IOSUD – Politehnica University Timișoara
Doctoral School for Engineering Studies

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

PhD thesis - Summary

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 Supervisor: Prof.Dr.Habil.Eng. Mihai Victor Micea

September, 2024

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 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 human-in-the-loop recommender 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.*

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.

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.

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 above, 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].

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, 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, ”Multi-core 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.

A brief summary of the key findings from each chapter follows.

Chapter 2 provides a critical review of related literature in the field of 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.

From this 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. 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 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 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. The chapter encapsulates the multi-recommender framework presented in [11].

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.

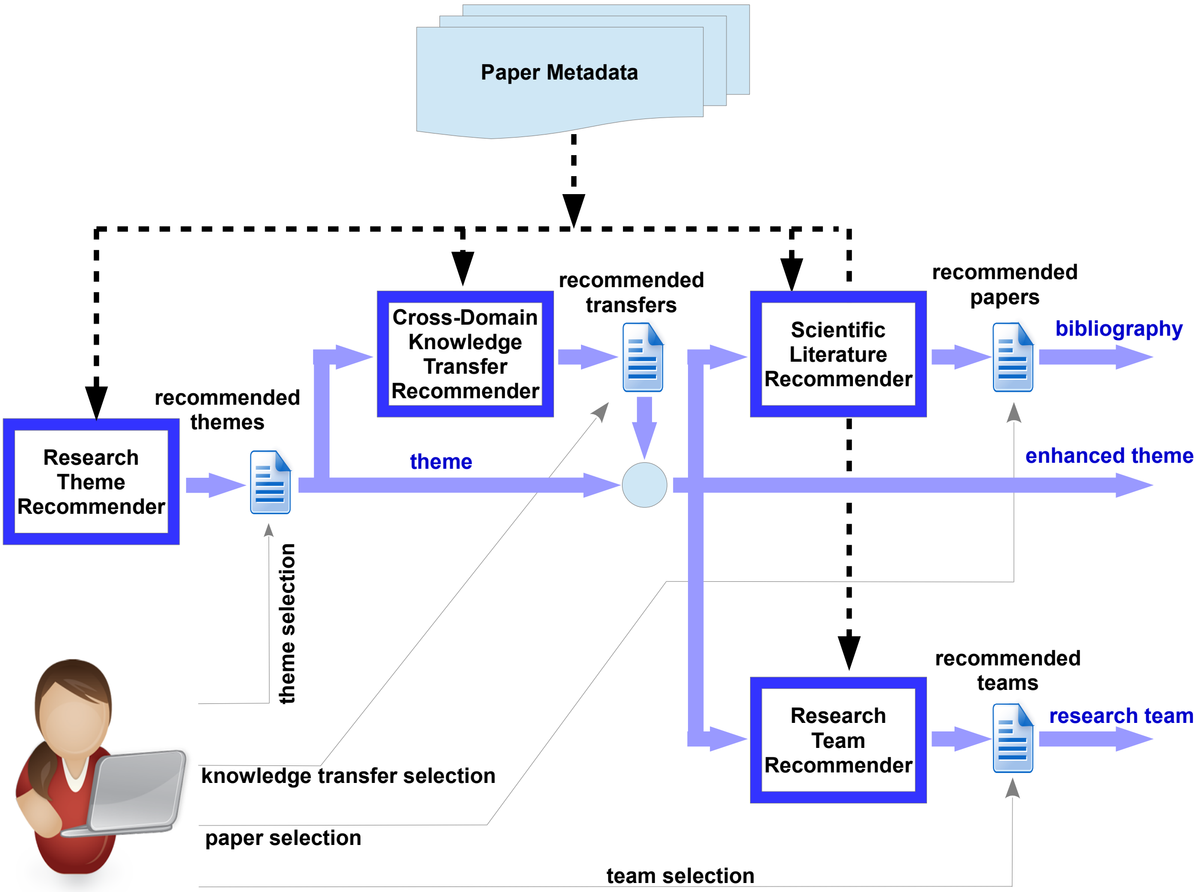

Figure 1: Multi-recommender system architecture

Our proposed multi-recommender system framework is made up of four recommender modules that may act either interconnected as in Figure 1 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.

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. Part of the results of this chapter was reported in [17] and offered as a public dataset on Mendeley Data [16].

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. The chapter encapsulates the methods and results presented in our papers [12, 13, 14, 15].

Following this strategy, we designed and built a human-in-the-loop recommender system having the architecture presented in Figure 2, including the needed human interventions during execution.

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).

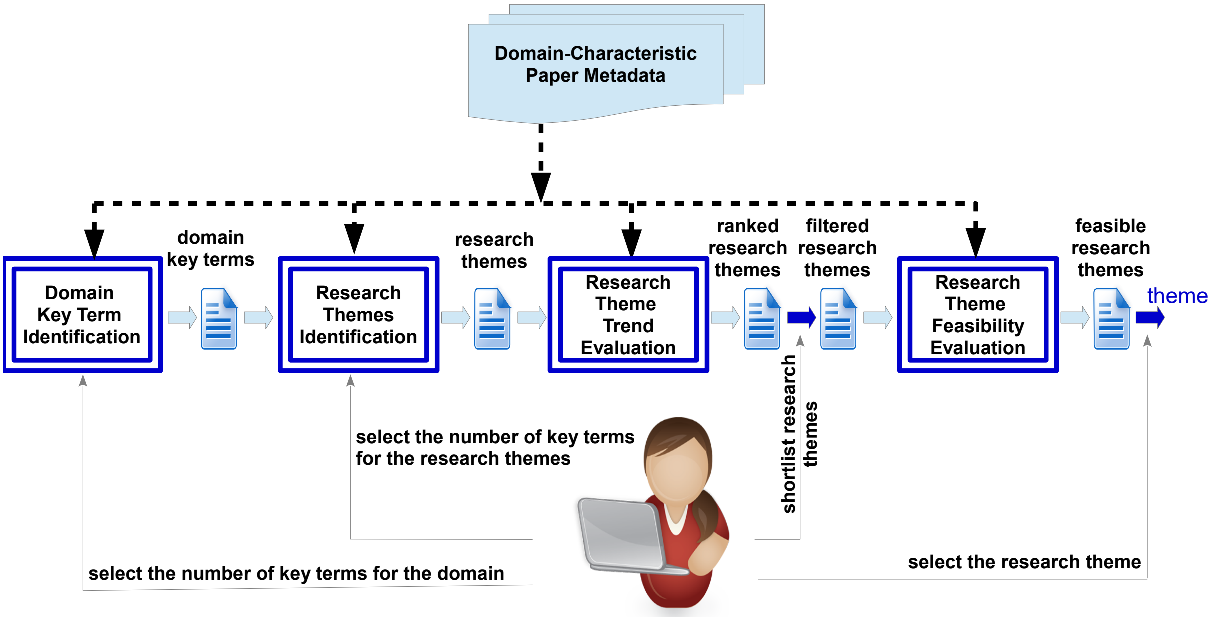
Domain-Characteristic Paper Metadata
Domain Key Term Identification
domain key terms
Research Themes Identification
research themes
Research Theme Trend Evaluation
ranked research themes
filtered research themes
Research Theme Feasibility Evaluation
feasible research themes
theme
shortlist research themes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

Figure 2: Research themes recommender system architecture

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.

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

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 3 and contains four functional steps.

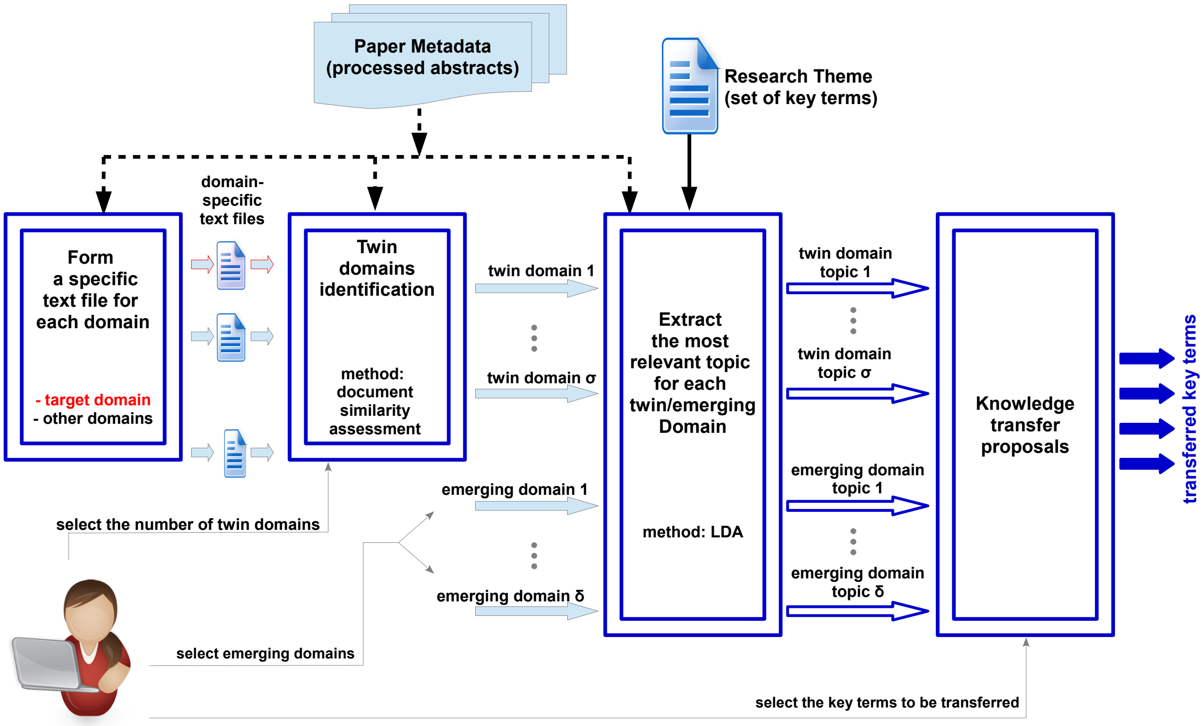

Figure 3: Cross-domain knowledge transfer recommender system architecture

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.

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 [19]. This method includes two stages:

a) *tf-idf* vectorization of the document corpus [20] 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.

b) evaluate similarities between pairs of target and source domain-characteristic text files using the cosine similarity metric [21].

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. 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.

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.

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 4. 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 Researcher's General Expertise (RGE), Researcher's Collaboration Ability (RCA), and Interpersonal Collaborations Inside Specified Groups (ICISG) indices are not project-dependent. To obtain the Researcher's Level of Expertise in a Given Area (RLEGA) indicator, we have to focus only on the set of key terms that precisely characterize the 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 the Researcher's Areas of Expertise (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 team 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,

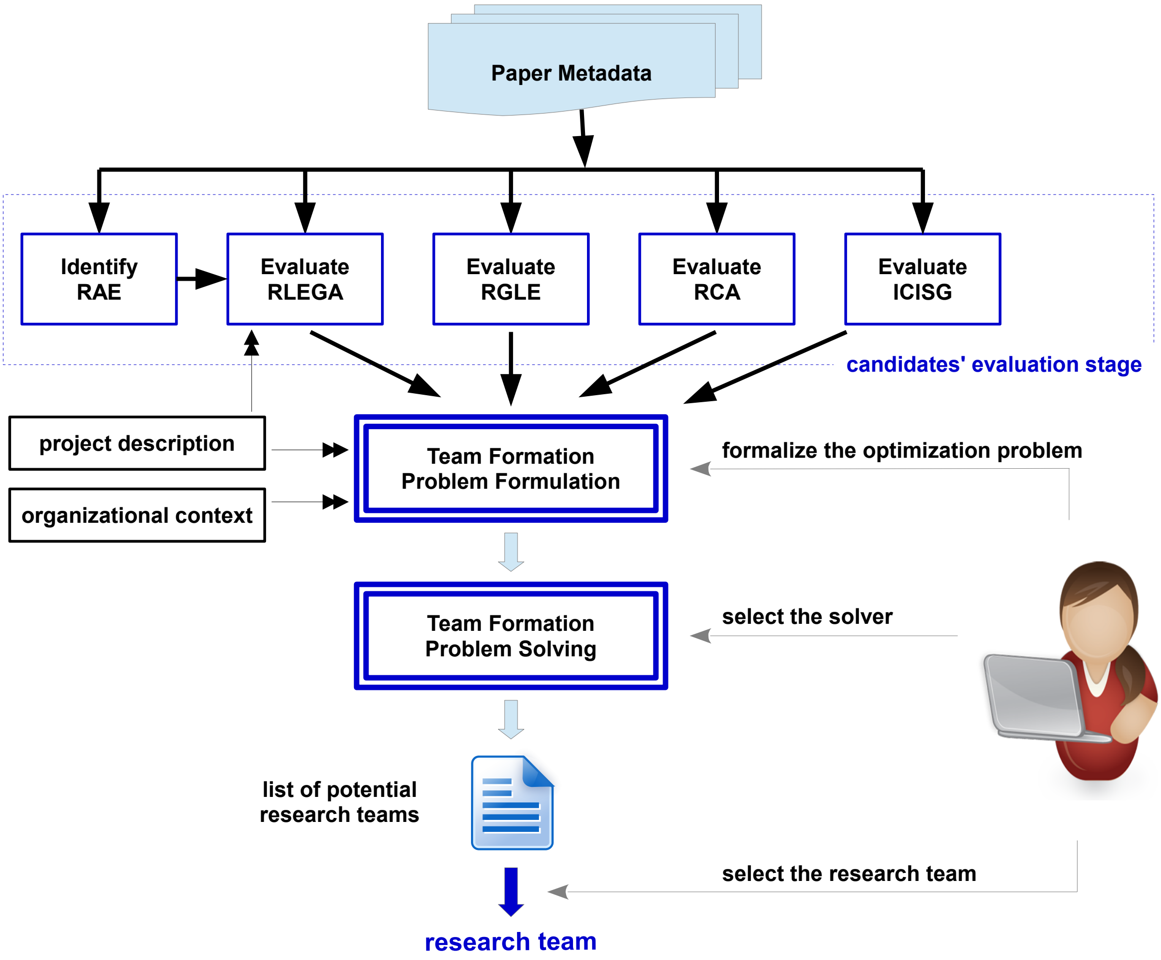

Figure 4: Bibliographic data-driven research team recommender framework [18]

modifying the problem formulation by reshaping the objective functions or the constraints, or choosing another solver).

As it may be noticed from Figure 4, 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.

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

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.

References

[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, *[accepted for publication].*

[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] 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.

[20] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[21] G. Salton, *Automatic text processing: The transformation, analysis, and retrieval of*. Addison-Wesley, Boston, USA, 1989, vol. 169.