Towards an Open Science — an Academic Recommendation Cloud Platform

Anna Kobusińska[1][0000−0002−3501−2840], Damian Tabaczyński[1], and Victor Chang[2][0000−0002−8012−5852]

[1] Faculty of Computing and Telecommunications, Poznan University of Technology, Poznań, Poland
{Anna.Kobusinska,Damian.Tabaczynski}@cs.put.poznan.pl
[2] Aston Business School, Aston University, Birmingham, UK
v.chang1@aston.ac.uk

Abstract. This paper provides a cloud-based academic recommender system that integrates multiple heterogeneous big data academic resources to deliver personalized recommendations within the academic networks. The proposed system introduces a hybrid recommendation mechanism combining content-based and collaborative filtering within a graph-based relationship model.

Keywords: scholarly datasets · hybrid recommendation mechanism · cloud computing · serverless

1 Introduction

Nowadays, the number of new publications is reaching millions per year worldwide [4]. The available, overwhelming amount of academic data makes it difficult for researchers to find relevant literature effectively. Existing traditional search engines and databases often provide keyword-based searches without personalization and context awareness, thus not meeting the requirements of researchers. This challenge requires the development of intelligent recommendation systems that utilize big data, cloud computing and advanced filtering techniques. Existing solutions such as Google Scholar, Microsoft Academic Graph and OpenAIRE Research Graph provide partial solutions but do not offer a fully integrated, personalized recommender system. For example, Google Scholar relies mainly on ranking-based citations but does not consider user preferences. The Microsoft Academic Graph provides extensive bibliometric data, but lacks advanced recommendation mechanisms. In turn, the OpenAIRE Research Graph integrates multiple repositories but cannot offer personalized recommendations. Therefore, this paper addresses these gaps by proposing a scalable, adaptive, and user-oriented academic recommendation platform. The solution proposes an approach integrating heterogeneous academic databases to improve recommendation coverage and completeness. In addition, it introduces a recommender algorithm that enables personalized recommendations based on the user's search history while working with the system and the user's scientific relationships. To

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_34

this end, it leverages content-based and collaborative filtering within a graph-based relationship model. Finally, the proposed system optimizes performance by leveraging cloud-based environments, ensuring scalability and responsiveness. The proposed solution allows for context-aware and highly relevant suggestions, significantly enhancing the research discovery process.

The structure of the paper is as follows: Section 2 provides a brief overview of the related works. Section 3 presents a general idea of the proposed solution. The relationship graph model and the recommendation mechanism are proposed in Section 4 and Section 5, respectively. Finally, the conclusions are discussed in Section 6.

2 Related Work

Academic recommendation systems have been widely explored to improve research discovery. Existing solutions can be divided into content-based filtering, collaborative filtering, and hybrid approaches. Among the available platforms and tools to facilitate the discovery of academic literature is Google Scholar, which uses a ranking based on citations but lacks personalized recommendations. Microsoft Academic Graph provides extensive bibliometric metadata, but does not offer real-time recommendations. Semantic Scholar, on the other hand, implements recommendations based on artificial intelligence but relies on NLP techniques without a strong graph-based model. Connected Papers, on the other hand, generates graph-based visualizations but requires manual input of article identifiers. Finally, ArnetMiner uses social network analysis to classify authors but has limited scalability for real-time recommendations. Each system has limitations regarding scalability, integration of heterogeneous data sources, and personalization mechanisms. Our approach addresses these gaps by incorporating graph-based modelling with hybrid recommendation techniques to provide a more comprehensive academic discovery experience.

Among studies dedicated to academic recommender systems, [2] can be mentioned, where the authors investigated citation recommendations based on deep learning in scientific networks. In turn, [1] presents a comprehensive study of article recommendation techniques, including collaborative filtering and content-based methods. In [14] digital libraries were analyzed, demonstrating increased search efficiency. A knowledge graph-based academic recommender system integrating citation networks and semantic analysis has been proposed in [11]. Article [15] proposes a contextual recommender framework for digital research repositories. Much of this research focuses on improving specific aspects of recommender systems, such as natural language processing (NLP) for article recommendations, modelling user behavior, or citation network analysis. However, few works address integrating multiple data sources, personalization, and scalable cloud deployment within a single framework. The proposed solution builds on these elements, offering an improved graph that optimizes the performance and adaptability of discoveries and a cloud-native recommendation engine.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_34

3 The general idea of the proposed solution

This paper proposes a recommendation platform, called *Academic Advisor*, designed to assist researchers by suggesting relevant publications, datasets, organizations, and scientific collaborations. The system leverages a novel recommendation algorithm that establishes relationships among academic entities, enhancing contextual discovery. The solution is user-oriented and designed to require minimal preliminary information input. The platform generates recommendations by comparing the retrieved data with its knowledge base. This eliminates users' need to input detailed preferences manually, streamlining the recommendation process.

To enhance personalization, users can rate recommendations, influencing future results. This feedback loop allows dynamic adaptation of the algorithm, refining the accuracy and relevance of recommendations. Additionally, an Academic Advisor can suggest publications, authors, institutions, and projects, recognizing that academic collaboration extends beyond literature.

The system architecture comprises three main components: clients (who interact with the system through a web application), application servers (which handle clients' requests, retrieve data, and process recommendations), and data source modules (which integrate data from multiple scholarly data sources). The high-level abstraction diagram of the solution with marked cloud parts is shown in Fig. 1. To generate recommendations, the client sends a request to the appli-

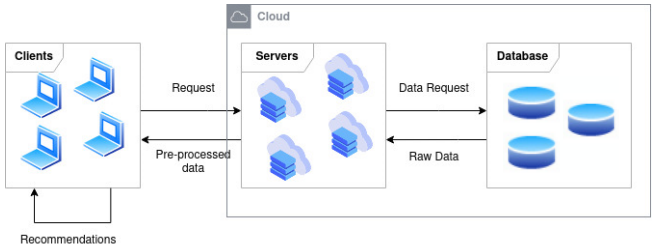

Fig. 1: Solution scheme

cation server. Before responding to the client, the server retrieves the necessary data from distributed academic sources by sending a data request to these repositories. To avoid sending massive amounts of data, the data is filtered explicitly on the data source side. The returned data set contains raw, filtered data and has to undergo pre-processing, ensuring that only the most relevant and filtered information is used, thus minimizing the data transmission load. After receiving the data, the client executes the recommendation algorithm in the Edge Worker, a component specifically designed for local computing. Edge Worker, which is part of the client's application, refines and classifies the recommendations before presenting the results via a graphical user interface. This approach optimizes performance by reducing backend server load and shortening response times.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_34

To ensure scalability, the system is deployed in a serverless cloud environment, eliminating the need for manual infrastructure management while providing automated scaling. The use of cloud computing also enhances system availability, security, and performance. Data is stored in replicated cloud databases, ensuring fault tolerance and fast access. Additionally, internal cloud networking reduces latency and improves security by avoiding exposure to external networks. This hybrid cloud-edge approach optimizes computational efficiency while maintaining a seamless user experience.

4 Graph-Based Relationship Model

The Academic Advisor uses a relationship graph to model academic units and their interrelationships. This model represents **Users** (authors, their profiles and interests), **Organizations** (authors' affiliations, e.g. universities or research institutes), **Projects** (understood as funded research initiatives and their results) and **Resources** (publications, data sets and software) as nodes. The following relationships between nodes, represented by edges, are considered: **Authorship** models a bidirectional relation between User and Resource node. This indicates that the given user is the author of the given resource. **Membership** is a bidirectional relation between the Organization and User node, and identifies a specific user as a member of a particular organization. **Origin** represents a bidirectional relation describing the origin of a particular resource. It models a connection between the Organization and Resource node. **Outcome** is a bidirectional Project–Resource relation that connects this specific project to its outcome nodes. Finally, **Like/Dislike** unidirectional relation between a specific user and any other node (even other users) specifies if users like or dislike this node as a recommendation. In the recommendation algorithm, this is the only relationship that does not influence connections.

The combination of three well-known representative academic networks: OpenAire Graph [10], Microsoft Academic Graph [9], and Google Scholar [5] is considered in the proposed solution as a data source. The OpenAIRE Research Graph is an open resource system that aggregates a collection of research data properties (metadata, links) available within the OpenAIRE Open Science infrastructure [8, 10, 13]. It contains information about organizations, funders, funding streams, projects, communities, data sources, and scientific products, including literature, datasets, software, etc. The Microsoft Academic Graph (MAG) is a heterogeneous graph containing scientific publications, citation relationships between publications, and information regarding authors, institutions, journals, conferences, and fields of study, among others [3, 12]. The MAG is marked by its completeness and stability. The third considered solution, Google Scholar network [6, 7], was chosen due to the system's and its users' constant updates. It is used as an additional set to extend the knowledge provided by the MAG and the OpenAIRE environments.

An example of a graph with a high degree of abstraction is shown in Figure 2. Such a graph structure enables effective browsing and filtering of information based on user preferences and relevance metrics derived from the system.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_34

5 Hybrid Recommendation Mechanism

The decision–making process for selecting the best recommendations is carried out by a dedicated edge worker on the client's behalf. As the specific set of data on which the worker operates is determined in advance, the worker is not extensively involved in the pre-processing process. The application server is responsible for selecting and preprocessing the data before transferring them to the final algorithm. In contrast, the data sources are responsible for data filtering. Ultimately, the client side (edge worker) implements the recommendation mechanism and visually presents the algorithm results.

The recommendation engine integrates content-based and collaborative filtering to provide precise and dynamic suggestions. The proposed algorithm is divided into two phases:

1. **Phase 1 Node Usability Function**: each entity in the graph is assigned a usability score that evaluates its relevance based on intrinsic attributes and network influence. The function value is calculated differently, based on the node type:

$$f(u) = \boldsymbol{\alpha * V_n} + \boldsymbol{\beta * V_r} + \boldsymbol{\gamma * V_l} \tag{1}$$

where $\boldsymbol{u}$ is the node for which calculations are performed, $\boldsymbol{\alpha, \beta, \gamma} \in [0, 1]$ are the factors of specific part of formula. There are three main components to this formula:

V_n node value — represents node-specific attributes (e.g., citation count for papers, impact score for authors), and indicates the evaluation of the node itself. The value $\boldsymbol{V_n} \in [0, 1]$ is strongly related to the node information, thus it tends to remain relatively constant over time. A node's value is calculated in two ways. First, if a node does not contain citation information, then:

$$\boldsymbol{V_n} = (base_x + C)/2 \tag{2}$$

where $base_x \in [0, 1]$ is the constant base value for a specific type of node, while $C \in [0, 1]$ is constant compensation for lack of information about citation number. If a node contains citation information, its value is calculated as follows:

- for User:

$$\boldsymbol{V_n} = (base_u + min(cit_u/T_u, 1))/2 \tag{3}$$

  where cit_u denominates number of citations for user and T_u user citations threshold.
- for Organization:

$$\boldsymbol{V_n} = (base_o + min(Avg_{cit}/T_o, 1))/2 \tag{4}$$

  where Avg_{cit} stands for an average number of citations of 10% the most cited users in this organization, and T_o is the organization citations threshold.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_34

– for Project:

$$\boldsymbol{V_n} = 1, \tag{5}$$

A project value is always equal to one because, from a logical standpoint, this type of node aggregates other results. As a result, it should be prioritized and the value should be the maximum, so 1.

– for Resource:

$$\boldsymbol{V_n} = (base_r + min(cit_r/T_r, 1))/2 \tag{6}$$

where cit_r denominates number of citations for resource and T_r resource citations threshold.

V_r reputation value — captures reputation from the academic community. The value $\boldsymbol{V_r} \in [-1, 1]$ represents the average values from Like relationships that end at this vertex, and varies greatly over time. A strong correlation exists between this indicator and the actions taken by the users of the entire system. The formula is presented as follows:

$$\boldsymbol{V_r} = (Avg_{likes}/Max_{like}) * min(likes/T_{likes}, 1) \tag{7}$$

where Avg_{likes} is an average value of ratings (Like relationships) from all users from the system, Max_{like} is the maximum value of the rating, $likes$ is the number of Like/Dislike relationships and T_{likes} likes count threshold.

V_l like value — integrates personalized user preferences (e.g., prior likes/dislikes). The value $\boldsymbol{V_l} \in [-1, 1]$ refers to the user's evaluation of a particular node, and varies considerably over time. Since only one "Like" relation may exist for a given user to a given node, it is a single number value originating from the Like relation. The formula is as follows:

$$\boldsymbol{V_l} = v_{like}/Max_{like} \tag{8}$$

where v_{like} is a value of the Like relationship and Max_{like} maximum value of rating.
Ultimately, the usability function for a node is the following:

$$f(u) = \boldsymbol{\alpha} * \begin{cases} \boldsymbol{(base_x + C)/2,} \text{ if citations undefined} \\ \boldsymbol{(base_u + min(cit_u/T_u, 1))/2,} \text{ if node=User} \\ \boldsymbol{(base_o + min(Avg_{cit}/T_o, 1))/2),} \text{ if node=Organization+1 or node=Project} \\ \boldsymbol{(base_r + min(cit_r/T_r, 1))/2,} \text{ if node=Resource} \end{cases} + \tag{9}$$

$$+\boldsymbol{\beta} * \boldsymbol{(Avg_{likes}/Max_{like})} * \boldsymbol{min(likes/T_{likes}, 1)}$$

$$+\boldsymbol{\gamma} * \boldsymbol{(v_{like}/Max_{like})}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_34

2. **Phase 2 Longest Path Sorting**: the topological sorting algorithm was applied to the directed acyclic graph (DAG) to determine the most influential academic units. The priority of the recommendations is based on the weighted longest path, ensuring the effective disclosure of high-impact research. DAG is a subgraph of the relation graph contained in the data source. Selecting the longest route can begin when all vertices have been sorted according to their topological position and values calculated. For each vertex, the longest path is calculated relative to the starting vertex, which is the user for whom the recommendations are made. Since edges in the graph do not have their values, but nodes do, the edges are assigned the values of the nodes. More precisely, the value of the edge is equal to the vertex value it leads to. *Calculation route* is a path calculated by summing the

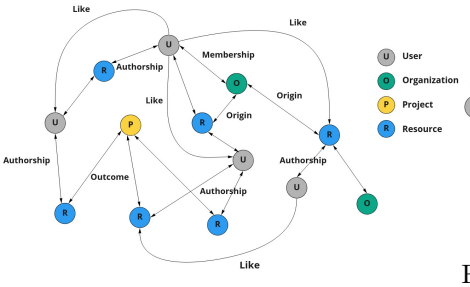

Fig. 2: Exemplary relationship graph

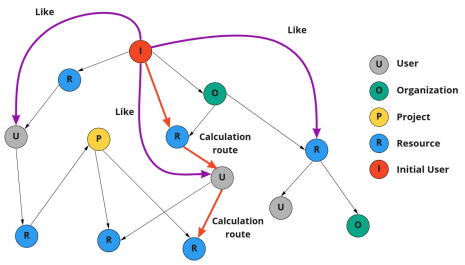

Fig. 3: Exemplary subgraph with parameters and factors for relationship graph in Fig. 2

vertices values of all the vertices on the path. The calculation is performed once and individually for each node. Calculations are performed in topological sort order for optimization purposes, and previous calculations are utilized to prevent redundant calculations. As part of the calculation route, all relationships except for Like relationships are considered. Specifically, this relation indicates only the value for the first phase of the algorithm. During the second phase of the recommendation mechanism, it does not affect any vertices. The result of each calculation route gives the final value of the recommender system for node (item), which can be viewed in the user interface. Figure 3 shows an example of the calculation route.

6 Conclusion

This paper proposes an academic recommendation system to address challenges in scholarly resource discovery. The idea is based on a hybrid filtering approach that combines collaborative and content-based recommendation techniques within a graph structure. The graph integrates multiple academic databases (OpenAIRE, Microsoft Academic Graph, Google Scholar) for comprehensive recommendations. Furthermore, the serverless cloud infrastructure ensures scalability and performance, optimizing resource utilization. Ongoing research is focused

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_34

on experimentally validating the system. The tests include a comprehensive performance evaluation in which recommendation accuracy, computational efficiency, and scalability will be assessed under real–world conditions. Future work also plans to explore improved personalization techniques, integration with additional academic data sources, and reinforcement learning-based recommendation strategies. Future work includes refining recommendation algorithms using reinforcement learning, expanding entity types, and improving integration with external repositories. Moreover, the experimental evaluation is planned.

References

1. J. Beel, B. Gipp, S. Langer, and C. Breitinger. Research-paper recommender systems: a literature survey. *Int. J. Digit. Libr.*, 17(4):305–338, 2016.
2. J. Choi, J. Lee, J. Yoon, S. Jang, J. Kim, and S. Choi. A two-stage deep learning-based system for patent citation recommendation. *Scientometrics*, 127(11):6615–6636, 2022.
3. M. Färber and L. Ao. The microsoft academic knowledge graph enhanced: Author name disambiguation, publication classification, and embeddings. *Quant. Sci. Stud.*, 3(1):51–98, 2022.
4. M. Fire and C. Guestrin. Over-optimization of academic publishing metrics: Observing goodhart's law in action. *GigaScience*, 8, 06 2019.
5. Google. Google Scholar. https://scholar.google.com/intl/en/scholar/about.html.
6. S. Hacohen, O. Medina, and S. Shoval. Autonomous driving: A survey of technological gaps using google scholar and web of science trend analysis. *IEEE Trans. Intell. Transp. Syst.*, 23(11):21241–21258, 2022.
7. G. Kalhor, A.A. Sarijalou, N.S. Sadr, and B. Bahrak. A new insight to the analysis of co-authorship in google scholar. *Appl. Netw. Sci.*, 7(1):21, 2022.
8. K.Vichos, M.De Bonis, I. Kanellos, S. Chatzopoulos, C. Atzori, N. Manola, P. Manghi, and T. Vergoulis. A preliminary assessment of the article deduplication algorithm used for the openaire research graph. In *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24-25, 2022 (hybrid event)*, volume 3160 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.
9. Microsoft. Microsoft Academic Graph. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/.
10. OpenAIRE. Openaire research graph. https://graph.openaire.eu/.
11. B. Padmaja, G. Sucharitha, and E. Krishna Rao Patro. *KGRecSys: Knowledge graph-based recommendation systems: A comprehensive overview.* 2025.
12. Hongwu Qin, Juntao Zeng, and Xiuqin Ma. Trend analysis of research direction in computer science based on microsoft academic graph. In *CONF-CDS 2021: The 2nd International Conference on Computing and Data Science, Stanford, CA, USA, January 28-30, 2021*, pages 18:1–18:4. ACM, 2021.
13. J. Schirrwagen, A.Bardi, A. Czerniak, A. Loehden, N. Rettberg, M. Mertens, and P. Manghi. Data sources and persistent identifiers in the open science research graph of openaire. *Int. J. Digit. Curation*, 15(1):1–5, 2020.
14. Ch. Troussas, A. Krouska, A. Koliarakis, and C. Sgouropoulou. Harnessing the power of user-centric artificial intelligence: Customized recommendations and personalization in hybrid recommender systems. *Computers*, 12(5), 2023.
15. J. Yang, X. Cheng, and W. Zhou. Research review and progress on practice of the resource recommendation research based on context-awareness. *Journal of Modern Information*, 40(2):153–159 and 167, 2020.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_34