International Journal on Digital Libraries
https://doi.org/10.1007/s00799-022-00339-w

Scientific paper recommendation systems: a literature review of recent publications

Christin Katharina Kreutz[1] · **Ralf Schenkel**[2]

Received: 23 December 2021 / Revised: 7 September 2022 / Accepted: 16 September 2022

Abstract
Scientific writing builds upon already published papers. Manual identification of publications to read, cite or consider as related papers relies on a researcher's ability to identify fitting keywords or initial papers from which a literature search can be started. The rapidly increasing amount of papers has called for automatic measures to find the desired *relevant* publications, so-called paper recommendation systems. As the number of publications increases so does the amount of paper recommendation systems. Former literature reviews focused on discussing the general landscape of approaches throughout the years and highlight the main directions. We refrain from this perspective, instead we only consider a comparatively small time frame but analyse it fully. In this literature review we discuss used methods, datasets, evaluations and open challenges encountered in all works first released between January 2019 and October 2021. The goal of this survey is to provide a comprehensive and complete overview of current paper recommendation systems.

Keywords Paper recommendation system · Publication suggestion · Literature review

1 Introduction

The rapidly increasing number of publications leads to a large quantity of possibly relevant papers [6] for more specific tasks such as finding related papers [28], finding ones to read [109] or literature search in general to inspire new directions and understand the state-of-the-art approaches [46]. Overall researchers typically spend a large amount of time on searching for relevant related work [7]. Keyword-based search options are insufficient to find relevant papers [9,52,109], they require some form of initial knowledge about a field. Oftentimes, users' information needs are not explicitly specified [56] which impedes this task further.

To close this gap, a plethora of paper recommendation systems have been proposed recently [37,39,88,104,117]. These systems should fulfil different functions: for junior researchers systems should recommend a broad variety of papers, for senior ones the recommendations should align more with their already established interests [9] or help them discover relevant interdisciplinary research [100]. In general paper recommendation approaches positively affect researchers' professional lives as they enable finding relevant literature more easily and faster [50].

As there are many different approaches, their objectives and assumptions are also diverse. A simple problem definition of a paper recommendation system could be the following: given one paper recommend a list of papers fitting the source paper [68]. This definition would not fit all approaches as some specifically do not require any initial paper to be specified but instead observe a user as input [37]. Some systems recommend sets of publications fitting the queried terms only if these papers are all observed together [60,61], most of the approaches suggest a number of single publications as their result [37,39,88,117], such that any single one of these papers satisfies the information need of a user fully. Most approaches assume that all required data to run a system is present already [37,117] but some works [39,88] explicitly crawl general publication information or even abstracts and keywords from the web.

In this literature review we observe papers recently published in the area of scientific paper recommendation between

✉ Christin Katharina Kreutz
christin.kreutz@th-koeln.de

Ralf Schenkel
schenkel@uni-trier.de

[1] Cologne University of Applied Sciences, Cologne, Germany

[2] Trier University, Trier, Germany

Published online: 05 October 2022

and including January 2019 and October 2021[1]. We strive to give comprehensive overviews on their utilised methods as well as their datasets, evaluation measures and open challenges of current approaches. Our contribution is fourfold:

- We propose a current multidimensional characterisation of current paper recommendation approaches.
- We compile a list of recently used datasets in evaluations of paper recommendation approaches.
- We compile a list of recently used evaluation measures for paper recommendation.
- We analyse existing open challenges and identify current novel problems in paper recommendation which could be specifically helpful for future approaches to address.

In the following Sect. 2 we describe the general problem statement for paper recommendation systems before we dive into the literature review in Sect. 3. Section 4 gives insight into datasets used in current work. In the following Sect. 5 different definitions of relevance, relevance assessment as well as evaluation measures are analysed. Open challenges and objectives are discussed in detail in Sect. 7. Lastly Sect. 8 concludes this literature review.

2 Problem statement

Over the years different formulations for a problem statement of a paper recommendation system have emerged. In general they should specify the input for the recommendation system, the type of recommendation results, the point in time when the recommendation will be made and which specific goal an approach tries to achieve. Additionally, the target audience should be specified.

As *input* we can either specify an initial paper [28], keywords [117], a user [37], a user and a paper [5] or more complex information such as user-constructed knowledge graphs [109]. Users can be modelled as a combination of features of papers they interacted with [19,21], e.g. their clicked [26] or authored publications [22]. Papers can for example be represented by their textual content [88].

As types of recommendation we could either specify single (independent) papers [37] or a set of papers which is to be observed completely to satisfy the information need [61]. A study by Beierle et al. [18] found that existing digital libraries recommend between three and ten single papers, in their case the optimal number of suggestions to display to users was five to six.

As for the *point in time*, most work focuses on immediate recommendation of papers. Only a few approaches also consider delayed suggestion[2] via newsletter for example [56].

In general, recommended papers should be relevant in one way or another to achieve certain *goals*. The intended goal of authors of papers could, e.g. either be to recommend papers which should be read [109] by a user or recommend papers which are simply somehow related to an initial paper [28], by topic, citations or user interactions.

Different *target audiences*, for example junior or senior researcher, have different demands from paper recommendation systems [9]. Usually paper recommendation approaches target single users but there are also works which strive to recommend papers for sets of users [110,111].

3 Literature review

In this chapter we first clearly define the scope of our literature review (see Sect. 3.1) before we conduct a meta-analysis on the observed papers (see Sect. 3.2). Afterwards our categorisation or lack thereof is discussed in depth (see Sect. 3.3), before we give short overviews of all paper recommendation systems we found (see Sect. 3.5) and some other relevant related work (see Sect. 3.6).

3.1 Scope

To the best of our knowledge the literature reviews by Bai et al. [9], Li and Zou [58] and Shahid et al. [92] are the most recent ones targeting the domain of scientific paper recommendation systems. They were accepted for publication or published in 2019 so they only consider paper recommendation systems up until 2019 at most. We want to bridge the gap between papers published after their surveys were finalised and current work so we only focus on the discussion of publications which appeared between January 2019 and October 2021 when this literature search was conducted.

We conducted our literature search on the following digital libraries: ACM[3], dblp[4], GoogleScholar[5] and Springer[6]. Titles of considered publications had to contain either *paper*, *article* or *publication* as well as some form of *recommend*.

[1] The most recent surveys [9,58,92] focusing on scientific paper recommendation appeared in 2019 such that this time frame is not yet covered.

[2] Non-immediate variants allow using methods which require more time to compute recommendations. Temporal patterns of user behaviour could be incorporated in the recommendation process to identify a fitting moment to present new recommendations to a user. The moment a recommendation is presented to a user influences their interest, as the delayed recommendation might no longer be relevant or does not fit the current task of a user.

[3] https://dl.acm.org/.

[4] https://dblp.uni-trier.de/.

[5] https://scholar.google.com/.

[6] https://link.springer.com/.

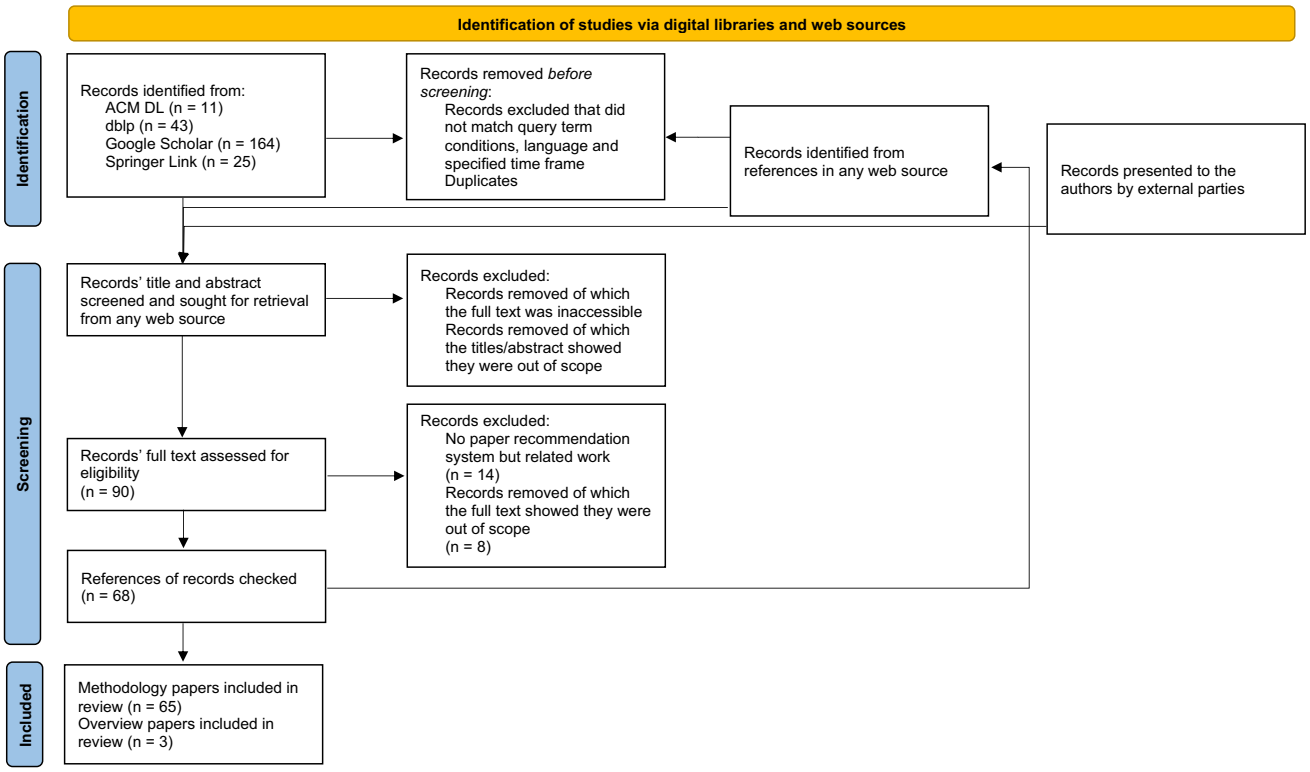

Fig. 1 PRISMA workflow of our literature review process

Papers had to be written in English to be observed. We judged relevance of retrieved publications by observing titles and abstracts if the title alone did not suffice to assess their topical relevance. In addition to these papers found by systematically searching digital libraries, we also considered their referenced publications if they were from the specified time period and of topical fit. For all papers their date of first publication determines their publication year which decides if they lie in our time observed time frame or not. For example, for journal articles we consider the point in time when they were first published online instead of the date on which they were published in an issue, for conference articles we consider the date of the conference instead a later date when they were published online. Figure 1 depicts the PRISMA [79] workflow for this study.

We refrain from including works in our study which do not identify as scientific paper recommendation systems such as Wikipedia article recommendation [70,78,85] or general news article recommendation [33,43,103]. Citation recommendation systems [72,90,124] are also out of scope of this literature review. Even though citation and paper recommendation can be regarded as analogous [45], we argue the differing functions of citations [34] and tasks of these recommendation systems [67] should not be mixed with the problem of paper recommendation. Färber and Jatowt [32] also support this view by stating that both are disjunctive, with paper recommendation pursuing the goal of providing papers to read and investigate while incorporating user interaction data and citation recommendation supporting users with finding citations for given text passages.[7] We also consciously refrain from discussing the plethora of more area-independent recommender systems which could be adopted to the domain of scientific paper recommendation.

Our literature research resulted in 82 relevant papers. Of these, three were review articles. We found 14 manuscripts which do not present paper recommendation systems but are relevant works for the area nonetheless, they are discussed in Sect. 3.6. This left 65 publications describing paper recommendation systems for us to analyse in the following.

3.2 Meta analysis

For papers within our scope, we consider their publication year as stated in the citation information for this meta-analysis. This could affect the publication year of papers compared to the former definition of which papers are included in this survey. For example, for journal articles we do not set the publication year as the point in time when they were first published online, instead for consistency (this

[7] For a survey of current trends in citation recommendation refer to Färber and Jatowt [32].

Table 1 Top most common venues where relevant papers were published together with their type and number of papers (#p). Other venues had only one associated paper

Type	Venue	#p
Journal	IEEE Access	5
Journal	Scientometrics	2
Journal	PeerJ CS	2
Conference	WWW	2
Conference	ChineseCSCW	2
Conference	CSCWD	2

data is present in the citation information of papers) for this analysis we use the year the issue was published in which the article is contained. Of the 65 relevant system papers, 21 were published in 2019, 23 were published in 2020 and 21 were published in 2021. On average each paper has 4.0462 authors (std. dev. = 1.6955) and 12.4154 pages (std. dev. = 9.2402). 35 (53.85%) of the papers appeared as conference papers, 27 (41.54%) papers were published in journals and there were two preprints (3.08%) which have not yet been published otherwise. There has been one master's thesis (1.54%) within scope. The most common venues for publications were the ones depicted in Table 1. Some papers [74–76,93,94] described the same approach without modification or extension of the actual paper recommendation methodology, e.g. by providing evaluations[8]. This left us with 62 different paper recommendation systems to discuss.

3.3 Categorisation

3.3.1 Former categorisation

The already mentioned three most recent [9,58,92] and one older but highly influential [16] literature reviews in scientific paper recommendation utilise different categorisations to group approaches. Beel et al. [16] categorise observed papers by their underlying recommendation principle into stereotyping, content-based filtering, collaborative filtering, co-occurrence, graph-based, global relevance and hybrid models. Bai et al. [9] only utilise the classes content-based filtering, collaborative filtering, graph-based methods, hybrid methods and other models. Li and Zou [58] use the categories content-based recommendation, hybrid recommendation, graph-based recommendation and recommendation based on deep learning. Shahid et al. [92] label approaches by the criterion they identify relevant papers with: content, metadata, collaborative filtering and citations.

The four predominant categories thus are content-based filtering, collaborative filtering, graph-based and hybrid systems. Most of these categories are defined precisely but graph-based approaches are not always characterised concisely: *Content-based filtering* (CBF) methods are said to be ones where user interest is inferred by observing their historic interactions with papers [9,16,58]. Recommendations are composed by observing features of papers and users [5]. In *collaborative filtering* (CF) systems the preferences of users similar to a current one are observed to identify likely relevant publications [9,16,58]. Current users' past interactions need to be similar to similar users' past interactions [9,16]. *Hybrid* approaches are ones which combine multiple types of recommendations [9,16,58].

Graph-based methods can be characterised in multiple ways. A very narrow definition only encompasses ones which observe the recommendation task as a link prediction problem or utilise random walk [5]. Another less strict definition identifies these systems as ones which construct networks of papers and authors and then apply some graph algorithm to estimate relevance [9]. Another definition specifies this class as one using graph metrics such as random walk with restart, bibliographic coupling or co-citation inverse document frequency [106]. Li and Zhou [58] abstain from clearly characterising this type of systems directly but give examples which hint that in their understanding of graph-based methods somewhere in the recommendation process, some type of graph information, e.g. bibliographic coupling or co-citation strength, should be used. Beel et al. [16] as well as Bai et al. [9] follow a similar line, they characterise graph-based methods broadly as ones which build upon the existing connections in a scientific context to construct a graph network.

When trying to classify approaches by their recommendation type, we encountered some problems:

1. We have to refrain from only utilising the labels the works give themselves (see Table 2 for an overview of self-labels of works which do classify themselves). Works do not necessarily (clearly) state, which category they belong to [28,49,60]. Another problem with self-labelling is authors' individual definitions of categories while disregarding all possible ones (as e.g. seen with Afsar et al. [1] or Ali et al. [5]). Mis-definition or omitting of categories could lead to an incorrect classification.
2. When considering the broadest definition of graph-based methods many recent paper recommendation systems tend to belong to the class of hybrid methods. Most of the approaches [5,46,48,49,57,88,105,117] utilise some type of graph structure information as part of the approach which would classify them as graph-based but as they

[8] These papers could either be a demo paper and a later published full paper or the conference and journal version of the same approach, which is then slightly extended by more experiments. These paper clusters are no exact duplicates or fraudulent publications.

Table 2 Indications as what type of paper recommendation system works describe themselves with indication if the description is a common used label (c)

Work	Label	c
[1]	Knowledge-based	×
[3]	Hybrid	✓
[4]	Deep learning-based	✓
[5]	Unified model	×
[19]	Graph-based	✓
[21]	User-specific	×
[24]	Hybrid	✓
[29]	Graph-based	✓
[30]	Active one-shot learning	×
[37]	Collaborative filtering	✓
[39]	Hybrid	✓
[41]	Hybrid	✓
[44]	Hybrid	✓
[45]	Hybrid	✓
[46]	hybrid	✓
[55]	Hybrid	✓
[57]	Network-based	×
[59]	Content-based	✓
[61]	Graph-based	✓
[62]	Neuro-collaborative filtering	×
[63]	Meta-path based	×
[64]	Heterogeneous graph representation based	×
[65]	Social network-based	×
[69]	Hybrid	✓
[71]	Content-based	✓
[74–76]	Content-based	✓
[84]	Hybrid	✓
[86]	Content-based	✓
[89]	Collaborative filtering	✓
[88]	Hybrid	✓
[93,94]	In-text citation frequencies-based	×
[96]	Hybrid	✓
[98]	content-based	✓
[104]	Hybrid	✓
[106]	Graph-based	✓
[108]	Hybrid	✓
[113]	Knowledge-aware path recurrent network	×
[109]	Graph-based	✓
[110]	Hybrid	✓
[111]	Hybrid	✓
[117]	Hybrid	✓
[118]	Network	×
[123]	Hybrid	✓

also utilise historic user-interaction data or descriptions of paper features (see, e.g. Li et al. [57] who describe their approach as network-based while using a graph structure, textual components and user profiles) which would render them as either CF or CBF also.

Thus we argue the former categories do not suffice to classify the particularities of current approaches in a meaningful way. So instead, we introduce more dimensions by which systems could be grouped.

3.3.2 Current categorisation

Recent paper recommendation systems can be categorised in 20 different dimensions by general information on the approach (G), already existing data directly taken from the papers used (D) and methods which might create or (re-)structure data, which are part of the approach (M):

- (G) Personalisation (person.): The approach produces personalised recommendations. The recommended items depend on the person using the approach, if personalisation is not considered, the recommendation solely depends on the input keywords or paper. This dimension is related to the existence of user profiles.
- (G) Input: The approach requires some form of input, either a paper (p), keywords (k), user (u) or something else, e.g. an advanced type of input (o). Hybrid forms are also possible. In some cases the input is not clearly specified throughout the paper so it is unknown (?).
- (D) Title: The approach utilises titles of papers.
- (D) Abstract (abs.): The approach utilises abstracts of papers.
- (D) Keyword (key.): The approach utilises keywords of papers. These keywords are usually explicitly defined by the authors of papers, contrasting key phrases.
- (D) Text: The approach utilises some type of text of papers which is not clearly specified as titles, abstracts or keywords. In the evaluation this approach might utilise specified text fragments of publications.
- (D) Citation (cit.): The approach utilises citation information, e.g. numbers of citations or co-references.
- (D) Historic interaction (inter.): The approach uses some sort of historic user-interaction data, e.g. previously authored, cited or liked publications. An approach can only include historic user-interaction data if it also somehow contains user profiles.
- (M) User profile (user): The approach constructs some sort of user profile or utilises profile information. Most approaches using personalisation also construct user profiles but some do not explicitly construct profiles but rather encode user information in the used structures.

- (M) Popularity (popul.): The approach utilises some sort of popularity indication, e.g. CORE rank, numbers of citations[9] or number of likes.
- (M) Key phrase (KP): The approach utilises key phrases. Key phrases are not explicitly provided by authors of papers but are usually computed from the titles and abstracts of papers to provide a descriptive summary, contrasting keywords of papers.
- (M) Embedding (emb.): The approach utilises some sort of text or graph embedding technique, e.g. BERT or Doc2Vec.
- (M) Topic model (TM): The approach utilises some sort of topic model, e.g. LDA.
- (M) Knowledge graph (KG): The approach utilises or builds some sort of knowledge graph. This dimension surpasses the mere incorporation of a graph which describes a network of nodes and edges of different types. A knowledge graph is a sub-category of a graph.
- (M) Graph: The approach actively builds or directly uses a graph structure, e.g. a knowledge graph or scientific heterogeneous network. Utilisation of a neural network is not considered in this dimension.
- (M) Meta-path (path): The approach utilises meta-paths. They usually are composed from paths in a network.
- (M) Random Walk (with Restart) (RW): The approach utilises Random Walk or Random Walk with Restart.
- (M) Advanced machine learning (AML): The approach utilises some sort of advanced machine learning component in its core such as a neural network. Utilisation of established embedding methods which themselves use neural networks (e.g. BERT) are not considered in this dimension. We do not consider traditional and simple ML techniques such as k means in this dimension but rather mention methods explicitly defining a loss function, using multi-layer perceptrons or GCNs.
- (M) Crawling (crawl.): The approach conducts some sort of web crawling step.
- (M) Cosine similarity (cosine): The approach utilises cosine similarity at some point.

Of the observed paper recommendation systems, six were general systems or methods which were only applied on the domain of paper recommendation [3,4,24,60,118,121]. Two were targeting explicit set-based recommendation of publications where only all papers in the set together satisfy users' information needs [60,61], two recommend multiple papers [42,71] (e.g. on a path [42]), all the other approaches focused on recommendation of k single papers. Only two approaches focus on recommendation of papers to user groups instead of single users [110,111]. Only one paper [56] supports subscription-based recommendation of papers, all other approaches solely regarded a scenario in which papers were suggested straight away.

Table 3 classifies the observed approaches according to the afore discussed dimensions.

[9] The number of citations can be regarded both as an input data as well as a method to denote popularity.

3.4 Comparison of paper recommendation systems in different categories

In this Section, we describe the scientific directions associated with the categories we presented in the previous section as the 65 relevant publications. We focus only on the methodological categories and describe how they are incorporated in the respective approaches.

3.4.1 User profile

32 approaches construct explicit user profiles. They utilise different components to describe users. We differentiate between profiles derived from user interactions and ones derived from papers.

Most user profiles are constructed from *users' actual interactions*: unspecified historical interaction [30,37,56,57,64,118], the mean of the representation of interacted with papers [19], time decayed interaction behaviour [62], liked papers [69,123], bookmarked papers [84,119], read papers [111,113], rated papers [3,4,110], clicked on papers [24,26,49], categories of clicked papers [1], features of clicked papers [104], tweets [74–76], social interactions [65] and explicitly defined topics of interest tags [119].

Some approaches derived user profiles from *users' written papers*: authored papers [5,21,22,55,63,74–76,116], a partitioning of authored papers [27], research fields of authored papers [41] and referenced papers [116].

3.4.2 Popularity

We found 13 papers using some type of popularity measure. Those can be defined on authors, venues or papers.

For *author-based popularity* measures we found unspecified ones [65] such as authority [116] as well as ones regarding the citations an author received: citation count of papers [22,96,108,119], change in citation count [25,26], annual citation count [26], number of citations related to papers [59], h-index [26]. We found two definitions of author's popularity using the graph structure of scholarly networks, namely the number of co-authors [41] and a person's centrality [108].

For *venue-based popularity* measures, we found an unspecific reputation notion [116] as well as incorporation of the impact factor [26,117].

For *paper-based popularity* measures we encountered some citation-based definitions such as vitality [117], citation

Table 3 Indications whether works utilise the specific data or methods. Papers describing the same approach without extension of the methodology (e.g. only describing more details or an evaluation) are regarded in combination with each other

Work	General		Data						Methods											
	Person.	Input	Title	Abs.	Key.	Text	Citat.	Inter.	User	Popul.	KP	Emb.	TM	KG	Graph	Path	RW	AML	Crawl.	Cosine
[1]	•	u	•					•	•					•	•			•		
[2]		p	•	•			•				•				•				•	•
[3]	•	u				•		•	•				•					•		
[4]	•	u				•		•	•									•		
[5]	•	pu				•	•	•	•			•	•		•			•		•
[19]	•	u				•		•	•			•			•			•		
[22]	•	u	•		•		•	•	•	•										
[21]	•	u	•	•	•				•			•								•
[25]		k			•	•	•			•										
[24]	•	k			•	•	•		•	•										
[26]	•	ku	•					•	•			•	•					•		
[27]		u				•		•	•				•							
[28]		p	•	•							•	•								•
[29]		p				•	•					•			•		•	•		•
[30]	•	pu				•		•	•			•						•		•
[37]	•	u	•	•				•	•			•						•		
[38]		p				•	•												•	
[39]		p	•	•			•												•	•
[41]	•	u					•	•	•	•		•			•			•		•
[42]		p				•									•	•	•	•		•
[44]		p				•	•					•			•		•			
[45]		p	•	•	•		•					•			•					
[46]		p				•	•				•				•				•	•
[48]		p				•	•					•			•		•	•		•
[49]	•	u						•	•						•			•		
[56]	•	u	•	•	•		•	•	•	•		•		•	•			•		•
[55]	•	k	•	•	•			•	•			•			•					•
[57]	•	u				•	•	•	•						•	•	•	•		
[59]		k			•		•			•										•
[60]		k			•		•								•					•
[61]		k			•		•								•					
[62]	•	u	•	•				•	•			•						•		
[63]	•	u						•	•						•	•	•	•		

Table 3 continued

Work	General		Data						Methods											
	Person.	Input	Title	Abs.	Key.	Text	Citat.	Inter.	User	Popul.	KP	Emb.	TM	KG	Graph	Path	RW	AML	Crawl.	Cosine
[64]	•	u	•	•	•			•	•			•			•	•		•		•
[65]	•	pu	•		•		•	•	•	•					•		•			
[69]	•	u	•	•				•	•									•		
[71]	•	p	•									•						•		
[74–76]	•	u				•			•											•
[84]	•	u	•	•	•			•	•	•										•
[86]		p			•	•					•									•
[89]		p					•								•					
[88]		p	•	•	•		•								•				•	•
[93,94]		p				•	•												•	
[95]		p				•								•	•					•
[96]		p	•	•			•			•					•			•		
[98]		k				•							•						•	•
[104]	•	u	•	•	•		•	•	•			•		•	•			•		
[106]		p				•	•								•		•			
[107]		p				•	•					•	•		•		•			•
[108]		p					•			•					•					
[113]	•	u			•	•	•	•	•			•		•	•			•		•
[109]	•	ko	•	•	•							•		•	•				•	•
[110]	•	u				•		•	•									•	•	•
[111]	•	u			•			•	•									•		•
[115]	•	ku			•	•	•	•	•	•		•						•		
[116]		pk	•	•	•		•						•					•		
[117]		k	•	•			•			•			•		•		•			•
[118]	•	u						•	•			•						•		
[119]	•	o			•		•	•	•	•		•								
[121]		?	•	•			•								•		•			•
[122]		?	•	•			•								•		•			•
[123]	•	u	•	•				•	•			•						•		

count of papers [22] and theirs centrality [96] in the citation network. Additionally, some approaches incorporated less formal interactions: number of downloads [56], social media mentions [119] and normalised number of bookmarks [84].

3.4.3 Key phrase

Only four papers use key phrases in some shape or form: Ahmad and Afzal [2] construct key terms from preprocessed titles and abstracts using tf-idf to represent papers. Collins and Beel [28] use the Distiller Framework [12] to extract uni-, bi- and tri-gram key phrase candidates from tokenised, part-of-speech tagged and stemmed titles and abstracts. Key phrase candidates were weighted and the top 20 represent candidate papers. Kang et al. [46] extract key phrases from CiteSeer to describe the diversity of recommended papers. Renuka et al. [86] apply rapid automatic keyword extraction.

In summary, different length key phrases usually get constructed from titles and abstracts with automatic methods such as tf-idf or the Distiller Framework to represent the most important content of publications.

3.4.4 Embedding

We found a lot of approaches utilising some form of embedding based on existing document representation methods. We distinguish by embedding of papers, users and papers and sophisticated embedding from the proposed approaches.

Among the most common methods was their application on *papers*: in an unspecified representation [30,119], Word2Vec [19,37,44,45,55,104,113], Word2Vec of LDA top words [24,107], Doc2vec [21,28,48,62,63,107], Doc2Vec of word pairs [109], BERT [123] and SBERT [5,19]. Most times these approaches do not mention which part of the paper to use as input but some specifically mention the following parts: titles [37], titles and abstracts [28,45], titles, abstracts and bodies [48], keywords and paper [119].

Few approaches observed *user profiles and papers*, here Word2Vec [21] and NPLM [29] embeddings were used.

Several approaches embed the information in their own model embedding: a heterogeneous information network [5], a two-layer NN [37], a scientific social reference network [41], the TransE model [56], node embeddings [63], paper, author and venue embedding [116], user and item embedding [118], a GRU and association rule mining model [71], a GCN embedding of users [104] and an LSTM model [113].

3.4.5 Topic model

Eight approaches use some topic modelling component. Most of them use LDA to represent papers' content [3,5,24,27,107,117]. Only two of them do not follow this method: Subathra and Kumar [98] use LDA on papers to find their top n words, then they use LDA again on these words' Wikipedia articles. Xie et al. [115] use a hierarchical LDA adoption on papers, which introduces a discipline classification.

3.4.6 Knowledge graph

Only six of the observed papers incorporate knowledge graphs. Only one uses a predefined one, the Watson for Genomics knowledge graph [95]. Most of the approaches build their own knowledge graphs, only one *asks users to construct* the graphs: Wang et al. [109] build two knowledge graphs, one in-domain and one cross-domain graph. The graphs are user-constructed and include representative papers for the different concepts.

All other approaches *do not rely on users* building the knowledge graph: Afsar et al. [1] utilise an expert-built knowledge base as a source for their categorisation of papers, which are then recommended to users. Li et al. [56] employ a knowledge graph-based embedding of authors, keywords and venues. Tang et al. [104] link words with high tf-idf weights from papers to LOD and then merge this knowledge graph with the user-paper graph. Wang et al. [113] construct a knowledge graph consisting of users and papers.

3.4.7 Graph

In terms of graphs, we found 33 approaches explicitly mentioning the graph structure they were utilising. We can describe which graph structure is used and which algorithms or methods are applied on the graphs.

Of the observed approaches, most specify some form of (heterogeneous) *graph structure*. Only a few of them are unspecific and mention an undefined heterogeneous graph [63–65] or a multi-layer [48] graph. Most works clearly define the type of graph they are using: author-paper-venue-label-topic graph [5], author-paper-venue-keyword graph [56,57], paper-author graph [19,29,55,104], paper-topic graph [29], author-paper-venue graph [42,121,122], author graph [41], paper-paper graph [42,49], citation graph [2,44–46,88,89,106,108,117] or undirected citation graph [60,61]. Some approaches specifically mention usage of co-citations [2,45], bibliographic coupling or both [88,89,96,108].

As for *algorithms or methods used on these graphs*, we encountered usage of centrality measures in different graph types [41,96,108], some use knowledge graphs (see Sect. 3.4.6), some using meta-paths (see Sect. 3.4.8), some using random walks e.g. in form of PageRank or hubs and authorities (see Sect. 3.4.9), construction of Steiner trees [61], usage of the graph as input for a GCN [104], BFS [113], clustering [117] or calculation of a closeness degree [117].

3.4.8 Meta-path

We found only four approaches incorporating meta-paths. Hua et al. [42] construct author-paper-author and author-paper-venue-paper-author paths by applying beam search. Papers on the most similar paths are recommended to users. Li et al. [57] construct meta-paths of a max length between users and papers and use random walk on these paths. Ma et al. [63,64] use meta-paths to measure the proximity between nodes in a graph.

3.4.9 Random walk (with restart)

We found twelve approaches using some form of random walk in their methodology. We differentiate between ones using random walk, random walk with restart and algorithms using a random walk component.

Some methods use *random walk* on heterogeneous graphs [29,65] and weighted multi-layer graphs [48]. A few approaches use random walk to identify [42,57] or determine the proximity between [64] meta-paths.

Three approaches explicitly utilise *random walk with restart*. They determine similarity between papers [106], identify papers to recommend [44] or find most relevant papers in clusters [117].

Some approaches use algorithms which *incorporate a random walk component*: PageRank [107] and the identifications of hubs and authorities [122] with PageRank [121].

3.4.10 Advanced machine learning

29 approaches utilised some form of advanced machine learning. We encountered different methods being used and some papers specifically presenting novel machine learning models. All of these papers surpass mere usage of a topic model or typical pre-trained embedding method.

We found a multitude of *machine learning methods* being used, from multi armed bandits [1], LSTM [24,37,113], multi-layer perceptrons [62,96,104], (bi-)GRU [37,69,71,123], matrix factorisation [4,62,69,110,111], gradient ascent or descent [41,57,63,116], some form of simple neural network [30,37,56], some form of graph neural network [19,49,104], autoencoder [4], neural collaborative filtering [62], learning methods [30,123] to DTW [48]. Three approaches ranked the papers to recommend [56,57,118] with, e.g. Bayesian Personalized Ranking. Two of the observed papers proposed topic modelling approaches [3,115].

Several papers proposed *models*: a bipartite network embedding [5], heterogeneous graph embeddings [29,42,48,63], a scientific social reference network [41], a paper-author-venue embedding [116] and a relation prediction model [64].

3.4.11 Crawling

We found nine papers incorporating a crawling step as part of their approach. PDFs are oftentimes collected from CiteSeer [38,46] or CiteSeerX [2,93,94], in some cases [39,88,110] the sources are not explicitly mentioned. Fewer used data sources are Wikipedia for articles explaining the top words from papers [98] or papers from ACM, IEEE and EI [109]. Some approaches explicitly mention the extraction of citation information [2,38,39,46,88,93,94] e.g. to identify co-citations.

3.4.12 Cosine similarity

Some form of cosine similarity was encountered in most (31) paper recommendation approaches. It is often applied between papers, between users, between users and papers and in other forms.

For application *between papers* we encountered the possibility of using unspecified embeddings: *unspecified word or vector representations* of papers [30,48,107,110], papers' key terms or top words [2,98] and key phrases [46]. We found some approaches using *vector space model* variants: unspecified [59], tf vectors [39,88], tf-idf vectors [42,95,111], dimensionality reduced tf-idf vectors [86] and lastly, tf-idf and entity embeddings [56]. Some approaches incorporated more advanced embedding techniques: SBERT embeddings [5], Doc2Vec embeddings [28], Doc2Vec embeddings with incorporation of their emotional score [109] and NPLM representations [29].

Cosine similarity was used *between preferences or profiles of users and papers* in the following ways: unspecified representations [63,84,113,115], Boolean representation of users and keywords [60], tf-idf vectors [21,74–76], cf-idf vectors [74–76] and hcf-idf vectors [74–76].

For *between users* application of cosine similarity, we found unspecified representations [41] and time-decayed Word2Vec embeddings of users' papers' keyword [55].

Other applications include the usage between input keywords and paper clusters [117] and between nodes in a graph represented by their neighbouring nodes [121,122].

3.5 Paper recommendation systems

The 65 relevant works identified in our literature search are described in this section. We deliberately refrain from trying to structure the section by classifying papers by an arbitrary dimension and instead point to Table 3 to identify those dimensions in which a reader is interested to navigate the following short descriptions. The works are ordered by the surname of the first author and ascending publication year. An exception to this rule are papers presenting extensions of

previous approaches with different first authors. These papers are ordered to their preceding approaches.

Afsar et al. [1] propose KERS, a multi-armed bandit approach for patients to help with medical treatment decision making. It consists of two phases: first an exploration phase identifies categories users are implicitly interested in. This is supported by an expert-built knowledge base. Afterwards an exploitation phase takes place where articles from these categories are recommended until a user's focus changes and another exploitation phase is initiated. The authors strive to minimise the exploration efforts while maximising users' satisfaction.

Ahmedi et al. [3] propose a personalised approach which can also be applied to more general recommendation scenarios which include user profiles. They utilise Collaborative Topic Regression to mine association rules from historic user interaction data.

Alfarhood and Cheng [4] introduce Collaborative Attentive Autoencoder, a deep learning-based model for general recommendation targeting the data sparsity problem. They apply probabilistic matrix factorisation while also utilising textual information to train a model which identifies latent factors in users and papers.

Ali et al. [5] construct PR-HNE, a personalised probabilistic paper recommendation model based on a joint representation of authors and publications. They utilise graph information such as citations as well as co-authorships, venue information and topical relevance to suggest papers. They apply SBERT and LDA to represent author embeddings and topic embeddings respectively.

Bereczki [19] models users and papers in a bipartite graph. Papers are represented by their contents' Word2Vec or BERT embeddings, users' vectors consist of representations of papers they interacted with. These vectors are then aggregated with simple graph convolution.

Bulut et al. [22] focus on current user interest in their approach which utilises k-Means and KNN. Users' profiles are constructed from their authored papers. Recommended papers are the highest cited ones from the cluster most similar to a user. In a subsequent work they extended their research group to again work in the same domain. Bulut et al. [21] again focus on users' features. They represent users as the sum of features of their papers. These representations are then compared with all papers' vector representations to find the most similar ones. Papers can be represented by TF-IDF, Word2Vec or Doc2Vec vectors.

Chaudhuri et al. [25] use indirect features derived from direct features of papers in addition to direct ones in their paper recommendation approach: keyword diversification, text complexity and citation analysis. In an extended group Chaudhuri et al. [26] later propose usage of more indirect features such as quality in paper recommendation. Users' profiles are composed of their clicked papers. Subsequently they again worked on an approach in the same area but in a slightly smaller group. Chaudhuri et al. [24] propose the general Hybrid Topic Model and apply it on paper recommendation. It learns users' preferences and intentions by combining LDA and Word2Vec. They compute user's interest from probability distributions of words of clicked papers and dominant topics in publications.

Chen and Ban [27] introduce CPM, a recommendation model based on topically clustered user interests mined from their published papers. They derive user need models from these clusters by using LDA and pattern equivalence class mining. Candidate papers are then ranked against the user need models to identify the best-fitting suggestions.

Collins and Beel [28] propose the usage of their paper recommendation system Mr. DLib as a recommender as-a-service. They compare representing papers via Doc2Vec with a key phrase-based recommender and TF-IDF vectors.

Du et al. [29] introduce HNPR, a heterogeneous network method using two different graphs. The approach incorporates citation information, co-author relations and research areas of publications. They apply random walk on the networks to generate vector representations of papers.

Du et al. [30] propose Polar++, a personalised active one-shot learning-based paper recommendation system where new users are presented articles to vote on before they obtain recommendations. The model trains a neural network by incorporating a matching score between a query article and the recommended articles as well as a personalisation score dependant on the user.

Guo et al. [37] recommend publications based on papers initially liked by a user. They learn semantics between titles and abstracts of papers on word- and sentence-level, e.g. with Word2Vec and LSTMs to represent user preferences.

Habib and Afzal [38] crawl full texts of papers from CiteSeer. They then apply bibliographic coupling between input papers and a clusters of candidate papers to identify the most relevant recommendations. In a subsequent work Afzal again used a similar technique. Ahmad and Afzal [2] crawled papers from CiteSeerX. Cosine similarity of TF-IDF representations of key terms from titles and abstracts is combined with co-citation strength of paper pairs. This combined score then ranks the most relevant papers the highest.

Haruna et al. [39] incorporate paper-citation relations combined with contents of titles and abstracts of papers to recommend the most fitting publications for an input query corresponding to a paper.

Hu et al. [41] present ADRCR, a paper recommendation approach incorporating author-author and author-paper citation relationships as well as authors' and papers' authoritativeness. A network is built which uses citation information as weights. Matrix decomposition helps learning the model.

Hua et al. [42] propose PAPR which recommends relevant paper sets as an ordered path. They strive to overcome rec-

ommendation merely based on similarity by observing topics in papers changing over time. They combine similarities of TF-IDF paper representations with random-walk on different scientific networks.

Jing and Yu [44] build a three-layer graph model which they traverse with random-walk with restart in an algorithm named PAFRWR. The graph model consists of one layer with citations between papers' textual content represented via Word2Vec vectors, another layer modelling co-authorships between authors and the third layer encodes relationships between papers and topics contained in them.

Kanakia et al. [45] build their approach upon the MAG dataset and strive to overcome the common problems of scalability and cold-start. They combine TF-IDF and Word2Vec representations of the content with co-citations of papers to compute recommendations. Speedup is achieved by comparing papers to clusters of papers instead of all other single papers.

Kang et al. [46] crawl full texts of papers from CiteSeer and construct citation graphs to determine candidate papers. Then they compute a combination of section-based citation and key phrase similarity to rank recommendations.

Kong et al. [48] present VOPRec, a model combining textual components in form of Doc2vec and Paper2Vec paper representations with citation network information in form of Struc2vec. Those networks of papers connect the most similar publications based on text and structure. Random walk on these graphs contributes to the goal of learning vector representations.

L et al. [49] base their recommendation on lately accessed papers of users as they assume future accessed papers are similar to recently seen ones. They utilise a sliding window to generate sequences of papers, on those they construct a GNN to aggregate neighbouring papers to identify users' interests.

Li et al. [56] introduce a subscription-based approach which learns a mapping between users' browsing history and their clicks in the recommendation mails. They learn a re-ranking of paper recommendations by using its metadata, recency, word representations and entity representations by knowledge graphs as input for a neural network. Their defined target audience are new users.

Li et al. [55] present HNTA a paper recommendation method utilising heterogeneous networks and changing user interests. Paper similarities are calculated with Word2Vec representations of words recommended for each paper. Changing user interest is modelled with help of an exponential time decay function on word vectors.

Li et al. [57] utilise user profiles with a history of preferences to construct heterogeneous networks where they apply random walks on meta-paths to learn personalised weights. They strive to discover user preference patterns and model preferences of users as their recently cited papers.

Lin et al. [59] utilise authors' citations and years they have been publishing papers in their recommendation approach. All candidate publications are matched against user-entered keywords, the two factors of authors of these candidate publications are combined to identify the overall top recommendations.

Liu et al. [60] explicitly do not require all recommended publications to fit the query of a user perfectly. Instead they state the set of recommended papers fulfils the information need only in the complete form. Here they treat paper recommendation as a link prediction problem incorporating publishing time, keywords and author influence. In a subsequent work, part of the previous research group again observes the same problem. In this work Liu et al. [61] propose an approach utilising numbers of citations (author popularity) and relationships between publications in an undirected citation graph. They compute Steiner trees to identify the sets of papers to recommend.

Lu et al. [62] propose TGMF-FMLP, a paper recommendation approach focusing on the changing preferences of users and novelty of papers. They combine category attributes (such as paper type, publisher or journal), a time-decay function, Doc2Vec representations of the papers' content and a specialised matrix factorisation to compute recommendations.

Ma et al. [64] introduce HIPRec, a paper recommendation approach on heterogeneous networks of authors, papers, venues and topics specialised on new publications. They use the most interesting meta-paths to construct significant meta-paths. With these paths and features from these paths they train a model to identify new papers fitting users. Together with another researcher Ma further pursued this research direction. Ma and Wang [63] propose HGRec, a heterogeneous graph representation learning-based model working on the same network. They use meta-path-based features and Doc2Vec paper embeddings to learn the node embeddings in the network.

Manju et al. [65] attempt to solve the cold-start problem with their paper recommendation approach coding social interactions as well as topical relevance into a heterogeneous graph. They incorporate believe propagation into the network and compute recommendations by applying random walk.

Mohamed Hassan et al. [69] adopt an existing tag prediction model which relies on a hierarchical attention network to capture semantics of papers. Matrix factorisation then identifies the publications to recommend.

Nair et al. [71] propose C-SAR, a paper recommendation approach using a neural network. They input GloVe embeddings of paper titles into their Gated Recurrent Union model to compute probabilities of similarities of papers. The resulting adjacency matrix is input to an association rule mining a priori algorithm which generates the set of recommendations.

Nishioka et al. [74,75] state serendipity of recommendations as their main objective. They incorporate users' tweets to construct profiles in hopes to model recent interests and developments which did not yet manifest in users' papers. They strive to diversity the list of recommended papers. In more recent work Nishioka et al. [76] explained their evaluation more in depth.

Rahdari and Brusilovsky [84] observe paper recommendation for participants of scientific conferences. Users' profiles are composed of their past publications. Users control the impact of features such as publication similarity, popularity of papers and its authors to influence the ordering of their suggestions.

Renuka et al. [86] propose a paper recommendation approach utilising TF-IDF representations of automatically extracted keywords and key phrases. They then either use cosine similarity between vectors or a clustering method to identify the most similar papers for an input paper.

Sakib et al. [89] present a paper recommendation approach utilising second-level citation information and citation context. They strive to not rely on user profiles in the paper recommendation process. Instead they measure similarity of candidate papers to an input paper based on co-occurred or co-occurring papers. In a follow-up work with a bigger research group Sakib et al. [88] combine contents of titles, keywords and abstracts with their previously mentioned collaborative filtering approach. They again utilise second-level citation relationships between papers to find correlated publications.

Shahid et al. [94] utilise in-text citation frequencies and assume a reference is more important to a referencing paper the more often it occurs in the text. They crawl papers from CiteSeerX to retrieve the top 500 citing papers. In a follow-up work with a partially different research group Shahid et al. [93] evaluate the previously presented approach with a user study.

Sharma et al. [95] propose IBM PARSe, a paper recommendation system for the medical domain to reduce the number of papers to review for keeping an existing knowledge graph up-to-date. Classifiers identify new papers from target domains, named entity recognition finds relevant medical concepts before papers' TF-IDF vectors are compared to ones in the knowledge graph. New publications most similar to already relevant ones with matching entities are recommended to be included in the knowledge base.

Subathra and Kumar [98] constructed an paper recommendation system which applies LDA on Wikipedia articles twice. Top related words are computed using pointwise mutual information before papers are recommended for these top words.

Tang et al. [104] introduce CGPrec, a content-based and knowledge graph-based paper recommendation system. They focus on users' sparse interaction history with papers and strive to predict papers on which users are likely to click. They utilise Word2Vec and a Double Convolutional Neural Network to emulate users' preferences directly from paper content as well as indirectly by using knowledge graphs.

Tanner et al. [106] consider relevance and strength of citation relations to weigh the citation network. They fetch citation information from the parsed full texts of papers. On the weighted citation networks they run either weighted co-citation inverse document frequency, weighted bibliographic coupling or random walk with restart to identify the highest scoring papers.

Tao et al. [107] use embeddings and topic modelling to compute paper recommendations. They combine LDA and Word2Vec to obtain topic embeddings. Then they calculate most similar topics for all papers using Doc2Vec vector representations and afterwards identify the most similar papers. With PageRank on the citation network they re-rank these candidate papers.

Waheed et al. [108] propose CNRN, a recommendation approach using a multilevel citation and authorship network to identify recommendation candidates. From these candidate papers ones to recommend are chosen by combining centrality measures and authors' popularity. Highly correlated but unrelated Shi et al. [96] present AMHG, an approach utilising a multilayer perceptron. They also construct a multilevel citation network as described before with added author relations. Here they additionally utilise vector representations of publications and recency.

Wang et al. [113] introduce a knowledge-aware path recurrent network model. An LSTM mines path information from the knowledge graphs incorporating papers and users. Users are represented by their downloaded, collected and browsed papers, papers are represented by TF-IDF representations of their keywords.

Wang et al. [109] require users to construct knowledge graphs to specify the domain(s) and enter keywords for which recommended papers are suggested. From the keywords they compute initially selected papers. They apply Doc2Vec and emotion-weighted similarity between papers to identify recommendations.

Wang et al. [110] regard paper recommendation targeting a group of people instead of single users and introduce GPRAH_ER. They employ a two-step process which first individually predicts papers for users in the group before recommended papers are aggregated. Here users in the group are not considered equal, different importance and reliability weights are assigned such that important persons' preferences are more decisive of the recommended papers. Together with a different research group two authors again pursued this definition of the paper recommendation problem. Wang et al. [111] recommend papers for groups of users in an approach called GPMF_ER. As with the previous approach they compute TF-IDF vectors of keywords of

papers to calculate most similar publications for each user. Probabilistic matrix factorisation is used to integrate these similarities in a model such that predictive ratings of all users and papers can be obtained. In the aggregation phase the number of papers read by a user is determined to replace the importance component.

Xie et al. [116] propose JTIE, an approach incorporating contents, authors and venues of papers to learn paper embeddings. Further, directed citation relations are included into the model. Based on users' authored and referenced papers personalised recommendations are computed. They consider explainability of recommendations. In a subsequent work part of the researchers again work on this topic. Xie et al. [115] specify on recommendation of papers from different areas for user-provided keywords or papers. They use hierarchical LDA to model evolving concepts of papers and citations as evidence of correlation in their approach.

Yang et al. [117] incorporate the age of papers and impact factors of venues as weights in their citation network-based approach named PubTeller. Papers are clustered by topic, the most popular ones from the clusters most similar to the query terms are recommendation candidates. In this approach, LDA and TF-IDF are used to represent publications.

Yu et al. [118] propose ICMN, a general collaborative memory network approach. User and item embeddings are composed by incorporating papers' neighbourhoods and users' implicit preferences.

Zavrel et al. [119] present the scientific literature recommendation platform Zeta Alpha, which bases their recommended papers on examples tagged in user-defined categories. The approach includes these user-defined tags as well as paper content embeddings, social media mentions and citation information in their ensemble learning approach to recommend publications.

Zhang et al. [121] propose W-Rank, a general approach weighting edges in a heterogeneous author, paper and venue graph by incorporating citation relevance and author contribution. They apply their method on paper recommendation. Network- (via citations) and semantic-based (via AWD) similarity between papers is combined for weighting edges between papers, harmonic counting defines weights of edges between authors and papers. A HITS-inspired algorithm computes the final authority scores. In a subsequent work in a slightly smaller group they focus on a specialised approach for paper recommendation. Here Zhang et al. [122] strive to emulate a human expert recommending papers. They construct a heterogeneous network with authors, papers, venues and citations. Citation weights are determined by semantic- and network-level similarity of papers. Lastly, recommendation candidates are re-ranked while combining the weighted heterogeneous network and recency of papers.

Zhao et al. [123] present a personalised approach focusing on diversity of results which consists of three parts. First LFM extracts latent factor vectors of papers and users from the users' interactions history with papers. Then BERT vectors are constructed for each word of the papers, with those vectors as input and the latent factor vectors as label a BiGRU model is trained. Lastly, diversity and a user's rating weights determine the ranking of recommended publications for the specific user.

3.6 Other relevant work

We now briefly discuss some papers which did not present novel paper recommendation approaches but are relevant in the scope of this literature review nonetheless.

3.6.1 Surrounding paper recommendation

Here we present two works which could be classified as ones to use on top of or in combination with existing paper recommendation systems: Lee et al. [51] introduce LIMEADE, a general approach for opaque recommendation systems which can for example be applied on any paper recommendation system. They produce explanations for recommendations as a list of weighted interpretable features such as influential paper terms.

Beierle et al. [18] use the recommendation-as-a-service provider Mr. DLib to analyse choice overload in user evaluations. They report several click-based measures and discuss effects of different study parameters on engagement of users.

3.6.2 (R)Evaluations

The following four works can be grouped as ones which provide (r)evaluations of already existing approaches. Their results could be useful for the construction of novel systems: Ostendorff [77] suggests considering the context of paper similarity in background, methodology and findings sections instead of undifferentiated textual similarity for scientific paper recommendation.

Mohamed Hassan et al. [68] compare different text embedding methods such as BERT, ELMo, USE and InferSent to express semantics of papers. They perform paper recommendation and re-ranking of recommendation candidates based on cosine similarity of titles.

Le et al. [50] evaluate the already existing paper recommendation system Mendeley Suggest, which provides recommendations with different collaborative or content-based approaches. They observe different usage behaviours and state utilisation of paper recommendation systems does positively effect users' professional lives.

Barolli et al. [11] compare similarities of paper pairs utilising n-grams, tf-idf and a transformer based on BERT. They

model cosine similarities of these pairs into a paper connection graph and argue for the combination of content-based and graph-based methods in the context of COVID-19 paper recommendation systems.

3.6.3 Living labs

Living labs help researchers conduct meaningful evaluations by providing an environment, in which recommendations produced by experimental systems are shown to real users in realistic scenarios [14]. We found three relevant works for the area of scientific paper recommendation: Beel et al. [14] proposed a living lab for scholarly recommendation built on top of Mr. DLib, their recommender-as-a-service system. They log users' actions such as clicks, downloads and purchases for related recommended papers. Additionally, they plan to extend their living lab to also incorporate research grant or research collaborator recommendation.

Gingstad et al. [36] propose ArXivDigest, an online living lab for explainable and personalised paper recommendations from arXiv. Users can either be suggested papers while browsing their website or via email as a subscription-type service. Different approaches can be hooked into ArXivDigest, the recommendations generated by them can then be evaluated by users. A simple text-based baseline compares user-input topics with articles. Target values of evaluations are users' clicked and saved papers.

Schaer et al. [91] held the Living Labs for Academic Search (LiLAS) where they hosted two shared tasks: dataset recommendation for scientific papers and ad-hoc multilingual retrieval of most relevant publications regarding specific queries. To overcome the gap between real-world and lab-based evaluations they allowed integrating participants' systems into real-world academic search systems, namely LIVIO and GESIS Search.

3.6.4 Multilingual/cross-lingual recommendation

The previous survey by Li and Zhou [58] identifies cross-language paper recommendation as a future research direction. The following two works could be useful for this aspect: Keller and Munz [47] present their results of participating on the CLEF LiLAS challenge where they tackled recommendation of multilingual papers based on queries. They utilised a pre-computed ranking approach, Solr and pseudo-relevance feedback to extend queries and identify fitting papers.

Safaryan et al. [87] compare different already existing techniques for cross-language recommendation of publications. They compare word by word translation, linear projection from a Russian to an English vector representation, VecMap alignment and MUSE word embeddings.

3.6.5 Related recommendation systems

Some recommendation approaches are slightly out of scope of pure paper recommendation systems but could still provide inspiration or relevant results: Ng [73] proposes CBRec, a children's book recommendation system utilising matrix factorisation. His goal is to encourage good reading habits of children. The approach combines readability levels of users and books with TF-IDF representations of books to find ones which are similar to ones which a child may have already liked.

Patra et al. [80] recommend publications relevant for datasets to increase reusability. Those papers could describe the dataset, use it or be related literature. The authors represent datasets and articles as vectors and use cosine similarity to identify the best fitting papers. Re-ranking them with usage of Word2Vec embeddings results in the final recommendation.

4 Datasets

As the discussed paper recommendation systems utilise different inputs or components of scientific publications and pursue slightly different objectives, datasets to experiment on are also of diverse nature. We do not consider datasets of approaches which do not contain an evaluation [60,119] or do not evaluate the actual paper recommendation [2,25,38,84,86] such as the cosine similarity between a recommended and an initial paper [2,86], the clustering quality on the constructed features [25] or the Jensen Shannon Divergence between probability distributions of words between an initial and recommended papers [38]. We also do not discuss datasets where only the data sources are mentioned but no remarks are made regarding the size or composition of the dataset [21,104] or ones where we were not able to identify actual numbers [65]. Table 4 gives an overview of datasets used in the evaluation of the considered discussed methods. Many of the datasets are unavailable only few years after publication of the approach. Most approaches utilise their own modified version of a public dataset which makes exact replication of experiments hard. In the following the main underlying data sources and publicly available datasets are discussed. Non-publicly available datasets are briefly described in Table 5.

4.1 dblp-based datasets

The dblp computer science bibliography (dblp) is a digital library offering metadata on authors, papers and venues from the area of computer science and adjacent fields [54]. They

Table 4 Overview of datasets utilised in most recent related work with (unofficial) names, public availability of the possibly modified dataset which was used (A?), and a list of papers it was used in. Datasets are grouped by their underlying data source if possible

Name	A?	Used by
DBLP + Citations v1 [105]	✓	[117]
DBLP + Citations v8 [105]	×	[63,64]
DBLP + Citations v11	✓	[5]
dblp + IEEE + ACM + Pubmed	×	[22]
DBLP paths	×	[42]
DBLP-Citation-network f. AMiner	×	[44]
dblp	×	[57]
DBLP-REC	×	[96]
dblp + AMiner KG	×	[113]
dblp + AMiner + venue	×	[116]
SPRD_Senior	✓	[27]
SPRD [101]	✓	[39,88,89]
Citeulike-a [112]	✓	[3,4,37,49,69,104,118,123]
Citeulike-t [112]	✓	[4]
Citeulike_huge	×	[62]
Citeulike_medium	×	[110]
Citeulike_tiny	×	[111]
ACM paths	×	[42]
ACM citation network V8	×	[74–76]
Scopus_tiny	×	[24,26]
ScienceDirect+Scopus	×	[56]
Scopus	×	[115]
AMiner	×	[57]
AMiner + Wanfang	×	[29]
AMiner_tiny	×	[30]
AMiner_huge	×	[108]
ACM C-D	×	[115]
AAN_original [83]	✓	[71]
AAN_modified	×	[5,49]
AAN_tiny	×	[106]
Sowiport	×	[28]
RARD_tiny	×	[30]
CiteSeer	×	[46]
CiteSeer_tiny	×	[94]
CiteSeer_medium	×	[92]
Patents_tiny	×	[30]
Patents	×	[116]
ACM H-I	×	[115]
Hep-TH graph	×	[61]
arXiv Hep-TH	×	[121]
MSA	×	[117]
MAG 2017	×	[121]
MAG 2018	×	[45]
BBC	✓	[1]
PRSDataset	✓	[37,49]
Physical review A	×	[48]
ACL selection network	×	[107]
Prostate cancer	×	[1]
Peltarion	×	[19]
Jabref	×	[28]
DM	×	[41]
Graphs	×	[49]
SCHOLAT	×	[55]
IEEE Xplore	×	[59]
KGs	×	[109]
Wanfang	×	[46]
Watson™for Genomics	×	[95]
Wikipedia	×	[98]
LibraryThing	×	[123]

provide publicly available short-time stored daily and longer-time stored monthly data dumps[10].

The *dblp + Citations v1* dataset [105] builds upon a dblp version from 2010 mapped on AMiner. It contains 1,632,442 publications with 2,327,450 citations.

The *dblp + Citations v11* dataset[11] builds upon dblp. It contains 4,107,340 papers, 245,204 authors, 16,209 venues and 36,624,464 citations

These datasets do not contain supervised labels provided by human annotators even though the citation information could be used as interaction data.

4.2 SPRD-based datasets

The Scholarly Paper Recommendation Dataset (abbreviation: SPRD)[12] was constructed by collecting publications written by 50 researchers of different seniority from the area of computer science which are contained in dblp from 2000 to 2006 [58,101,102]. The dataset contains 100,351 candidate papers extracted from the ACM Digital Library as well as citations and references for papers. Relevance assessments of papers relevant to their current interests of the 50 researchers are also included.

A subset of SPRD, *SPRD_Senior*, which contains only the data of senior researchers can also be constructed [99].

[10] https://dblp.uni-trier.de/xml/.

[11] https://www.aminer.org/citation.

[12] (shortened) http://shorturl.at/cIQR1.

Table 5 Description of private datasets utilised in most recent related work with (unofficial) names. Datasets are grouped by their underlying data source if possible

Name	Used by	Description
DBLP + Citations v8 [105]	[63,64]	2,133 *p* from 20 *v* from 2000 to 2016, 39,530 *a*, 15,708 *p* topics
dblp + IEEE + ACM + Pubmed	[22]	Sources: dblp, IEEE, ACM, Pubmed. 3,394,616 *p* (titles), *a*, publication years, keywords, *r*
DBLP paths	[42]	1,782,700 *p* (titles, abstracts, keywords), 2,052,414 *a*, 18,936 *v*, 100,000 *t*, 9,590,600 *i*
DBLP-Citation-network f. AMiner	[44]	63,469 *p* from 2013 to 2019, 152,586 *a*
dblp	[57]	2,126,267 *p*, 8686 *v*, 1,221,259 *a*, 256,214 *t*, 3765 *u* relations
DBLP-REC	[96]	DBLP-Citation-network v11 + ScienceDirect + IEEE, 3,590,853 *p*, 3,276,803 *a*, 35,254,530 *c*
dblp + AMiner KG	[113]	KG with 223,431 *a*, 337,561 *p*, 5578 *v*, 1179 keyword nodes, 16,328,642 *c*
dblp + AMiner + venue	[116]	3,056,388 *p* (titles, abstracts, keywords), 1,752,401 *a*, 354,693 keywords, 11,397 *v*, *c*, discipline labels
Citeulike_huge	[62]	210,137 *p*, 3,039 *u*, 284,960 *u-p i* from Nov 2004 to Dec 2007
Citeulike_medium	[110]	2,065 users, 718 groups, 85,542 *p*
Citeulike_tiny	[111]	1,659 users, 718 groups, 82,376 *p*, 198,744 *i*
ACM paths	[42]	2,385,057 *p* (titles, abstracts, keywords), 2,004,398 *a*, 269,467 *v*, 61,618 *t*, 12,048,682 *i*
ACM citation network V8	[74–76]	1,669,237 *p* (titles, abstracts), *v*, *a*
Scopus_tiny	[24,26]	2,000 *p*
ScienceDirect + Scopus	[56]	*u*'s browsed *p* prior to first email from ScienceDirect, *p* metadata from Scopus, 4,392 recommendation sessions (emails with clicks on *p*, *u*' browsing history)
Scopus	[115]	528,224 *p*, *a*, *r*, discipline tags
Scopus + venue	[116]	1,304,907 *p* (titles, abstracts, keywords), 482,602 *a*, 127,630 keywords, 7653 *v*, *c*, discipline labels
AMiner	[57]	2,070,699 *p*, 263,250 *v*, 1,557,147 *a*, 735,059 *t*, 9398 *u* relations
AMiner + Wanfang	[29]	4 mio *p*. 3 sets: data from 2018 and 2019 (221,076 *p*, 503,945 *a*), mathematical analysis (98,702 *p*, 117,183 *a*), image processing (49,098 *p*, 107,290 *a*)
AMiner_tiny	[30]	188 input *p*, 10 candidate *p* for each input
AMiner_huge	[108]	2,092,356 *p*, 1,712,433 *a*, 8,024,869 *c*, 4,258,615 co-autorships
ACM C-D	[115]	43,380 *p* from AMiner, *a*, ACM CSS tags
AAN_modified	[5,49]	21,455 *p* from 312 *v* from NLP, 17,342 *a*, 113,367 *c*
AAN_tiny	[106]	2082 *p* (ids, titles, publication year), 8194 *c*, avg. 7.87 *c* per *p*, *a*, *v*
Sowiport	[28]	*u i* data from Mar 2017 to Oct 2018, 0.1% click-through rate
RARD_tiny	[30]	800 input *p* from Related-Article Recommendation Dataset from Sowiport [13]
CiteSeer	[46]	1,100 *p*, 10 sets of relevant *p*
CiteSeer_tiny	[94]	400 *c*-pairs, 1,230 *c* contexts
CiteSeer_medium	[92]	10 *p*, 226 *c*-pairs
Patents_tiny	[30]	67 input patents, 20 candidate patents for each input
Patents	[116]	182,260 patents, 73,974 *a*
ACM H-I	[115]	70,090 patents with ownership from 2017, *r*, ACM CSS tags
Hep-TH graph	[61]	graph with 8,721 *p* (keywords)
arXiv Hep-TH	[121]	~29,000 *p*, 350,000 *c*, 14,909 *a*, 428 journals

Table 5 continued

Name	Used by	Description
MSA	[117]	101,205 *p*, 190,146 *c* in 300 conferences
MAG 2017	[121]	Based on data until 2017, area: intrusion detection in cyber security, 6428 *p*, 94,887 *c*, 18,890 *a*, 6428 journals
MAG 2018	[45]	Based on MAG Azure database from Oct 2018, 206,676,892 *p*
Physical Review A	[48]	393 *p* from 2007 to 2009 with 2,664 *c* from American Physical Society
ACL selection network	[107]	18,718 *p* (titles, summaries) from ACL proceedings
prostate cancer	[1]	500 *p* tagged with 5 categories
Peltarion	[19]	290 *p*, *u i* from Dec 2018 to May 2021 of *u* of Peltarion Knowledge Center who have read ≥ 5 *p*
Jabref	[28]	*u i* data from Mar 2017 to Oct 2018, 0.22% click-through rate
DM	[41]	8,301 *p* from journals: DMKD, TKDE + conferences: KDD, ICDM, SDM
Graphs	[49]	Cora (1 graph, 2.7k nodes), TU-IMDB (1.5k graphs, 13 nodes each), TU-MUTAG (188 molecules, 18 nodes)
SCHOLAT	[55]	34,518 *p* (titles, abstracts, keywords), *a*
IEEE Xplore	[59]	3 *p* (keywords), *r*, *a* appeared in IEEE between 2010 and 2017
KGs	[109]	Knowledge graphs, 600 *p* from information retrieval + machine learning
Wanfang	[46]	500 *p*, 5 sets of relevant *p*
Watson™for Genomics	[95]	15,320 *p* from top 10 percentile genomics journals from Jun 2016
Wikipedia	[98]	1000 *p* from Wikipedia, 20 topics
LibraryThing	[123]	120,150 books (titles, abstracts), *u*, 185,210 favourites records, 150,216 ratings, 139,530 reviews of 12,350 *u*

We used the following abbreviations: user(s) *u*, paper(s) *p*, interaction(s) *i*, author(s) *a*, venue(s) *v*, reference(s) *r*, citation(s) *c*, term(s) *t*

These datasets specifically contain supervised labels provided by human annotators in the form of sets of papers, which researchers found relevant for themselves.

4.3 CiteULike-based datasets

CiteULike [20] was a social bookmarking site for scientific papers. It contained papers and their metadata. Users were able to include priorities, tags or comments for papers on their reading list. There were daily data dumps available from which datasets could be constructed.

Citeulike-a [112][13] contains 5,551 users, 16,980 papers with titles and abstracts from 2004 to 2006 and their 204,986 interactions between users and papers. Papers are represented by their title and abstract.

Citeulike-t [112][14] contains 7,947 users, 25,975 papers and 134,860 user-paper interactions. Papers are represented by their pre-processed title and abstract.

These datasets contain labelled data as they build upon CiteULike, which provides bookmarked papers of users.

4.4 ACM-based datasets

The ACM Digital Library (ACM) is a semi-open digital library offering information on scientific authors, papers, citations and venues from the area of computer science[15]. They offer an API to query for information. Datasets building upon this source do not contain supervised labels provided by annotators even though the citation information could be used as interaction data.

4.5 Scopus-based datasets

Scopus is a semi-open digital library containing metadata on authors, papers and affiliations in different scientific areas[16]. They offer an API to query for data. Datasets building upon this source usually do not contain labels provided by annotators.

4.6 AMiner-based datasets

ArnetMiner (AMiner) [105] is an open academic search system modelling the academic network consisting of authors,

[13] https://github.com/js05212/citeulike-a.

[14] https://github.com/js05212/citeulike-t.

[15] https://dl.acm.org/.

[16] https://www.scopus.com/home.uri.

papers and venues from all areas[17]. They provide an API to query for information. Datasets building upon this source usually do not contain labelled user interaction data.

4.7 AAN-based datasets

The ACL Anthology Network (AAN) [81–83] is a networked database containing papers, authors and citations from the area of computational linguistics[18]. It consists of three networks representing paper-citation relations, author-collaboration relations and the author-citation relations. The original dataset contains 24,766 papers and 124,857 citations [71]. Datasets building upon this source usually do not contain labelled user interaction data even though the paper-citation, author-collaboration or author-citation relationships could be utilised to replace this data.

4.8 Sowiport-based datasets

Sowiport was an open digital library containing information on publications from the social sciences and adjacent fields [15,40]. The dataset linked papers by their attributes such as authors, publishers, keywords, journals, subjects and citation information. Via author names, keywords and venue titles the network could be traversed by triggering them to start a new search [40]. Sowiport cooperated with the recommendation-as-a-service system Mr. DLib [28]. Datasets building upon this source usually contain labelled user interaction data, the clicked papers of users.

4.9 CiteSeerX-based datasets

CiteSeerX [35,114] is a digital library focused on metadata and full-texts of open access literature[19]. It is the overhauled form of the former digital library CiteSeer. Datasets building upon this source usually do not inherently contain labelled user interaction data.

4.10 Patents-based datasets

The Patents dataset provides information on patents and trademarks granted by the United States Patent and Trademark Office[20]. Datasets building upon this source usually do not contain labelled user interaction data.

4.11 Hep-TH-based datasets

The original unaltered *Hep-TH* [53] dataset[21] stems from the area of high energy physics theory. It contains papers in a graph which were published between 1993 and 2003. It was released as part of KDD Cup 2003. Datasets building upon this source usually do not contain labelled user interaction data.

4.12 MAG-based datasets

The Microsoft Academic Graph (MAG) [97] was an open scientific network containing metadata on academic communication activities[22]. Their heterogeneous graph consists of nodes representing fields of study, authors, affiliations, papers and venues. Datasets building upon this source usually do not contain labelled user interaction data besides citation information.

4.13 Others

The following datasets have no common underlying data source: The *BBC*[23] dataset contains 2,225 BBC news articles which stem from 5 topics. This dataset does not contain labelled user interaction data.

PRSDataset[24] contains 2,453 users, 21,940 items and 35,969 pairs of users and items. This dataset contains user-item interactions.

5 Evaluation

The performance of a paper recommendation system can be quantified by measuring how well a target value has been approximated by the recommended publications. Relevancy estimations of papers can come from different sources, such as human ratings or datasets. Different interactions derived from clicked or liked papers determine the target values which a recommendation system should approximate. The quality of the recommendation can be described by evaluation measures such as precision or MRR. For example, a dataset could provide information on clicked papers, that are then deemed relevant. The target value which should be approximated with the recommender system are those clicked papers, and the percentage of the recommendations which are contained in the clicked papers could then be reported as the system's precision.

[17] https://www.aminer.org/.

[18] https://aan.how/download/.

[19] https://citeseerx.ist.psu.edu/index.

[20] https://bulkdata.uspto.gov/.

[21] https://snap.stanford.edu/data/cit-HepTh.html.

[22] (shortened) http://shorturl.at/orwXY.

[23] http://mlg.ucd.ie/datasets/bbc.html.

[24] https://sites.google.com/site/tinhuynhuit/dataset.

Due to the vast differences in approaches and datasets used to apply the methods, there is also a spectrum of used evaluation measures and objectives. In this section, first we observe different notions of relevance of recommended papers and individual assessment strategies for relevance. Afterwards we analyse commonly used evaluation measures and list ones which are only rarely encountered in evaluation of paper recommendation systems. Lastly we shed light on the different types of evaluation which authors conducted.

In this discussion we again only consider paper recommendation systems which also evaluate their actual approach. We disregard approaches which do evaluate other properties [2,25,38,84,86,122] or contain no evaluation [60,119]. Thus we observe 54 different approaches in this analysis.

5.1 Relevance and assessment

Relevance of recommended publications can be evaluated against multiple target values: clicked papers [24,56, 104], references [44,115], references of recently authored papers [57], papers an author interacted with in the past [49], degree-of-relevancy which is determined by citation strength [94], a ranking based on future citation numbers [121] as well as papers accepted [26] or deemed relevant by authors [39,88].

Assessing the relevance of recommendations can also be conducted in different ways: the top n papers recommended by a system can be judged by either a referee team [109] or single persons [26,74,75]. Other options for relevance assessment are the usage of a dataset with user ratings [39,88] or emulation of users and their interests [1,57].

Table 6 holds information on utilised relevance indicators and target values which indicate relevance for the 54 discussed approaches. *Relevancy* describes the method that defines which of the recommended papers are relevant:

- Human rating: The approach is evaluated using assessments of real users of results specific to the approach.
- Dataset: The approach is evaluated using some type of assessment of a target value which is not specific to the approach but from a dataset. The assessment was either conducted for another approach and re-used or it was collected independent of an approach.
- Papers: The approach is evaluated by some type of assessment of a target value which is directly generated from the papers contained in the dataset such as citations or their keywords.

The *target values* in Table 6 describe the entities which the approach tried to approximate:

- Clicked: The approximated target value is derived from users' clicks on papers.
- Read: The approximated target value is derived from users' read papers.
- Cited: The approximated target value is derived from cited papers.
- Liked: The approximated target value is derived from users' liked papers.
- Relevancy: The approximated target value is derived from users' relevance assessment of papers.
- Other user: The approximated target value is derived from other entities associated with a user input, e.g. acceptance of users, users' interest and relevancy of the recommended papers' topics.
- Other automatic: The approximated target value is automatically derived from other entities, e.g. user profiles, papers with identical references, degree-of-relevancy, keywords extracted from papers, papers containing the query keywords in the optimal Steiner tree, neighbouring (cited and referencing) papers, included keywords, the classification tag, future citation numbers and an unknown measure derived from a dataset. We refrain from trying to introduce sub-categories for this broad field.

Only three approaches evaluate against multiple target values [21,30,104]. Six approaches (11.11%) utilise clicks of users, only one approach (1.85%) uses read papers as target value. Even though cited papers are not the main objective of paper recommendation systems but rather citation recommendation systems, this target was approximated by 13 (24.07%) of the observed systems. Ten approaches (18.52%) evaluated against liked papers, 15 (27.78%) against relevant papers and 13 (24.07%) against some other target value, either user input (three, 5.55%) or automatically derived (ten, 18.52%).

5.2 Evaluation measures

We differentiate between commonly used and rarely used evaluation measures for the task of scientific paper recommendation. They are described in the following sections. Table 6 holds indications of utilised evaluation measures for the 54 discussed approaches. *Measures* are the methods used to evaluate the approach's ability to approximate the target value which can be of type precision, recall, f1 measure, nDCG, MRR, MAP or another one.

Out of the observed systems, twelve[25] approaches [1,28, 30,49,59,64,69,71,74–76,107,115,116] (22.22%) only report one single measure, all others report at least two different ones.

[25] One approach is described in three papers.

Table 6 Indications whether approaches utilise the specified relevancy definitions, target values of evaluations and evaluation measures

Work	Relevancy			Target value							Measures						
	Human rating	Dataset	Papers	Clicked	Read	Cited	Liked	Relevancy	Other user	Other automatic	Precision	Recall	F1	nDCG	MRR	MAP	Other
[1]			•	•													•
[3]		•					•					•			•		
[4]		•					•					•					•
[5]		•								•		•			•	•	
[19]		•			•										•	•	
[22]	•							•			•	•	•				
[21]	•					•		•			•	•	•				•
[24]	•								•		•						•
[26]	•			•							•	•				•	•
[27]		•						•						•	•		
[28]	•			•													•
[29]			•							•		•			•	•	
[30]	•	•		•				•						•			
[37]		•					•				•			•	•		•
[39]		•						•			•	•	•		•	•	
[41]			•			•											•
[42]	•							•			•	•					
[44]			•			•						•		•			
[45]	•							•			•			•			
[46]	•		•					•				•					•
[48]			•			•					•	•	•	•			
[49]		•								•							•
[55]		•		•							•					•	
[56]			•							•	•	•	•				
[57]			•			•					•	•					
[59]			•			•										•	
[61]			•							•	•	•	•				•
[62]		•					•							•	•		•
[63]			•			•											•
[64]			•			•					•	•	•				

Table 6 continued

Work	Relevancy			Target value							Measures						
	Human rating	Dataset	Papers	Clicked	Read	Cited	Liked	Relevancy	Other user	Other automatic	Precision	Recall	F1	nDCG	MRR	MAP	Other
[65]	•							•			•			•	•		•
[69]		•					•					•					
[71]			•			•											•
[74–76]	•								•								•
[89]		•						•			•	•	•		•	•	
[88]		•						•			•	•	•		•	•	
[93,94]	•		•					•		•				•			•
[95]		•						•			•	•	•				
[96]	•							•			•				•	•	
[98]			•						•		•	•					
[104]		•		•			•				•	•	•				•
[106]			•							•	•	•	•				
[107]			•			•											•
[108]	•							•						•	•		•
[113]			•			•								•			•
[109]	•	•								•					•		•
[110]		•					•				•	•				•	
[111]		•					•				•	•			•	•	
[115]			•			•								•			
[116]		•								•				•			
[117]			•			•					•	•					
[118]		•					•							•			•
[121]			•							•							•
[123]		•					•				•	•	•				•

Table 7 Common evaluation measures and percentage of observed evaluations of paper recommendation systems in which they were applied

	P	*R*	*F*1	nDCG	MRR	MAP
%	48.15	24.07	50	25.92	27.78	22.22

Percentages are rounded to two decimal places

5.2.1 Commonly used evaluation measures

Bai et al. [9] identify *precision* (P), *recall* (R), *F1*, *nDCG*, *MRR* and *MAP* as evaluation features which have been used regularly in the area of paper recommendation systems. Table 7 gives usage percentages of each of these measures in observed related work.

Alfarhood and Cheng [4] argue against the use of precision when utilising implicit feedback. If a user gives no feedback for a paper it could either mean disinterest or that a user does not know of the existence of the specific publication.

5.2.2 Rarely used evaluation measures

We found a plethora of rarer used evaluation measures which have either been utilised only by the work they were introduced in or to evaluate few approaches. Our analysis in this aspect might be highly influenced by the narrow time frame we observe. Novel measures might require more time to be adopted by a broader audience. Thus we differentiate between novel rarely used evaluation measures and ones where authors do not explicitly claim they are novel. A list of rare but already defined evaluation measures can be found in Table 8. In total 25 approaches (46.3%) did use an evaluation measure not considered common.

Novel rarely used Evaluation Measures. In our considered approaches we only encountered three novel evaluation measures: *Recommendation quality* as defined by Chaudhuri et al. [26] is the acceptance of recommendations by users rated on a Likert scale from 1 to 10.

TotNP_EU is a measure defined by Manju et al. [65] specifically introduced for measuring performance of approaches regarding the cold start problem. It indicates the number of new publications suggested to users with a prediction value above a certain threshold.

TotNP_AVG is another measure defined by Manju et al. [65] for measuring performance of approaches regarding the cold start problem. It indicates the average number of new publications suggested to users with a prediction value above a certain threshold.

5.3 Evaluation types

Evaluations can be classified into different categories. We follow the notion of Beel and Langer [17] who differentiate between user studies, online evaluations and offline evaluations. They define *user studies* as ones where users' satisfaction with recommendation results is measured by collecting explicit ratings. *Online evaluations* are ones where users do not explicitly rate the recommendation results; relevancy is derived from e.g. clicks. In *offline evaluations* a ground truth is used to evaluate the approach.

From the 54 observed approaches we found four using multiple evaluation types [29,46,92,94,109]. Twelve (22.22%) were conducting user studies which describe the size and composition of the participant group.[26] Only two approaches [28,65] (3.7%) in the observed papers were evaluated with an online evaluation. We found 44 approaches (81.48%) providing an offline evaluation. Offline evaluations being the most common form of evaluation is unsurprising as this tendency has also been observed in an evaluation of general scientific recommender systems [23]. Offline evaluations are fast and do not require users [23]. Nevertheless the margin by which this form of evaluation is conducted could be rather surprising.

A distinction in *lab-based* vs. *real world* user studies can be conducted [16,17]. User studies where participants rate recommendations according to some criteria and are aware of the study are lab-based, all others are considered real-world studies. Living labs [14,36,91] for example enable real-world user studies. On average the lab-based user studies were conducted with 17.83 users. Table 9 holds information on the number of participants for all studies as well as the composition of groups in terms of seniority.

For offline evaluation, they can either be ones with an *explicit* ground truth given by a dataset containing user rankings, *implicit* ones by deriving user interactions such as liked or cited papers or *expert* ones with manually collected expert ratings [17]. We found 22 explicit offline evaluations (40.74%) corresponding to ones using datasets to estimate relevance (see Table 6) and 21 implicit offline evaluations (38.89%) corresponding to ones using paper information to identify relevant recommendations (see Table 6). We did not find any expert offline evaluations.

6 Changes compared to 2016

This chapter briefly summarises some of the changes in the set of papers we observed when compared to the study by Beel et al. [16]. Before we start the comparison, we want

[26] Shi et al. [96] also conduct a user study but do not describe their participants.

Table 8 Overview of rare existing measures used in evaluations of observed approaches

Measure	Used by	Description
Average precision	[108]	Area under precision-recall curve
Receiver operating characteristic	[121]	Plot of true positives against false positives
AUC	[37,104]	Area under receiver operating characteristic curve
Computation time	[26,61]	Time to compute recommendation list
DCG	[4]	Summed up relevancy divided by logarithm of rank + 1
Click-through-rates	[24,28]	percentage of Clicks on recommendations
Reward	[1,36]	Weighted sum of interactions of users with recommendations, e.g. clicked and saved papers
Spearman correlation coefficient	[45,121]	Correlation between ranks of paper lists
Hit ratio	[62,113,118]	Percentage of relevant items in top k recommendations
Accuracy	[21,64,92]	Percentage of relevant papers which the approach identified
Specificity	[21]	True negative rate
Mean absolute error	[41]	Average difference between real and predicted values
Root mean square error	[41]	Expected squared difference between real and predicted values
Fallout	[65]	Percentage of irrelevant recommendations out of all irrelevant papers
Support	[71]	Frequency of occurrences of set
TopN	[109]	Probability that target keywords are encountered in first n recommended papers
FindN	[109]	Number of target keywords which are encountered in first n recommended papers
Coverage	[123]	Method's ability to discover the long tail of papers
Popularity	[123]	Average logarithm of the number of ratings of papers in recommendation, indicates novelty of results
Average paper popularity	[61]	Paper popularity divided by number of recommendations
Intra-list similarity	[123]	Dissimilarity between recommended papers, smaller value indicates more diverse recommendation
Serendipity score	[74–76]	Summed up usefulness divided by unexpectedness of recommended papers
Success rate	[61]	Number of recommendations $< 2\times$ number of keywords
Number of recommended papers	[61]	Size of set of recommended papers

to point to the fact that we observed papers from two years in which the publication process could have been massively affected by the COVID-19 pandemic.

6.1 Number of papers per year and publication medium

Beel et al. [16] studied works between and including 1998 and 2013 while we observed works which appeared between January 2019 and October 2021. While the previous study did include all 185 papers (of which 96 were paper recommendation approaches) in their discussion of papers per year which were published in the area of the topic paper or citation recommendation but later on only studied 62 papers for an in-depth review, we generally only studied 65 publications which present novel paper recommendation approaches (see Sect. 3.5) in this aspect. Compared to the time frame observed in this previous literature review, we encountered fewer papers being published on the actual topic of scientific paper recommendation per year. In the former work, the published number of papers was rising and hitting 40 in 2013. We found this number being stuck on a constant level between 21 and 23 in the three years we observed. This could hint at differing interest in this topic over time, with a current demise or the trend to work in this area having surpassed its zenith.

While Beel et al. [16] found 59% of conference papers and 16% of journal articles, we found 54.85% of conference papers and 41.54% of journal articles. The shift to journal articles could stem from a general shift towards journal articles in computer science[27].

[27] Compare the 99.363 journal articles and 151.617 conference papers published in 2013 to the 187.263 journal articles and 157.460 conference articles in 2021 in dblp.

Table 9 For all observed works with user studies we list their number of participants (# P) and their composition

Work	# P	Composition
Bulut et al. [22]	50	PhD students studying in Turkey in 2019
Bulut et al. [21]	10 + 30	Researchers
Chaudhuri et al. [24]	50	NA
Chaudhuri et al. [26]	45	From 9 different areas, different seniority levels: 12 faculty members, 20 postgraduate students, 13 undergraduate students
Du et al. [30]	NA	College students or patent analysis experts
Hua et al. [42]	10	Experts
Kanakia et al. [45]	40	Full-time computer science researchers at Microsoft Research
Kang et al. [46]	12	Postgraduates
Nishioka et al. [74–76]	22	Seniority based on highest degree: 2 Master's, 13 PhD, 7 lecturers/professors; 2 female, 20 male; 17 working in academia, 3 working in industry
Shahid et al. [93]	20	Post-graduate students
Waheed et al. [108]	20	Researchers
Wang et al. [109]	5	1 doctoral supervisor, 2 master supervisors, 2 graduate students

NA indicates that #P or compositions were not described in a specific user study

Table 10 Percentage of studies using the different methods. Some studies utilised multiple methods, thus the percentages do not add up to 100%

	Offline	Online	User quant.	User qual.
[16]	71	7	25	3
Current	81.48	3.7	24.07	0

6.2 Classification

While Beel et al. [16] found 55% of their studied 62 papers applying methods from content-based filtering, we found only found 7.69% (5) of our 65 papers identifying as content-based approaches. Beel et al. [16] report 18% of approaches applied collaborative filtering. We encountered 4.62% (three) having this component as part of their self-defined classification. As for graph-based recommendation approach, Beel et al. [16] found 16% while we only encountered 7.69% (five) of papers with this description. In terms of hybrid approaches, Beel et al. [16] encountered five (8.06%) truly hybrid ones. In our study, we found 18 approaches (27.69%) labelling themselves as hybrid recommendation systems.[28]

6.3 Evaluation

Table 10 shows the comparison of the distributions of the different types of evaluations between our study observing 54 papers with evaluations and the one conducted by Beel et al. [16], which regards 75 papers for this aspect. The percentage of quantitative user studies (User quant) is comparable for both studies. A peculiar difference is the percentage of offline evaluations, which is much higher in our current time frame.

When observing the evaluation measures, we found some differences compared to the previous study. While 48.15% of papers with an evaluation report precision in our case, in Beel et al.'s [16] 72% of approaches with an evaluation report this value. As a contrast, we found 50% of papers reporting F1 while only 11% of papers reported this measure according to Beel et al. [16]. This might hint at a shift away from precision (which Beel et al. [16] did describe as a problematic measure) to focus more on also incorporating recall into the quality assessment of recommendation systems.

6.4 Discussion

In general, the two reviews regard different time frames. We encounter non-marginal differences in the three dimensions discussed in this Section. A more concise comparison could be made if a time slice would be regarded for both studies, such that the research output and shape could be observed from three years each. We cannot clearly identify emerging trends (as with the offline evaluation) as we do not know if it has been conducted in this percentage of papers since the 2010s or if it only just picked up to be a more wide-spread evaluation form.

7 Open challenges and objectives

All paper recommendation approaches which were considered in this survey could have been improved in some

[28] Note that not all approaches classified their type of paper recommendation and several papers did not classify themselves in the wide-spread categorisation (see Sect. 3.3.1).

way or another. Some papers did not conduct evaluations which would satisfy a critical reader, others could be more convincing if they compared their methods to appropriate competitors. The possible problems we encountered within the papers can be summarised in different open challenges, which papers should strive to overcome. We separate our analysis and discussion of open challenges in those which have already been described by previous literature reviews (see Sect. 7.1) and ones we identify as new or emerging problems (see Sect. 7.2). Lastly we briefly discuss the presented challenges (see Sect. 7.3).

7.1 Challenges highlighted in previous works

In the following we will explain possible shortcomings which were already explicitly discussed in previous literature reviews [9,16,92]. We regard these challenges in light of current paper recommendation systems to identify problems which are nowadays still encountered.

7.1.1 Neglect of user modelling

Neglect of user modelling has been described by Beel et al. [16] as identification of target audiences' information needs. They describe the trade-off between specifying keywords which brings recommendation systems closer to search engines and utilising user profiles as input.

Currently only some approaches consider users of systems to influence the recommendation outcome, as seen with Table 3 users are not always part of the input to systems. Instead many paper recommendation systems assume that users do not state their information needs explicitly but only enter keywords or a paper. With paper recommendation systems where users are not considered, the problem of neglecting user modelling still holds.

7.1.2 Focus on accuracy

Focus on accuracy as a problem is described by Beel et al. [16]. They state putting users' satisfaction with recommendations on a level with accuracy of approaches does not depict reality. More factors should be considered.

Only over one fourth of current approaches do not only report precision or accuracy but also observe more diversity focused measures such as MMR. We also found usage of less widespread measures to capture different aspects such as popularity, serendipity or click-through-rate.

7.1.3 Translating research into practice

The missing translation of research into practice is described by Beel et al. [16]. They mention the small percentage of approaches which are available as prototype as well as the discrepancy between real world systems and methods described in scientific papers.

Only five of our observed approaches definitively must have been available online at any point in time [28,45,65,84,119]. We did not encounter any of the more complex approaches being used in widespread paper recommendation systems.

Table 11 Overview of research groups with multiple papers

Group	Papers
Capital University of Science and Technology	[2,38]
Fırat University	[21,22]
IIT Kharagpur	[24–26]
Qufu Normal University	[60,61]
Kyoto-Kiel-Essex	[74–76]
University of Malaya-Bayero University	[88,89]
Pakistan	[93,94]
Hefei University of Technology	[110,111]
Shandong University	[115,116]
Australia	[121,122]

Table 12 Percentage of the 64 considered papers with different numbers of authors (#). Publications with 1 and 10 authors were encountered only once (1.56% each)

#	2	3	4	5	6	7	8
%	14.06	31.25	14.06	23.44	7.81	3.13	3.13

7.1.4 Persistence and authority

Beel et al. [16] describe the lack of persistence and authority in the field of paper recommendation systems as one of the main reasons why research is not adapted in practice.

The analysis of this possible shortcoming of current work could be highly affected by the short time period from which we observed works. We found several groups publishing multiple papers as seen in Table 11 which corresponds to 29.69% of approaches. The most papers a group published was three so this amount still cannot fully mark a research group as authority in the area.

7.1.5 Cooperation

Problems with cooperation are described by Beel et al. [16]. They state even though approaches have been proposed by multiple authors building upon prior work is rare. Corporations between different research groups are also only encountered sporadically.

Here again we want to point to the fact that our observed time frame of less than three years might be too short to

make substantive claims regarding this aspect. Table 12 holds information on the different numbers of authors for papers and the percentage of papers out of the 64 observed ones which are authored by groups of this size. We only encountered little cooperation between different co-author groups (see Haruna et al. [39] and Sakib et al. [88] for an exception). There were several groups not extending their previous work [121,122]. We refrain from analysing citations of related previous approaches as our considered period of less than three years is too short for all publications to have been able to be recognised by the wider scientific community.

7.1.6 Information scarcity

Information scarcity is described by Beel et al. [16] as researchers' tendency to only provide insufficient detail to re-implement their approaches. This leads to problems with reproducibility.

Many of the approaches we encountered did not provide sufficient information to make a re-implementation possible: with Afsar et al. [1] it is unclear how the knowledge graph and categories were formed, Collins and Beel [28] do not describe their Doc2Vec enough, Liu et al. [61] do not specify the extraction of keywords for papers in the graph and Tang et al. [104] do not clearly describe their utilisation of Word2Vec. In general oftentimes details are missing [3,4,60,117]. Exceptions to these observations are e.g. found with Bereczki [19], Nishioka et al. [74–76] and Sakib et al. [88].

We did not find a single paper's code e.g. provided as a link to GitHub.

7.1.7 Cold start

Pure collaborative filtering systems encounter the cold start problem as described by Bai et al. [9] and Shahid et al. [92]. If new users are considered, no historical data is available, they cannot be compared to other users to find relevant recommendations.

While this problem still persists, most current approaches are no pure collaborative filtering based recommendation systems (see Sect. 3.3.1). Systems using deep learning could overcome this issue [58]. There are approaches specifically targeting this problem [59,96], some [59] also introduced specific evaluation measures (totNP_EU and avgNP_EU) to quantify systems' ability to overcome the cold start problem.

7.1.8 Sparsity or reduce coverage

Bai et al. [9] state the user-paper-matrix being sparse for collaborative filtering based approaches. Shahid et al. [92] also mention this problem as the *reduce coverage problem*. This trait makes it hard for approaches to learn relevancy of infrequently rated papers.

Again, while this problem is still encountered, current approaches mostly are no longer pure collaborative filtering-based systems but instead utilise more information (see Sect. 3.3.1). Using deep learning in the recommendation process might reduce the impact of this problem [58].

7.1.9 Scalability

The problem of scalability was described by Bai et al. [9]. They state paper recommendation systems should be able to work in huge, ever expanding environments where new users and papers are added regularly.

A few approaches [38,46,88,109] contain a web crawling step which directly tackles challenges related to outdated or missing data. Some approaches [26,61] evaluate the time it takes to compute paper recommendations which also indicates their focus on this general problem. But most times scalability is not explicitly mentioned by current paper recommendation systems. There are several works [42,45,96, 108,116] evaluating on bigger datasets with over 1 million papers and which thus are able to handle big amounts of data. Sizes of current relevant real-world data collections exceed this threshold many times over (see, e.g. PubMed with over 33 million papers[29] or SemanticScholar with over 203 million papers[30]). Kanakia et al. [45] explicitly state scalability as a problem their approach is able to overcome. Instead of comparing each paper to all other papers they utilise clustering to reduce the number of required computations. They present the only approach running on several hundred million publications. Nair et al. [71] mention scalability issues they encountered even when only considering around 25,000 publications and their citation relations.

7.1.10 Privacy

The problem of privacy in personalised paper recommendation is described by Bai et al. [9]. Shahid et al. [92] also mention this as a problem occurring in collaborative filtering approaches. An issue is encountered when sensitive information such as habits or weaknesses that users might not want to disclose is used by a system. This leads to users' having negative impressions of systems. Keeping sensitive information private should therefore be a main goal.

In the current approaches, we did not find a discussion of privacy concerns. Some approach even explicitly utilise likes [84] or association rules [3] of other users while failing to mention privacy altogether. In approaches not incorporating any user data, this issue does not arise at all.

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

[30] https://www.semanticscholar.org/product/api.

7.1.11 Serendipity

Serendipity is described by Bai et al. [9] as an attribute often encountered in collaborative filtering [16]. Usually paper recommender systems focus on identification of relevant papers even though also including not obviously relevant ones might enhance the overall recommendation. Junior researchers could profit from stray recommendations to broaden their horizon, senior researchers might be able to gain knowledge to enhance their research. The ratio between clearly relevant and serendipitous papers is crucial to prevent users from losing trust in the recommender system.

A main objective of the works of Nishioka et al. [74–76] is serendipity. Other approaches do not mention this aspect.

7.1.12 Unified scholarly data standards

Different data formats of data collections is mentioned as a problem by Bai et al. [9]. They mention digital libraries containing relevant information which needs to be unified in order to use the data in a paper recommendation system. Additionally the combination of datasets could also lead to problems.

Many of the approaches we observe do not consider data collection or preparation as part of the approach, they often only mention the combination of different datasets as part of the evaluation (see e.g. Du et al. [29], Li et al. [56] or Xie et al. [115]). An exception to this general rule are systems which contain a web crawling step for data (see e.g. Ahmad and Afzal [2] or Sakib et al. [88]). Even with this type of approaches the combination of datasets and their diverse data formats is not identified as a problem.

7.1.13 Synonymy

Shahid et al. [92] describe the problem of synonymy encountered in collaborative filtering approaches. They define this problem as different words having the same meaning.

Even though there are still approaches (not necessarily CF ones) utilising basic TF-IDF representations of papers [2,42,86,95], nowadays this problem can be bypassed by using a text embedding method such as Doc2Vec or BERT.

7.1.14 Gray sheep

Gray sheep is a problem described by Shahid et al. [92] as an issue encountered in collaborative filtering approaches. They describe it as some users not consistently (dis)agreeing with any reference group.

We did not find any current approach mentioning this problem.

7.1.15 Black sheep

Black sheep is a problem described by Shahid et al. [92] as an issue encountered in collaborative filtering approaches. They describe it as some users not (dis)agree-ing with any reference group.

We did not find any current approach mentioning this problem.

7.1.16 Shilling attack

Shilling attacks are described by Shahid et al. [92] as a problem encountered in collaborative filtering approaches. They define this problem as users being able to manually enhance visibility of their own research by rating authored papers as relevant while negatively rating any other recommendations.

Although we did not find any current approach mentioning this problem we assume maybe it is no longer highly relevant as most approaches are no longer pure collaborative filtering ones. Additionally from the considered collaborative filtering approaches no one explicitly stated to feed relevance ratings back into the system.

7.2 Emerging challenges

In addition to the open challenges discussed in former literature reviews by Bai et al. [9], Beel et al. [16] and Shahid et al. [92] we identified the following problems and derive desirable goals for future approaches from them.

7.2.1 User evaluation

Paper recommendation is always targeted at human users. But oftentimes an evaluation with real users to quantify users' satisfaction with recommended publications is simply not conducted [84]. Conducting huge user studies is not feasible [38]. So sometimes user data to evaluate with is fetched from the presented datasets [39,88] or user behaviour is artificially emulated [1,19,57]. Noteworthy counter-examples[31] are the studies by Bulut et al. [22] who emailed 50 researchers to rate relevancy of recommended articles or Chaudhuri et al. [26] who asked 45 participants to rate their acceptance of recommended publications. Another option to overcome this issue is utilisation of living labs as seen with ArXivDigest [36], Mr. DLib's living lab [14] or LiLAS for the related tasks of dataset recommendation for scientific publications and multi-lingual document retrieval [91].

Desirable goal Paper recommendation systems targeted at users should always contain a user evaluation with a description of the composition of participants.

[31] For a full list of approaches conducting user studies see Table 9.

7.2.2 Target audience

Current works mostly fail to clearly characterise the intended users of a system altogether and the varying interests of different types of users are not examined in their evaluations. There are some noteworthy counter-examples: Afsar et al. [1] mention cancer patients and their close relatives as intended target audience. Bereczki [19] identifies new users as a special group they want to recommend papers to. Hua et al. [42] consider users who start diving into a topic which they have not yet researched before. Sharma et al. [95] name subject matter experts incorporating articles into a medical knowledge base as their target audience. Shi et al. [96] clearly state use cases for their approach which always target users which are unaware of a topic but already have one interesting paper from the area. They strive to recommend more papers similar to the first one.

User characteristics such as registration status of users are already mentioned by Beel et al. [16] as a factor which is disregarded in evaluations. We want to extend on this point and highlight the oftentimes missing or inadequate descriptions of intended users of paper recommendation systems. Traits of users and their information needs are not only important for experiments but should also be regarded in the construction of an approach. The targeted audience of a paper recommendation system should influence its suggestions. Bai et al. [9] highlight different needs of junior researchers which should be recommended a broad variety of papers as they still have to figure out their direction. They state recommendations for senior researchers should be more in line with their already established interests. Sugiyama and Kan [100] describe the need to help discover interdisciplinary research for this experienced user group. Most works do not recognise possible different functions of paper recommendation systems for users depending on their level of seniority. If papers include an evaluation with real persons, they e.g. mix Master's students with professors but do not address their different goals or expectations from paper recommendation [74]. Chaudhuri et al. [26] have junior, experienced and expert users as participants of their study and give individual ratings but do not calculate evaluation scores per user group. In some studies the exact composition of test users is not even mentioned (see Table 9).

Desirable goal Definition and consideration of a specific target audience for an approach and evaluation with members of this audience. If there is no specific person group a system should suit best, this should be discussed, executed and evaluated accordingly.

7.2.3 Recommendation scenario

Suggested papers from an approach should either be ones to read [44,109], to cite or fulfil another specified information need such as help patients in cancer treatment decision making [1]. Most work does not clearly state which is the case. Instead recommended papers are only said to be related [4,28], relevant [4,5,26,27,38,42,45,48,56,57,105,115,117], satisfactory [42,61], suitable [21], appropriate and useful [22,88] or a description which scenario is tackled is skipped altogether [3,37,39,84].

In rare cases if the recommendation scenario is mentioned there is the possibility of it not perfectly fitting the evaluated scenario. This can, e.g. be seen in the work of Jing and Yu [44] where they propose paper recommendation for papers to read but evaluate papers which were cited. Cited papers should always be ones which have been read beforehand but the decision to cite papers can be influenced by multiple aspects [34].

Desirable goal The clear description of the recommendation scenario is important for comparability of approaches as well as the validity of the evaluation.

7.2.4 Fairness/diversity

Anand et al. [8] define fairness as the balance between relevance and diversity of recommendation results. Only focusing on fit between the user or input paper and suggestions would lead to highly similar results which might not be vastly different from each other. Having diverse recommendation results can help cover multiple aspects of a user query instead of only satisfying the most prominent feature of the query [8]. In general more diverse recommendations provide greater utility for users [76]. Ekstrand et al. [31] give a detailed overview of current constructs for measuring algorithmic fairness in information access and describe possibly arising problems in this context.

Most of the current paper recommendation systems do not consider fairness but some approaches specifically mention diversity [26,74–76] while striving to recommend relevant publications. Thus these systems consider fairness.

Over one fourth of considered approaches with an evaluation report MMR as a measure of their system's quality. This at least seems to show researchers' awareness of the general problem of diverse recommendation results.

Desirable Goal Diversification of suggested papers to ensure fairness of the approach.

7.2.5 Complexity

Paper recommendation systems tend to become more complex, convoluted or composed of multiple parts. We observed this trend by regarding the classification of current systems compared to previous literature reviews (see Sect. 3.3.1). While systems' complexity increases, users' interaction with the systems should not become more complex. If an approach requires user interaction at all, it should be as simple as pos-

sible. Users should not be required to construct sophisticated knowledge graphs [109] or enter multiple rounds of keywords for an approach to learn their user profile [24].

Desirable Goal Maintain simplicity of usage even if approaches become more complex.

7.2.6 Explainability

Confidence in the recommendation system has already been mentioned by Beel et al. [16] as an example of what could enhance users' satisfaction but what is overlooked in approaches in favour of accuracy. This aspect should be considered with more vigour as the general research area of explainable recommendation has gained immense traction [120]. Gingstad et al. [36] regard explainability as a core component of paper recommendation systems. Xie et al. [116] mention explainability as a key feature of their approach but do not state how they achieve it or if their explanations satisfy users. Suggestions of recommendation systems should be explainable to enhance their trustworthiness and make them more engaging [66]. Here, different explanation goals such as effectiveness, efficiency, transparency or trust and their influence on each other should be considered [10]. If an approach uses neural networks [24,37,49,56] it is oftentimes impossible to explain why the system learned, that a specific suggested paper might be relevant.

Lee et al. [51] introduce a general approach which could be applied to any paper recommendation system to generate explanations for recommendations. Even though this option seems to help solve the described problem it is not clear how valuable post-hoc explanations are compared to systems which construct them directly.

Desirable Goal The conceptualisation of recommendation systems which comprehensibly explain their users why a specific paper is suggested.

7.2.7 Public dataset

Current approaches utilise many different datasets (see Table 4). A large portion of them are built by the authors such that they are not publicly available for others to use as well [1,30,111]. Part of the approaches already use open datasets in their evaluation but a large portion still does not seem to regard this as a priority (see Table 5). Utilisation of already public data sources or construction of datasets which are also published and remain available thus should be a priority in order to support reproducibility of approaches.

Desirable Goal Utilisation of publicly available datasets in the evaluation of paper recommendation systems.

7.2.8 Comparability

From the approaches we observed, many identified themselves as paper recommendation ones but only evaluated against systems, which are more general recommendation systems or ones utilising some same methodologies but not from the sub-domain of paper recommendation (seen with e.g. Guo et al [37], Tanner et al. [106] or Yang et al. [117]). While some of the works might claim to only be applied on paper recommendation and be of more general applicability (see, e.g. the works by Ahmedi et al. [3] or Alfarhood and Cheng [4]) we state that they should still be compared to ones, which mainly identify as paper recommendation systems as seen in the work of Chaudhuri et al. [24]. Only if a more general approach is compared to a paper recommendation approach, its usefulness for the area of paper recommendation can be fully assessed.

Several times, the baselines to evaluate against are not even other works but artificially constructed ones [2,38] or no other approach at all [22].

Desirable Goal Evaluation of paper recommendation approaches, even those which are applicable in a wider context, should always be against at least one paper recommendation system to clearly report relevance of the proposed method in the claimed context.

7.3 Discussion and outlook

From the already existing problems, several of them are still encountered in current paper recommendation approaches. Users are not always part of the approaches so users are not always modelled but this also prevents privacy issues. Accuracy seems to still be the main focus of recommendation systems. Novel techniques proposed in papers are not available online or applied by existing paper recommendation systems. Approaches do not provide enough details to enable re-implementation. Providing the code online or in a living lab environment could help overcome many of these issues.

Other problems mainly encountered in pure collaborative filtering systems such as the cold start problem, sparsity, synonymy, gray sheep, black sheep and shilling attacks do not seem to be as relevant anymore. We observed a trend towards hybrid models, this recommendation system type can overcome these issues. These hybrid models should also be able to produce serendipitous recommendations.

Unifying data sources is conducted often but nowadays it does not seem to be regarded as a problem. With scalability we encountered the same. Approaches are oftentimes able to handle millions of papers, here they do not specifically mention scalability as a problem they overcome but they also mostly do not consider huge datasets with several hundreds of millions of publications.

Due to the limited scope of our survey we are not able to derive substantive claims regarding cooperation and persistence. We found around 30% of approaches published by groups which authored multiple papers and very few collaborations between different author groups.

As for the newly introduced problems, part of the observed approaches conducted evaluations with users, on publicly available datasets and against other paper recommendation systems. Many works considered a low complexity for users. Even though user evaluations are desirable, they come with high costs. Usage of evaluation datasets with real human annotations could help overcome this issue partially, another straightforward solution would be the incorporation in a living lab. The second option would also help with comparability of approaches. Usage of available datasets can become increasingly complicated if approaches use new data which is currently not contained in existing datasets.[32]

Target audiences in general were rarely defined, the recommendation scenario was mostly not described. Diversity was considered by few. Overall the explainability of recommendations was dismissed. The first two of these issues are ones which could be comparatively easily fixed or addressed in the papers without changing the approach. As for diversity and explainability, the approaches would need to be modelled specifically such that these attributes could be satisfied.

To conclude, there are many challenges which are not constantly considered by current approaches. They define the requirements for future works in the area of paper recommendation systems.

8 Conclusion

This literature review of publications targeting paper recommendation between January 2019 and October 2021 provided comprehensive overviews of their methods, datasets and evaluation measures. We showed the need for a richer multi-dimensional characterisation of paper recommendation as former ones no longer seem sufficient in classifying the increasingly complex approaches. We also revisited known open challenges in the current time frame and highlighted possibly under-observed problems which future works could focus on.

Efforts should be made to standardise or better differentiate between the varying notions of relevancy and recommendation scenarios when it comes to paper recommendation. Future work could try revaluate already existing methods with real humans and against other paper recommendation systems. This could for example be realised in an extendable paper recommendation benchmarking system similar to the in a living lab environments ArXivDigest [36], Mr. DLib's living lab [14] or LiLAS [91] but with the additional property that it also provides build-in offline evaluations. As fairness and explainability of current paper recommendation systems have not been tackled widely, those aspects should be further explored. Another direction could be the comparison of multiple rare evaluation measures on the same system to help identify those which should be focused on in the future. As we observed a vast variety in datasets utilised for evaluation of the approaches (see Table 4), construction of publicly available and widely reusable ones would be worthwhile.

[32] We did not encounter many papers utilising types of data as part of their approach, which is not typically included in existing datasets; one of the noteworthy exceptions could be the approach by Nishioka et al. [74–76], which utilised Tweets of users.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Afsar, M.M., Crump, T., Far, B.H.: An exploration on-demand article recommender system for cancer patients information provisioning. In: FLAIRS Conference'21 (2021). https://doi.org/10.32473/flairs.v34i1.128339
2. Ahmad, S., Afzal, M.T.: Combining metadata and co-citations for recommending related papers. Turkish J. Electr. Eng. Comput. Sci. **28**(3), 1519–1534 (2020). https://doi.org/10.3906/elk-1908-19
3. Ahmedi, L., Rexhepi, E., Bytyçi, E.: Using association rule mining to enrich user profiles with research paper recommendation. Int. J. Comput. Dig Syst. (2021). https://doi.org/10.12785/ijcds/110192
4. Alfarhood, M., Cheng, J.: Collaborative attentive autoencoder for scientific article recommendation. In: ICMLA'19, pp. 168–174. IEEE (2019). https://doi.org/10.1109/ICMLA.2019.00034
5. Ali, Z., Qi, G., Muhammad, K., Ali, B., Abro, W.A.: Paper recommendation based on heterogeneous network embedding. Knowl. Based Syst. **210**, 106438 (2020). https://doi.org/10.1016/j.knosys.2020.106438
6. Alzoghbi, A., Ayala, V.A.A., Fischer, P.M., Lausen, G.: PubRec: Recommending Publications Based on Publicly Available Meta-Data. In: LWA'15, CEUR workshop proceedings. vol. 1458, pp. 11–18. CEUR-WS.org (2015). http://ceur-ws.org/Vol-1458/D01_CRC69_Alzoghbi.pdf
7. Amami, M., Faiz, R., Stella, F., Pasi, G.: A graph based approach to scientific paper recommendation. In: WI'17, pp. 777–782. ACM (2017). https://doi.org/10.1145/3106426.3106479

8. Anand, A., Chakraborty, T., Das, A.: FairScholar: balancing relevance and diversity for scientific paper recommendation. In: ECIR'17, LNCS. **10193**, 753–757 (2017). https://doi.org/10.1007/978-3-319-56608-5_76
9. Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., Xia, F.: Scientific paper recommendation: a survey. IEEE Access **7**, 9324–9339 (2019). https://doi.org/10.1109/ACCESS.2018.2890388
10. Balog, K., Radlinski, F.: Measuring recommendation explanation quality: the conflicting goals of explanations. In: SIGIR'20, pp. 329–338. ACM (2020). https://doi.org/10.1145/3397271.3401032
11. Barolli, L., Di Cicco, F., Fonisto, M.: An Investigation of Covid-19 Papers for a Content-Based Recommendation System. In: 3PGCIC'22, pp. 156–164. Springer (2022). https://doi.org/10.1007/978-3-030-89899-1_16
12. Basaldella, M., Nart, D.D., Tasso, C.: Introducing distiller: a unifying framework for knowledge extraction. In: IT@LIA@AI*IA'15, CEUR Workshop Proceedings, vol. 1509. CEUR-WS.org (2015). http://ceur-ws.org/Vol-1509/ITALIA2015_paper_4.pdf
13. Beel, J., Carevic, Z., Schaible, J., Neusch, G.: RARD: the related-article recommendation dataset. D Lib Mag. (2017). https://doi.org/10.1045/july2017-beel
14. Beel, J., Collins, A., Kopp, O., Dietz, L.W., Knoth, P.: Online Evaluations for Everyone: Mr. DLib's Living Lab for Scholarly Recommendations. In: ECIR'19, LNCS, **11438**, 213–219. Springer (2019). https://doi.org/10.1007/978-3-030-15719-7_27
15. Beel, J., Dinesh, S., Mayr, P., Carevic, Z., Jain, R.: Stereotype and most-popular recommendations in the digital library Sowiport. In: ISI'17, Schriften zur Informationswissenschaft, **70**, 96–108. Verlag Werner Hülsbusch (2017). https://doi.org/10.18452/1441
16. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: a literature survey. Int. J. Digit. Libr. **17**(4), 305–338 (2016). https://doi.org/10.1007/s00799-015-0156-0
17. Beel, J., Langer, S.: A Comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In: TPDL'15, LNCS, **9316**, 153–168. Springer (2015). https://doi.org/10.1007/978-3-319-24592-8_12
18. Beierle, F., Aizawa, A., Collins, A., Beel, J.: Choice overload and recommendation effectiveness in related-article recommendations. Int. J. Digit. Libr. **21**(3), 231–246 (2020). https://doi.org/10.1007/s00799-019-00270-7
19. Bereczki, M.: Graph neural networks for article recommendation based on implicit user feedback and content. Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS) (2021)
20. Bogers, T., van den Bosch, A.: recommending scientific articles using citeulike. In: RecSys'08, pp. 287–290. ACM (2008). https://doi.org/10.1145/1454008.1454053
21. Bulut, B., Gündoğan, E., Kaya, B., Alhajj, R., Kaya, M.: User's research interests based paper recommendation system: a deep learning approach. pp. 117–130. Springer (2020). https://doi.org/10.1007/978-3-030-33698-1_7
22. Bulut, B., Kaya, B., Kaya, M.: A paper recommendation system based on user interest and citations. In: UBMYK'19, pp. 1–5 (2019). https://doi.org/10.1109/UBMYK48245.2019.8965533
23. Champiri, Z.D., Asemi, A., Salim, S.S.B.: Meta-analysis of evaluation methods and metrics used in context-aware scholarly recommender systems. Knowl. Inf. Syst. **61**(2), 1147–1178 (2019). https://doi.org/10.1007/s10115-018-1324-5
24. Chaudhuri, A., Samanta, D., Sarma, M.: Modeling user behaviour in research paper recommendation system (2021). arXiv:2107.07831
25. Chaudhuri, A., Sarma, M., Samanta, D.: Advanced feature identification towards research article recommendation: a machine learning based approach. In: TENCON'19, pp. 7–12. IEEE (2019). https://doi.org/10.1109/TENCON.2019.8929386
26. Chaudhuri, A., Sinhababu, N., Sarma, M., Samanta, D.: Hidden features identification for designing an efficient research article recommendation system. Int. J. Digit. Libr. **22**(2), 233–249 (2021). https://doi.org/10.1007/s00799-021-00301-2
27. Chen, J., Ban, Z.: Academic Paper Recommendation Based on Clustering and Pattern Matching, pp. 171–182. Springer, Cham (2019). https://doi.org/10.1007/978-981-32-9298-7_14
28. Collins, A., Beel, J.: Document embeddings vs. keyphrases vs. terms for recommender systems: a large-scale online evaluation. In: JCDL'19, pp. 130–133. IEEE (2019). https://doi.org/10.1109/JCDL.2019.00027
29. Du, N., Guo, J., Wu, C.Q., Hou, A., Zhao, Z., Gan, D.: Recommendation of academic papers based on heterogeneous information networks. In: AICCSA'20, pp. 1–6. IEEE (2020). https://doi.org/10.1109/AICCSA50499.2020.9316516
30. Du, Z., Tang, J., Ding, Y.: POLAR++: active one-shot personalized article recommendation. IEEE Trans. Knowl. Data Eng. **33**(6), 2709–2722 (2021). https://doi.org/10.1109/TKDE.2019.2953721
31. Ekstrand, M.D., Das, A., Burke, R., Diaz, F.: Fairness and discrimination in information access systems (2021). arXiv:2105.05779
32. Färber, M., Jatowt, A.: Citation recommendation: approaches and datasets. Int. J. Digit. Libr. **21**(4), 375–405 (2020). https://doi.org/10.1007/s00799-020-00288-2
33. Feng, S., Meng, J., Zhang, J.: News recommendation systems in the era of information overload. J. Web Eng. **20**(2), 459–470 (2021). https://doi.org/10.13052/jwe1540-9589.20210
34. Garfield, E.: Can citation indexing be automated? (1964)
35. Giles, C.L., Bollacker, K.D., Lawrence, S.: CiteSeer: an automatic citation indexing system. In: ACM DL'98, pp. 89–98. ACM (1998). https://doi.org/10.1145/276675.276685
36. Gingstad, K., Jekteberg, Ø., Balog, K.: ArXivDigest: a living lab for personalized scientific literature recommendation. In: CIKM'20, pp. 3393–3396. ACM (2020). https://doi.org/10.1145/3340531.3417417
37. Guo, G., Chen, B., Zhang, X., Liu, Z., Dong, Z., He, X.: Leveraging title-abstract attentive semantics for paper recommendation. In: IAAI'20, pp. 67–74. AAAI Press (2020). https://aaai.org/ojs/index.php/AAAI/article/view/5335
38. Habib, R., Afzal, M.T.: Sections-based bibliographic coupling for research paper recommendation. Scientometrics **119**(2), 643–656 (2019). https://doi.org/10.1007/s11192-019-03053-8
39. Haruna, K., Ismail, M.A., Qazi, A., Kakudi, H.A., Hassan, M., Muaz, S.A., Chiroma, H.: Research paper recommender system based on public contextual metadata. Scientometrics **125**(1), 101–114 (2020). https://doi.org/10.1007/s11192-020-03642-y
40. Hienert, D., Sawitzki, F., Mayr, P.: Digital library research in action: supporting information retrieval in sowiport. D Lib. Mag. (2015). https://doi.org/10.1045/march2015-hienert
41. Hu, D., Ma, H., Liu, Y., He, X.: Scientific paper recommendation using author's dual role citation relationship. In: Intelligent information processing'20, IFIP advances in information and communication technology. **581**, 121–132. Springer (2020). https://doi.org/10.1007/978-3-030-46931-3_12
42. Hua, S., Chen, W., Li, Z., Zhao, P., Zhao, L.: Path-Based Academic Paper Recommendation. In: WISE'20, LNCS, **12343**, 343–356. Springer (2020). https://doi.org/10.1007/978-3-030-62008-0_24
43. Ji, Z., Wu, M., Yang, H., Armendáriz-Iñigo, J.E.: Temporal sensitive heterogeneous graph neural network for news recommendation. Future Gener. Comput. Syst. **125**, 324–333 (2021). https://doi.org/10.1016/j.future.2021.06.007
44. Jing, S., Yu, S.: Research of paper recommendation system based on citation network model. In: ML4CS'20, LNCS, **12488**,

237–247. Springer (2020). https://doi.org/10.1007/978-3-030-62463-7_22
45. Kanakia, A., Shen, Z., Eide, D., Wang, K.: A Scalable hybrid research paper recommender system for microsoft academic (2019). arXiv:1905.08880
46. Kang, Y., Hou, A., Zhao, Z., Gan, D.: A hybrid approach for paper recommendation. IEICE Trans. Inform. Syst. **E104D**(8), 1222–1231 (2021). https://doi.org/10.1587/transinf.2020BDP0008
47. Keller, J., Munz, L.P.M.: TEKMA at CLEF-2021: BM-25 based rankings for scientific publication retrieval and data set recommendation. In: CLEF'21, CEUR Workshop Proceedings. **2936**, 1700–1711. CEUR-WS.org (2021). http://ceur-ws.org/Vol-2936/paper-144.pdf
48. Kong, X., Mao, M., Wang, W., Liu, J., Xu, B.: VOPRec: vector representation learning of papers with text information and structural identity for recommendation. IEEE Trans. Emerg. Top. Comput. **9**(1), 226–237 (2021). https://doi.org/10.1109/TETC.2018.2830698
49. L, H., Liu, S., Pan, L.: Paper recommendation based on author-paper interest and graph structure. In: CSCWD'21, pp. 256–261. IEEE (2021). https://doi.org/10.1109/CSCWD49262.2021.9437743
50. Le, M., Kayal, S., Douglas, A.: The impact of recommenders on scientific article discovery: the case of mendeley suggest. In: ImpactRS@RecSys'19, CEUR Workshop Proceedings, vol. 2462. CEUR-WS.org (2019). http://ceur-ws.org/Vol-2462/paper5.pdf
51. Lee, B.C.G., Lo, K., Downey, D., Weld, D.S.: Explanation-based tuning of opaque machine learners with application to paper recommendation (2020). arXiv:2003.04315
52. Lee, J., Lee, K., Kim, J.G.: Personalized academic research paper recommendation system (2013). arXiv:1304.5457
53. Leskovec, J., Kleinberg, J.M., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: SIGKDD'05, pp. 177–187. ACM (2005). https://doi.org/10.1145/1081870.1081893
54. Ley, M.: DBLP: some lessons learned. Proc VLDB Endow. **2**(2), 1493–1500 (2009). https://doi.org/10.14778/1687553.1687577
55. Li, W., Chang, C., He, C., Wu, Z., Guo, J., Peng, B.: Academic paper recommendation method combining heterogeneous network and temporal attributes. In: Chinese CSCW'21, pp. 456–468. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-2540-4_33
56. Li, X., Chen, Y., Pettit, B., de Rijke, M.: Personalised reranking of paper recommendations using paper content and user behavior. ACM Trans. Inform. Syst. **37**(3), 1–23 (2019). https://doi.org/10.1145/3312528
57. Li, Y., Wang, R., Nan, G., Li, D., Li, M.: A personalized paper recommendation method considering diverse user preferences. Decis. Support Syst. **146**, 113546 (2021). https://doi.org/10.1016/j.dss.2021.113546
58. Li, Z., Zou, X.: A review on personalized academic paper recommendation. Comput. Inf. Sci. **12**(1), 33–43 (2019). https://doi.org/10.5539/cis.v12n1p33
59. Lin, S.j., Lee, G., Peng, S.L.: Academic article recommendation by considering the research field trajectory. pp. 447–454. Springer (2021). https://doi.org/10.1007/978-3-030-65407-8_39
60. Liu, H., Kou, H., Chi, X., Qi, L.: Combining time, keywords and authors information to construct papers correlation graph (S). In: SEKE'19, pp. 11–19. KSI Research Inc. and Knowledge Systems Institute Graduate School (2019). https://doi.org/10.18293/SEKE2019-161
61. Liu, H., Kou, H., Yan, C., Qi, L.: Keywords-driven and popularity-aware paper recommendation based on undirected paper citation graph. Complex (2020). https://doi.org/10.1155/2020/2085638
62. Lu, Y., He, Y., Cai, Y., Peng, Z., Tang, Y.: Time-aware neural collaborative filtering with multi-dimensional features on academic paper recommendation. In: CSCWD'21, pp. 1052–1057. IEEE (2021). https://doi.org/10.1109/CSCWD49262.2021.9437673
63. Ma, X., Wang, R.: Personalized scientific paper recommendation based on heterogeneous graph representation. IEEE Access **7**, 79887–79894 (2019). https://doi.org/10.1109/ACCESS.2019.2923293
64. Ma, X., Zhang, Y., Zeng, J.: Newly published scientific papers recommendation in heterogeneous information networks. Mob. Networks Appl. **24**(1), 69–79 (2019). https://doi.org/10.1007/s11036-018-1133-9
65. Manju, G., Abhinaya, P., Hemalatha, M.R., Manju, G.G., Manju, G.G.: Cold start problem alleviation in a research paper recommendation system using the random walk approach on a heterogeneous user-paper graph. Int. J. Intell. Inf. Technol. **16**(2), 24–48 (2020). https://doi.org/10.4018/IJIIT.2020040102
66. McInerney, J., Lacker, B., Hansen, S., Higley, K., Bouchard, H., Gruson, A., Mehrotra, R.: Explore, exploit, and explain: personalizing explainable recommendations with bandits. In: RecSys'18, pp. 31–39. ACM (2018). https://doi.org/10.1145/3240323.3240354
67. Medic, Z., Snajder, J.: A survey of citation recommendation tasks and methods. J. Comput. Inf. Technol. **28**(3), 183–205 (2020)
68. Mohamed Hassan, H.A., Sansonetti, G., Gasparetti, F., Micarelli, A., Beel, J.: BERT, ELMo, USE and InferSent sentence encoders: the panacea for research-paper recommendation? In: RecSys'19, CEUR Workshop Proceedings. **2431**, 6–10. CEUR-WS.org (2019). http://ceur-ws.org/Vol-2431/paper2.pdf
69. Mohamed Hassan, H.A., Sansonetti, G., Micarelli, A.: Tag-aware document representation for research paper recommendation (2020). https://www.researchgate.net/publication/343319230_Tag-Aware_Document_Representation_for_Research_Paper_Recommendation
70. Moskalenko, O., Sáez-Trumper, D., Parra, D.: Scalable recommendation of wikipedia articles to editors using representation learning. In: RecSys'20, CEUR workshop proceedings. **2697**. CEUR-WS.org (2020). http://ceur-ws.org/Vol-2697/paper1_complexrec.pdf
71. Nair, A.M., Benny, O., George, J.: Content based scientific article recommendation system using deep learning technique. In: Inventive systems and control, pp. 965–977. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-1395-1_70
72. Ng, Y.: Research paper recommendation based on content similarity, peer reviews, authority, and popularity. In: ICTAI'20, pp. 47–52. IEEE (2020). https://doi.org/10.1109/ICTAI50040.2020.00018
73. Ng, Y.K.: CBRec: a book recommendation system for children using the matrix factorisation and content-based filtering approaches. Int. J. Bus. Intell. Data Mining **16**(2), 129–149 (2020). https://doi.org/10.1504/IJBIDM.2020.104738
74. Nishioka, C., Hauke, J., Scherp, A.: Research paper recommender system with serendipity using tweets vs. diversification. In: ICADL'19, LNCS. **11853**, 63–70. Springer (2019). https://doi.org/10.1007/978-3-030-34058-2_7
75. Nishioka, C., Hauke, J., Scherp, A.: Towards serendipitous research paper recommender using tweets and diversification. In: TPDL'19, LNCS. **11799**, 339–343. Springer (2019). https://doi.org/10.1007/978-3-030-30760-8_29
76. Nishioka, C., Hauke, J., Scherp, A.: Influence of tweets and diversification on serendipitous research paper recommender systems. PeerJ Comput. Sci. **6**, e273 (2020). https://doi.org/10.7717/peerj-cs.273
77. Ostendorff, M.: Contextual document similarity for content-based literature recommender systems (2020). arXiv:2008.00202

78. Ostendorff, M., Breitinger, C., Gipp, B.: A qualitative evaluation of user preference for link-based vs. text-based recommendations of wikipedia articles (2021). arXiv:2109.07791
79. Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., Moher, D.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ **372** (2021). https://www.bmj.com/content/372/bmj.n71
80. Patra, B.G., Maroufy, V., Soltanalizadeh, B., Deng, N., Zheng, W.J., Roberts, K., Wu, H.: A content-based literature recommendation system for datasets to improve data reusability: a case study on gene expression omnibus (geo) datasets. J. Biomed. Inform. **104**, 103399 (2020). https://doi.org/10.1016/j.jbi.2020.103399
81. Radev, D.R., Joseph, M.T., Gibson, B., Muthukrishnan, P.: A bibliometric and network analysis of the field of computational linguistics. J. Am. Soc. Inform. Sci. Technol. (2009). https://doi.org/10.1002/asi.23394
82. Radev, D.R., Muthukrishnan, P., Qazvinian, V.: The ACL anthology network corpus. In: NLPIR4DL'09 (2009). https://doi.org/10.5555/1699750.1699759
83. Radev, D.R., Muthukrishnan, P., Qazvinian, V., Abu-Jbara, A.: The ACL anthology network corpus. Lang. Resour. Eval. (2013). https://doi.org/10.1007/s10579-012-9211-2
84. Rahdari, B., Brusilovsky, P.: User-controlled hybrid recommendation for academic papers. In: IUI Companion'19, pp. 99–100. ACM (2019). https://doi.org/10.1145/3308557.3308717
85. Rahdari, B., Brusilovsky, P., Thaker, K., Barria-Pineda, J.: Knowledge-driven wikipedia article recommendation for electronic textbooks. In: EC-TEL'20, LNCS. **12315**, 363–368. Springer (2020). https://doi.org/10.1007/978-3-030-57717-9_28
86. Renuka, S., Raj Kiran, G.S.S., Rohit, P.: An unsupervised content-based article recommendation system using natural language processing. Data Intell. Cogn. Inform. (2021). https://doi.org/10.1007/978-981-15-8530-2_13
87. Safaryan, A., Filchenkov, P., Yan, W., Kutuzov, A., Nikishina, I.: Semantic recommendation system for bilingual corpus of academic papers. In: AIST'20, Communications in Computer and Information Science. **1357**, 22–36. Springer (2020). https://doi.org/10.1007/978-3-030-71214-3_3
88. Sakib, N., Ahmad, R.B., Ahsan, M., Based, M.A., Haruna, K., Haider, J., Gurusamy, S.: A hybrid personalized scientific paper recommendation approach integrating public contextual metadata. IEEE Access **9**, 83080–83091 (2021). https://doi.org/10.1109/ACCESS.2021.3086964
89. Sakib, N., Ahmad, R.B., Haruna, K.: A collaborative approach toward scientific paper recommendation using citation context. IEEE Access **8**, 51246–51255 (2020). https://doi.org/10.1109/ACCESS.2020.2980589
90. Samad, A., Islam, M.A., Iqbal, M.A., Aleem, M.: Centrality-based paper citation recommender system. EAI Endorsed Trans. Ind. Networks Intell. Syst. **6**(19), e2 (2019). https://doi.org/10.4108/eai.13-6-2019.159121
91. Schaer, P., Breuer, T., Castro, L.J., Wolff, B., Schaible, J., Tavakolpoursaleh, N.: Overview of LiLAS 2021—living labs for academic search (extended overview). In: CLEF'21, CEUR Workshop Proceedings. **2936**, 1668–1699. CEUR-WS.org (2021). http://ceur-ws.org/Vol-2936/paper-143.pdf
92. Shahid, A., Afzal, M.T., Abdar, M., Basiri, M.E., Zhou, X., Yen, N.Y., Chang, J.: Insights into relevant knowledge extraction techniques: a comprehensive review. J. Supercomput. **76**(3), 1695–1733 (2020). https://doi.org/10.1007/s11227-019-03009-y
93. Shahid, A., Afzal, M.T., Alharbi, A., Aljuaid, H., Al-Otaibi, S.: In-text citation's frequencies-based recommendations of relevant research papers. PeerJ Comput. Sci. **7**, e524 (2021). https://doi.org/10.7717/peerj-cs.524
94. Shahid, A., Afzal, M.T., Saleem, M.Q., Idrees, M.S.E., Omer, M.K.: Extension of direct citation model using in-text citations. Comput. Mater. Continua **66**(3), 3121–3138 (2021). https://doi.org/10.32604/cmc.2021.013809
95. Sharma, B., Willis, V.C., Huettner, C.S., Beaty, K., Snowdon, J.L., Xue, S., South, B.R., Jackson, G.P., Weeraratne, D., Michelini, V.: Predictive article recommendation using natural language processing and machine learning to support evidence updates in domain-specific knowledge graphs. JAMIA Open **3**(3), 332–337 (2020). https://doi.org/10.1093/jamiaopen/ooaa028
96. Shi, H., Ma, W., Zhang, X., Jiang, J., Liu, Y., Chen, S.: A hybrid paper recommendation method by using heterogeneous graph and metadata. In: IJCNN'20, pp. 1–8. IEEE (2020). https://doi.org/10.1109/IJCNN48605.2020.9206733
97. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.P., Wang, K.: An overview of microsoft academic service (MAS) and applications. In: WWW'15, pp. 243–246. ACM (2015). https://doi.org/10.1145/2740908.2742839
98. Subathra, P., Kumar, P.N.: Recommending research article based on user queries using latent dirichlet allocation. in: soft computing and signal processing, pp. 163–175. Springer Singapore (2020). https://doi.org/10.1007/978-981-15-2475-2_15
99. Sugiyama, K., Kan, M.: Scholarly paper recommendation via user's recent research interests. In: JCDL'10, pp. 29–38. ACM (2010). https://doi.org/10.1145/1816123.1816129
100. Sugiyama, K., Kan, M.: Serendipitous recommendation for scholarly papers considering relations among researchers. In: JCDL'11, pp. 307–310. ACM (2011). https://doi.org/10.1145/1998076.1998133
101. Sugiyama, K., Kan, M.: Exploiting potential citation papers in scholarly paper recommendation. In: JCDL'13, pp. 153–162. ACM (2013). https://doi.org/10.1145/2467696.2467701
102. Sugiyama, K., Kan, M.: A comprehensive evaluation of scholarly paper recommendation using potential citation papers. Int. J. Digit. Libr. **16**(2), 91–109 (2015). https://doi.org/10.1007/s00799-014-0122-2
103. Symeonidis, P., Kirjackaja, L., Zanker, M.: Session-based news recommendations using SimRank on multi-modal graphs. Expert Syst. Appl. **180**, 115028 (2021). https://doi.org/10.1016/j.eswa.2021.115028
104. Tang, H., Liu, B., Qian, J.: Content-based and knowledge graph-based paper recommendation: exploring user preferences with the knowledge graphs for scientific paper recommendation. Concurr. Comput. Pract. Exp. (2021). https://doi.org/10.1002/cpe.6227
105. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In: SIGKDD'08, pp. 990–998. ACM (2008). https://doi.org/10.1145/1401890.1402008
106. Tanner, W., Akbas, E., Hasan, M.: Paper recommendation based on citation relation. In: Big Data'19, pp. 3053–3059. IEEE (2019). https://doi.org/10.1109/BigData47090.2019.9006200
107. Tao, M., Yang, X., Gu, G., Li, B.: Paper recommend based on LDA and PageRank, pp. 571–584. Springer (2020). https://doi.org/10.1007/978-981-15-8101-4_51
108. Waheed, W., Imran, M., Raza, B., Malik, A.K., Khattak, H.A.: A hybrid approach toward research paper recommendation using centrality measures and author ranking. IEEE Access **7**, 33145–33158 (2019). https://doi.org/10.1109/ACCESS.2019.2900520
109. Wang, B., Weng, Z., Wang, Y.: A novel paper recommendation method empowered by knowledge graph: for research beginners (2021). arXiv:2103.08819

110. Wang, G., Wang, H., Yang, Y., Xu, D., Yang, J., Yue, F.: Group article recommendation based on ER rule in scientific social networks. Appl. Soft Comput. **110**, 107631 (2021). https://doi.org/10.1016/j.asoc.2021.107631
111. Wang, G., Zhang, X., Wang, H., Chu, Y., Shao, Z.: Group-oriented paper recommendation with probabilistic matrix factorization and evidential reasoning in scientific social network. IEEE Trans. Syst. Man Cybern. Syst. (2021). https://doi.org/10.1109/TSMC.2021.3072426
112. Wang, H., Chen, B., Li, W.: Collaborative topic regression with social regularization for tag recommendation. In: IJCAI'13, pp. 2719–2725. IJCAI/AAAI (2013). http://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/7006
113. Wang, X., Xu, H., Tan, W., Wang, Z., Xu, X.: Scholarly paper recommendation via related path analysis in knowledge graph. In: ICSS'20, pp. 36–43. IEEE (2020). https://doi.org/10.1109/ICSS50103.2020.00014
114. Wu, J., Kim, K., Giles, C.L.: CiteSeerX: 20 years of service to scholarly big data. In: AIDR'19, pp. 1:1–1:4. ACM (2019). https://doi.org/10.1145/3359115.3359119
115. Xie, Y., Sun, Y., Bertino, E.: Learning domain semantics and cross-domain correlations for paper recommendation. In: SIGIR'21, pp. 706–715. ACM (2021). https://doi.org/10.1145/3404835.3462975
116. Xie, Y., Wang, S., Pan, W., Tang, H., Sun, Y.: Embedding based personalized new paper recommendation. In: Chinese CSCW'21, pp. 558–570. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-2540-4_40
117. Yang, Q., Li, Z., Liu, A., Liu, G., Zhao, L., Zhang, X., Zhang, M., Zhou, X.: A novel hybrid publication recommendation system using compound information. World Wide Web **22**(6), 2499–2517 (2019). https://doi.org/10.1007/s11280-019-00687-9
118. Yu, M., Hu, Y., Li, X., Zhao, M., Xu, T., Liu, H., Xu, L., Yu, R.: Paper recommendation with item-level collaborative memory network. In: KSEM'19, LNCS. **11775**, 141–152. Springer (2019). https://doi.org/10.1007/978-3-030-29551-6_13
119. Zavrel, J., Grotov, A., Mitnik, J.: Building a platform for ensemble-based personalized research literature recommendations for AI and data science at zeta alpha. pp 536–537. Association for Computing Machinery (2021). https://doi.org/10.1145/3460231.3474619
120. Zhang, Y., Chen, X.: Explainable recommendation: a survey and new perspectives. Found. Trends Inf. Retr. **14**(1), 1–101 (2020). https://doi.org/10.1561/1500000066
121. Zhang, Y., Wang, M., Gottwalt, F., Saberi, M., Chang, E.: Ranking scientific articles based on bibliometric networks with a weighting scheme. J. Inform. **13**(2), 616–634 (2019). https://doi.org/10.1016/j.joi.2019.03.013
122. Zhang, Y., Wang, M., Saberi, M., Chang, E.: Towards expert preference on academic article recommendation using bibliometric networks. In: PAKDD'20, pp. 11–19. Springer (2020). https://doi.org/10.1007/978-3-030-60470-7_2
123. Zhao, X., Kang, H., Feng, T., Meng, C., Nie, Z.: A hybrid model based on LFM and BiGRU toward research paper recommendation. IEEE Access **8**, 188628–188640 (2020). https://doi.org/10.1109/ACCESS.2020.3031281
124. Zhu, Y., Lin, Q., Lu, H., Shi, K., Qiu, P., Niu, Z.: Recommending scientific paper via heterogeneous knowledge embedding based attentive recurrent neural networks. Knowl. Based Syst. **215**, 106744 (2021). https://doi.org/10.1016/j.knosys.2021.106744