IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 June 2024, accepted 11 August 2024, date of publication 27 August 2024, date of current version 6 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3450632

RESEARCH ARTICLE

Alleviating Cold Start in the EOSC Recommendations: Extended Page Rank Algorithm

MARCIN WOLSKI[1], ANTONI KLOREK[1], AND ANNA KOBUSINSKA[2], (Member, IEEE)
[1]Poznań Supercomputing and Networking Center, 61-139 Poznań, Poland
[2]Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland

Corresponding author: Marcin Wolski (marcin.wolski@man.poznan.pl)

This work was supported by the EOSC Future Project which is co-funded by European Union Horizon 2020 Programme Call INFRAEOSC-03-2020 under Grant 101017536.

ABSTRACT Recommender systems are becoming crucial in academia, where the number of available scientific resources is continuously increasing. One of the main challenges of such systems is a cold start problem, which often occurs when new users have no preference for any items or recommend new items that no community user has recommended yet. In the case of academic systems, where researchers are usually reluctant to express their explicit feedback or scientific interests, the cold start problem has a considerable and long–term impact on the recommendation algorithms. To alleviate this problem, this paper discusses a graph–based recommendation approach extending the Page Rank algorithm by using a co–authorship network. The proposed approach aims to enhance existing recommendation capabilities in the European Open Science Cloud (EOSC). The first results of the evaluation indicate that the proposed recommendation model is promising and reduces the cold start user–side problem in the academic domain.

INDEX TERMS Academia, cold start, graph–based recommendations, recommendation system, scholarly data.

I. INTRODUCTION

Recommender systems (RSs) have recently become an integral part of academic information systems, replacing keyword-based search techniques [1]. Among all types of academic recommenders, the most popular are those that primarily recommend scientific papers [2]. Paper recommender systems answer a researchers' typical need to filter a substantial number of academic articles in order to find those relevant to their research [3], [4]. However, in recent years, more and more researchers have also been interested in other scientific information, such as finding other researchers working in a similar field, searching for authors with a significant citation index, or trawling information on shared facilitated computing resources.

The European Open Science Cloud (EOSC) is an ongoing effort to address, among other things, the above challenges. For this purpose, the EOSC connects existing European e-infrastructures, integrates cloud solutions and provides a coherent point of access to various public and commercial services and information systems in the field of academic research [5].

As a result, EOSC resources comprise, among others, massive amounts of research effort results, such as published articles or datasets. As a primary dataset of scientific resources for the EOSC [6], the OpenAIRE Research Graph (OARG) dataset was used [7], [8], [9]. However, EOSC also contains software and e–infrastructure services, such as computational power, storage, and networks to support scientific experiments [10], which distinguishes EOSC from other environments.

To facilitate the findability and discoverability of the available EOSC resources, various recommendation

The associate editor coordinating the review of this manuscript and approving it for publication was Mansoor Ahmed.

mechanisms for EOSC (EOSC–RS) have been proposed [11], [12]. However, historical data on EOSC user (researcher) activity and use of scientific resources reveal that most visitors to the EOSC environment work as unauthenticated users. Moreover, users who register with the EOSC rarely provide information directly about their interests when accessing scientific resources. As a consequence of the lack of history of user interaction with the EOSC system, the recommender approaches proposed for EOSC face a cold start problem, which occurs when the recommender system lacks sufficient information to form reliable predictions or suggestions. This results in a reduced performance of the EOSC solutions proposed so far. Another problem that occurs in the EOSC is the so–called data sparsity problem. This problem arises because the EOSC is an example of an environment with many resources that significantly exceed the number of users so far. [13].

Problems of cold start and data sparsity have been extensively discussed in the existing literature [14], [15], [16]. Both issues typically relate to the initial phase of RS operation. The recommendation algorithms are expected to start working correctly once the system has reached a sufficient critical mass of active users for rating items, commenting, or ordering resources. Unfortunately, the direct use of the recommendation algorithms known from the literature in the context of academic platforms, particularly the EOSC, faces several problems stemming from their specificity. They include the following determinants of a cold start: a lack of information about user preferences, a lack of active users, and incomplete or constantly changing datasets. Moreover, they are most often present throughout the life cycle of these platforms, not only in the early phase of their operation. Consequently, to solve the cold start problem in academic systems, existing recommendation approaches must be specially tailored to the assumptions and characteristics of these systems.

Therefore, in this paper, we propose a solution that addresses the cold start problem in the EOSC environment. The proposed approach takes into account assumptions derived from the characteristics of the EOSC, assuming that recommendations will be generated in an environment where massive amounts of scholarly data are stored and, on the other hand, information about user activity is limited [17].

Although some attempts have been made to address the cold start and data sparsity problems in academic recommender systems, the most commonly used approach is the graph–based method, which uses various scientific and social networks. Among these, some of the most important and frequently used are citation networks (citation–citation relations) or networks that consider social relationships. Unfortunately, such information is not available in the EOSC environment, so it cannot be directly applied to make recommendations in the EOSC.

The contributions of this paper are fourfold:

- To deal with cold start in a scientific paper recommendation for EOSC, this paper develops a model based on a co–authorship network (i.e., author–author relationships). The proposed solution extends the concept used in the Page Rank (PR) algorithm, using the proposed author-author relationship.
- To improve the recommendation quality, the paper introduces a novel idea of personalization by modifying the PR algorithm to use the user-selected papers as starting nodes. Furthermore, following the results of the existing research on PR [18], [19], [20], the proposed solution adds the weights to the edges of the graph (representing selected attributes of the article) utilised by PR.
- In the context of big scholarly data in EOSC, the existing records of publications were enriched to create more credible data sources for recommendation algorithms.
- The experiments on existing, actual data collected and discovered in EOSC were conducted to evaluate the applicability of the proposed solution using co–authorship relationships. The paper presents and analyses the promising results.

The paper is structured as follows. Section II introduces the general idea of the EOSC and describes various statistics related to the use of this system. Next, related works are discussed in Section III, which includes the existing approaches addressing the recommendation methods used in academia. The general idea of the basic Page Rank algorithm is discussed in Section IV. Section V presents the co–authorship graph considered in the paper, and Section VI outlines the details of the proposed solution. The evaluation of the proposed PR algorithm, using offline and online methods, is presented in Section VII. The discussion of the proposed solution is provided in Section VIII. Finally, the conclusions and the future works are provided in Sections IX and X, respectively.

II. EUROPEAN OPEN SCIENCE CLOUD

EOSC is a key acronym for various European R&D projects related to Open Science at the national, regional, and European levels [21], [22], [23], [24]. The EOSC Portal, introduced as a universal hub for EOSC users, provides access to 4M+ assets, including also datasets, publications, software, trainings and more. The portal facilitates the collaboration of 300+ content providers, which have exposed 400+ services, ensuring a diverse range of resources for researchers and practitioners [25] The scientific resources are available now as a part of the EU EOSC Node Hub.[1]

The EOSC users include researchers, resource providers, technology enablers, trainers, and policymakers who provide and exploit the EOSC resources. The estimated size of the EOSC target population is approximately 2 million, including 1.7 million researchers covering all major scientific fields and levels of seniority [26]. The OpenAire Research Graph (OARG) is the primary provider of Research Products for the EOSC Platform, exposing publications,

[1] https://open-science-cloud.ec.europa.eu/resources

datasets, research–supporting software, configurations, and other products.

As of today, more than 4,000 users are registered on the EOSC Portal. Their activity in the EOSC portal is summarised in Table 1.

TABLE 1. Users' activity in the EOSC portal, as of February 2024.

Metric	Count
Number of registered users	4,333
Number of users who updated their portal profile*	169
Number of users with at least one project setup	424
Number of users with at least one resource liked**	3
Number of users with at least one resource disliked	40
Users with ORCID	488

Unfortunately, only about 3% of EOSC users have updated their profiles by stating their scientific interests or categories or created at least one project using EOSC resources [27]. There is no other way to derive, e.g., the user's research domain, scientific background, or resource preferences. Additionally, EOSC users can add a resource to their favourites, which could be seen as a positive vote, or disliked a resource, which means a negative vote. However, both options are visible only through a single EOSC component, which is a user dashboard. Thus, the use of explicit high–quality feedback, including detailed input on user interests, is significantly limited.

Anonymous users are responsible for the vast majority of interactions with the system, and they generate more than 90% of the traffic, e.g. browsing EOSC resources [28]. Logged–in users are, on average, more active than non-logged–in users, taking into account the interaction with the system expressed in terms of the number of events generated.

Table 2 shows the number of actions performed by logged–in and anonymous users (source: EOSC Preprocessor [11]). These values are related only to the publications.

TABLE 2. Users' actions on publications, as of June 2023.

	Type of user	
Statistics	Logged-in	Anonymous
# of all visited publications	62	1693
Avg # of visited publications by a single user	4.4	1.94
Min # of visited publications among users	1	1
Max # of visited publications among users	42	59

The analysed data of the EOSC system reveal that EOSC users use the provided resources anonymously, in a single session, searching for a few items, rather than using the EOSC for more complex scientific activities. Such an approach is not exceptional — it is common in publication catalogues or search engines (such as Google Scholar), where scientists reach out to find specific resources.

Therefore, the only way to determine user preferences in the EOSC is to monitor their activity. A resource selection in the EOSC portal is treated as if it interests the user. This is an example of implicit feedback provided by the end user, which the recommender system collects and processes to infer the user's preference [11].

III. RELATED WORK

Academic recommender systems use a wide range of recommendation methods. A simple approach to calculating usability can determine the relevance of an item based on overall popularity (i.e. global relevance [29]). For example, the system could recommend the most frequently downloaded scientific articles from the repository or those with the highest average ratings. In another scenario, the system could present the user with a list of datasets and software, usually required with the selected articles, which is called a co–occurrence recommendation. To indicate co–occurrence recommendations, the recommended items are those that are often found together with some source items [29].

Content–based filtering (CB) is one of the most widely used and researched recommendation approaches in academia [1], [29]. On the contrary, collaborative filtering (CF) recommendation systems are limited in this field [3]. CF algorithms can work effectively when the number of users is greater than or equal to the number of articles in the digital library [29].

Both CB and CF are often combined in a single recommendation engine. For example, a content–based method can obtain a user profile based on the content of articles in which users are interested. The profile of a researcher can be enriched with information about his previous research interests contained in his/her previous publications [3]. Collaborative filtering techniques can use global relevance attributes to rank candidates and graph methods to expand or restrict potential candidates for recommendation. These approaches, called hybrid, consider various aspects of users, items and the relationships between them.

Graph–based techniques are of great interest in the area of academic RS [29]. They exploit the inherent connections that commonly exist in academia through collaborative research. In these methods, recommendations are translated into graph–search tasks. The basic recommendation generation algorithms are based solely on graph search and do not take into account the content of the articles or the user profile (except when combining different approaches).

Graph-based techniques are commonly used to model social relationships between users. The so–called social recommender systems (SRS) have been proposed to enhance the capabilities of basic recommendation methods. They are based on the user and his/her friend having similar interests and tastes. Early approaches assumed the same degree of friendship for all users in a friend group. More recent approaches assume a variable degree of friendship to adequately model that users live in groups of highly dispersed people, such as family, neighbours, or classmates. Some of these people share common interests, while others may even disagree with the user [30].

Explicit social relationships are directly established by users (through the use of social platform features), and they can directly enhance the collaborative filtering process. In turn, implicit relationships among users can be inferred on the basis of their historical ratings of items. These inferred relationships can enhance the performance of social recommendation systems, particularly when explicit social connections between users have not been established [31].

In recent years, solutions offering different recommendation mechanisms have focused on applying machine learning/deep learning (ML/DL) techniques to recommend items. In particular, deep neural networks (DNNs) are the most commonly used approaches in recommendation systems. These methods outperform state–of–the–art algorithms using CB, CF, or graph–based methods in individual cases. They also aim to overcome some known problems that the basic methods solve only partially or with limitations [32]. For example, a hybrid deep learning recommendation system [33] has been proposed to fill the gaps in ensemble filtering systems. The developed solution uses embedding to represent users and items to learn non–linear latent factors. Another solution uses collaborative filtering based on neural networks [34] to replace the internal product of the interaction between the user and the element with a neural architecture. In [35], the authors developed a collaborative deep learning algorithm (CDL) that jointly performs deep learning of the representation of content information and collaborative filtering of the evaluation matrix (feedback). Furthermore, reinforcement learning (RL) is a popular approach to predict user behaviour by analysing browsing history on the portal [36].

The recommendation methods based on artificial intelligence have produced competitive results on known data sets (such as MovieLens, NetFlix or BookCrossing). Contrary to popular belief, it is not always the case that DNNs can handle all fundamental problems in the recommender systems area. Deep learning is known to be data intensive because it requires enough data to support rich parameterization fully. In many fields, including academia, no rich datasets enable DL approaches to be used effectively [37]. Furthermore, the working phase of DL algorithms usually requires substantial computational resources, which can be problematic for companies or researchers who do not have easy access to such hardware.

Since this paper uses a graph–based approach to recommend scientific papers, the following section will outline the related work proposed so far using this method. To the best of our knowledge, graph–based approaches have been widely applied in the domain of e–science in two main scenarios: to enhance the capabilities of hybrid systems and to overcome the cold start and data sparsity problems.

A. GRAPH–BASED APPROACHES

The basic relations depicted in the graph are usually twofold: user–item and item–item relations. The user–based methods use the ratings of similar users on the same item to make predictions. Although such methods were initially quite popular, they are not easily scalable and sometimes inaccurate. The subsequently proposed item–based methods compute predicted ratings as a function of the ratings of the same user on similar items. Item–based approaches provide more accurate but less diverse recommendations [38].

In the academic field, item-item relationships translate into a few types of scientific network [39], including the most widely used: (1) co–authorship (network) to model author–publication relationships, and (2) citation network to model citation relationships (publication A cites publication B).

Co–authorship is one of the most tangible and well–documented forms of scientific collaboration through published papers [40]. Co–author networks have been studied extensively. In these networks, two scholars are normally connected if they have co–authored one or more papers together. Collaboration benefits the productivity of scientific research and is positively correlated with the number of articles written by an author and the number of times the given article has been cited [41].

A method called CARE incorporates author relations and historical preferences to recommend scholarly articles [42]. This method assumes that some researchers prefer to search through articles published by the same authors to find articles that interest them. As a result, in the CARE method, a graph is built on the basis of the co–authors' relatedness information. Then, a random walk with a restart is employed to generate a recommendation list. CARE is based on user preferences and favourite documents, which cannot be applied to the EOSC. Furthermore, the CARE algorithm does not consider additional attributes that affect the reputation of an article.

Furthermore, the prediction of new collaboration opportunities with the use of a co–author network has been a popular topic of scientific effort. A recommender system has been proposed by scientists to find possible collaborators with respect to their research interests to support novice researchers and increase their publishing activity [43]. The recommendation problem was formulated as a link prediction within the co–authorship network. The network was derived from the bibliographic database and enriched with information from research papers.

Another method named MVCWalker aims to recommend potential collaborators based on a co–authorship network [40]. To improve the quality and accuracy of the recommendation, the authors defined the importance of the link in the graph by exploiting a several additional factors, such as the order of the co–authors or the latest collaboration time.

Notwithstanding the promising results achieved by the two cases mentioned above, these could be applied in EOSC only when the number of active users increases considerably.

Regarding citation networks, they rely on citation–related connections within the scientific literature. Recommendation systems that are built upon these connections compute relatedness among academic papers. During computation, usually two factors (co–citation and bibliographic coupling) are determined to measure relevance by focusing on neighbours [3], [39]. Some authors use bibliometrics to improve the results of recommender systems in digital libraries [44]. In addition, multilevel citation networks can be used to recommend research, empirical and exploratory papers [4].

Both types of scientific networks (citation and coauthorship) can be combined and form a hybrid recommendation system. An interesting example of a hybrid algorithm for recommending research articles uses the citation network of articles and the author Relationship Network (CNRN) [45]. To choose the best-matched papers published by key authors, the algorithm assesses the importance of the papers. The most relevant papers from the citation network are taken to create the co–authorship network. On the edges between authors, the frequencies of collaboration were listed. The author analysis could also be deepened with the use of social networks [46]. In both cases, a citation count and centrality measures, such as closeness centrality, betweenness centrality, or eigenvector centrality, were incorporated to choose the key authors. Setting additional weights for previously determined key authors seems to be a promising approach for the future development of the PR algorithm for EOSC. However, this method could be applied in later stages of EOSC, that is, to recommend papers for registered users who would have a long history of clicked (visited) publications.

B. COLD START

The cold start problem can be considered in the context of items and users [47]. The cold start of items occurs when the number of ratings previously submitted for these items prevents recommending them properly to users. In turn, the cold start of users happens when a new user who has already joined the online environment has no or a very poor record of interaction with the system (e.g. only a few reviews). In such a situation, there is no interaction between the new user and others. Therefore, no similarity between them can be measured. As a result, the recommender systems cannot make reliable recommendations.

Academic RSs usually deal with a vast number of items (e.g., millions of publications), yet users typically rate only a small fraction of these items. In such an environment, data sparsity significantly affects the accuracy of recommendations due to the scarcity of ratings. Various approaches have been proposed to deal with data sparsity. They mostly rely on different imputation methods to fill in the missing values in the set of available ratings. In such a case, the reliability of predictions often comes as an issue, which in turn can affect the reliability of provided recommendations [48]. Researchers have proposed various approaches to increase the reliability of recommender systems [48], [49], mainly by increasing the system's confidence in its recommendations — the more reliable a prediction, the less likely it is to be wrong [50].

In real-world applications, the collaborative filtering method, which suffers the most from data sparsity, is used only as a complementary technique to other approaches. It is effective only when the system overcomes the cold start problem and reaches a sufficient level of maturity represented by a minimum number of users who rate items, interact with resources, or provide comments. Importantly, in the EOSC Portal, cold start persists throughout the entire operational period, not just during the early-stages.

This problem can be mitigated to some extent by content–based recommender algorithms that can predict an item's importance even without its previous evaluations. On the other hand, CF approaches suffer from the cold start of new users' problems. They cannot recommend items to new users without a history of previous user interactions with the system. Interestingly, graph–based approaches, especially those based on item relationships such as co–author networks or citation networks, do not suffer from the aforementioned cold start problem and can thus be applied to real academic platforms.

In general, an operational recommender system takes advantage of different basic recommender techniques, mainly to overcome the typical drawbacks of a single approach and to adapt the system to specific conditions (such as the availability of datasets in actual implementations). Hybrid solutions constitute the most common approach to dealing with cold starts.

Hybrid RSs designed to address the cold start issues often combine the outputs of single CF and BF methods [51], [52], and others enhance the basic CF and BF algorithms by incorporating social networks, for example [47]. Also, the users' demographic information (such as gender or age) can be applied to enhance the rating profiles used by CF [53]. The recommendation context, such as location, time, or social information, not only helped alleviate cold starts but can also significantly improve the recommendation process [49], [54]. In [55], a hybrid recommender system for research documents that replaced academic search engines was introduced. The proposed solution is based on different input sources such as citation networks, author analysis, ratings, or text mining. In addition, the algorithm considers the reputation of the article, represented by an impact factor or an H-index. Although this solution has many advantages, it is a standalone application that processes the full papers uploaded by the user.

Implicit ratings are often utilised by hybrid solutions to overcome the cold start. Implicit ratings can be extracted from user–object interactions, for example the more pages with a document they read, the more it is assumed that they would like to get the documents. Some of the interactions between the user and the resource, such as downloading the paper, adding it to the researcher's profile, editing the paper's details, and viewing its bibliography, can also be seen as positive votes [29].

Another group of hybrid solutions dealing with cold start problems takes advantage of DNNS, using ancillary information such as item–content information [35]. Depending on the availability of data, additional information about users (e.g. age, occupation, location) and items (e.g. year of publication, title) is passed to a deep neural network [33].

Despite the popularity of hybrid approaches and their effectiveness in alleviating cold start, it is still unclear which of the proposed directions used in these solutions is the most promising, mainly due to the reproducibility problem [29]. Many hybrid solutions combine information from one source - for example, user profiles in the web portal with another source, such as a dataset with publications. This feature may not be applicable in a real-world environment due to inconsistencies in the format or content of data, missing items or incomplete descriptions, which is also true for EOSC.

Moreover, hybrid systems designed for commercial purposes often include additional techniques when there is insufficient information for generating recommendations. The most popular ones include recommending: a) random items to new users or new items to random users (random strategy), b) popular items to new users or new items to the majority of active users (maximum waiting strategy), and c) a set of different items to new users or a new item to a set of different users (exploratory strategy). These strategies may not be effective in academia. The focus interviews with EOSC users, which were conducted as part of the EOSC project, revealed negative reactions of respondents to recommendations based on the general population [26]. Many respondents declared that they would not use a system that provided unreliable recommendations, and the researchers frequently questioned popular objects.

IV. PAGE RANK ALGORITHM—THE GENERAL OVERVIEW

The PageRank (PR) algorithm was originally proposed to determine the relevance of a given website based on outbound and inbound links. The algorithm takes advantage of the link structure on the Web to produce the ranking of every web page.

The algorithm was formally formulated as a method for assigning a universal rank to web pages based on weight propagation [56]. A page obtains a high rank if the sum of the ranks of its links is high. The PageRank of page p is given as:

$$PR(p) = \frac{(1-d)}{N} + d\sum_{i=1}^{k} \frac{PR(p_i)}{C(p_i)}$$

where N is the total number of pages on the web, d is damping factor, p_i is the page that links to p, and $C(p_i)$ is the number of outlinks of p_i. PageRank of a page is conceived as the probability of a web surfer visiting the page after clicking on many links. Dumping factor d is the probability of a surfer going to the net page, and $(1-d)$ is the probability of a random jump [57].

PR has been successfully applied in RSs where a network of related linked records occurs [58]. Google's PageRank is an example approach for recommending research papers applied in Google Scholar. PageRank measures the authority of a paper and ranks it based on the number of citations it receives from other academic articles [59]. The weighted PageRank for academic recommendations has gained significant interest from researchers. For example, the version of weighted PageRank from [60] gives more weight to co–author ties with fewer co–authors than those with large numbers of co–authors. The authors from [61] substituted the $C(p)^{-1}$ to the fraction of the journal's PageRank transferred to the journals it cites. Another approach integrates an author's community impact and academic impact with the use of citation and co–authorship network topology [62]. Compared to other weighted PR algorithms, the authors focused on the random surfing aspect $\frac{(1-d)}{N}$ and developed it into citation ratios. In the case of EOSC, these are publications linked by a common author (or by references, if such a graph can be created). Based on information on the author of a publication, it is possible to create a graph of related publications (co–authorship network), in which the vertices denote publications and the edge between two given vertices indicates a common author. The purpose of the recommendation is to suggest a relevant publication of interest to the user.

PR is the most typical random–walk algorithm in the field of computer science. The procedure utilizes a so–called "Random Surfer Model" that traverses through links between sites and once in a time jumps to a random page chosen based on some probability distribution [63]. Researchers have developed a number of PR variants, such as personalized PageRank (PPR). PPR is widely applied to diverse domains, such as information retrieval, recommendations, and knowledge discovery, due to its theoretical simplicity and flexibility [64]. The PPR algorithm is often used to recommend friends, using a method to assess the importance of nodes in a graph with respect to a query node. Researchers are also improving the original random walk rules and proposing new algorithms, such as a random walk with restart (RWR) [65] or lazy random walk [66], among others.

PaperRank is an extension of the PR algorithm for evaluating scientific documents, which considers the indirect relationships between these documents [3]. Although PaperRank is a suitable method to determine the importance of an article, it tends to rank documents based primarily on the number of citations. As a result, recent articles are usually ranked low, even when the article is known to be outstanding literature. This is an important limitation because recent articles may be significant to researchers interested in the latest scientific findings and the direction of scientific research.

V. CO–AUTHORSHIP GRAPH IN THE EOSC

In order to model a co–authorship graph in EOSC, the subset of OpenAIRE Research Graph resources was used (EOSC–OARG). The subset is currently deployed as an operational data source for the EOSC Portal [25]. It contains more than 2 million publications (source: EOSC–OARG 5.0, at the time of writing the paper, this particular dataset is not publicly available, whereas only the full OARG is exposed). There are approximately 1.57 million unique authors within the set of publications. The average number of authors per publication is 2.14, indicating that many authors are likely to contribute more than one article.

Figure 1 presents a graph structure in EOSC–OARG, built upon the co-authorship relations for a single publication. The vertices of the graph represent publications, and the edges connect the publications written by the same author(s). This figure depicts the first (selected) publication as a blue node. This node is connected with all other articles written by the first, the second, or both authors of this publication (light blue edges). This results in a relatively small but dense graph, as many attached publications are interconnected.

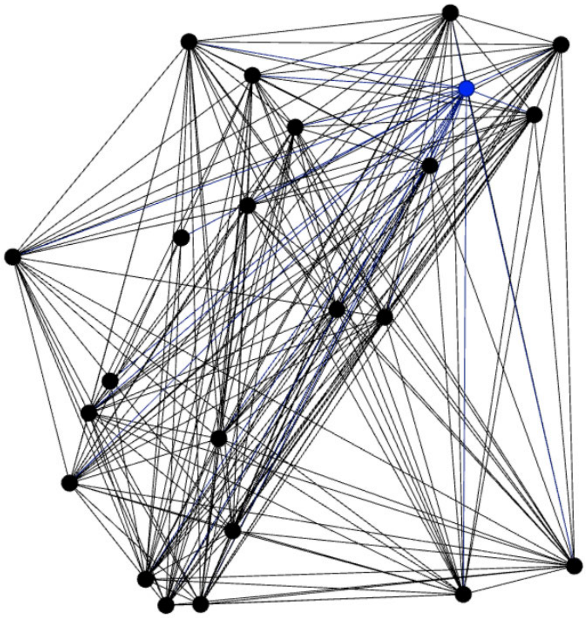

FIGURE 1. Co-authorship graph for a selected EOSC publication created with the NetworkX tool. A blue node represents a starting point, i.e., the publication visited by a user. The rest of the graph nodes are publications written by the same author(s). Edges indicate that two connected publications are co-authored.

There are 23 vertices in total in the graph considered 1, which means that the selected publication is associated with 22 other works that share the same author(s).

In some cases, the structure of vertices and edges constitutes a complete graph, i.e. all nodes are connected with one another. Such a situation occurs when a source publication has only one author: then all other connected publications have the same author. This assumption is correct when connections are at a single depth, meaning that there are only edges outgoing from the source publication. The graph can be extended by connecting the found entities with other items from the database, which indicates adding second, third and next depth to the graph.

An alternative option to the co–authorship graph used in the EOSC was the citation graph, where "cites", and conversely "cited" relations are used. The main source of information used to build the citation graph is references, i.e., a list of cited articles appearing at the end of academic papers. Unfortunately, in the present data set for EOSC, only about 15% of the publications have information on citations. It is often not fulfilled entirely, e.g., a publication actually cites 10 other works, but only a single one is present. Thus, proceeding with this method would result in many missing connections and, consequently, an empty set of recommendations for most of the publications.

VI. EXTENDED PAGE RANK ALGORITHM

The first step towards applying the PR algorithm in EOSC was to determine the set of attributes that might be used as weights for the co–authorship graph built on top of existing EOSC–OARG records with publications. Unfortunately, the EOSC–OARG data source contained missing or incomplete metadata fields, and its use as a single dataset for recommendation algorithms would require extensive improvement effort to clean up the records [67]. Table 3 presents selected fields available as publications metadata. As can be seen, crucial information is absent in the majority of entities.

TABLE 3. Missing values in EOSC-OARG (publications), as of Feb 2024.

Statistic	Count	Percentage to total (%)
Number of publications	2,376,573	100.00
Publications with affiliation	384,141	16.16
Publications with subject	512,070	21.55
Publications with keywords	1,818,948	76.53
Publications with DOI	2,150,949	90.52
Publications with relations	916,893	38.57

The most effective way to overcome this drawback was to link the EOSC–OARG records with other scholarly graphs to enrich its content. OpenAlex (OA) was selected as a fully open catalogue of the global research system [68] and eventually linked with OARG [67].

As a result, 813541 publications were retrieved from OA and matched with EOSC–OARG, which constituted more than 40% of the total available in the EOSC–OARG dataset. Several metadata fields that may be useful in the current recommendation approach were extracted. These were, among others:

- *Cited–By–Count* indicating how many citations a publication has received, 237062 publications were enriched.
- *Cited–Works* representing a list of other works that refer to the publication, 344067 publications were enriched.
- *Related–Works* presenting a list of similar publications, 384643 were enriched.

It is worth mentioning that a single publication might have multiple references or related items. Moreover, during the linking process, it turned out that OA had many publications that were not present in the EOSC–OARG. Thus, only those present in both datasets eventually remained.

A subset of rich records was created by combining information from both scholarly datasets [69]. These records, which indicate a more credible data source than the original EOSC–OARG, were assumed to be used primarily for the recommendation algorithm. In other words, the usage of these rich records would take precedence over the original records to recommend publications of better quality.

Eventually, two selected attributes of a publication were selected as weights in order to be assigned to the (co–authorship) graph edges:

- *Number of citations* (by other publications) — the number indicates how many times a publication has been cited. Such information can be used to determine the popularity of a publication by adding the popularity value as weight/cost to vertices in a graph of connected publications.
- *Number of downloads/views* — the number representing the global popularity of a resource, is calculated as the number of downloads/views of an article. This information is present in 83.37% of OARG records (in theory, downloads may be more important than views, often indicating that the user was more interested in the paper).

Other existing attributes, such as *item references between publications* or *item related work*, were also considered to be applied in the PR algorithm. However, they would introduce an additional layer of information in the original algorithm (such as content–based features), which would, in turn, change the pure graph–based approach into a hybrid one. The hybrid–based method remains a promising direction for the future.

After weighting the graph edges, additional modifications were applied to the algorithm's parameters. In particular, we aimed to introduce personalization into the original algorithm by influencing its behaviour to be more aware of publications the user had visited. In the context of EOSC, the publications visited are treated as those that the user likes II. However, it must be stated that such an assumption of positive implicit feedback through visited ("clicked") publications might not always be true [70].

In the proposed algorithm, the "random surfer model" was incorporated with precisely stated jump destinations for the surfer. The surfer does not teleport to a random vertex selected from a uniform probability distribution but to the vertex chosen from the ones that represent the previously clicked user's publications. These destinations do not have to be equally probable; we can modify the probabilities based on the order of the user's actions (a publication visited more recently will have a higher probability than the one clicked many sessions ago).

A. ALGORITHM DETAILS

The input data for the algorithm have the form of a graph defined as follows: G(V,E), where

- V = vertices indicating publications that were clicked by the user or that were put into the graph by having common author(s) with any of the clicked ones.
- E = edges representing connections between publications that have at least one common author.

It is a directed graph where there are always two edges between the vertices, one in each direction. In the proposed algorithm, we introduce the personalization concept. The two mechanisms described below influence the algorithm's behaviour.

1. Assign weights to the edges. The weights are influenced by:

- The number of downloads and views (i.e., publication's popularity). Although this number is linked to a publication, weights are assigned to all edges entering this vertex.
- The number of citations — also reflects the publication's popularity, and weights are assigned to the entering edges.

2. Defining random jumps.

- Choosing the vertices to which the surfer might teleport.
- Defining vertices' probabilities based on user clicks' order.

Taking into account the previously adopted assumptions and definitions, the next steps of the proposed PR algorithm are as follows:

1) The algorithm starts at a random vertex selected from the ones that reflect the user's clicked publication.
2) With probability α, a neighbouring vertex is chosen.
3) With probability 1-α, a random jump is performed to an independent vertex. The vertex is chosen from the set of vertices that reflect the user's clicks. The choice is made based on the probability distribution among these nodes.
4) The algorithm continues until it reaches the maximum number of iterations. Each iteration terminates when the algorithm reaches an error smaller than the given tolerance.

The output of the proposed algorithm is vertices with results obtained by running the algorithm. These vertices that reflect the already-clicked publications were excluded. Then, the resulting items were sorted in a descending order, and the first n items were chosen as recommendations to the user.

Figure 2 presents the steps of the proposed recommendation approach.

As the graph is created based on co–author relationships, the highest results are usually returned for publications written by the same authors as in the user click. However, it is also possible to recommend articles authored by other people when a person is a co–author of a publication and this co–author wrote the article clicked by the user. Longer chains are also possible.

B. ALGORITHM'S EFFICIENCY

Although widely applied, recommendation algorithms based on graph processing and random walk algorithms (such as PR) have some well–known drawbacks [65]. The most commonly motioned issues indicate that there is a need to

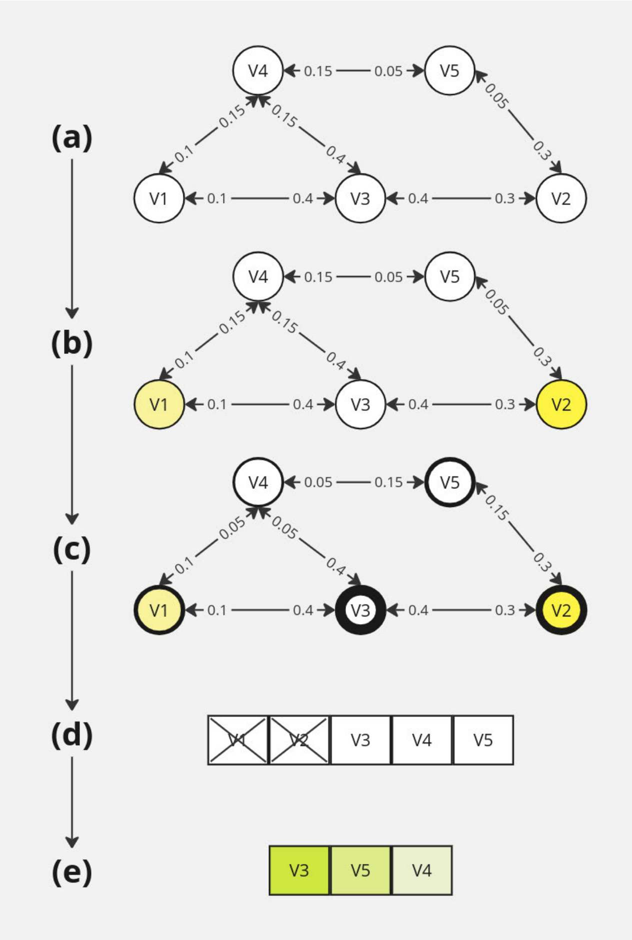

FIGURE 2. Overview of the proposed recommendation PR approach: (a) assigning weights to edges taking into account the number of downloads and views and the number of citations; (b) defining random jumps vertices of the PR algorithm; (c) running the personalized PR; (d) excluding vertices indicating the already visited publications; and (e) sorting the results and choosing the first n items to recommend.

generate ranking scores on all candidate items at each step for each user, which might lead to low efficiency,

A state of the EOSC environment was simulated to assess how the proposed algorithm would handle large–scale systems. In the experiment, the presence of both active and less active users was assumed in EOSC. A subset of the events (user actions) was generated for each simulated user, indicating that the user had selected or opened a random publication.

The characteristics of the simulated EOSC users and their activities were as follows:

- 400 active users
  - Between 5 and 15 random publication visits per user.
  - Publications may be repeated between users.
- 2000 occasional users
  - Between 1 and 5 random publication visits per user.
  - Publications may be repeated between users.

For the simulated data, the co–authorship graph with weights was created and the algorithm VI-A was applied. The aim of this effort was to investigate time efficiency, i.e., how long it takes to deliver recommendations for selected simulated users, and coverage, i.e., what percentage of users will receive recommendations.

The results of the experiment showed coverage of 97.4%, which indicates that the recommendations were produced for almost all users. Users with fewer visits may face no recommended items when an author of a visited publication wrote only a single paper (Section VII-A provides a more detailed analysis of this issue),

Regarding time efficiency, the time required to generate recommendations for simulated users took less than a second. This response time has an acceptable rate. Moreover, the configuration of PR parameters may influence the operating time. Changing the maximum number of iterations and a criterion to stop an iteration affects the time needed to provide the output of the algorithm. However, at the same time, lowering the PR algorithm parameters may result in less accurate and diverse outcomes. What is even more important, compared to an algorithm for scoring Internet websites, the recommendation algorithm does not need to traverse the entire graph. The smaller number of connections and random returns to starting vertices result in more centred activity.

VII. EVALUATION

The evaluation approaches for a recommendation system utilise both online and offline methods [38]. The offline testing of a RS typically incorporates the previously collected datasets (such as ACM, DBLP or CiteSeer in the academic context) as a standardised benchmark to compare algorithms across various settings [1]. Given the EOSC constraints, including the historical data about users activity (presented in Table 2 and data quality issues (described in Table 3), there was no direct way to benchmark the proposed algorithm or determine its performance against other published models.

In the case of an online approach, the user reactions or subjective assessments concerning the presented recommendations are measured. Interactions with real users can provide additional information about the system's performance and usability. A user study is conducted by recruiting a set of test subjects and asking them to perform several tasks that require an interaction with the recommendation system [71].

A. OFFLINE EVALUATION

The main goal of the offline testing was to determine if the proposed algorithm achieved its primary purpose, i.e., whether it alleviated the cold start problem in EOSC by generating recommendations for each user visiting the EOSC Portal.

A sample of real data representing user activity was taken in the EOSC portal to evaluate the proposed method. This sample was processed, and only the user activities (UA) that represented a specific action called "visiting the paper summary page" were extracted. As a result, the 1963 users who had visited one or more papers were determined. Data were sampled from a seven–month period from January 2023 to August 2023. Among all extracted users (representing both registered and anonymous users), there were 1560 (79.47%) people with single visits, 192 (9.78%) with two clicks, and 211 (10.75%) with three or more viewed publications.

The general evaluation procedure was as follows: for a user taken from the sample above (or sub–sample), every visited paper of this user was taken to build the co–authorship graph. The algorithm was then run for the built graph. Finally, the number of recommended (similar) publications for the user was calculated.

In the first iteration, the algorithm was run for the whole sample regardless of the number of clicks by each user (1963 users). The recommendations were offered to 1,296 users, which is more than 66%. The moderate result is mainly due to the fact that the majority of users visited only one publication, and for some of these publications, there were no other papers in the EOSC-OARG dataset written by the same author(s).

Next, we ran an algorithm on the subsamples for the users categorised by the number of visited resources: single publication visits, two publication visits, and three or more visits. The detailed results are presented in the Table 4.

We also show the number of generated recommendations to analyse the algorithm's potential and present that usually, more than a single recommendation is created. The results indicate that more active users (the ones who viewed more articles) are more likely to receive more recommendations. This is because more clicks result in more authors and a bigger co–authorship graph. Therefore, the probability of the presence of the author(s) of the visited publication(s) is higher.

TABLE 4. Evaluation details.

	Number of recommendations		
Number of visited publications	**At least 1**	**At least 3**	**At least 5**
1	59.23%	36.67%	29.94%
2	87.5%	69.27%	59.90%
3+	96.68%	82.94%	78.20%

B. EXPERT EVALUATION

Two selected research communities from the fields of medicine and computer science were invited to participate in the evaluation of generated recommendations. They received an evaluation survey.[2] with a detailed description of the whole evaluation process.

In the first step, a researcher (expert in a given field) who decided to participate in the evaluation was asked to provide a several selected publications from the EOSC Portal concerning his/her research interests. Based on this, the top 5 recommendations for similar publications were generated using the PR algorithm described in the paper. In the second step, the 5 recommended papers were sent back to the expert, who was asked to evaluate their relevance. The scores were binary, which means that a "1" was given if the recommended publication was adequate and a "0" if it was not. This simple approach for assessing a publication's relevance followed a well-known pattern (i.e. like/dislike of a resource), and it was also consistent with the EOSC portal feature. In the future, a 5-rating scale is going to be introduced to assess the level of experts' feelings about the recommended paper in a more nuanced way.

Eventually, the answers from 7 SMEs were used to aggregate the scores and calculate the Average Precision@K (AP@K) for K=1, K=3 and K=5 for each expert and the Mean Average Precision@K (MAP@K) together for all 7 evaluations for K=3 and K=5. The AP@K measurements are presented in table 5. MAP@3 is equal to 0.68 and MAP@5 to 0.69.

TABLE 5. Evaluation metrics.

Expert	AP@1	AP@3	AP@5
Expert1	1	0.66	0.4
Expert2	0	0.33	0.2
Expert3	1	0.66	0.4
Expert4	0	0.33	0.6
Expert5	1	0.66	0.4
Expert6	1	0.66	0.8
Expert7	1	0.66	0.6

The results of the experts' evaluation highlight a number of conclusions. Firstly, the algorithm generated at least five recommendations for each expert based only on their single action. Furthermore, the top recommendation presented to experts was satisfactory in five out of seven cases, which is a very promising observation. The problem was recommending a single item that would meet the user's requirements. Most of the further recommendations were also relevant, as most experts scored two out of three recommendations as accurate. The experts originate from two distinctive research areas, and even for less common scientific domains, the algorithm was able to generate recommendations.

VIII. DISCUSSION

As the graph is created based on co–authorship, the publications with the highest results will usually be those where the author is the same as in the clicked publications. For a user whose clicked publications do not have a connection, we cannot recommend another publication based on the co–author network. In this case a user may receive general recommendations, e.g., frequently clicked items, or based on another feature, such as publications in the same field (if they exist and if the field can be identified). It is assumed that there should not be many such users in a real application because a user usually clicks on more than one publication, and most authors have another article in the dataset.

We can minimise the likelihood of the situation described above by extending the original co–authorship graph with all

[2] https://forms.office.com/e/y30RqWawEe

publications recorded for each author presented in the graph (that means we create an additional layer of the graph).

Adding weights affects the results of recommendations but not as much as the distance from the initial vertices; in other words, even if you add a very large weight to a node much farther away from the algorithm's restart locations, close publications are usually still more likely to be recommended. However, the weights allow you to influence the performance of the algorithm and may contribute to favouring more popular publications.

If a small number of publications appear on the graph, there may be a situation where a specific publication visited by a user is not connected by an edge with any other publication. In this case, this publication is not considered in the construction of the graph, thus the PR algorithm does not return any recommendations. The evaluation conducted indicates that such a situation occurs for about 34% of cases where a user has visited only one paper so far.

In some cases, a separate subgraph may be created using one of the clicked publications — publications in this small subgraph may be more likely to get a recommendation.

In Figure 3, the black nodes are the starting points (visited by the user) of the publications. The darker the green, the higher the probability of a recommendation. In the small subgraph (3 nodes), the connected nodes have a higher probability than those in the larger part of the graph (even though they are at the same distance from the clicked publication, for example, one edge). Those in the small subgraph got a recommendation probability of 0.14, and those in the large subgraph got 0.02.

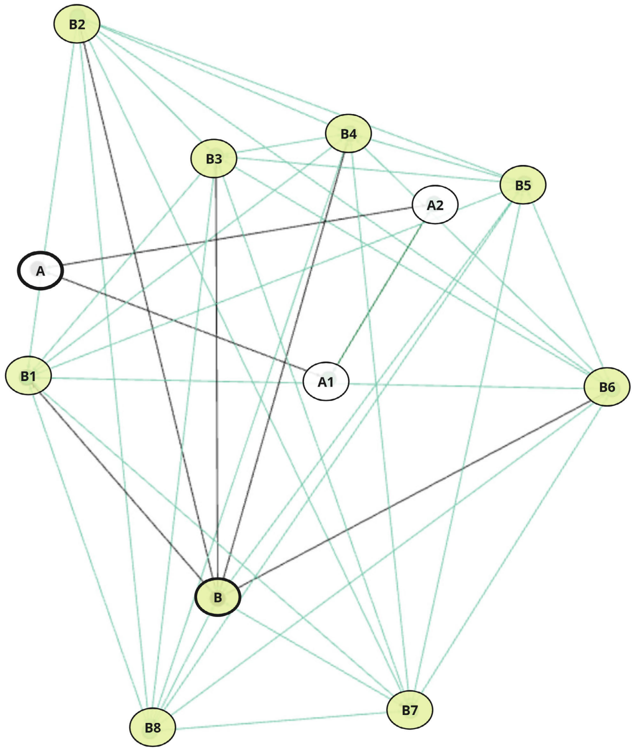

FIGURE 3. Separate subgraphs based on co-authorship relations. Nodes A and B are starting points; they represent publications selected by a user. Nodes Ax have a higher recommendation probability than Bx, as Ax belong to the smaller subgraph.

Finally, unambiguous identification of the author is problematic — there are often authors with the same name and surname. There is an identifier called ORCID (''ORCID provides a persistent digital identifier (an ORCID ID) that you own and control and that distinguishes you from all other researchers''), which can be used as an author ID. Unfortunately, in a used dataset, authors rarely have an ORCID assigned, or it is not present in the metadata of a publication. Also, a problem occurs when a single author has ORCID entered in some publications but not in others. As a result, such an author can be classified as two different people.

In the EOSC-OARG dump (version 4.2), there are 194762 authors with ORCID; the total number of authors is approximately 1.52 million (approximately 12.8%). In the EOSC, only for about 13% of the authors, the ORCID is stated. When analysing the ORCIDs assigned to the authors in the dataset, it turned out that some values occur surprisingly often. Table 6 shows the most common ones.

TABLE 6. ORCID Occurrences.

ORCID	Number of occurrences
0000-0002-1825-0097	64239
0000-0001-6651-4424	6209
0000-0003-0981-0061	3791
0000-0003-3819-8537	3352
0000-0003-2443-5271	3105

The first author is a suspicious case. ORCID refers to Josiah Carberry, described as follows: ''Josiah Stinkney Carberry is a fictional professor, created as a joke in 1929. He is said to still teach at Brown University and is known for his work in ''psychoceramics'', the supposed study of ''cracked pots''.'' [72].

Other authors are well-known researchers and scientists. The second ORCID refers to David Nielsen from Arizona State University, the third to Milton Love from the University of California, the fourth to Jonathan Prag from the University of Oxford and the fifth to Richard Robbins from Walter Reed Army Institute of Research. These four individuals have authored and co-authored numerous publications and boast many citations.

IX. CONCLUSION

Notwithstanding this high use of RS, its proper design, development, and operation remain challenging. The concepts and approaches of the recommendations strongly depend on their application [2], and certain challenges can have different impacts in different domains [44].

As far as the academic world is concerned, the reasons for such challenges in applying common recommendations, methods, and techniques are twofold. Firstly, open academic datasets (such as OAG, Oarg, Aminer and DBLP) with millions of author records, documents, citations, figures, or tables are the basis for providing recommendation services on academic platforms, including EOSC. These datasets may often contain incomplete, inaccurate, or unsubstantiated data [67]. Secondly, on academic platforms, information about user activity and, therefore, user preferences can often be obtained only from an analysis of user behaviour — academics are reluctant to provide their interests, preferences, or academic domain, especially when they do not see the benefit of doing so.

Recommendation algorithms proposed for academia usually consider the researcher's scientific field, activities, and connections, such as co–author relationships or citation results suggesting relevant items. However, these methods rarely take into account the quality of the data. For example, the use of a citation network, a common approach to increasing academic RS capacity, is often limited by incomplete information on references between articles. Social networks, which are used to resolve more accurate links between users, have been widely proposed to improve RS performance. However, its use to deal with cold starts is questionable. Firstly, building trust between the new platform and users will attract their attention and activity. Secondly, matching a user's activity on one platform with their activity on another, such as a social academic network, requires explicit user consent.

To conclude, we argue that in the case of cold start and data sparsity scenarios, where the availability and quality of information are the key factors, using the co–authorship network is a promising solution, despite some issues with proper author identification. The extended PR is an example of such an algorithm to generate item–based recommendations. These recommendations can be shown along with the item (publication) that a user visits.

The method proposed in this paper uses relationships between the authors of the publications. In addition, it utilises other available attributes of the articles that improve the recommendations. These are citations, downloads, and view numbers that, among others, indicate the popularity of the publication. The initial evaluation of our algorithm indicates that it considerably alleviates the problem of cold start in EOSC.

X. SUMMARY AND FUTURE WORKS

The recommendation algorithm proposed in this article will further enhance the intelligent recommendation features currently available in EOSC–RS and may facilitate the introduction of new capabilities, such as suggesting research collaborations.

This paper presents the extended Page Rank algorithm to alleviate the cold start and data sparsity problems in EOSC. A cold start is a situation in which the recommendation system cannot provide suggestions for users who have just joined it or cannot recommend a resource that has been added.

We implemented an extended Page Rank algorithm based on common author relationships. The evaluation results confirmed its applicability to recommend papers on academic platforms for various scientific communities, with users occasionally using available resources. Furthermore, the method proposed in this article can be adapted to other domains that struggle with the number of active users, have frequent data updates, and do not have the option to collect rich information to recommend items apart from tracking user behaviour (representing implicit feedback).

Future work will include combining the co–authorship graph with additional information derived from publication records (such as the similarity between papers) to offer recommendations in case of missing authorship connections in the graph. In parallel, we will conduct an in-depth usability evaluation of the recommendation algorithms with real users.

REFERENCES

[1] Z. Zhang, B. G. Patra, A. Yaseen, J. Zhu, R. Sabharwal, K. Roberts, T. Cao, and H. Wu, "Scholarly recommendation systems: A literature survey," *Knowl. Inf. Syst.*, vol. 65, no. 11, pp. 4433–4478, Nov. 2023, doi: 10.1007/s10115-023-01901-x.

[2] F. Beierle, J. Tan, and K. Grunert, "Analyzing social relations for recommending academic conferences," in *Proc. 8th ACM Int. Workshop Hot Topics Planet-Scale Mobile Comput. Social Netw.*, Paderborn, Germany, Jul. 2016, pp. 37–42.

[3] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong, and F. Xia, "Scientific paper recommendation: A survey," *IEEE Access*, vol. 7, pp. 9324–9339, 2019.

[4] J. Son and S. B. Kim, "Academic paper recommender system using multilevel simultaneous citation networks," *Decis. Support Syst.*, vol. 105, pp. 24–33, Jan. 2018.

[5] P. Budroni, J. Claude-Burgelman, and M. Schouppe, "Architectures of knowledge: The European open science cloud," *ABI Technik*, vol. 39, no. 2, pp. 130–141, Jul. 2019.

[6] A. V. D. Almeida, M. M. Borges, and L. Roque, "The European open science cloud: A new challenge for Europe," in *Proc. 5th Int. Conf. Technol. Ecosyst. Enhancing Multiculturality*, Oct. 2017, p. 6.

[7] P. Manghi, N. Houssos, M. Mikulicic, and B. Jörg, "The data model of the openaire scientific communication e-infrastructure," in *Proc. 6th Res. Conf. Metadata Semantics Res. (MTSR)*, in Communications in Computer and Information Science, vol. 343, Cádiz, Spain. Cham, Switzerland: Springer, 2012, pp. 168–180.

[8] A.-M. Mugabushaka, M. Baglioni, A. Bardi, and P. Manghi, "Scholarly outputs of EU research funding programs: Understanding differences between datasets of publications reported by grant holders and OpenAIRE research graph in H2020," 2021, *arXiv:2109.10638*.

[9] A. Bardi and L. Benassi, "Boosting open science in the IPERION HS research infrastructure with openaire," *ERCIM News*, vol. 2023, no. 133, p. 2, 2023.

[10] T. Ferrari, D. Scardaci, and S. Andreozzi, "The open science commons for the European research area," in *Earth Observation Open Science and Innovation* (ISSI Scientific Report Series), vol. 15. Cham, Switzerland: Springer, 2018, pp. 43–68, doi: 10.1007/978-3-319-65633-5_3.

[11] M. Wolski, K. Martyn, M. Łabędzki, M. Xydas, T. Zamani, B. Walter, J. W. Shepherdson, M. Kolomanski, and N. Triantafyllis, "AI/ML recommender system interoperability guideline (1.0.0)," Zenodo, EOSC Future Consortium, Greece, 2023, doi: 10.5281/zenodo.7849178.

[12] M. Wolski, K. Martyn, and B. Walter, "A recommender system for EOSC. Challenges and possible solutions," in *Proc. Int. Conf. Res. Challenges Inf. Sci.*, 2022, pp. 70–87.

[13] M. Singh, "Scalability and sparsity issues in recommender datasets: A survey," *Knowl. Inf. Syst.*, vol. 62, no. 1, pp. 1–43, Jan. 2020.

[14] E. Çano and M. Morisio, "Hybrid recommender systems: A systematic literature review," *Intell. Data Anal.*, vol. 21, no. 6, pp. 1487–1524, Nov. 2017.

[15] D. Roy and M. Dutta, "A systematic review and research perspective on recommender systems," *J. Big Data*, vol. 9, no. 1, p. 59, Dec. 2022.

[16] D. K. Panda and S. Ray, "Approaches and algorithms to mitigate cold start problems in recommender systems: A systematic literature review," *J. Intell. Inf. Syst.*, vol. 59, no. 2, pp. 341–366, Oct. 2022.

[17] P. Manghi, A. Bardi, C. Atzori, M. Baglioni, N. Manola, J. Schirrwagen, P. Principe, M. Artini, A. Becker, and M. De Bonis, ''The OpenAIRE research graph data model,'' Zenodo, Athena Res. Innov. Center, Greece, 2019.

[18] F. Jiang and Z. Wang, ''Pagerank-based collaborative filtering recommendation,'' in *Information Computing and Applications*, R. Zhu, Y. Zhang, B. Liu, and C. Liu, Eds., Berlin, Germany: Springer, 1007, pp. 597–604.

[19] A. Vellino, ''Recommending journal articles with pagerank ratings,'' *Recommender Syst.*, vol. 2009, p. 17, Jan. 2009.

[20] M. Gori and A. Pucci, ''ItemRank: A random-walk based scoring algorithm for recommender engines,'' in *Proc. 20th Int. Jt. Conf. Artif. Intell. (IJCAI)*, Jan. 2007, pp. 2766–2771.

[21] D. Castelli, ''EOSC as a game-changer in the social sciences and humanities research activities,'' in *Proc. Workshop Lang. Resour. SSH Cloud*, D. Broeder, M. Eskevich, and M. Monachini, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 37–38. [Online]. Available: https://aclanthology.org/2020.lr4sshoc-1.7

[22] J.-W. Boiten et al., ''EOSC-LIFE WP4 TOOLBOX: Toolbox for sharing of sensitive data—A concept description,'' 2022, *arXiv:2207.14296*.

[23] F. Zamorano, M. Reyes, G. Espinoza, and E.-P. Reyes, ''Design of an eOSCE for the Chilean healthcare context,'' *Int. J. Online Biomed. Eng.*, vol. 18, no. 13, pp. 56–72, Oct. 2022. [Online]. Available: https://online-journals.org/index.php/i-joe/article/view/32767

[24] B. Mons, ''Comments to jean-claude Burgelman's article politics and open science: How the European open science cloud became reality (the untold story)—'EOSC is a bigger ME' and the dunning Kruger effect,'' *Data Intell.*, vol. 3, no. 1, pp. 32–39, 2021, doi: 10.1162/dint_a_00074. [Online]. Available: https://direct.mit.edu/dint/article/3/1/32/94905/Comments-to-Jean-Claude-Burgelman-s-article

[25] *EOSC Future Marketplace*. Accessed: Dec. 18, 2023. [Online]. Available: https://marketplace.eosc-portal.eu/services

[26] B. W. C. A. Hienola and J. Shepherdson, ''D5.2A EOSC front-office requirements analysis,'' EOSC Future Consortium, Greece, Tech. Rep. D5.2, 2022. [Online]. Available: https://eoscfuture.eu/wp-content/uploads/2024/03/EOSC-Future-WP5-ICOS-D5.2b-Front-Office-Requirement-Analysis-2022-09-28-1.pdf

[27] Accessed: Dec. 18, 2023. [Online]. Available: https://marketplace.eosc-portal.eu

[28] Access Ltd. Accessed: Dec. 18, 2023. *EOSC Metrics Service*. [Online]. Available: https://rseval.eosc.grnet.gr

[29] J. Beel, B. Gipp, S. Langer, and C. Breitinger, ''Research paper recommender systems: A literature survey,'' *Int. J. Digit. Libraries*, pp. 1–34, Jun. 2015.

[30] P.-Y. Chen, E. Hayes, V. Larivière, and C. R. Sugimoto, ''Social reference managers and their users: A survey of demographics and ideologies,'' *PLoS ONE*, vol. 13, no. 7, Jul. 2018, Art. no. e0198033.

[31] S. Ahmadian, M. Meghdadi, and M. Afsharchi, ''Incorporating reliable virtual ratings into social recommendation systems,'' *Appl. Intell.*, vol. 48, no. 11, pp. 4448–4469, Nov. 2018.

[32] X. Zhao, L. Zhang, L. Xia, Z. Ding, D. Yin, and J. Tang, ''Deep reinforcement learning for list-wise recommendations,'' 2017, *arXiv:1801.00209*.

[33] P. Kumar and B. Bhasker, ''DNNRec: A novel deep learning based hybrid recommender system,'' *Expert Syst. Appl.*, vol. 144, Apr. 2020, Art. no. 113054. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419307717

[34] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, ''Neural collaborative filtering,'' in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 173–182.

[35] H. Wang, N. Wang, and D.-Y. Yeung, ''Collaborative deep learning for recommender systems,'' in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, p. 1950.

[36] M. M. Afsar, T. Crump, and B. Far, ''Reinforcement learning based recommender systems: A survey,'' *ACM Comput. Surv.*, vol. 55, no. 7, pp. 1–38, Jul. 2023.

[37] S. Zhang, L. Yao, A. Sun, and Y. Tay, ''Deep learning based recommender system: A survey and new perspectives,'' *ACM Comput. Surv.*, vol. 52, no. 1, pp. 1–38, Feb. 2019, doi: 10.1145/3285029.

[38] C. C. Aggarwal, *Recommender Systems*, vol. 1. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-29659-3.

[39] X. Kong, Y. Shi, S. Yu, J. Liu, and F. Xia, ''Academic social networks: Modeling, analysis, mining and applications,'' *J. Netw. Comput. Appl.*, vol. 132, pp. 86–103, Apr. 2019.

[40] F. Xia, Z. Chen, W. Wang, J. Li, and L. T. Yang, ''MVCWalker: Random walk-based most valuable collaborators recommendation exploiting academic factors,'' *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 364–375, Sep. 2014.

[41] R. Zhao, Z. Liu, X. Zhang, J. Wang, and Z. Zhang, ''Mapping world scientific collaboration on the research of COVID-19: Authors, journals, institutions, and countries,'' 2021, doi: 10.24251/HICSS.2021.059.

[42] F. Xia, H. Liu, I. Lee, and L. Cao, ''Scientific article recommendation: Exploiting common author relations and historical preferences,'' *IEEE Trans. Big Data*, vol. 2, no. 2, pp. 101–112, Jun. 2016.

[43] V. Tishin, A. Sosedka, P. Ibragimov, and V. Porvatov, ''Citation network applications in a scientific co-authorship recommender system,'' in *Proc. Int. Conf. Anal. Images, Social Netw. Texts*, in Lecture Notes in Computer Science, vol. 13217, E. Burnaev, D. I. Ignatov, S. Ivanov, M. Y. Khachay, O. Koltsova, A. Kutuzov, S. O. Kuznetsov, N. V. Loukachevitch, A. Napoli, A. Panchenko, P. M. Pardalos, J. Saramäki, A. V. Savchenko, E. Tsymbalov, and E. Tutubalina, Eds., Tbilisi, Georgia. Cham, Switzerland: Springer, 2021, pp. 293–299, doi: 10.1007/978-3-031-16500-9.

[44] J. Beel and S. Dinesh, ''Real-world recommender systems for academia: The pain and gain in building, operating, and researching them,'' in *Proc. 5th Workshop Bibliometric-Enhanced Inf. Retr. (BIR)*, vol. 1823, Aberdeen, U.K., Apr. 2017, pp. 6–17.

[45] W. Waheed, M. Imran, B. Raza, A. K. Malik, and H. A. Khattak, ''A hybrid approach toward research paper recommendation using centrality measures and author ranking,'' *IEEE Access*, vol. 7, pp. 33145–33158, 2019.

[46] M. K. Pandia and A. Bihari, ''Important author analysis in research professionals' relationship network based on social network analysis metrics,'' in *Computational Intelligence in Data Mining*, vol. 3, L. C. Jain, H. S. Behera, J. K. Mandal, and D. P. Mohapatra, Eds., New Delhi, India: Springer, 2015, pp. 185–194.

[47] V. A. Rohani, Z. M. Kasirun, S. Kumar, and S. Shamshirband, ''An effective recommender algorithm for cold-start problem in academic social networks,'' *Math. Problems Eng.*, vol. 2014, pp. 1–11, Mar. 2014, Art. no. 123726.

[48] S. Ahmadian, M. Afsharchi, and M. Meghdadi, ''A novel approach based on multi-view reliability measures to alleviate data sparsity in recommender systems,'' *Multimedia Tools Appl.*, vol. 78, no. 13, pp. 17763–17798, Jul. 2019.

[49] S. Ahmadian, N. Joorabloo, M. Jalili, and M. Ahmadian, ''Alleviating data sparsity problem in time-aware recommender systems using a reliable rating profile enrichment approach,'' *Expert Syst. Appl.*, vol. 187, Jan. 2022, Art. no. 115849. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421012100

[50] J. Bobadilla, A. Gutiérrez, F. Ortega, and B. Zhu, ''Reliability quality measures for recommender systems,'' *Inf. Sci.*, vols. 442–443, pp. 145–157, May 2018, doi: 10.1016/j.ins.2018.02.030.

[51] R. Torres, S. M. McNee, M. Abel, J. A. Konstan, and J. Riedl, ''Enhancing digital libraries with TechLens+,'' in *Proc. 4th ACM/IEEE-CS Joint Conf. Digit. Libraries*, Jun. 2004, pp. 228–236.

[52] C. Yang, B. Wei, J. Wu, Y. Zhang, and L. Zhang, ''CARES: A ranking-oriented CADAL recommender system,'' in *Proc. 9th ACM/IEEE-CS Joint Conf. Digit. Libraries*, Jun. 2009, pp. 203–212.

[53] F. Tahmasebi, M. Meghdadi, S. Ahmadian, and K. Valiallahi, ''A hybrid recommendation system based on profile expansion technique to alleviate cold start problem,'' *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 2339–2354, Jan. 2021.

[54] G. Adomavicius and A. Tuzhilin, ''Context-aware recommender systems,'' in *Recommender Systems Handbook*. Cham, Switzerland: Springer, 2010, pp. 217–253.

[55] B. Gipp, J. Beel, and C. Hentschel, ''Scienstein: A research paper recommender system,'' in *Proc. Int. Conf. Emerg. trends Comput. (ICETiC)*, 2009, pp. 309–315.

[56] S. Brin and L. Page, ''The anatomy of a large-scale hypertextual web search engine,'' *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 107–117, Apr. 1998.

[57] N. Ma, J. Guan, and Y. Zhao, ''Bringing PageRank to the citation analysis,'' *Inf. Process. Manage.*, vol. 44, no. 2, pp. 800–810, Mar. 2008.

[58] M. Gori and A. Pucci, ''Research paper recommender systems: A random-walk based approach,'' in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, Dec. 2006, pp. 778–781.

[59] J. Beel and B. Gipp, ''Google scholar's ranking algorithm: An introductory overview,'' in *Proc. 12th Int. Conf. Scientometr. Informetr. (ISSI)*, Rio de Janeiro, Brazil, 2009, pp. 230–241.

[60] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel, "Co-authorship networks in the digital library research community," *Inf. Process. Manage.*, vol. 41, no. 6, pp. 1462–1480, Dec. 2005.

[61] J. Bollen, M. A. Rodriquez, and H. Van de Sompel, "Journal status," *Scientometrics*, vol. 69, pp. 669–687, Jan. 2006.

[62] E. Yan and Y. Ding, "Discovering author impact: A PageRank perspective," *Inf. Process. Manage.*, vol. 47, no. 1, pp. 125–134, Jan. 2011.

[63] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," in *Proc. Web Conf.*, 1999, p. 19134. [Online]. Available: https://api.semanticscholar.org/CorpusID:1508503

[64] S. Park, W. Lee, B. Choe, and S.-G. Lee, "A survey on personalized PageRank computation algorithms," *IEEE Access*, vol. 7, pp. 163049–163062, 2019.

[65] F. Xia, J. Liu, H. Nie, Y. Fu, L. Wan, and X. Kong, "Random walks: A review of algorithms and applications," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 2, pp. 95–107, Apr. 2020.

[66] J. Shen, Y. Du, W. Wang, and X. Li, "Lazy random walks for superpixel segmentation," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1451–1462, Apr. 2014.

[67] M. Wolski, A. Klorek, C. Mazurek, and A. Kobusinska, "Linking scholarly datasets—The EOSC perspective," in *Proc. 23rd Int. Conf. Comput. Sci. (ICCS)*, in Lecture Notes in Computer Science, vol. 14073, Prague, Czech Republic. Cham, Switzerland: Springer, 2023, pp. 1–16.

[68] J. Priem, H. Piwowar, and R. Orr, "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," 2022, *arXiv:2205.01833*.

[69] A. Klorek and M. Wolski. *Dataset: OpenAire Research Graph Linked With OpenAlex*. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.13365368

[70] Z. Wang, Q. Xu, Z. Yang, X. Cao, and Q. Huang, "Implicit feedbacks are not always favorable: Iterative relabeled one-class collaborative filtering against noisy interactions," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3070–3078.

[71] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: Introduction and challenges," in *Recommender Systems Handbook*. Cham, Switzerland: Springer, 2015, pp. 1–34.

[72] *A Bogus Reserchers With ORCID*. [Online]. Available: https://orcid.org/0000-0002-1825-0097

MARCIN WOLSKI received the Graduate degree from the Institute of Computer Science, Poznań University of Technology. He is currently a Senior Researcher with Poznań Supercomputing and Networking Centre. He has over 15 years of experience developing advanced information services, database technologies, and software development. Within the confines of the EOSC future project, he has been coordinating the development and implementation of the EOSC recommendation system. For many years, he has been responsible for introducing a structured approach to software development processes with GEANT. Facilitator of software best practices and quality assurance of open–source software. The author or co-author of numerous publications concerning software management, software quality, and recommendation systems.

ANTONI KLOREK was born in Poznań, Poland, in 2000. He received the B.S. degree in artificial intelligence from Poznań University of Technology, Poznań, in 2023, where he is currently pursuing the M.S. degree in artificial intelligence. Since 2022, he has been a Machine Learning Engineer with Poznań Supercomputing and Networking Center, Poznań. His research interests include natural language processing, large language models, and retrieval augmented generation.

ANNA KOBUSINSKA (Member, IEEE) received the Ph.D. and Habilitation degrees in computer science from Poznań University of Technology. She is currently the Head of the Distributed Systems Group, Faculty of Informatics and Telecommunications, Poznań University of Technology, and an Adjunct Professor with the Graduate Institute of Artificial Intelligence, Chang Gung University, Taiwan. Her research interests include large-scale distributed systems, service-oriented systems, cloud computing, and edge computing. In her work, she focuses on distributed algorithms, big data analysis, and reliability of distributed processing, specifically consistency models and replication techniques, challenges associated with the use of blockchain technology, and edge intelligence. She has served and is currently serving as a GC, PC, and TPC member for several international conferences and workshops. She has authored many publications in high-quality, peer-reviewed international conferences and journals. She has also participated in various research projects supported by national and EC organizations in collaboration with academic institutions and industrial partners.