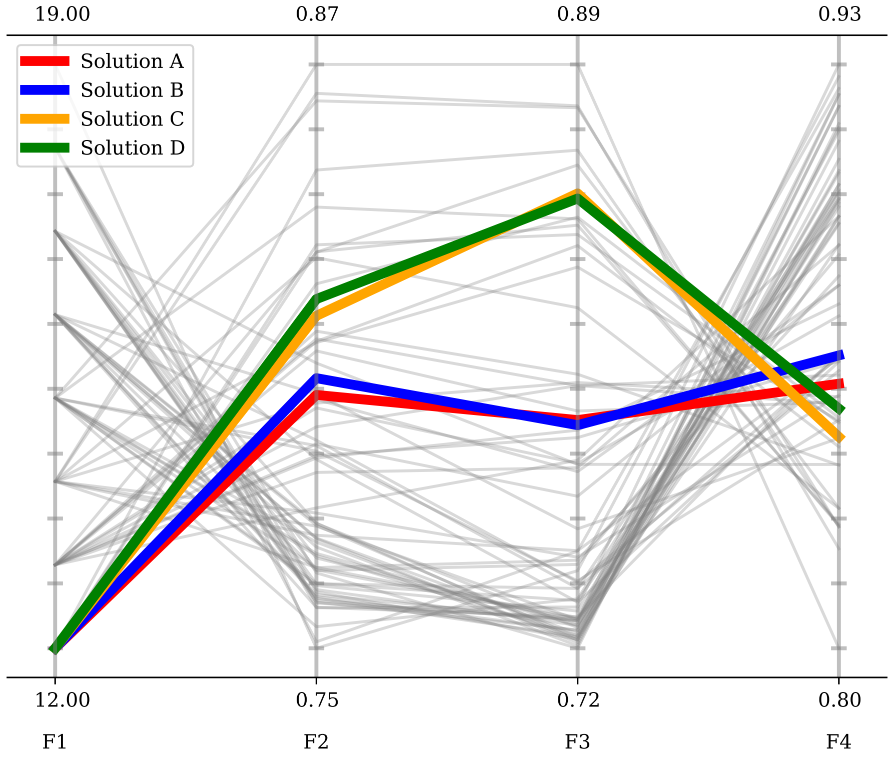


Figure 7.10: Optimization results as parallel coordinate plot [18]

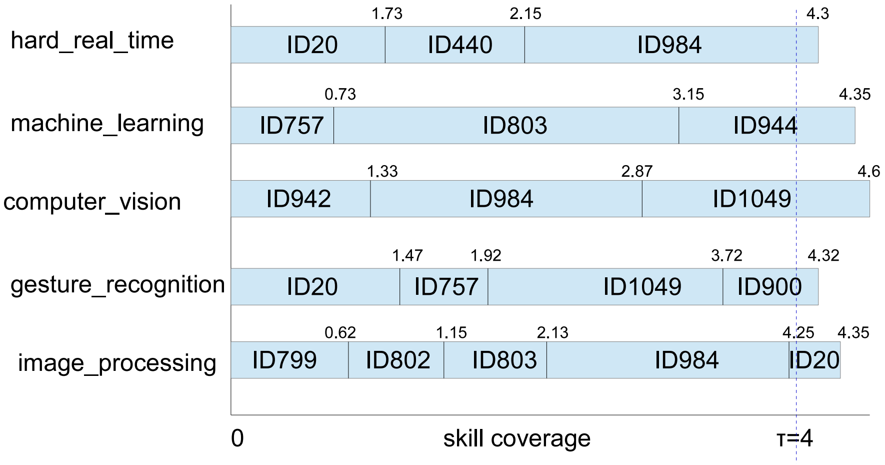


Figure 7.11: Coverage of the scientific areas in the case of the best team [18]

simple record structure that is not appropriate for a deep and systematic examination of candidates' technical and teamwork skills.

Even though some works employ other bibliographic databases they use the same limited number of record fields (i.e., 'title' and 'authors') thus not taking full advantage of a more complex metadata structure [46, 175, 176]. By adding additional critical insights that bibliographic information may provide, our technique provides more relevant and accurate expert teams, this being a combined result of two aspects:

(1) a more comprehensive multi-objective optimization model that lies behind the research team formation problem is formalized; compared with the state-of-the-art methods that contain no more than three optimization objectives [174]), our model has four objectives.

(2) our method examines a larger set of metadata fields to assess the personal and interpersonal attributes of candidates, thus allowing for capturing some insights never used before. To exemplify, employing all three fields that summarize a publication (i.e., 'title', 'keywords', and 'abstract') for key terms discovery instead of only the 'title' field as all of the prior works do, allows for a higher term granularity offering more control over the entire data-driven research team formation procedure. Supplementary, utilizing the metadata fields containing the number of citations and the number of downloads alongside the number of publications, we are able to provide a better categorization of the researcher's general or domain-specific expertise. Moreover, inspecting the 'affiliation' field, to identify existing collaborations not only inside the researcher's organization but also with researchers from other organizations, may more accurately reveal candidates' teamwork capabilities.

To assess the overall influence of assuming information from the 'title', 'abstract', and 'keywords' metadata fields on the establishment of a research team, we considered four distinct scenarios described in Table 7.5. It is evident that when taking into account all three of the aforementioned data fields, we provide a far more complete picture of every retrieved publication than when using only one or two of these fields (the number of unique key terms is far higher when all the three fields are considered). It is important to highlight that all state-of-the-art techniques rely on datasets taken from the DBLP bibliographic database, which only contain the 'title' field and lack the 'abstract' and 'keywords' fields.

Moreover, we may observe that if we only use the 'title' or 'keyword' fields, no team will be generated. This is expressed in the five τ_{max} components, one for every key term describing the research theme, which must be higher than the considered redundancy (i.e., $\tau_{max,i} > 4$ with i=1...5). This occurs because at least one of the key terms related to candidates do not appear in the considered metadata fields (i.e., 'title' and 'keywords', respectively).

Table 7.5: Using diverse metadata fields to derive the key terms [18]

Metadata fields	**Key Terms**	τ_{maxHRT}	τ_{maxML}	τ_{maxCV}	τ_{maxGR}	τ_{maxIP}	**Shortlisted Candidates**	**Team Members**
’title’	1844	6.5105	0.7678	3.8345	4.4193	3.6055	11	no team
’keywords’	1254	0	18.0555	21.9328	7.6421	15.8878	56	no team
’title’,’keywords’	2651	6.5105	18.0555	21.9328	7.6851	17.5061	61	15
’title’,’keywords’,’abstract’	6493	6.5105	29.8370	30.1965	12.3179	27.5543	84	12

Chapter 8

Conclusion

In any research domain, identifying potential high-reward research themes becomes increasingly challenging mainly due to the difficulties in evaluating the state of research and its trends which derive from an explosive rise in the number of publications. To cover this gap, we designed a human-in-the-loop multi-recommender system framework meant to frame new research themes and facilitate their starting by using an ensemble of AI techniques. Employing information encapsulated in bibliographic metadata, our semi-automatic recommender framework evaluates the research trends to identify popular sets of terms, helps researchers in discovering feasible research gaps, and formalizes the proposed research themes as undirected graphs of terms. Moreover, the system gathers useful theme-related knowledge, suggests knowledge transfers from twin and emerging domains, and helps the research team formation.

The current abundance of available data complicates human decision-making to an unprecedented level. Due to their promise to process large amounts of information and to extract personalized suggestions, recommender systems powered by artificial intelligence are progressively being used for a variety of tasks, thus changing the way we make decisions in a variety of domains. The need for such tools is increasing, especially to assist decisions that rely on tedious and time-consuming exploratory activities. In this category, there is intriguing promise in the framing of new research themes, which are not only based on researchers' expertise and interests but also driven by research trends, technological and scientific novelties, and team formation-related challenges.

Any attempt to frame new research themes generally starts with a literature review. Since a comprehensive and systematic review of an exponentially increasing body of work becomes harder, researchers have applied diverse filters to narrow the amount of information taken into consideration and by this to reduce the time needed to identify potential research ideas. For example, such filters are used when searching for relevant publications and may include: (a) considering only papers published in top-tier journals and conferences; (b) preferring authors or groups of authors based on their affiliation; (c) considering publications from a reduced set of databases; or (d) focusing on popular work within the community. With the growing number of publications, the filter-based strategy could become less effective (as filters must be increasingly selective, thus arguably limiting the number of relevant research themes that can be discovered). This situation calls for the need for automatic or semi-automatic research theme recommender systems. In this context, our system provides a convenient and promising solution.

Besides the benefits that were already mentioned in this thesis, including the possibility of investigating a large body of information and evaluating ongoing research trends, our multi-recommender system framework also offers three other advantages:

- Compared to the corresponding manual procedure, our multi-recommender system can be configured to be almost immune to user's subjectivity (e.g., inherent fears of novelty and uncertainty, concerns regarding the long time and effort needed for a researcher's preparation for a new theme, or worries that the projected results will not materialize).
- The multi-recommender can be easily extended either by adding new information sources (e.g., patents or surveys within the scientific community), or by incorporating other AI/ML methods to process information and derive the recommendations.
- A recommender system is useless if the intended user does not trust it [177]. The trust can be established in the case of a recommender system by clearly explaining the logic behind it, including the way it generates recommendations, and the reasons a given item is recommended. This condition is met in our case since the underlying mechanism of the proposed multi-recommender is simple to understand and generally reflects the way researchers are manually framing their new research themes.

We also identified some **limitations of our work** that are rooted either in the specific nature of the solved problem or in the way the set of recommenders are implemented:

- The proposed recommender system is hard to systematically evaluate in real-life scenarios. There are three reasons for this: (a) the researchers rarely and sporadically change their research theme, so gathering a comprehensive dataset for evaluation purposes is hard; (b) the manual procedures for research theme framing, cross-domain knowledge transfer, bibliography identification or team formation usually do not cover all possible cases, thus the evaluation dataset tends to be unbalanced; and (c) evaluating the plausibility of each recommendation requires skilled expertise, which might be difficult to access. The lack of evaluation is not uncommon for recommender systems [178] and can be partially mitigated by making the logic behind the system to be available and clear to the user.
- Since the access to the publication databases is often not free of charge and moreover, the format in which publications are stored in these databases is not standardized, including all relevant publications in the process is challenging.

- There are certain publication types (i.e., books, book chapters, scientific reports, patents) that were not considered as inputs for our multi-recommender system, mainly because extracting their extended abstracts (obtained by concatenating abstract, title, and keywords) is not straightforward.
- Our implementation does not consider the language as a parameter. Thus, our multi-recommender system covers only publications and research themes written in English.

8.1 Contributions

The thesis proposes several new approaches and techniques for aiding researchers when intending to start working on new research projects. The original contributions of this work are briefly presented below.

- ✦ A comprehensive and detailed analysis of the existing computer-based recommender methods and systems for aiding research efforts.
- ✦ A multi-recommender framework to derive hot and customized research theme proposals alongside recommendations regarding adequate cross-domain knowledge transfers, initial bibliography to start with, and research team selection. This architecture takes bibliographic metadata as inputs and incorporates four independent recommender modules.
- ✦ A practical methodology to customize the TagMe entity linking method according to the needs of a given scientific domain using a preliminary list of user-defined domain-specific key terms.
- ✦ A research theme recommender system that, based on bibliographic records, identifies the research themes that characterize a given scientific domain and evaluates their hotness and feasibility.
- ✦ A technique to investigate the popularity of research themes within the scientific community. Modeling the research theme as a finite set of key terms facilitated a multivariate time series trend analysis by employing a variant of the Mann-Kendall test.
- ✦ A method to extract research trends from paper metadata, when considering the publication and indexing latency, using the auto-ARIMA prediction method and Mann-Kendall test.
- ✦ A statistical double-threshold technique to assess the feasibility of a research theme by considering novelty and success-related aspects.

- ✦ A cross-domain knowledge transfer recommender system that, based on bibliographic metadata, identifies the pieces of knowledge from twin and emerging domains to be transferred and customized.
- ✦ A method to identify the twin domains of a given scientific domain from where the knowledge transfers are more likely to occur using an NLP approach based on document similarity evaluation.
- ✦ A general team formation optimization model able to be tailored to suit various data-driven team formation processes, derived from the set cover optimization problem.
- ✦ A set of four new synthetic indicators to effectively describe experts' knowledge and their collaborative prospects from the bibliographic metadata.
- ✦ A novel multi-objective team formation optimization model that adequately includes the candidate-related indicators derived from bibliographic records.

8.2 Perspectives

Since in our opinion this work is among the first to tackle the problem of research theme framing and addressing using bibliographic records, the research area is wide open. To further improve the proposed human-in-the-loop multi-recommender system five research directions are worth mentioning: (a) automating the selection and fine-tuning of the parameters used by employed AI techniques; (b) including new sources of information regarding scientific research (e.g., databases containing research projects like CORDIS or software repositories like GitHub); (c) coping with fake and bogus scientific publications; (d) validating the proposed multi-recommender framework on other relevant bibliographic/bibliometric databases, like PubMed, Scopus, Web of Science or Scopus, and analyzing how information acquired from various bibliographic sources can enhance the accuracy of the proposed approach; (e) designing of an effective and more customer-oriented system for scientific literature recommendations, this problem being only tangentially tackled inside this thesis.

List of Publications

The following publications are stemming from the research carried out by the author during his PhD studies:

Web of Science Journal Papers:

1. **C.-D. Curiac**, and A. Doboli, "Combining informetrics and trend analysis to understand past and current directions in electronic design automation", Scientometrics, vol. 127(10), pp. 5661-5689, Springer, 2022.

2. T. Andreica, **C.-D. Curiac**, C. Jichici and B. Groza, "Android head units vs. in-vehicle ECUs: performance assessment for deploying in-vehicle intrusion detection systems for the CAN bus", IEEE Access, vol. 10, pp. 95161-95178, IEEE, 2022.

3. **C.-D. Curiac**, O. Banias, and M. Micea, "Evaluating research trends from journal paper metadata, considering the research publication latency", Mathematics, vol. 10(2), 233, MDPI, 2022.

4. **C.-D. Curiac**, A. Doboli, and D.-I. Curiac, "Co-occurrence-based double thresholding method for research topic identification", Mathematics, vol. 10(17), 3115, MDPI, 2022.

5. **C.-D. Curiac**, and M. Micea, "Identifying hot information security topics using LDA and multivariate Mann-Kendall test", IEEE Access, vol. 11, pp. 18374-18384, IEEE, 2023.

6. **C.-D. Curiac**, M. Micea, T.-R. Plosca, D.-I. Curiac, and A. Doboli "Dataset for bibliometric data-driven research team formation: case of Politehnica University of Timisoara scholars for the interval 2010-2022", Data in Brief, vol. 53, 110275, Elsevier, 2024.

7. T.-R. Plosca, **C.-D. Curiac**, and D.-I. Curiac. "Investigating semantic differences in user-generated content by cross-domain sentiment analysis means", Applied Sciences, vol. 14(6), 2421, MDPI, 2024.

8. **C.-D. Curiac**, M. Micea, T.-R. Plosca, D.-I. Curiac, and A. Doboli "Optimized interdisciplinary research team formation using a genetic algorithm and extended bibliometric data". *[under review]*

Book Chapters:

1. **C.-D. Curiac**, M. Micea, T.-R. Plosca, D.-I. Curiac, S. Doboli and A. Doboli “Towards automating new research problem framing and exploration based on symbolic-numerical knowledge extracted from bibliometric data”, in Bibliometrics - An Essential Methodological Tool for Research Projects. IntechOpen, London, UK, 2024.

Peer-Reviewed Conference Papers:

1. **C.-D. Curiac**, and M. Micea, "Evaluating research trends using key term occurrences and multivariate Mann-Kendall test", in Proceedings of the International Symposium on Electronics and Telecommunications (ISETC 2022), pp. 1–4, IEEE, 2022.

2. M.D. Baciu, E.A. Capota, C.S. Stângaciu, **C.-D. Curiac**, and M. Micea, "Multicore time-triggered OCBP-based scheduling for mixed criticality periodic task systems", in Proceedings of the International Symposium on Electronics and Telecommunications (ISETC 2022), pp. 1–4, IEEE, 2022.

Public Datasets:

1. **C.-D. Curiac**, M. Micea, T.-R. Plosca, D.-I. Curiac, and A. Doboli "Dataset for bibliometric data-driven research team formation", Mendeley Data, version 1, doi: 10.17632r4vrvhb23h.1, 2023.

Bibliography

[1] G. Wisskirchen, B. Biacabe, U. Bormann, A. Muntz, G. Niehaus, G. Soler, and B. von Brauchitsch, "Artificial intelligence and robotics and their impact on the workplace," *IBA Global Employment Institute*, vol. 11, no. 5, pp. 49–67, 2017.

[2] J. Andreu-Perez, F. Deligianni, D. Ravi, and G.-Z. Yang, "Artificial intelligence and robotics," https://arxiv.org/ftp/arxiv/papers/1803/1803.10813.pdf, 2016, accessed: 2024-04-25.

[3] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research paper recommender systems: A literature survey," *International Journal on Digital Libraries, Springer*, vol. 17, pp. 305–338, 2016.

[4] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[5] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems: Survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

[6] K. Wei, J. Huang, and S. Fu, "A survey of e-commerce recommender systems," in *Proc. of the International Conference on Service Systems and Service Management*. IEEE, Chengdu, China, Jul. 2007, pp. 1–5.

[7] M. Schedl, P. Knees, and F. Gouyon, "New paths in music recommender systems research," in *Proc. of the Eleventh ACM Conference on Recommender Systems*. ACM, Como, Italy, 2017, p. 392–393.

[8] H. Lee and P. Kang, "Identifying core topics in technology and innovation management studies: A topic model approach," *The Journal of Technology Transfer*, vol. 43, pp. 1291–1317, 2018.

[9] H. Small, K. Boyack, and R. Klavans, "Identifying emerging topics in science and technology," *Research Policy*, vol. 43, no. 8, pp. 1450–1467, 2014.

[10] T. Kuhn, “The structure of scientific revolutions,” *University of Chicago press*, 1962.

[11] C.-D. Curiac, M. Micea, T.-R. Plosca, D.-I. Curiac, S. Doboli, and A. Doboli, “Towards automating new research problem framing and exploration based on symbolic-numerical knowledge extracted from bibliometric data,” in *Bibliometrics - An Essential Methodological Tool for Research Projects*. IntechOpen, London, UK, 2024.

[12] C.-D. Curiac, O. Banias, and M. Micea, “Evaluating research trends from journal paper metadata, considering the research publication latency,” *Mathematics*, vol. 10, no. 2, p. 233, 2022.

[13] C.-D. Curiac and M. Micea, “Evaluating research trends using key term occurrences and multivariate Mann-Kendall test,” in *2022 International Symposium on Electronics and Telecommunications (ISETC)*. IEEE, Timișoara, Romania, 2022, pp. 1–4.

[14] C.-D. Curiac and M. V. Micea, “Identifying hot information security topics using LDA and multivariate Mann-Kendall test,” *IEEE Access*, vol. 11, pp. 18 374–18 384, 2023.

[15] C.-D. Curiac, A. Doboli, and D.-I. Curiac, “Co-occurrence-based double thresholding method for research topic identification,” *Mathematics*, vol. 10, no. 17, p. 3115, 2022.

[16] C.-D. Curiac, M. Micea, T.-R. Plosca, D.-I. Curiac, and A. Doboli, “Dataset for bibliometric data-driven research team formation,” *Mendeley Data*, 2023, doi: 10.17632/r4vrvhb23h.1.

[17] C.-D. Curiac, M. Micea, T.-R. Plosca, D.-I. Curiac, and A. Doboli, “Dataset for bibliometric data-driven research team formation: Case of Politehnica University of Timisoara scholars for the interval 2010-2022,” *Data In Brief*, vol. 53, p. 110275, 2024.

[18] C.-D. Curiac, M. Micea, T.-R. Plosca, D.-I. Curiac, and A. Doboli, “Optimized interdisciplinary research team formation using a genetic algorithm and publication metadata records,” *[under review]*.

[19] F. Ricci, L. Rokach, and B. Shapira, “Recommender systems: Techniques, applications, and challenges,” *Recommender Systems Handbook*, pp. 1–35, 2021.

[20] R. Burke, A. Felfernig, and M. H. Göker, “Recommender systems: An overview,” *AI Magazine*, vol. 32, no. 3, pp. 13–18, 2011.

[21] L. Lü, M. Medo, C. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou, “Recommender systems.” *Physics Reports*, vol. 519, no. 1, pp. 1–49, 2012.

[22] B. Smith and G. Linden, “Two decades of recommender systems at amazon.com,” *IEEE Internet Computing*, vol. 21, no. 3, pp. 12–18, 2017.

[23] L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter, “Phoaks: A system for sharing recommendations,” *Communications of the ACM*, vol. 40, no. 3, pp. 59–62, 1997.

[24] H. Ko, S. Lee, Y. Park, and A. Choi, “A survey of recommendation systems: recommendation models, techniques, and application fields,” *Electronics*, vol. 11, no. 1, p. 141, 2022.

[25] J. L. Ortega, *Academic search engines: A quantitative outlook.* Elsevier, Oxford, UK, 2014.

[26] A. Polonioli, “The ethics of scientific recommender systems,” *Scientometrics*, vol. 126, no. 2, pp. 1841–1848, 2021.

[27] T. Yoneya and H. Mamitsuka, “PURE: A PubMed article recommendation system based on content-based filtering,” *Genome Informatics*, vol. 18, pp. 267–276, 2007.

[28] B. Gipp, J. Beel, and C. Hentschel, “Scienstein: A research paper recommender system,” in *Proc. of the International Conference on Emerging Trends in Computing (ICETiC’09).* IEEE,Virudhunagar, India, 2009, pp. 309–315.

[29] T. Achakulvisut, D. E. Acuna, T. Ruangrong, and K. Kording, “Science concierge: A fast content-based recommendation system for scientific publications,” *PloS One*, vol. 11, no. 7, p. e0158423, 2016.

[30] G. Guo, B. Chen, X. Zhang, Z. Liu, Z. Dong, and X. He, “Leveraging title-abstract attentive semantics for paper recommendation,” in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01. AAAI, New York, USA, 2020, pp. 67–74.

[31] K. Haruna, M. A. Ismail, A. Qazi, H. A. Kakudi, M. Hassan, S. A. Muaz, and H. Chiroma, “Research paper recommender system based on public contextual metadata,” *Scientometrics*, vol. 125, pp. 101–114, 2020.

[32] N. Sakib, R. B. Ahmad, M. Ahsan, M. A. Based, K. Haruna, J. Haider, and S. Gurusamy, “A hybrid personalized scientific paper recommendation approach integrating public contextual metadata,” *IEEE Access*, vol. 9, pp. 83 080–83 091, 2021.

[33] Á. Tejeda-Lorente, C. Porcel, E. Peis, R. Sanz, and E. Herrera-Viedma, "A quality based recommender system to disseminate information in a university digital library," *Information Sciences*, vol. 261, pp. 52–69, 2014.

[34] J. Serrano-Guerrero, E. Herrera-Viedma, J. A. Olivas, A. Cerezo, and F. P. Romero, "A Google wave-based fuzzy recommender system to disseminate information in university digital libraries 2.0," *Information Sciences*, vol. 181, no. 9, pp. 1503–1516, 2011.

[35] M. Färber and A. Jatowt, "Citation recommendation: Approaches and datasets," *International Journal on Digital Libraries*, vol. 21, no. 4, pp. 375–405, 2020.

[36] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, "Context-aware citation recommendation," in *Proc. of the 19th International Conference on World Wide Web*. ACM, Raleigh, NC, USA, 2010, pp. 421–430.

[37] H.-C. Wang, J.-W. Cheng, and C.-T. Yang, "Sentcite: a sentence-level citation recommender based on the salient similarity among multiple segments," *Scientometrics*, vol. 127, no. 5, pp. 2521–2546, 2022.

[38] L. Yang, Y. Zheng, X. Cai, H. Dai, D. Mu, L. Guo, and T. Dai, "A LSTM based model for personalized context-aware citation recommendation," *IEEE Access*, vol. 6, pp. 59 618–59 627, 2018.

[39] C. Jeong, S. Jang, E. Park, and S. Choi, "A context-aware citation recommendation model with BERT and graph convolutional networks," *Scientometrics*, vol. 124, pp. 1907–1922, 2020.

[40] H. Huang, H. Wang, and X. Wang, "An analysis framework of research frontiers based on the large-scale open academic graph." *Proceedings of the Association for Information Science and Technology*, vol. 57, no. 1, p. e307, 2020.

[41] M. Linnenluecke, M. Marrone, and A. Singh, "Conducting systematic literature reviews and bibliometric analyses," *Australian Journal of Management*, vol. 45, no. 2, pp. 175–194, 2020.

[42] M. Cornolti, P. Ferragina, and M. Ciaramita, "A framework for benchmarking entity-annotation systems," in *Proc. of the 22nd International Conference on World Wide Web*. ACM, Rio de Janeiro, Brazil, 2013, pp. 249–260.

[43] M. Marrone, "Application of entity linking to identify research fronts and trends," *Scientometrics*, vol. 122, no. 1, pp. 357–379, 2020.

[44] H. Wang, T. Hsu, and Y. Sari, "Personal research idea recommendation using research trends and a hierarchical topic model," *Scientometrics, Springer*, vol. 121, no. 3, pp. 1385–1406, 2019.

[45] T. Lappas, K. Liu, and E. Terzi, "Finding a team of experts in social networks," in *Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, USA, 2009, pp. 467–476.

[46] B. Srivastava, T. Koppel, S. Paladi, S. Valluru, R. Sharma, and O. Bond, "ULTRA: A data-driven approach for recommending team formation in response to proposal calls," in *Proc. 2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, Orlando, FL, USA, 2022, pp. 1002–1009.

[47] S. Kaw, Z. Kobti, and K. Selvarajah, "Transfer learning with graph attention networks for team recommendation," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Gold Coast, Australia, 2023, pp. 1–8.

[48] P. Leedy and J. Ormrod, *Practical research.* Pearson Education, Saddle River, NJ, 2005, vol. 108.

[49] P. Vugteveen, R. Lenders, and P. Van den Besselaar, "The dynamics of interdisciplinary research fields. the case of river research," *Scientometrics, Springer*, vol. 100, no. 1, pp. 73–96, 2014.

[50] J. Luo and G. Knoblich, "Studying insight problem solving with neuroscientific methods," *Methods, Elsevier*, vol. 42, no. 1, pp. 77–86, 2007.

[51] Z. Li and X. Zou, "A review on personalized academic paper recommendation." *Computer and Information Science*, vol. 12, no. 1, pp. 33–43, 2019.

[52] R. Sharma, D. Gopalani, and Y. Meena, "An anatomization of research paper recommender system: Overview, approaches and challenges," *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105641, 2023.

[53] M. McTear, Z. Callejas, and D. Griol, *The conversational interface: talking to smart devices.* Springer, Switzerland, 2016.

[54] C. Xiong, J. Callan, and T.-Y. Liu, "Bag-of-entities representation for ranking," in *Proc. of the ACM International Conference on the Theory of Information Retrieval.* ACM, Newark, DE, Sep. 2016, pp. 181–184.

[55] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques and solutions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, pp. 443–460, 2014.

[56] P. Ferragina and U. Scaiella, "Tagme: On-the-fly annotation of short text fragments (by wikipedia entities)," in *International Conference on Information and Knowledge Management.* ACM, Toronto, Canada, Oct. 2010, pp. 1625–1628.

[57] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in *Proc. of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Sapporo, Japan, Jul. 2003, pp. 216–223.

[58] M. Dojchinovski, D. Reddy, T. Kliegr, T. Vitvar, and H. Sack, "Crowdsourced corpus with entity salience annotations," in *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), Portorož, Slovenia, May 2016, pp. 3307–3311.

[59] J. L. Martinez-Rodriguez, A. Hogan, and I. Lopez-Arevalo, "Information extraction meets the semantic web: A survey," *Semantic Web*, vol. 11, no. 2, pp. 255–335, 2020.

[60] F. Hasibi, K. Balog, and S. Bratsberg, "On the reproducibility of the tagme entity linking system," in *European Conference on Information Retrieval Research*. Springer, Padua, Italy, 2016, pp. 446–449.

[61] I. Witten and D. Milne, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," in *Proc. of the AAAI WikiAI Workshop*. AAAI, Chicago, IL, 2008, pp. 25–30.

[62] O. Medelyan, I. Witten, and D. Milne, "Topic indexing with Wikipedia," in *Proc. of the AAAI WikiAI Workshop*. AAAI, Chicago, IL, 2008, pp. 19–24.

[63] "Electronic design automation - Wikipedia, the free encyclopedia," https://en.wikipedia.org/wiki/Electronic_design_automation, 2024, accessed: 2024-04-25.

[64] M. Daradkeh, L. Abualigah, S. Atalla, and W. Mansoor, "Scientometric analysis and classification of research using convolutional neural networks: A case study in data science and analytics," *Electronics*, vol. 11, no. 13, p. 2066, 2022.

[65] D. Gal, B. Thijs, W. Glänzel, and K. R. Sipido, "Hot topics and trends in cardiovascular research," *European Heart Journal*, vol. 40, no. 28, pp. 2363–2374, 2019.

[66] D. Zhao and A. Strotmann, "The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis," *Journal of the Association for Information Science and Technology*, vol. 65, no. 5, pp. 995–1006, 2014.

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

[68] B. Wang, S. Liu, K. Ding, Z. Liu, and J. Xu, "Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: A case study in lte technology," *Scientometrics*, vol. 101, pp. 685–704, 2014.

[69] D. Fang, H. Yang, B. Gao, and X. Li, "Discovering research topics from library electronic references using latent Dirichlet allocation," *Library Hi Tech*, vol. 36, no. 3, pp. 400–410, 2018.

[70] L. Lei, Y. Deng, and D. Liu, "Examining research topics with a dependency-based noun phrase extraction method: A case in accounting," *Library Hi Tech*, vol. 41, no. 2, pp. 570–582, 2023.

[71] A. Barua, S. W. Thomas, and A. E. Hassan, "What are developers talking about? An analysis of topics and trends in stack overflow," *Empirical Software Engineering*, vol. 19, pp. 619–654, 2014.

[72] J. Swacha, "Topic evolution in the research on educational gamification," *Education Sciences*, vol. 12, no. 10, p. 640, 2022.

[73] C.-D. Curiac and A. Doboli, "Combining informetrics and trend analysis to understand past and current directions in electronic design automation," *Scientometrics*, vol. 127, no. 10, pp. 5661–5689, 2022.

[74] D. Sharma, B. Kumar, and S. Chand, "A trend analysis of machine learning research with topic models and Mann-Kendall test," *International Journal of Intelligent Systems and Applications*, vol. 11, no. 2, pp. 70–82, 2019.

[75] M. Katsurai, "Bursty research topic detection from scholarly data using dynamic co-word networks: A preliminary investigation," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA).* IEEE, Beijing, China, 2017, pp. 115–119.

[76] S. Kumar, M. Marrone, Q. Liu, and N. Pandey, "Twenty years of the international journal of accounting information systems: A bibliometric analysis," *International Journal of Accounting Information Systems*, vol. 39, p. 100488, 2020.

[77] K. K. Mane and K. Börner, "Mapping topics and topic bursts in PNAS," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl_1, pp. 5287–5290, 2004.

[78] C. W. Cai, M. K. Linnenluecke, M. Marrone, and A. K. Singh, "Machine learning and expert judgement: Analyzing emerging topics in accounting and finance research in the Asia–Pacific," *Abacus*, vol. 55, no. 4, pp. 709–733, 2019.

[79] T. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. 1, pp. 5228–5235, 2004.

[80] M. Hoffman, D. Blei, C. Wang, and J. Paisley, “Stochastic variational inference,” *Journal of Machine Learning Research*, vol. 14, no. 5, pp. 1303–1347, 2013.

[81] G. Box, G. Jenkins, and G. Reinsel, *Time-series analysis: Forecasting and control*. Holden-Day San Francisco, USA, 1970.

[82] C. Chatfield, *Time-series forecasting*. CRC press, Boca Raton, FL, USA, 2000.

[83] P. Whittle, *Hypothesis testing in time-series analysis*. Almquist and Wiksell, Uppsala, Sweden, 1951.

[84] J. Cryer and K. Chan, *Time-series analysis with applications in R*. Springer Science & Business Media, New York, NY, 2008.

[85] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and practice*. OTexts, Melbourne, Australia, 2018.

[86] H. B. Mann, “Nonparametric tests against trend,” *Econometrica: Journal of the Econometric Society*, pp. 245–259, 1945.

[87] K. Kendall, “Thin-film peeling-the elastic term,” *Journal of Physics D: Applied Physics*, vol. 8, no. 13, p. 1449, 1975.

[88] D. P. Lettenmaier, “Multivariate nonparametric tests for trend in water quality 1,” *Journal of the American Water Resources Association*, vol. 24, no. 3, pp. 505–512, 1988.

[89] C. Libiseller and A. Grimvall, “Performance of partial Mann–Kendall tests for trend detection in the presence of covariates,” *Environmetrics: The Official Journal of the International Environmetrics Society*, vol. 13, no. 1, pp. 71–84, 2002.

[90] T. Pohlert, “Non-parametric trend tests and change-point detection,” https://CRAN.R-project.org/package=trend, 2023, accessed: 2024-04-25.

[91] A. Burauskaite-Harju, A. Grimvall, and C. v. Brömssen, “A test for network-wide trends in rainfall extremes,” *International Journal of Climatology*, vol. 32, no. 1, pp. 86–94, 2012.

[92] M. Hussain and I. Mahmud, “pymannkendall: a python package for non parametric Mann Kendall family of trend tests,” *Journal of Open Source Software*, vol. 4, no. 39, p. 1556, 2019.

[93] R. Vangara, M. Bhattarai, E. Skau, G. Chennupati, H. Djidjev, T. Tierney, J. P. Smith, V. G. Stanev, and B. S. Alexandrov, “Finding the number of latent topics with semantic non-negative matrix factorization,” *IEEE Access*, vol. 9, pp. 117 217–117 231, 2021.

[94] Ö. Bihrat and M. Bayazit, “The power of statistical tests for trend detection,” *Turkish Journal of Engineering and Environmental Sciences*, vol. 27, no. 4, pp. 247–251, 2003.

[95] F. Wang, W. Shao, H. Yu, G. Kan, X. He, D. Zhang, M. Ren, and G. Wang, “Re-evaluation of the power of the Mann-Kendall test for detecting monotonic trends in hydrometeorological time series,” *Frontiers in Earth Science*, vol. 8, p. 14, 2020.

[96] K. H. Hamed, “Trend detection in hydrologic data: the Mann–Kendall trend test under the scaling hypothesis,” *Journal of Hydrology*, vol. 349, no. 3-4, pp. 350–363, 2008.

[97] G. S. Marchini, K. V. Faria, F. L. Neto, F. C. M. Torricelli, A. Danilovic, F. C. Vicentini, C. A. Batagello, M. Srougi, W. C. Nahas, and E. Mazzucchi, “Understanding urologic scientific publication patterns and general public interests on stone disease: Lessons learned from big data platforms,” *World Journal of Urology*, vol. 39, pp. 2767–2773, 2021.

[98] C. Zou, “Analyzing research trends on drug safety using topic modeling,” *Expert Opinion on Drug Safety*, vol. 17, no. 6, pp. 629–636, 2018.

[99] X. Chen and H. Xie, “A structural topic modeling-based bibliometric study of sentiment analysis literature,” *Cognitive Computation*, vol. 12, pp. 1097–1129, 2020.

[100] X. Chen, H. Xie, G. Cheng, and Z. Li, “A decade of sentic computing: topic modeling and bibliometric analysis,” *Cognitive Computation*, vol. 14, no. 1, pp. 24–47, 2022.

[101] R. M. Hirsch and J. R. Slack, “A nonparametric trend test for seasonal data with serial dependence,” *Water Resources Research*, vol. 20, no. 6, pp. 727–732, 1984.

[102] A. L. O. King, F. N. Mirza, H. N. Mirza, N. Yumeen, V. Lee, and S. Yumeen, “Factors associated with the american academy of dermatology abstract publication: A multivariate analysis,” *Journal of the American Academy of Dermatology*, vol. 86, no. 6, pp. 1416–1419, 2022.

[103] R. M. Andrew, “Towards near real-time, monthly fossil CO2 emissions estimates for the European Union with current-year projections,” *Atmospheric Pollution Research*, vol. 12, no. 12, p. 101229, 2021.

[104] S. Yue and C. Wang, “The Mann-Kendall test modified by effective sample size to detect trend in serially correlated hydrological series,” *Springer Water Resources Management*, vol. 18, no. 3, pp. 201–218, 2004.

[105] K. H. Hamed and A. R. Rao, “A modified Mann-Kendall trend test for autocorrelated data,” *Journal of Hydrology*, vol. 204, no. 1-4, pp. 182–196, 1998.

[106] S. Seabold and J. Perktold, “Statsmodels: Econometric and statistical modeling with python,” in *Proc. of the 9th Python in Science Conference*, vol. 57, no. 61. Austin, TX, USA, 2010, pp. 10–25 080.

[107] S. Bird, E. Klein, and E. Loper, *Natural language processing with python: Analyzing text with the natural language toolkit.* O’Reilly, Sebastopol, CA, USA, 2009.

[108] N. Hardeniya, J. Perkins, D. Chopra, N. Joshi, and I. Mathur, *Natural language processing: python and NLTK.* Packt Publishing Ltd., Birmingham, UK, 2016.

[109] Q. He, “Knowledge discovery through co-word analysis,” *Library Trends*, vol. 48, no. 1, pp. 133–159, 1999.

[110] J. E. Grable, “Financial risk tolerance and additional factors that affect risk taking in everyday money matters,” *Journal of Business and Psychology*, vol. 14, no. 4, pp. 625–630, 2000.

[111] H. Nakamura, S. Ii, H. Chida, K. Friedl, S. Suzuki, J. Mori, and Y. Kajikawa, “Shedding light on a neglected area: A new approach to knowledge creation,” *Sustainability Science*, vol. 9, no. 2, pp. 193–204, 2014.

[112] T. Ogawa and Y. Kajikawa, “Generating novel research ideas using computational intelligence: A case study involving fuel cells and ammonia synthesis,” *Technological Forecasting and Social Change*, vol. 120, pp. 41–47, 2017.

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

[114] H. B. Kang, S. Mysore, K. Huang, H.-S. Chang, T. Prein, A. McCallum, A. Kittur, and E. Olivetti, “Augmenting scientific creativity with retrieval across knowledge domains,” *arXiv preprint arXiv:2206.01328*, 2022.

[115] D. Rotolo, D. Hicks, and B. R. Martin, “What is an emerging technology?” *Research Policy*, vol. 44, no. 10, pp. 1827–1843, 2015.

[116] H. Zhang, X. Kong, and Y. Zhang, “Cross-domain collaborative recommendation without overlapping entities based on domain adaptation,” *Multimedia Systems*, vol. 28, no. 5, pp. 1621–1637, 2022.

[117] H. Zhang, X. Kong, and Y. Zhang, “Cross-domain recommendation with multi-auxiliary domains via consistent and selective cluster-level knowledge transfer,” *Expert Systems with Applications*, vol. 223, p. 119861, 2023.

[118] D. Sailaja, M. Kishore, B. Jyothi, and N. Prasad, “An overview of pre-processing text clustering methods,” *International Journal of Computer Science and Information Technologies*, vol. 6, no. 3, pp. 3119–24, 2015.

[119] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[120] G. Salton, *Automatic text processing: The transformation, analysis, and retrieval of.* Addison-Wesley, Boston, USA, 1989, vol. 169.

[121] E. Babkin, N. Karpov, and O. Kozyrev, “Towards creating an evolvable semantic platform for formation of research teams,” in *Perspectives in Business Informatics Research: 12th International Conference*. Springer, Warsaw, Poland, 2013, pp. 200–213.

[122] S. Milojević, “Principles of scientific research team formation and evolution,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 11, pp. 3984–3989, 2014.

[123] G. D’Aniello, M. Gaeta, M. Lepore, and M. Perone, “Knowledge-driven fuzzy consensus model for team formation,” *Expert Systems with Applications*, vol. 184, p. 115522, 2021.

[124] S. Cavalcante, B. Gadelha, E. C. C. de Oliveira, and T. Conte, “How to better form software development teams? An analysis of different formation criteria,” in *Proc. of the 22nd International Conference on Enterprise Information Systems*. Scitepress, 2020, pp. 90–100.

[125] D. H. Prasad and M. H. Vasanth, “Value of bottom-up team formation for complex adaptive business systems,” in *2014 IEEE International Systems Conference Proc.* IEEE, Ottawa, ON, Canada, 2014, pp. 272–276.

[126] P. Zainal, D. Razali, and Z. Mansor, “Team formation for agile software development: a review,” *International Journal on Advanced Science, Engineering and Information Technology*, vol. 10, no. 2, pp. 555–561, 2020.

[127] Y. Mahajan and J.-H. Cho, “Prada-tf: Privacy-diversity-aware online team formation,” in *2021 IEEE International Conference on Web Services (ICWS).* IEEE, Chicago, IL, USA, 2021, pp. 493–499.

[128] L. S. Marcolino, A. X. Jiang, and M. Tambe, “Multi-agent team formation: Diversity beats strength?” in *Twenty-Third International Joint Conference on Artificial Intelligence*. Beijing, China, 2013.

[129] R. G. Askin and Y. Huang, “Forming effective worker teams for cellular manufacturing,” *International Journal of Production Research*, vol. 39, no. 11, pp. 2431–2451, 2001.

[130] M. Tavana, F. Azizi, F. Azizi, and M. Behzadian, “A fuzzy inference system with application to player selection and team formation in multi-player sports,” *Sport Management Review*, vol. 16, no. 1, pp. 97–110, 2013.

[131] E. Ozceylan, “A mathematical model using ahp priorities for soccer player selection: a case study,” *South African Journal of Industrial Engineering*, vol. 27, no. 2, pp. 190–205, 2016.

[132] E. Andrejczuk, R. Berger, J. A. Rodriguez-Aguilar, C. Sierra, and V. Marín-Puchades, “The composition and formation of effective teams: computer science meets organizational psychology,” *The Knowledge Engineering Review*, vol. 33, p. e17, 2018.

[133] E. Mourelatos, N. Giannakopoulos, and M. Tzagarakis, “Personality traits and performance in online labour markets,” *Behaviour & Information Technology*, vol. 41, no. 3, pp. 468–484, 2022.

[134] A. de Korvin, M. F. Shipley, and R. Kleyle, “Utilizing fuzzy compatibility of skill sets for team selection in multi-phase projects,” *Journal of Engineering and Technology Management*, vol. 19, no. 3-4, pp. 307–319, 2002.

[135] O. Hlaoittinun, E. Bonjour, and M. Dulmet, “A team building approach for competency development,” in *2007 IEEE International Conference on Industrial Engineering and Engineering Management*. IEEE, Singapore, 2007, pp. 1004–1008.

[136] A. Shah, R. Ganesan, S. Jajodia, H. Cam, and S. Hutchinson, “A novel team formation framework based on performance in a cybersecurity operations center,” *IEEE Transactions on Services Computing*, vol. 16, no. 4, pp. 2359–2371, 2023.

[137] E. L. Fitzpatrick and R. G. Askin, “Forming effective worker teams with multi-functional skill requirements,” *Computers & Industrial Engineering*, vol. 48, no. 3, pp. 593–608, 2005.

[138] S.-J. Chen and L. Lin, “Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering,” *IEEE Transactions on Engineering Management*, vol. 51, no. 2, pp. 111–124, 2004.

[139] A. Gajewar and A. Das Sarma, “Multi-skill collaborative teams based on densest subgraphs,” in *Proc. of the 2012 SIAM International Conference on Data Mining*. SIAM, Anaheim, CA, USA, 2012, pp. 165–176.

[140] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi, "Online team formation in social networks," in *Proc. of the 21st International Conference on World Wide Web*. ACM, New York, USA, 2012, pp. 839–848.

[141] Y. Han, Y. Wan, L. Chen, G. Xu, and J. Wu, "Exploiting geographical location for team formation in social coding sites," in *Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference*. Springer, Jeju, South Korea, 2017, pp. 499–510.

[142] L. Chen, Y. Ye, A. Zheng, F. Xie, Z. Zheng, and M. R. Lyu, "Incorporating geographical location for team formation in social coding sites," *World Wide Web*, vol. 23, pp. 153–174, 2020.

[143] G. K. Awal and K. K. Bharadwaj, "Team formation in social networks based on collective intelligence–an evolutionary approach," *Applied Intelligence*, vol. 41, pp. 627–648, 2014.

[144] A. Yadav, S. Mishra, and A. S. Sairam, "A multi-objective worker selection scheme in crowdsourced platforms using NSGA-II," *Expert Systems with Applications*, vol. 201, p. 116991, 2022.

[145] K. Selvarajah, P. M. Zadeh, Z. Kobti, Y. Palanichamy, and M. Kargar, "A unified framework for effective team formation in social networks," *Expert Systems with Applications*, vol. 177, p. 114886, 2021.

[146] J. Jiang, K. Di, B. An, Y. Jiang, Z. Bu, and J. Cao, "Batch crowdsourcing for complex tasks based on distributed team formation in e-markets," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 12, pp. 3600–3615, 2022.

[147] E. Salas, T. L. Dickinson, S. A. Converse, and S. I. Tannenbaum, "Toward an understanding of team performance and training," in *Teams: Their training and performance*. Ablex Publishing, 1992, pp. 3–29.

[148] J. Cannon, S. Tannenbaum, and E. Salas, "Defining team competencies and establishing team training requirements," in *Team Effectiveness and Decision Making in Organizations*. Jossey-Bass, San Francisco, CA, USA, 1995.

[149] E. Salas, D. E. Sims, and C. S. Burke, "Is there a "big five" in teamwork?" *Small Group Research*, vol. 36, no. 5, pp. 555–599, 2005.

[150] P. Ballesteros-Perez, M. C. González-Cruz, and M. Fernández-Diego, "Human resource allocation management in multiple projects using sociometric techniques," *International Journal of Project Management*, vol. 30, no. 8, pp. 901–913, 2012.

[151] M. Colenso, "How to accelerate team development and enhance team productivity," in *Kaizen Strategies for Improving Team Performance*. Prentice-Hall, London, UK, 2000.

[152] J. E. Galvin, V. R. McKinney, and K. M. Chudoba, "From me to we: The role of the psychological contract in team formation," in *Proc. of the 38th Annual Hawaii International Conference on System Sciences*. IEEE, Big Island, HI, USA, 2005, pp. 49c–49c.

[153] J. S. Hornsby, "Critical elements of team formation to enhance organizational innovation," in *The challenges of corporate entrepreneurship in the disruptive age*. Emerald Publishing Limited, 2018, vol. 28, pp. 123–140.

[154] S. S. Rangapuram, T. Bühler, and M. Hein, "Towards realistic team formation in social networks based on densest subgraphs," in *Proc. of the 22nd International Conference on World Wide Web*. ACM, New York, USA, 2013, pp. 1077–1088.

[155] M. Fathian, M. Saei-Shahi, and A. Makui, "A new optimization model for reliable team formation problem considering experts' collaboration network," *IEEE Transactions on Engineering Management*, vol. 64, no. 4, pp. 586–593, 2017.

[156] W. V. Cunningham and P. Villaseñor, "Employer voices, employer demands, and implications for public skills development policy connecting the labor and education sectors," *The World Bank Research Observer*, vol. 31, no. 1, pp. 102–134, 2016.

[157] C. R. Paris, E. Salas, and J. A. Cannon-Bowers, "Teamwork in multi-person systems: a review and analysis," *Ergonomics*, vol. 43, no. 8, pp. 1052–1075, 2000.

[158] M. Marrone, S. Lemke, and L. M. Kolbe, "Entity linking systems for literature reviews," *Scientometrics*, vol. 127, no. 7, pp. 3857–3878, 2022.

[159] L. C. Hon and B. Brunner, "Diversity issues and public relations," *Journal of Public Relations Research*, vol. 12, no. 4, pp. 309–340, 2000.

[160] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of computer computations*. Springer, Boston, MA, 1972, pp. 85–103.

[161] D. Knop, M. Kouteckỳ, and M. Mnich, "Combinatorial n-fold integer programming and applications," *Mathematical Programming*, vol. 184, no. 1-2, pp. 1–34, 2020.

[162] J. Arora, *Multi-objective optimum design concepts and methods. Introduction to optimum design*. Academic Press: Cambridge, MA, USA, 2017.

[163] K. Miettinen, *Nonlinear multiobjective optimization*. Springer Science & Business Media, Boston, USA, 1999, vol. 12.

[164] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[165] J. Blank and K. Deb, "Pymoo: Multi-objective optimization in python," *IEEE Access*, vol. 8, pp. 89 497–89 509, 2020.

[166] H. Wi, S. Oh, J. Mun, and M. Jung, "A team formation model based on knowledge and collaboration," *Expert Systems with Applications*, vol. 36, no. 5, pp. 9121–9134, 2009.

[167] H. Rahman, S. Thirumuruganathan, S. B. Roy, S. Amer-Yahia, and G. Das, "Worker skill estimation in team-based tasks," *Proceedings of the VLDB Endowment*, vol. 8, no. 11, pp. 1142–1153, 2015.

[168] L. Li, H. Tong, Y. Wang, C. Shi, N. Cao, and N. Buchler, "Is the whole greater than the sum of its parts?" in *Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, USA, 2017, pp. 295–304.

[169] N. Amin, K. U. Khan, B. Dolgorsuren, and Y.-K. Lee, "Extracting top-k interesting subgraphs with weighted query semantics," in *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, Jeju, South Korea, 2017, pp. 366–373.

[170] R. Hamidi Rad, H. Fani, E. Bagheri, M. Kargar, D. Srivastava, and J. Szlichta, "A variational neural architecture for skill-based team formation," *ACM Transactions on Information Systems*, vol. 42, no. 1, pp. 1–28, 2023.

[171] M.-C. Juang, C.-C. Huang, and J.-L. Huang, "Efficient algorithms for team formation with a leader in social networks," *The Journal of Supercomputing*, vol. 66, pp. 721–737, 2013.

[172] M. Kargar and A. An, "Discovering top-k teams of experts with/without a leader in social networks," in *Proc. of the 20th ACM International Conference on Information and Knowledge Management*. ACM, New York, USA, 2011, pp. 985–994.

[173] N. Berktaş and H. Yaman, "A branch-and-bound algorithm for team formation on social networks," *INFORMS Journal on Computing*, vol. 33, no. 3, pp. 1162–1176, 2021.

[174] M. Niveditha, G. Swetha, U. Poornima, and R. Senthilkumar, “A genetic approach for tri-objective optimization in team formation,” in *Proc. of the 2016 Eighth International Conference on Advanced Computing (ICoAC)*. IEEE, Chennai, India, 2017, pp. 123–130.

[175] Y. Mahajan, Z. Guo, J.-H. Cho, and I.-R. Chen, “Privacy-preserving and diversity-aware trust-based team formation in online social networks,” http://hdl.handle.net/10919/113973, 2023, accessed: 2024-04-25.

[176] M. Neshati, H. Beigy, and D. Hiemstra, “Expert group formation using facility location analysis,” *Information Processing & Management*, vol. 50, no. 2, pp. 361–383, 2014.

[177] R. Ponnusamy, W. A. Degife, and T. Alemu, “Recommender frameworks outline system design and strategies: A review,” *Knowledge Computing and its Applications*, pp. 261–285, 2018.

[178] C. Hayes, P. Massa, P. Avesani, and P. Cunningham, “An on-line evaluation framework for recommender systems,” in *AH’2002 Workshop on Recommendation and Personalization in E-Commerce*. Malaga, Spain, Jun. 2002, pp. 50–59.