International Journal on Digital Libraries
https://doi.org/10.1007/s00799-024-00398-1

Author name disambiguation literature review with consolidated meta-analytic approach

Natan S. Rodrigues[1] · **Ari M. Mariano**[2] · **Celia G. Ralha**[1]

Received: 30 April 2023 / Revised: 6 February 2024 / Accepted: 3 March 2024

Abstract
Name ambiguity is a common problem in many bibliographic repositories affecting data integrity and validity. This article presents an author name disambiguation (AND) literature review using the theory of the consolidated meta-analytic approach, including quantitative techniques and bibliometric aspects. The literature review covers information from 211 documents of the Web of Science and Scopus databases in the period 2003 to 2022. A taxonomy based on the literature was used to organize the identified approaches to solve the AND problem. We identified that the most widely used AND solving approaches are author grouping associated with similarity functions and clustering methods and some works using author assignment allied to classification methods. The countries that publish most in AND are the USA, China, Germany, and Brazil with 21%, 19%, 13% and 8% of the total papers, respectively. The review results provide an overview of AND state-of-the-art research that can direct further investigation based on the quantitative and qualitative information from the AND research history.

Keywords Name ambiguity · Author name disambiguation · Bibliographic repository · Theory of the consolidated meta-analytic approach · Literature review

1 Introduction

Scientific digital bibliographic repositories, such as DBLP [1], AMiner [2], and CiteSeerX [3], provide bibliographic information offering features that allow the identification of scientific research, authors, and their respective communities. Such repositories can list millions of bibliographic records presenting a vital source of information for academic communities and allowing relevant publication search in a centralized way [4]. In addition to the literature search facility, these digital libraries provide functional analysis and information used for decision-making by funding agencies and academic institutions for grants and individual promotion decisions [5].

Name ambiguity may arise in citation records when an author's name is not accurately identified. This situation can occur when an author is listed in the bibliography using different names or when two or more authors have the same name.

This problem can be due to many reasons, including name changes due to personal reasons, variations in the transliteration of non-Roman names, typographical errors, the absence of standard practices, and decentralized content generation, such as through automatic harvesting. These aspects have been discussed in various studies, as evidenced in previous literature [6–8].

Consequently, the effectiveness of the primary functions of digital bibliographic repositories, including searching, navigating, and suggesting content, can be significantly impacted by the uncertainty surrounding author names. These issues may impact significantly the accuracy and reliability of citation records, which can affect the quality of scientific research [4].

However, developing effective methods for AND is a challenging task. In the literature, there are different approaches to solving this problem, with techniques varying from

✉ Natan S. Rodrigues
natan5souza@gmail.com

Ari M. Mariano
arimariano@unb.br

Celia G. Ralha
ghedini@unb.br

[1] Computer Science Department, Exact Sciences Institute, University of Brasilia, Campus Universitário Darcy Ribeiro, Federal District, Brasília 70910-900, Brazil

[2] Production Engineering Department, Faculty of Technology, University of Brasilia, Campus Universitário Darcy Ribeiro, Federal District, Brasília 70910-900, Brazil

Published online: 10 April 2024

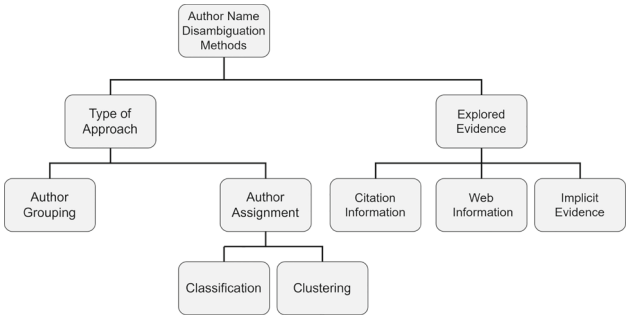

Fig. 1 The taxonomy used in this literature review [4, 11]

heuristic-based to more complex methods that leverage artificial intelligence with supervised and unsupervised learning [9]. This set of methods forms an area of study known as Author Name Disambiguation (AND) [10].

With the scientific interest and concern for AND, literature reviews presenting different methods are emerging. The review of [5] provides techniques available in the literature from 2010 to 2016, comparing them at an abstract level, discussing limitations, and classifying them into five categories. But these categories only classify the techniques, unlike other review approaches that classify the type of evidence explored. The authors in [11] presented a brief survey of AND automatic methods and proposed a taxonomy for classifying the techniques, including explored evidence types. In [4], the authors updated the review emphasizing the previous taxonomy and sorting automatic techniques for AND in bibliographic repositories (2003–2020).

In [12], the literature review focuses on approaches that applied AND methods in the PubMed bibliographic repository until 2019. The authors proposed a new taxonomy with a subdivision of the category author grouping based on the taxonomy presented in [4, 11] with similar evidence types explored. A recent review analyzes the development of incremental AND methods using similarity comparison strategies from 2011 to 2020 [13].

In this literature review work, we use the hierarchical taxonomy proposed by [4, 11] to classify AND approaches. The documents found during this review fit adequately in the taxonomy classifications according to the diagram shown in Fig. 1. In the sequence, AND methods and techniques of this taxonomy are detailed.

1.1 Type of approach

- Author Grouping aims to group references of the same author using a type of similarity by analyzing the attributes of these references. Usually, these methods use clustering techniques, pre-defined similarity functions, or machine learning techniques, extracting information from co-authorship relationships or a set of heuristic rules.
- Author Assignment methods assign each author record using the construction of a model that represents the author. These methods aim to directly attribute the authorship record to their respective authors, adopting some classification or clustering technique.

1.1.1 Explored evidence

- Citation Information extracts information directly from the citation records, such as author and co-author names, paper titles, year of publication, and other information. These attributes are the most commonly used AND methods available in the literature. However, sometimes they do not provide enough information about the approaches used.
- Web Information is extracted from the Web and used as supplementary information about an author's publication profile. This obtained information is used as attributes to calculate the authorship record similarity.
- Implicit Evidence is obtained from visible attribute elements, such as the latent topics of a citation which returns each topic probability given a particular citation. This value is used as an attribute or evidence to calculate the similarity between authorship records.

Analyzing the reviews presented, we note they classify AND approaches, describe the main characteristics, and propose a taxonomy for classification. In contrast, our review work imposes questions focusing on the most cited works, most relevant authors, most-used approaches, and pursued lines of research. To answer such questions, we use meta-analysis based on statistics to summarize the work results. The Theory of Consolidated Meta-analytic Approach (*Teoria do Enfoque Meta-analítico Consolidado* - TEMAC [14], in Portuguese) is used to collect data on AND area with quantitative information available in bibliographic repositories, such as the Web of Science (WoS) and Scopus. However, a systematic review answers research questions by collecting and summarizing empirical evidence that fits pre-specified eligibility criteria as suggested by [15, 16]. We think a systematic review would add investigation threads to our work in a complementary way but not replaceable. In addition, a meta-analytical AND review has not been conducted to date, creating a foundation for forthcoming research endeavors that will contribute to the existing body of knowledge built upon prior review studies [5, 11, 12].

The following sections present the literature review methodology, results and analysis, and conclusions with future work.

2 Methodology

Traditionally, systematic literature reviews focus on accessing many different digital bibliographic repositories to enrich the findings [15, 16]. In this work, a consolidated meta-analytic approach was chosen to carry out a literature review because of the methodological advantage, as its process reduces the access to private scientific digital bibliographic repositories minimizing bias and maximizing the coverage possibility. Specifically, considering the Brazilian Public Universities, there is a free access to major scientific repositories such as Scopus and WoS through the Periodicos portal of the Coordination for the Improvement of Higher Education Personnel (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - CAPES), institution linked to the Ministry of Education in Brazil.

The TEMAC [14] emerged as an exploratory solution, supported by prior strategies and grounded in bibliometric principles focusing on the need to unify various systematic methods with a meta-analytical framework with recent publications [17, 18]. Moreover, the AND meta-analytical review studies have not yet been conducted, starting a point for future research work adding to the body of knowledge of existing systematic review studies.

The TEMAC consists of three steps, the research preparation, data presentation and interrelation, detailing integration model, and validation by evidence as presented in the sequence.

2.1 Research preparation

The review preparation step is vital as wrong choices generate unsatisfactory results (e.g., inadequate search strings). The work of [19], using meta-analytic analysis, states that one of the essential stages in review studies is the reading of articles to define specific criteria to include and exclude studies. We defined specific criteria for inclusion (IC) and exclusion (EC) to guide the selection of academic works as follows:

- IC-1: primarily addressing AND as an integral component of the study.
- IC-2: published in peer-reviewed conferences or journals, available in major bibliographic repositories.
- EC-1: works that are not available in online repositories.
- EC-2: primarily associated with domains other than information systems, computer science, and engineering.
- EC-3: published beyond the timeframe of 2003–2022.

These choices were substantiated by addressing four fundamental questions during the selection process:

1. What is the search descriptor, string, or keyword?
   The string “author name disambiguation" is used without and/or connectives to include studies in different contexts and scientific bibliographic repositories.
2. Which are the databases?
   The WoS and Scopus bibliographic repositories, esteemed within numerous academic communities, were chosen due to the fact that the works within these databases emanate from peer-reviewed conferences or journals. This decision was grounded in WoS’s extensive temporal coverage and Scopus’s comprehensive scope of science and technology journals. These databases are complementary and widely employed in literature reviews.
3. What is the space-time field of the research?
   Temporal delineation is crucial, as databases have varying time coverage. These documents ranged from 2003, which marked the first work on AND on the web, to 2022.
4. Which are the knowledge areas?
   We determined the knowledge domains after examining the documents included in the WoS and Scopus databases. After this examination, the domains were categorized into Computer Science, Social and Information Sciences, Medicine, Engineering, and Mathematics. Table 1 presents AND knowledge areas in both databases.

Using a meta-analytic review approach establishes a foundation for conducting an exploratory study, ensuring the inclusion of relevant works for constructing up-to-date knowledge of the research AND area.

2.2 Data presentation and interrelation

This section covers the data presentation and interrelation aspects of TEMAC meta-analytical framework, including the laws and tools used in the exploratory study of the AND research area.

2.2.1 Laws

The TEMAC meta-analytical overall framework includes quantitative techniques and bibliometric aspects based on three laws.

- Bradford’s law [20] allows finding journals that publish the most on the topic. The scientific journals of an area should be ordered in a decreasing manner according to their productivity, generating nuclei where appear few journals usually account for a high share of total publications. While a high number of journals publish fewer articles in the area [21]. This law also measures bibliographic dispersion, how much knowledge is dispersed in journals. The Bradford’s law is computed from the journals n that have published the most articles on the subject, which would be the core. As one moves away from the

Table 1 Knowledge areas in WoS and Scopus databases

WoS	Scopus
Information Science & Library Science	Computer Science
Computer Science Information Systems	Social Sciences
Computer Science Interdisciplinary Applications	Mathematics
Computer Science Theory Methods	Engineering
Computer Science Artificial Intelligence	Decision Sciences
Engineering Electrical Electronic	Medicine
Computer Science Software Engineering	Multidisciplinary
Multidisciplinary Sciences	Business, Management and Accounting
Telecommunications	Arts and Humanities
Computer Science Hardware Architecture	Materials Sciences
Medical Informatics	Agricultural and Biological Sciences
Health Care Sciences Services	Biochemistry, Genetics and Molecular Biology
Mathematics Interdisciplinary Applications	Energy
Medicine General Internal	Neuroscience
Operations Research Management Science	Physics and Astronomy
Physics Multidisciplinary	
Business	
Cardiac Cardiovascular Systems	
Computer Science Cybernetics	
Education Educational Research	
Education Scientific Disciplines	
Engineering Mechanical	
Management	
Mathematical Computational Biology	
Medicine Research Experimental	
Physics of Fluids and Plasmas	
Physics Mathematical	
Regional Urban Planning	
Social Sciences Mathematical Methods	
Statistics Probability	

core, an increasing proportion of the articles in the subsequent zones is observed $1{:}n{:}n^2{:}n^3$. In the context of this study, Bradford's law facilitates citing a limited number of scientific journals in the AND area, which collectively account for a substantial portion of the total publications.

- The Elitism or Prince law is born from Lotka's Law [22], one of the most discussed models under bibliometrics, which states that the number of authors making n contributions is about $1/n^2$ of those making a single publication. The Elitism law seeks to reveal the most important (most-cited) authors and papers employing the square root of the total number of authors, unveiling what is considered an elite. If n represents the total number of authors, $\sqrt{n}$ would represent the elite of the studied area. In this study, the most cited authors reveal the most important authors and documents responsible for more than half of the contributions in the AND area.
- The 80/20 law (Pareto rule) [23] is inspired by information systems used in commerce and industry, where 80% of information demand is satisfied by 20% of the set of information sources. In this work, this law searches for more relevant journals, conferences, countries, and universities that publish the most in the AND area, and the choice of more representative keywords.

2.2.2 Tools

The data presentation and interrelation using the consolidated meta-analytic approach of TEMAC allows for a review of the most relevant authors and citations, journals, countries, orga-

nizations, or universities, and knowledge areas most related to the research field. To perform the data analysis review, we used the VOSviewer bibliometric tool [24], and the BiblioTools [25, 26].

The VOSviewer tool for visualizing and analyzing bibliometric networks provides insights into patterns, relationships, and trends in the research literature. The VOSviewer allowed the production of visual representations of bibliometric data. The co-authorship and co-citation networks' visualization with research clusters, influential authors, and new research directions help reveal relationships between authors and documents in review works. In summary, the VOSviewer tool helps to explore large bibliometric datasets, contributing to understanding the landscape of research fields.

BiblioTools is a suite of Python scripts for bibliometric analysis integrable to different digital repositories, with numerous functions, such as data mining, data processing, data analysis, keyword visualization, and automated report generation. BiblioTools makes it possible to refine and clean raw data with a preprocessing script preparing the dataset for analysis. We explored our data, producing a variety of co-occurrence networks, such as co-words, co-authors, and co-citations. The BiblioTools allows the visualization of bibliographic coupling networks and clusters, providing information about publications, authors, and research topic connections.

Using the BiblioTools and VOSviewer, we extracted information as follows.

1. An analysis of journals and conferences with the largest number of documents on the topic;
2. Journals with the largest number of documents;
3. Publications in journals and conferences per year;
4. Authors who published the most versus most cited authors;
5. The countries that published the most;
6. Organizations or Universities that published the most;
7. Knowledge areas that most publish;
8. Keyword frequency.

2.3 Detailing, integrating model and validation by evidence

In the third step, deeper analyses allow a better understanding of the topic, selecting principal authors, approaches, lines of research, and validation by evidence with a comparison of results from the different databases.

This evidence is obtained with the analysis of co-citations and bibliographic coupling maps. The co-citation method connects different authors and documents based on their appearance together in the lists of references obtained in the bibliographic repositories. On the other hand, the bibliographic coupling method connects authors and documents based on the number of references they share between them. In other words, while the co-citation presents works constantly cited together and may show similarities between studies, the coupling uses the premise that works that quote the same articles have similar contexts, but indicate the current research fronts using up-to-date space-time.

Co-citation and coupling analyses are commonly used in systematic reviews [27–31]. They fulfill the functions of revealing the main research approaches, establishing the fronts, and revealing future research directions [14]. By establishing a link between references (past) and the most prominent works (future), they fulfill the function of snowballing in an automated way. Thus, through co-citation, one can understand the main approaches of the past while the coupling identifies the primary current studies. Also, the keyword cloud is essential for revealing lines of research demonstrating the different applications in certain areas [19]. The keyword cloud is usually carried out using the frequency

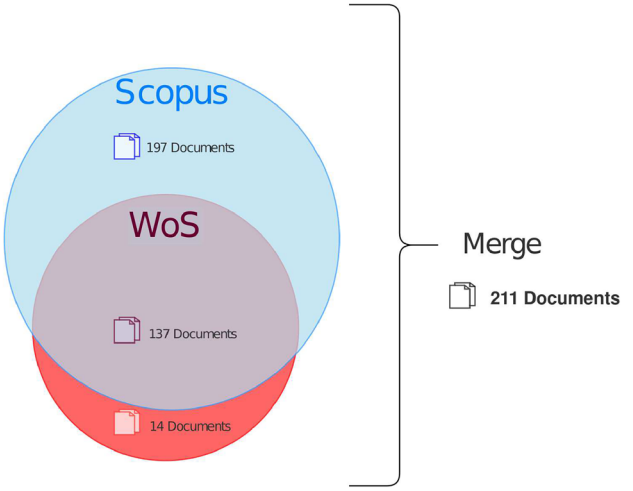

Fig. 2 Documents obtained in the Scopus, WoS, and the merged databases

Table 2 Document types in the databases

Document type	WoS	Scopus	Merged
Journal Article	81	90	98
Conference Article	62	85	87
Conference Review	0	14	14
Review	3	5	5
Book Chapter	0	1	1
Data Paper	1	1	1
Erratum	2	1	3
Early Access Article	2	0	2
Total	151	197	211

Table 3 Journals with the largest number of documents

Journal	h-index	SJR	WoS	Scopus	Merged
Scientometrics	123	0.929	21	19	21
Journal of the Association for Information Science and Technology	150	0.848	8	8	8
Journal of the American Society for Information Science and Technology	18		4	4	4
Journal of Information Science	69	0.761	3	3	3
Journal of Informetrics	77	1.437	4	3	4
IEEE Access	158	927	2	3	4

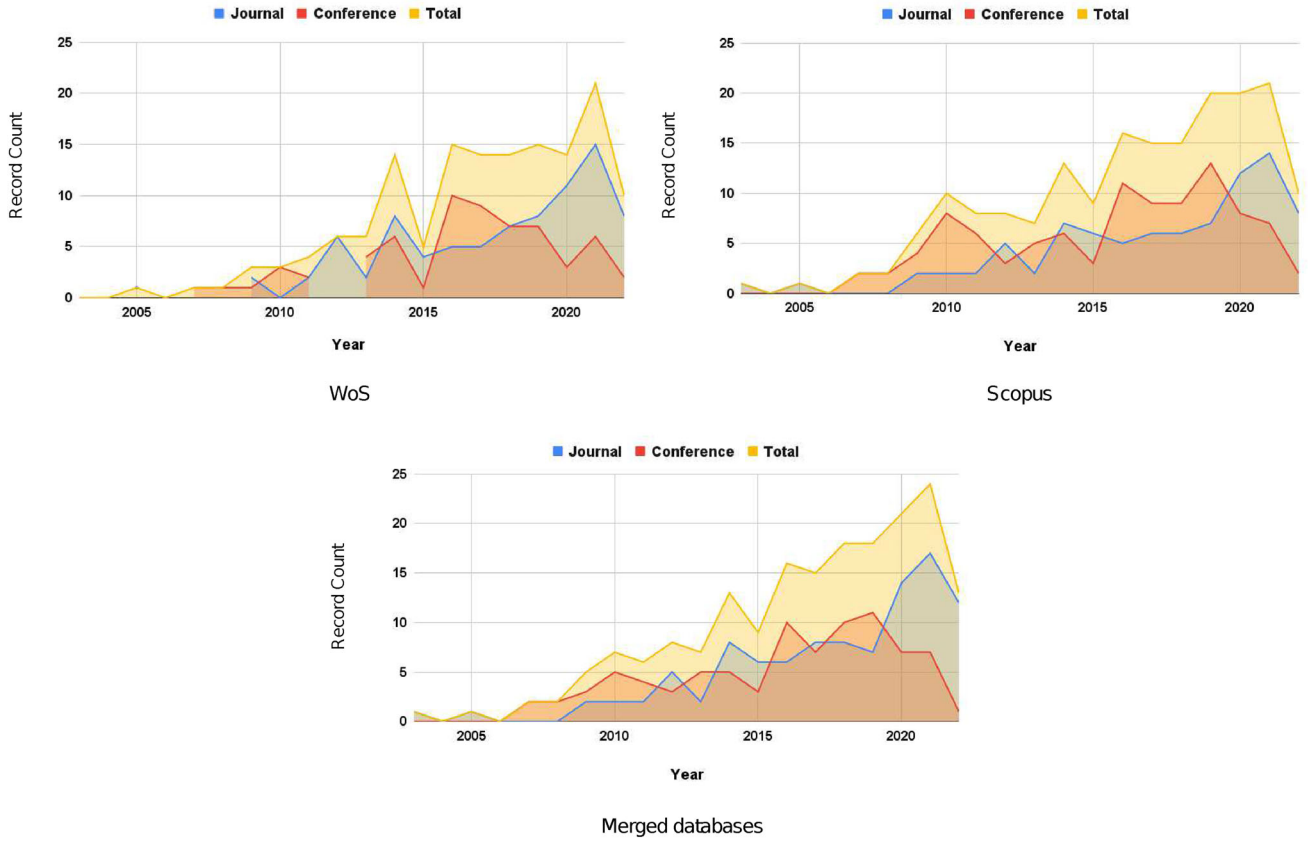

Fig. 3 Evolution of journal and conference publications per year

of keywords and can be enhanced by using the co-occurrence of these words.

3 Results and analysis

The database search results returned 197 documents in Scopus, where 137 were also in the WoS. The data export included the complete work records, including fields of *Author, Title, Abstract, Keywords, Addresses*, and *Cited References*. For a broad literature review investigation, we handled a merge with the documents obtained in the WoS and Scopus databases when 14 unique ones compose the WoS database. Figure 2 presents a Venn Diagram with 211 documents recovered in both bibliographic databases. Additional information on the literature review results and analysis using BiblioTools [25] is available.[1]

3.1 Data presentation and interrelation

In this step, we present the literature review interrelationship of data and quantitative information. Table 2 presents document types in WoS and Scopus databases. On the WoS, approximately 53% (81) are journal articles, 41% (62) conference articles, 2% (3) reviews, 2.6% (4) *erratum* and early

[1] https://adan2and.site/.

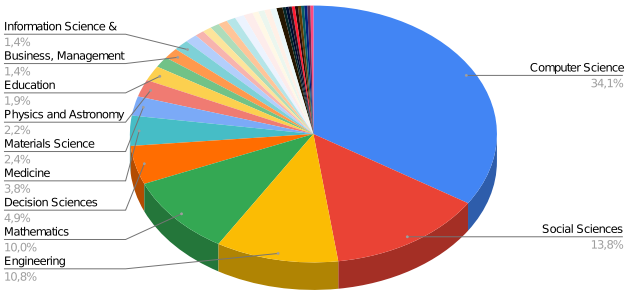

Fig. 4 Distribution of documents by knowledge area in the merged databases

Table 4 Number of Authors versus Number of Publications (merged database)

Number of documents	Number of authors
13	2
12	1
11	1
7	1
6	2
5	6
4	8
3	13
2	53
1	351

access, and 0.6% (1) data paper. According to the Scopus database, approximately 45.6% (90) of documents are journal articles, 43.1% (85) conference papers, and the remainder divided into conference reviews 7.1% (14), review 2.5% (5), book chapter 0.5% (1), data paper 0.5% (1), and *erratum* 0.5% (1). Considering the documents obtained from the databases merged, 46.4% (98) are journal articles, 41.2% (87) conference articles, 6.6% (14) conference reviews, 2.3% (5) reviews, 3.3% (7) are divided into book chapters, data paper, *erratum*, and early access.

According to Table 3, most documents are journal-type. The *Scientometrics*, *h-index* of 123, and *SJR* of 0.929 is the one with the most publications in the AND area with 21 publications in the WoS and the merged databases. The rest of the journals have 23 publications in the merged database. The IEEE Access has the highest *h-index* of 158 and an *SJR* of 927, but only four publications in the merged databases. It is also possible to verify that in this set of documents, the journals related to the Information Science area are in the majority.

The document distribution among the databases is similar, where we can see that most are journal and conference articles. Figure 3 presents the evolution of publications in journals and conferences per year. Note that in all cases, journals and conference documents have alternated over the years. However, considering recent years 2020 to 2022, the amount of journal articles has increased.

As shown in Fig. 4, the knowledge area of Computer Science (34.1%), Social Sciences (13.8%), and Engineering (10.8%) are related to more than half of the total number of AND documents in the WoS and Scopus merged database (58.7%). The leadership in Computer Science might be related to the fact that AND is an open problem in the area, triggering methods and approaches to solve it. Although some works use databases related to the Medicine domain (PubMed [32]), this area corresponds to 3.8% of all documents.

Based on the literature review method, it is possible to identify the authors with the most publications and the most cited ones. Considering the WoS, Scopus, and merged datasets, we empirically tested what number of publications would present an appropriate citation index to filter authors. As presented in Table 4, the number of documents decreases as the number of authors increases. We checked through statistical observation, that 97% of the authors have fewer than five publications. Using a well-established bibliometric principle (the Elitism or Prince law, presented in Sect. 2.2), a small number of authors contribute to a large number of publications. Thus, selecting authors with at least five publications, we focus on the most prolific authors and, presumably, the most influential in the field. With this filter, the WoS database returned eight authors, and Scopus and the merged returned 13. We analyzed these authors, comparing the number of documents and citations.

As shown in Table 5, the author with the highest number of documents from the WoS database was Jinseok Kim (12). However, the results for Scopus and merged databases showed that M. A. Gonçalves was the author with the highest number of documents (13). This author was the most cited in the WoS database (305). However, with authors not previously identified in the WoS database but found in Scopus and the merged databases presented Torvik's works [33–37] with more citations (549). In addition, in the merged database, Torvik's work [36] is the most cited one (189).

As shown in Fig. 5, we measured the number of documents and citations by countries using information from the WoS, Scopus, and merged databases. Following the specification of cited authors and document quantity per author (Table 5), we filtered the countries of the merged databases with five or more publications. Regarding documents by region, the USA, China, Germany and Brazil lead the list of countries that publish the most, considering the three databases. In the WoS and Scopus merged databases, for example, the USA has 45 (21.3%) documents, China 41 (19.4%), Germany 28 (13.2%), and Brazil 18 (8.5%).

According to the graph in Fig. 6, information about the number of citations changes compared to the number of documents. The USA, Brazil, and China present the most

Table 5 Most cited authors and authors' documents

Author	Doc. WoS	Citations WoS	Doc. Scopus	Citations Scopus	Doc. Merged	Citations Merged
Gonçalves, M. A. (Orcid: 0000-0002-2075-3363)	9	305	13	507	13	507
Kim, J. (Orcid: 0000-0001-6481-2065)	12	85	13	166	13	166
Ferreira, A. A (Orcid: 0000-0002-2487-6600)	8	302	12	500	12	500
Laender, A. H. F. (Orcid: 0000-0001-5032-2233)	7	297	11	499	11	499
Asghar, S. (Orcid: 0000-0001-6883-3584)	6	43	6	72	7	72
Hussain, I. (Orcid: 0000-0002-1586-1503)	6	43	6	72	6	72
Smalheiser, N. R. (Orcid: 0000-0003-1079-3406)	3	306	6	524	6	524
Chandra, J. (Orcid: 0000-0001-5994-9024)	5	13	5	18	5	18
Giles, C. L. (Orcid: 0000-0002-1931-585X)	4	35	5	168	5	168
Mondal, S. (Orcid: 0000-0002-2159-3410)	5	13	5	18	5	18
Torvik, V. I. (Orcid: 0000-0002-0035-1850)	3	340	5	549	5	549
Veloso, A. (Orcid: 0000-0002-9177-4954)	3	69	5	166	5	166
Zhang, L. (Orcid: 0000-0003-2104-0194)	2	5	5	10	5	10

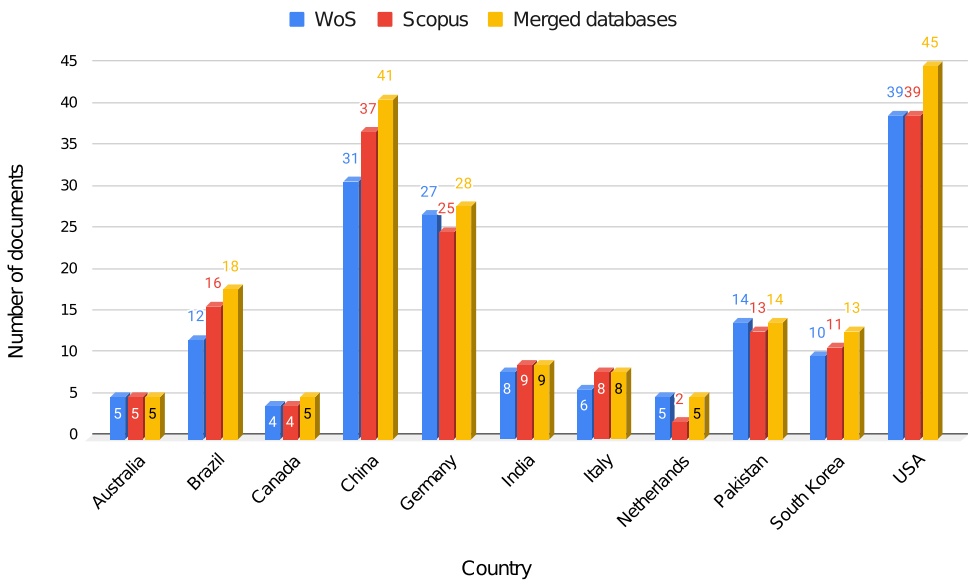

Fig. 5 Documents by country

citations. The USA leads comfortably with 1337 document citations. Also, it is possible to verify that Brazil (510) has more document citations than China (311), with a lower document number. China's citation number is very close to the amount from Germany (51). But Germany is the third country that publishes the most.

We filtered the number of publications per year on the three databases studied. As proposed in the literature review method, the selected documents were from 2003 to 2022. Analyzing Fig. 3 that uses the WoS, Scopus, and merged databases, it is possible to observe a publication increase in the AND area since 2003 but with a decrease in publications in 2015, considering that in 2014 there was growth. It is also important to note that even with a world pandemic scenario in 2020 and 2021, there was a growth in publications compared to previous years. In 2022, we did not obtain the total number of publications, requiring a new survey in 2023 to validate the annual growth.

We conducted the same analysis of publications and citations by organizations that publish studies in the AND area. We included organizations with more than 20 citations in each database (WoS and Scopus) as presented in Table 6. Considering WoS and Scopus merged databases, we found that North American and Brazilian organizations regularly publish with good document citation scores. We checked that most publications are done jointly in Brazil. The *Departamento de Ciência da Computação da Universidade Federal de Minas Gerais* and the *Departamento de Computação da Universidade Federal de Ouro Preto*. These two organizations have 18 publications and 682 citations. Unlike Brazilian organizations, North American organizations usually do not publish together. However, each organization has

Fig. 6 Citations by country

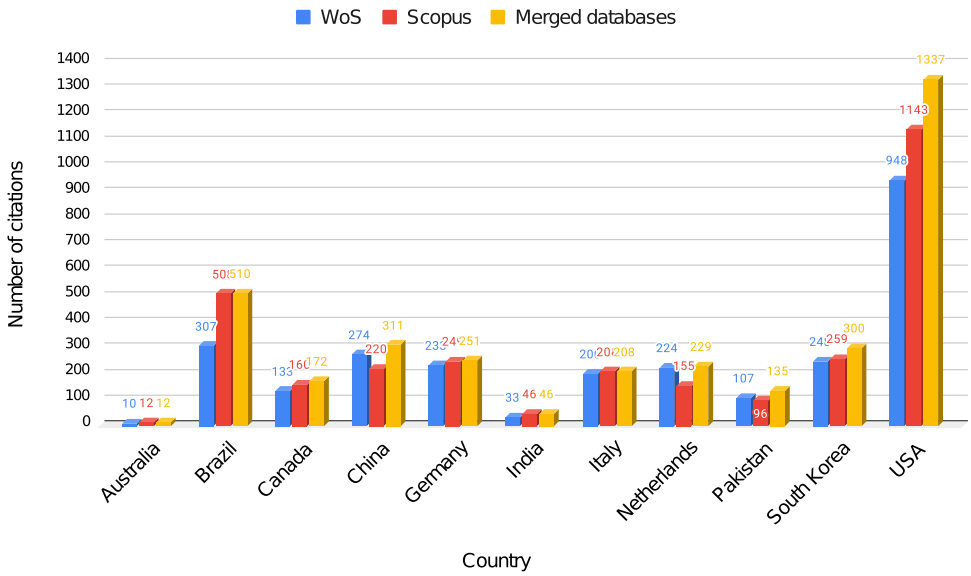

Table 6 Documents and citations by organizations

Organization	Country	Doc. WoS	Citation WoS	Doc. Scopus	Citation Scopus	Doc. Merged	Citation Merged
School of Information Sciences, University of Illinois at Urbana-Champaign	USA	11	482	2	58	11	498
Departamento de Ciência da Computação, Universidade Federal de Minas Gerais	Brazil	9	305	7	359	10	392
Departamento de Computação, Universidade Federal de Ouro Preto	Brazil	7	228	4	231	8	290
Institute for Research on Innovation & Science, University of Michigan	USA	12	176	6	76	12	192
School of Information Management, Wuhan University	China	4	65	4	28	7	115
Heidelberg Institute for Theoretical Studies (GGMBH)	Germany	4	58	4	65	4	65
Microsoft Research	USA	3	25	2	32	4	44
Mathematics Department, Fiz Karlsruhe, Berlin	Germany	2	36	2	41	2	41
Computer Science and Engineering, Pennsylvania State University, Univ. Park	USA	4	42	2	27	4	61

many publications. The Information Sciences School of Illinois University at Urbana-Champaign has 11 publications and 498 citations. The Institute for Research on Innovation & Science of Michigan University has 12 publications with 192 citations.

Using the document titles and abstracts in the WoS and Scopus merged databases, we generate a word cloud as shown in Fig. 7.The cloud included words related to the AND problem and solving approaches, such as data, clustering, information, learning, similarity, publication, model network, libraries, and graph. In the bibliographic co-citation and coupling analysis, such approaches validate the recurrent use of the methods in the AND area.

Fig. 7 Word cloud considering document titles and abstracts in the WoS and Scopus merged databases

Table 7 Correspondence references cited in Figs. 8 and 10

Co-citation analysis (Fig. 8)	Coupling analysis (Fig. 10)
Shin [38]—Cluster 1	Zhang [39]—Cluster 1
Ferreira [40]—Cluster 1	Kim [41]—Cluster 1
Kim [42]—Cluster 2	Xu [33]—Cluster 2
Kim [43]—Cluster 2	Colavizza [44]—Cluster 3
Levin [45]—Cluster 3	
Ferreira [11]—Cluster 3	
Cota [46]—Cluster 3	
Smalheiser [35]—Cluster 3	
Torvik [36]—Cluster 4	
Torvik [34]—Cluster 4	

3.2 Detailing, integrating model and validation by evidence

In this section, we present the third step of the review methodology, including the co-citation analysis, bibliographic coupling analysis, and overview of AND publication.

Table 7 provides a direct relationship between the references shown in the Figs. 8 and 10 and their corresponding representations in the text, which helps to improve the readability and clarity of our results.

3.2.1 Co-citation analysis

In the co-citation analysis of Fig. 8, we identify a similarity between the authors' contributions and their areas of study interest. There are four dark red spots representing co-citation cores. Below, we will detail the leading studies of each cluster and classify them according to the taxonomy used in this literature review (Fig. 1).

Cluster 1

This cluster consists of two works, Shin et al. [38] and Ferreira et al. [40], which use co-authorship information for disambiguation. This similarity indicates the proximity in the heat map and justifies the high co-citation of the cluster. However, the computational approach for disambiguation is different. The work of [38] has an approach based on graphs constructed with co-authorship relations to solve the AND problem using the DBLP and Arnetminer databases. According to the taxonomy, we can classify the type of approach as Author Grouping and the explored evidence as Citation Information and Web Information.

The work presented by Ferreira et al. [40] uses a three-step approach for AND. First, using a heuristic based on co-authorship makes the citations clustered. Using similarities, some of these clusters will be selected to become training data in the second step. In the third step, the selected clusters are added into an associative name disambiguator with self-training capabilities. This work is classified according to the taxonomy in the type of approach as Author Assignment and explored evidence as Citation Information and Web Information (i.e., data extracted from DBLP and BDBComp).

Cluster 2

While neither of the two works in this cluster suggests a direct solution for the AND problem, they do provide strategies that help resolution approaches. The cluster appearance is justified by their high co-citation in the literature, serving as a basis for other studies. Kim et al. [42] introduce a method for generating labeled data to compose machine learning approaches. With test runs, the proposal achieved high performance compared to works in the literature. Kim [43] implements a framework integrating five validation measures for AND approaches using clustering. This integration may help scholars in the AND area to compare the similarities and differences of the various validation measures before selecting the ones that best characterize the clustering performances of their AND methods.

Cluster 3

The authors in [35] present a brief literature review focusing on the definition and challenges of the AND problem. Ferreira et al. [11] conducted a literature review with approaches to AND resolution, suggesting a taxonomy for classifying these approaches. We observed that the two reviews are close compared to the whole heatmap, evidencing a large co-citation of these papers in the studied databases.

Two other works propose approaches to AND. First, Levin et al. [45] present a self-supervised algorithm that uses bootstrap techniques for clustering and a supervised training algorithm. The work uses information from the authors' citations and other attributes such as email, authors' names, and language. We classify this work in the type of approach as Author Grouping and the explored evidence as Citation Information.

The work presented by [46] uses a heuristic-based approach for AND with similarity functions of authorship

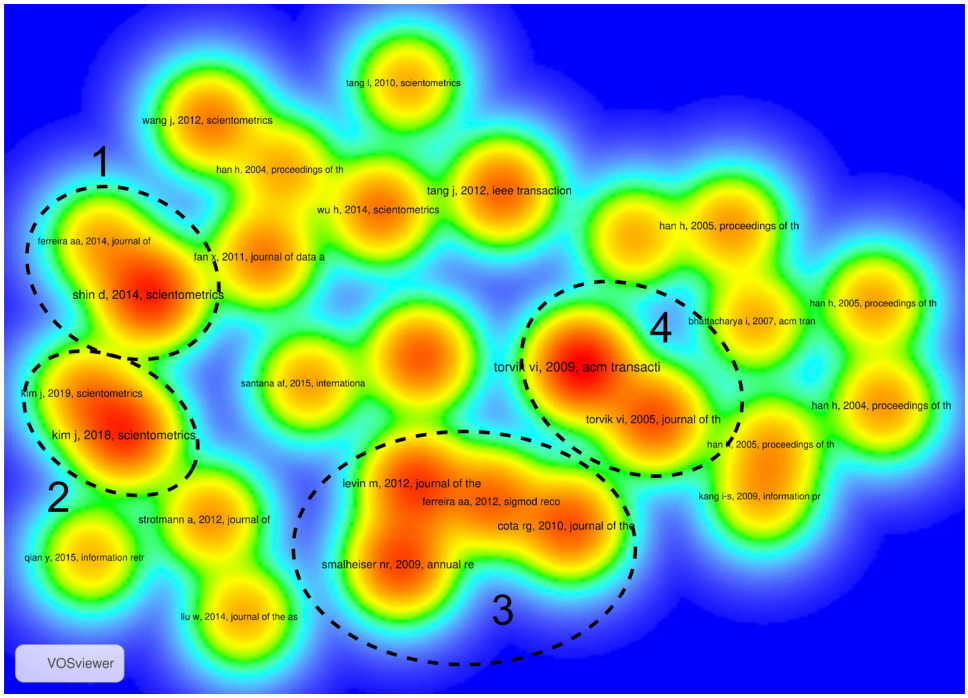

Fig. 8 Co-citation clusters under heat map analysis. The circular dotted lines with explicit numbered labels indicate each cluster

evidence records extracted from DBLP and BDBComp. According to the taxonomy, the type of approach is Author Grouping and explored evidence Citation Information.

Cluster 4

This cluster contains two papers by the same authors. The first one of [36], a probabilistic approach, named Authority, to solve the AND problem in the MEDLINE [32] database using information such as title, journal name, co-authorship, language, and other features. Authority computes the similarity between two articles by analyzing the authors' names and emails. The model also presents ways of automatically generating training sets, methods to estimate the probability between author names, and an agglomerative clustering algorithm based on maximum likelihood to compute clusters of articles that represent the authors studied.

The second work of [34] also uses a probabilistic model for AND, but it only used authors' names, discarding other information such as email addresses and affiliations. Thus, it is evident that the 2009 work is an evolution of the 2005 one. According to the taxonomy, both papers use the approach type as Author Grouping and as explored evidence Citation Information. The other clusters of co-citation presented by the heat map in Fig. 8 did not present patterns detected by this study.

Figure 9 presents another co-citation analysis using cluster density with clustering among all the co-citation papers and allows insight into other similarities among the documents in each group. Thus, we can analyze the other works not so evident in Fig. 8. Note there are three general clusters: green, blue, and red. A common characteristic is a space-time between the documents. The red has papers from 2005 to 2010 and the blue from 2009 to 2015. This space-time feature does not appear in the green cluster, as it is more diverse with documents from 2004 to 2019. Note there are works on the cluster edges, uniting groups based on the date characteristic, such as [11] that link the blue and red clusters.

The green cluster is quite diverse as there are various types of approaches, such as cognitive maps and network analysis [47], probabilistic models [48], heuristic-based models [49], agglomerative hierarchical clustering [50], and supervised learning [51, 52].

The red cluster cover works using clustering approaches [53–55]. The study conducted by [56] investigates the influence of co-authorship attributes for solving the AND problem. The blue cluster presents a similarity with research using cluster similarity and agglomerative clustering for AND [57, 58]. In contrast, Strotmann and Zhao [59] presents research that indicates the influence of AND on citation and bibliographic base analysis studies.

3.2.2 Bibliographic coupling analysis

The bibliographic coupling analysis allows insights into the current state of the AND research area. We present in this section the AND research fronts, including works from 2019 to 2022. The works are classified considering the AND approach using the taxonomy presented in Fig. 1.

Figure 10 presents a bibliographic coupling of works using a heat map for the merged WoS and Scopus merged databases. Note there are three clusters explicitly num-

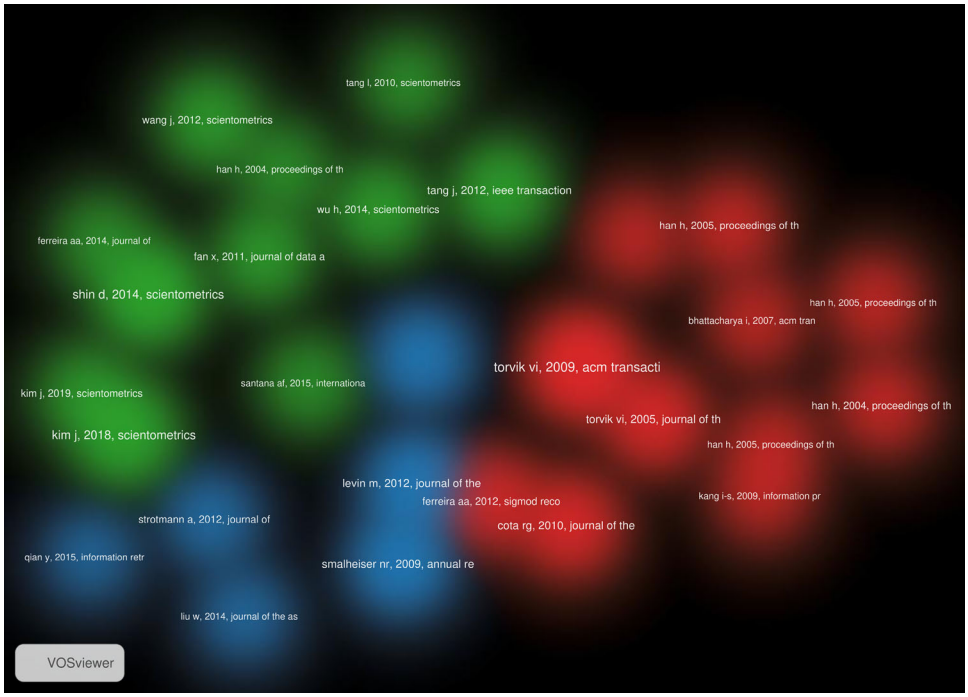

Fig. 9 Co-citation clusters under cluster density analysis

bered, highlighting the current AND research fronts. In the sequence, we present a summary of the works included in the three clusters.

Cluster 1

The authors in [39] used a graph node embedding approach to solve the AND problem. This type of solution is inspired by the word embedding model but adapted for a graph structure solution. A graph is constructed with co-authorship relationships, using the random walk method for learning graphs and assigning clusters to unique people in the real world. The approach used CiteSeerX data with results improved compared to similar approaches.

The authors in [41] propose a hybrid pairwise classification method for estimating the probability that an author record is correct in a bibliographic repository. This solution uses global features extracted from text using supervised training on a dataset of an author's citations. This text classification and the supervised training use word embedding methods such as Bag of Words and TF-IDF with data from PubMed and ArnetMiner. According to the taxonomy, [39, 41] can be classified as Author Grouping approaches because they use similarity calculation with training and machine learning. Authors use word embedding as basis for the AND method, justifying the proximity of the studies observed in the heat map presented in Fig. 10.

Cluster 2

In [33], the authors create a knowledge graph with information from the PubMed repository extracting bio-entities from abstracts. In this work, the authors do not propose a new approach to solving the AND problem. However, approaches already known in the literature were used together, such as Authority (uses a graph approach) and Semantic Scholar (uses a binary training classifier to join pairs of author names and create author clusters). The constructed knowledge graph allowed the creation of links between biological entities, articles, authors, and affiliations. In the AND step, the results achieved F1 scores of 98.09%. We can classify the approach of this work as Author Grouping and the explored evidence as Citation Information and Web Information.

Cluster 3

The work of [44] created an automatic system for data availability declarations in bibliographic repositories using PubMed. The authors compare first and last names with string similarity techniques. It appears in the heat map of the literature because it cites several influential AND works.

With the literature coupling analysis, we note a current use of grounded techniques for AND with Author Grouping and Clustering methods. The first article found in the dataset of this review dates back to 2003. Since the bibliographic coupling seeks to obtain current research fronts, works from 2019 to 2022 were selected for the bibliographic coupling analysis. Analysis was conducted to classify and present the most recent articles (2020–2022), classifying them according to the taxonomy in Table 8 and Fig. 11.

Based on the co-citation analysis performed across the studied time-space (2003–2022), the works present mainly the use of author grouping approaches. This analysis result is consistent with the findings of [4]. Moreover, the conducted coupling bibliographic analysis points to viable alternatives

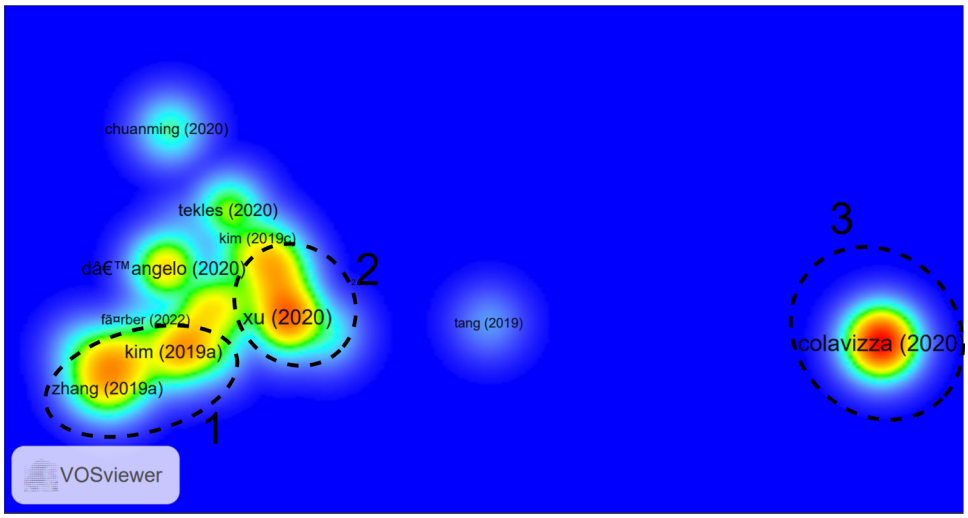

Fig. 10 Bibliographic coupling in the WoS and Scopus merged databases. The Circular dotted lines with explicit numbered labels indicate each cluster

to address AND-related problems, indicating a wide range of research in the area.

3.2.3 Overview of AND publications

In this section, we present an overview of recent AND works published between 2020 and 2022, arranged by the taxonomy of Fig. 1 [4, 11]. The taxonomy is also used in the co-citation (Sect. 3.2.1) and coupling analysis (Sect. 3.2.2) sections. This time frame was chosen as earlier reviews covered works until 2019 [11, 12]. The coupling analysis includes some works presented in this section.

The overview of current research works is essential to complete the meta-analytical analysis by citing the particular techniques, strategies, and emerging themes within the AND research area. To organize the works overview analysis in a concise way, we guided it by the book devoted to AND study in bibliographic repositories [4]. We present a synopsis of each work included in Table 8. The work synopsis presents the AND approaches, including author grouping, mainly through agglomerative clustering, standing out as prevalent methods in recent studies.

The author grouping approach, as defined in Sect. 1, is especially appropriate for datasets with lots of co-authorship data. Large bibliographic datasets can benefit from this approach, unlike the author assignment approach, as it does not depend on time-consuming manual labeling author annotations. Agglomerative clustering in author grouping is a common approach as it is simple to use and can produce hierarchical clusters to be analyzed at different granularities. This clustering method provides flexibility when creating author groups. Figure 11 shows how the works relate to each other in the used taxonomy.

We identified a set of five works using tree-based learning models, such as Gradient Boosting, Random Forest, and Decision Tree [60–64]. The works may use other Machine Learning techniques in conjunction with the mentioned, such as Naive Bayes [62], Logistic Regression [60, 61], and Network Graphs [63].

Some works address distinct supervised techniques, such as [10] that use transfer learning. The authors in [65] show that ORCID can validate the performance of supervised AND methods that use large-scale bibliographic data. The authors in [66] used the DBLP database with a neural network to learn the representations of coauthors and titles so AND could consider the similarity between these attributes.

Li et al. [67] present an algorithm with multiple similarity strategies for AND implementing using collaboration network calculations, affiliation, and publications attributes of authors. Another multi-strategy approach is presented by [68] using string comparison with Jaccard similarity, Levenshtein distance, and co-authorship network comparison. Waqas and Qadir [69] propose a multilayer heuristic with a clustering approach. The clustering uses attributes inherent to the author and publication, such as title, abstract, keywords, email, and affiliation. Word embedding Word2Vec is used to extract the attributes.

D'angelo and van Eck [70] use a rule-based scoring approach with author, publication, citation, and institution attributes. Clusters with meta-data allow indexing of a particular author to disambiguate. Zhang et al. [71] use heuristic rules combined with neural networks to analyze publication attributes, such as title and affiliation. An advantage of this method is the possibility of extending the method's application to other datasets.

Mozafari [72] proposes a genetic algorithm for determining the similarity coefficient between two authors for AND. The algorithm determines the importance of the attributes in the publications, electing an optimal coefficient for comparison between authors.

Table 8 Classification of papers from 2020 to 2022 according to the taxonomy of Fig. 1 [4, 11]

References	Type of approach			Explored evidence	DataSet	Bibliographic Citation Database
	Author grouping		Author assignment classification method			
	Similarity function	Clustering method				
Kim and Owen-Smith [10]	Transfer Learning	Agglomerative		Citation Information and Researchers' personal files from Web	DBLP, AMiner, KISTI, MEDLINE American Physical Society	Scopus WoS
Xu et al. [33]	Authority with a learn probabilistic metric; Semantic Scholar with error-drive and hank-based learning	Agglomerative		Medline Metadata	PubMed	Scopus WoS
D'Angelo and van Eck [70]	Ruled-Based Scoring	Agglomerative		Author's informations (name, academic rank, research fields and institutional affiliation)	Data Source from Italian Ministry of Education Universities and Research	Scopus WoS
Chuanming et al. [75]	Unsupervised Learning	Agglomerative		Citation Information	CiteSeerX, AMiner, DBLP	Scopus
Jhawa et al. [60]			Ensemble-based Classification with Random Forest and Gradient Boosted Tree	Medline Metadata	PubMed	Scopus
Li et al. [67]	Heuristic	Partitioning		XML files with author and publications attributes	Scopus WoS	Scopus
Ma et al. [76]	Graph Auto-Encoder and Graph Embedding with Word2Vec	Agglomerative		Citation Information and Researchers' personal files from Web	AMiner	Scopus
Jinqi et al. [73]	Maximum flow in network graph	Agglomerative		Citation Information and Researchers' personal files from Web	AMiner, Microsoft Academic	Scopus
Ma et al. [77]	Meta-path based algorithm with node embeddings in a homogeneous network	Agglomerative		Citation Information and Researchers' personal files from Web	AMiner	Scopus
Wang et al. [78]			Supervised classification technique with Random walk based model	DBLP data dump with author's informations	DBLP	Scopus WoS
Wang et al. [79]	Adversarial representation learning model with heterogeneous information network	Agglomerative		Author's name and citation information	AMiner	Scopus WoS
Zhang and Ban [64]	Rule-based disambiguation in a graph model	Agglomerative		Author's name and citation information	AMiner	Scopus
Zhang et al. [71]	Convolutional Neural Network to compare clusters of publications	Agglomerative		Citation Information and Researchers' personal files from Web	AMiner	Scopus

Table 8 continued

References	Type of approach					Bibliographic
	Author grouping		Author assignment classification method	Explored evidence	DataSet	Citation
	Similarity function	Clustering method				Database
Pooja et al. [80]	Graph with Edge Pruning-Based Approach	Agglomerative		Citation Information and Researchers' personal files from Web	AMiner, Scopus WoS	Scopus
Rodrigues et al. [68]	Multi-strategic approach with comparison of strings and author's networks	Agglomerative		Citation Information and Researchers' personal files from Web	DBLP	Scopus
Zhou et al. [74]	Graph similarity with Inverse Document Frequency	Partitioning		Author's name and citation information	AMiner	Scopus WoS
Firdaus et al. [81]			Naïve Bayes, Random Forest, Support Vector Machine and Deep Neural Network	DBLP data dump with author's information	DBLP	Scopus
Xiong et al. [82]	Unsupervised Learning with Variotinal AutoEncoder	Agglomerative		Author's name and citation information	AMiner, DBLP, CiteSeerX	Scopus
Kim & Owen-smith [65]	Authority similarities	Agglomerative		Medline Metadata	PubMed	Scopus WoS
Mozafari [72]	Genetic Algorithm to learn from the available samples	Agglomerative		Author's information (name, academic rank, research fields, and institutional affiliation)	Iranian Ministry of Science, Ministry of Health	Scopus
Mihaljević and Santamaría [63]	Supervised Learning with Decision Tree, Random Forest, and Histogram-based Gradient Boosting	Agglomerative		Author's name and documents	NASA/ADS	Scopus
Correia et al. [83]			Web page with form to crowdsourcing campaign			Scopus WoS
Zhang et al. [84]	Graph Attention Networks	Spectral Clustering		Author's name and citation information	AMiner	Scopus WoS
Pooja et al. [85]	Multi-dimensional representation learning based with meta-content and author similarity graphs	Agglomerative		Author's name and citation information	AMiner, DBLP, CiteSeer, Zbmath	Scopus
Zhang et al. [86]			Supervised Learning with Random Forest	Citation Information and Researchers' personal files from Web	PubMed, Microsoft Academic, Semantic Scholar	Scopus WoS

Table 8 continued

References	Type of approach			Explored evidence	DataSet	Bibliographic Citation Database
	Author grouping		Author assignment classification method			
	Similarity function	Clustering method				
Kim et al. [62]	Gradient Boosting, Logistic Regression, Naïve Bayes, and Random Forest			Author's name and citation information	KISTI, AMiner, GESIS, UM-IRIS	Scopus WoS
Waqas and Qadir [69]	Multilayer heuristics based clustering with Research2vec, and Cosine similarity	Agglomerative		Author's name and citation information	AMiner, BDBComp	Scopus WoS
Firdaus et al. [87]			Cost-Sensitive Deep Neural Network	Author's name and citation information	DBLP	Scopus
Rehs [61]	Random Forest and Logistic Regression	Partitioning		Author's name and documents	WoS	Scopus WoS
Färber and Lamprecht [88]	Ruled-based with Jaro-Winkler similarity	Agglomerative		XML files with author and publication attributes	OpenAire, WikiData	Scopus WoS
Pooja et al. [9]	Attention-Based Graph Convolution with a multihop neighborhood	Agglomerative		Author's name and citation information	AMiner	Scopus
Backes and Deitze [89]	Progressive block merging	Agglomerative		Author's name and documents	WoS	Scopus WoS
Manzoor et al. [90]	Convolutional Neural Network to classification	Agglomerative		Medline Metadata	PubMed	Scopus WoS
Boukhers and Asundi [66]	Neural network that learns author and co-authors representations	Agglomerative		Author's name and citation information	DBLP	Scopus WoS
Färber and Ao [91]	Unsupervised Approach with ruled-based classifier	Agglomerative		Author's name and documents	MAKG	Scopus WoS
Qiping et al. [92]	Network representation learning	Agglomerative		Author's name and citation information	AMiner, DBLP, CiteSeerX	Scopus WoS
Santini et al. [93]	Multimodal Knowledge Graph Embeddings	Agglomerative		Author's name and citation information	AMiner, ORCID	Scopus WoS
Waqas and Qadir [94]	Manually cross check and cosine similarity to detect ambiguities	Agglomerative		Citation Information and researchers' personal files from Web	Google Scholar, DBLP	Scopus WoS

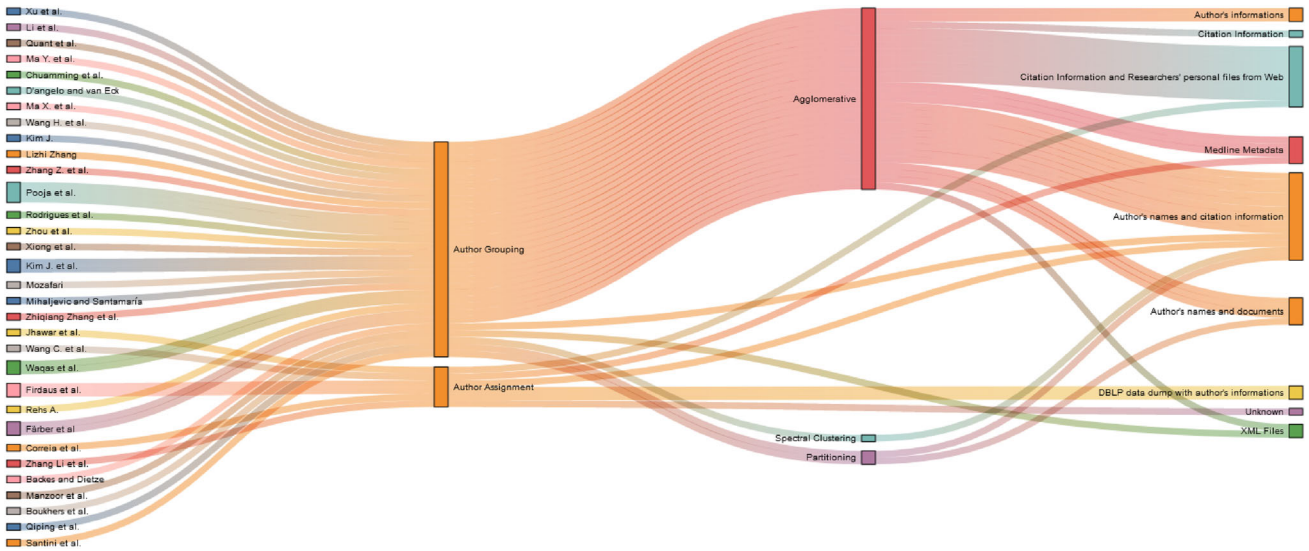

Fig. 11 Sankey diagram of approaches used for AND from 2020 to 2022

Jinqi et al. [73] propose an algorithm to put entities and resources into a network graph to set the resource node capacity-based sharing degree. The network graph uses relationships between the author and publication nodes to calculate the flow capacity between nodes which allows clustering of the graph.

Zhang and Ban [64] use publication relationships to construct the graph, with the strongly related publications grouped, forming atomic clusters and reducing the graph size. At another stage, a rule-based similarity algorithm analyzes and combines the feature information from the publication graph to perform AND.

Zhou et al. [74] present an approach with five graphs formed by publication attributes, co-authorship, location, title, keywords, and affiliation. Each attribute creates the node where the edges are the similarity weights between publication pairs. A fusion graph of the attributes is built. A random walk algorithm is applied to the graph to determine paths that represent the local node structural information. Then, a multilayer perceptron algorithm is applied to the graph structure.

Santini et al. [93] propose a Knowledge Graph Embeddings (KGE) using information from the AMiner database. The KGE has three parts: multimodal information extraction from the KGE, a blocking procedure, and hierarchical agglomerative clustering. Qiping et al. [92] use citation information to construct a heterogeneous information network. Representation learning for clustering the authors and disambiguation is applied, and cluster analysis with rule matching is performed.

Ma et al. [76] propose incorporating a Word2Vec model into a Ghaph-Based approach. The algorithm extracts attributes and the relationships between publications, authors, and co-authors. Word2Vec serves to obtain these features allowing the insertion of other features that may appear in the dataset. Subsequently, a graph with relationships between publications and authors is built. Then, an algorithm for clustering and similarity analysis between nodes and edges is applied.

Pooja et al. work presents solutions using a graph-based approach as a basis associated with other computational techniques (e.g., Clustering). The work [80] uses a graph-based clustering approach for AND. Jaccard and Cosine similarity characterizes the relationships between authors and publications in the graphs, and Web information refines the results. In [85], the authors use graphs with publication attributes and a Word2Vec embedding model to create vectors that will serve as input to an agglomerative hierarchical clustering (HAC), which is widely used in AND studies. In [9], the authors use a graph-based approach configured with multi-hop neighborhoods and apply HAC for AND in the final step of the algorithm. The work in [95] uses graphs to build the network of authors and publications and uses clustering for disambiguation. However, the differential of this new approach is the ability to work with online information from digital bibliographic repositories.

The author in [75, 79, 82, 84] use word embedding, graphs, and clustering respectively. Ma et al. [77] use the same approach applied to robotic literature consultants.

The work of [78] proposes a technique with partial classification in three steps to solve the AND problem. The first uses a probability propagation constraint to infer the distribution of a given author's name. In the second step, a portion of the author name in the documents is linked to their respective authoring if the model exhibits high confidence. In the last step, the initial classification algorithm parameters are updated.

Firdaus et al. [81] propose two methods for AND. In the first work, the technique uses four steps for disambiguation: data labeled, publication attributes extracted, deep neural network, random forest, naive Bayes, and SVM classification are done, and the validation of the result comparing the classification techniques. In the second work, Firdaus et al. [87] uses classification with deep neural network technique is increased with cost-sensitive learning considering cost variation from unclassified data.

Manzoor et al. [90] use convolutional neural networks for unbalanced and balanced dataset classification. According to the authors, the solution is flexible by learning the attributes without concatenating similarity measures. The same method is also Single Citation Based which preprocesses the dataset efficiently, decreasing computational costs.

Farber and Ao [91] do not propose a new approach but use an unsupervised rule-based classification method. The method does not require data training, adapted for the authors' proposal.

Correia et al. [83] propose a crowd-systems-based prototype allowing interaction and contribution from the Web for the general public correcting name ambiguities, missing data, and incorrect references in a digital bibliographic repository. Backes and Dietze [89] present a technique for progressive AND with lattice structures for name inclusion. Waqas and Qadir [94] do not propose an AND resolution but present a dataset to assist developers. The "CustAND" labeled dataset with 7886 publication records is presented using data from DBLP and Google Scholar.

4 Conclusion

This article presented AND literature review using the theory of the consolidated meta-analytic approach with the WoS, Scopus, and merged bibliographic repositories. A taxonomy was used to classify AND methods in the reviewed works. With the bibliometric laws of analysis, it was possible to present the most cited papers, authors, countries, organizations, knowledge areas, journals that publish the most documents, and the frequency of keywords, highlighting the evolution of AND from 2003 to 2022.

Summarising the key findings, we note that AND authors publish more in journals than conferences and book chapters (Table 2). The journal with the largest number of documents is the Scientometrics (Table 3). The evolution of journal and conference publications shows an increase over the years from 2003 to 2022 accentuated from 2016 (Fig. 3). The Computer Science knowledge area presents the highest contribution in AND with 34%, followed by Social Sciences ($\approx$ 14%) and Engineering ($\approx$ 11%) that together correspond more than half of the total works (Fig. 4). The countries that publish the most are the USA (21.3%), China (19.4%), Germany (13.2%), and Brazil (8.5%) (Fig. 5). Considering the number of citations in the merged databases (WoS and Scopus), the USA leads with 1337 document citations, Brazil with 510, and China with 311 (Fig. 6), reinforced by the organizations that regularly publish with more than 20 citations including the North American and Brazilian ones (Table 6).

During the co-citation and bibliographic coupling analyses, we identified four and three clusters, respectively. In the co-citation, four clusters show that graph, supervised learning, and heuristic-based approaches with probability applications for solving the AND problem are used. Furthermore, co-citation indicates the prevalence and effectiveness of author grouping techniques in current AND literature, particularly in addressing issues associated with large bibliographic databases. The bibliographic coupling indicates current research for AND with word embedding and supervised learning. We note that most of the approaches use AMiner and DBLP as bibliographic bases for information extraction.

Presenting this literature review of the AND panorama, we intend to help researchers direct current studies resulting in the creation of new techniques to solve the problem. However, the meta-analytic approach used in this literature review presents some limitations. The focus is an exploratory overview of the research area grounded by bibliometric principles, not including a protocol like a systematic literature review with specific research questions. However, the results of the meta-analytic approach are complementary to systematic literature review methods, adding knowledge of existing studies in the AND research area. Another limitation is related to the WoS and Scopus merged databases, which was done using a script in Python language since the VOSviewer tool allows only one database at a time.

In future work, we can use new bibliographic databases in complement to other literature review approaches, such as a systematic review. Additional WoS and Scopus database knowledge areas can be used, to enlarge the scope of research in the AND area, such as multidisciplinary sciences.

Acknowledgements Prof. Célia G. Ralha thanks the research productivity grant number 309688/2021-3 in the Computer Science area from the Brazilian National Council for Scientific and Technological Development (CNPq).

References

1. DBLP. Computer science bibliography. https://dblp.org/. Accessed 02 Mar 2023
2. ArnetMiner. Aminer. https://www.aminer.org/. Accessed 25 Apr 2023
3. CiteSeerX. An evolving scientific literature digital library and search engine. https://citeseerx.ist.psu.edu/. Accessed 25 Apr 2023
4. Ferreira, A.A., Gonçalves, M.A., Laender, A.H.F.: Automatic Disambiguation of Author Names in Bibliographic Repositories. Synthesis Lectures on Information Concepts, Retrieval, and Services (SLICRS), Springer, Cham (2020)
5. Hussain, I., Asghar, S.: A survey of author name disambiguation techniques: 2010–2016. Knowl. Eng. Rev. **32**, e22 (2017)
6. McKay, D., Sanchez, S., Parker, R.: What's my name again? Sociotechnical considerations for author name management in research databases, pp. 240–247 (2010)
7. Gomide, J., Kling, H., Figueiredo, D.: Name usage pattern in the synonym ambiguity problem in bibliographic data. Scientometrics **112**, 747–766 (2017)
8. Lagoze, C., Van de Sompel, H.: The open archives initiative: building a low-barrier interoperability framework, pp. 54–62 (2001)
9. Pooja, K.M., Mondal, S., Chandra, J.: Exploiting higher order multi-dimensional relationships with self-attention for author name disambiguation. ACM Trans. Knowl. Discov. Data **16**, 1–23 (2022)
10. Kim, J., Owen-Smith, J.: Model reuse in machine learning for author name disambiguation: an exploration of transfer learning. IEEE Access **8**, 188378–188389 (2020)
11. Ferreira, A.A., Gonçalves, M.A., Laender, A.H.: A brief survey of automatic methods for author name disambiguation. SIGMOD Rec. **41**, 15–26 (2012)
12. Sanyal, D.K., Bhowmick, P.K., Das, P.P.: A review of author name disambiguation techniques for the Pubmed bibliographic database. J. Inf. Sci. **47**, 227–254 (2021)
13. Cao Simeng, L.C.: Review of studies on incremental name disambiguation. Data Anal. Knowl. Discov. **6**, 10 (2022)
14. Mariano, A.M., Rocha, M.S.: Revisão da literatura: apresentação de uma abordagem integradora. In: Proceedings of XXVI AEDEM: Annual Meeting of the European Academy of Management and Business Economics, pp. 427–442. Springer (2017)
15. Kitchenham, B.: Procedures for performing systematic reviews. Keele, UK, Keele Univ. **33**, 1–26 (2004)
16. Kitchenham, B., et al.: Systematic literature reviews in software engineering—a systematic literature review. Inf. Softw. Technol. **51**, 7–15 (2009)
17. Vera-Olivera, H., et al.: Data modeling and NoSQL databases—a systematic mapping review. ACM Comput. Surv. **54**, 1–26 (2021)
18. Mariano, A.M., Reis, A.C.B., dos Santos Althoff, L., Barros, L. B.: Industrial engineering and operations management I, Ch. A Bibliographic Review of Software Metrics: Applying the Consolidated Meta-Analytic Approach, pp. 243–256. Springer (2019)
19. Correa, P.R., Cruz, R.G.: Meta-análisis sobre la implantación de sistemas de planificación de recursos empresariales (ERP). J. Inf. Syst. Technol. Manag. **2**, 245–273 (2005)
20. Brookes, B.C.: Bradford's law and the bibliography of science. Nature **224**, 953–956 (1969)
21. Heradio, R., Fernandez-Amoros, D., Cerrada, C., Cobo, M.J.: Group decision-making based on artificial intelligence: a bibliometric analysis. Mathematics **8**, 1566 (2020)
22. Lotka, A.J.: The frequency distribution of scientific productivity. J. Wash. Acad. Sci. **16**, 317–323 (1926)
23. Trueswell, R.L.: Some behavioral patterns of library users: The 80/20 rule (1969)
24. VOSviewer. Visualizing scientific landscapes. Centre for Science and Technology Studies, Leiden University, Netherlands. https://www.vosviewer.com/. Accessed 17 Nov 2022
25. Grauwin, S.: BiblioTools/BiblioMaps—a freely available set of scripts developed to create maps of science based on bibliographic data. http://www.sebastian-grauwin.com/bibliomaps/index.html. Accessed 28 Nov 2023
26. Grauwin, S., Jensen, P.: Mapping scientific institutions. Scientometrics **89**, 943–954 (2011)
27. Ankrah, J., Monteiro, A., Madureira, H.: Bibliometric analysis of data sources and tools for shoreline change analysis and detection. Sustainability **14**, 4895 (2022)
28. Crispim, R.T., Netto, C.O., Camboim, G.F., Camboim, F.F.: Capabilities for service innovation: bibliometric analysis and directions for future research. Rev. Adm. Mackenzie **23**, eRAMD220030 (2022)
29. Garakhanova, N.: Bibliometric analysis on digital diplomacy studies. Korkut Ata Türkiyat Araştırmaları Dergisi, pp. 1325–1338 (2023)
30. Müller, M.: Pyblionet-software for the creation, visualization and analysis of bibliometric networks. SoftwareX **24**, 101565 (2023)
31. Khider, H., Hammoudi, S., Meziane, A., Cuzzocrea, A.: BPM in the era of industry 4.0: a bibliometric analysis, pp. 651–659 (2023)
32. MEDLINE. Pubmed. https://pubmed.ncbi.nlm.nih.gov/. 2003–2022. Accessed 25 Apr 2023
33. Xu, J., et al.: Building a PubMed knowledge graph. Sci. Data **7**, 1–15 (2020)
34. Torvik, V.I., Weeber, M., Swanson, D.R., Smalheiser, N.R.: A probabilistic similarity metric for Medline records: a model for author name disambiguation. J. Am. Soc. Inf. Sci. Technol. **56**, 140–158 (2005)
35. Smalheiser, N.R., Torvik, V.I.: Author name disambiguation. Ann. Rev. Inf. Sci. Technol. **43**, 1–43 (2009)
36. Torvik, V.I., Smalheiser, N.R.: Author name disambiguation in MEDLINE. ACM Trans. Knowl. Discov. Data **3**, 1–29 (2009)
37. Torvik, V.I., Weeber, M., Swanson, D.R., Smalheiser, N.R.: A probabilistic similarity metric for Medline records: a model for author name disambiguation. In: AMIA Annual Symposium Proceedings, 1033 (2003)
38. Shin, D., Kim, T., Choi, J., Kim, J.: Author name disambiguation using a graph model with node splitting and merging based on bibliographic information. Scientometrics **100**, 15–50 (2014)
39. Zhang, W., Yan, Z., Zheng, Y.: Author name disambiguation using graph node embedding method. In: Proceedings of IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 410–415. IEEE (2019)
40. Ferreira, A.A., Veloso, A., Gonçalves, M.A., Laender, A.H.F.: Self-training author name disambiguation for information scarce scenarios. J. Assoc. Inf. Sci. Technol. **65**, 1257–1278 (2014)
41. Kim, K., Rohatgi, S., Giles, C.L.: Hybrid deep pairwise classification for author name disambiguation. In: Proceedings of 28th ACM International Conference on Information and Knowledge Management (CIKM), pp. 2369–2372. ACM (2019)
42. Kim, J., Kim, J., Owen-Smith, J.: Generating automatically labeled data for author name disambiguation: an iterative clustering method. Scientometrics **118**, 253–280 (2019)
43. Kim, J.: A fast and integrative algorithm for clustering performance evaluation in author name disambiguation. Scientometrics **120**, 661–681 (2019)
44. Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., McGillivray, B.: The citation advantage of linking publications to research data. PLoS ONE **15**, 1–18 (2020)

45. Levin, M., Krawczyk, S., Bethard, S., Jurafsky, D.: Citation-based bootstrapping for large-scale author disambiguation. J. Am. Soc. Inf. Sci. Technol. **63**, 1030–1047 (2012)
46. Cota, R.G., Ferreira, A.A., Nascimento, C., Gonçalves, M.A., Laender, A.H.F.: An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. J. Am. Soc. Inf. Sci. Technol. **61**, 1853–1870 (2010)
47. Tang, L., Walsh, J.: Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. Scientometrics **84**, 763–784 (2010)
48. Tang, J., Fong, A.C., Wang, B., Zhang, J.: A unified probabilistic framework for name disambiguation in digital library. IEEE Trans. Knowl. Data Eng. **24**, 975–987 (2012)
49. Santana, A.F., Gonçalves, M.A., Laender, A.H., Ferreira, A.A.: On the combination of domain-specific heuristics for author name disambiguation: the nearest cluster method. Int. J. Digit. Libr. **16**, 229–246 (2015)
50. Wu, H., Li, B., Pei, Y., He, J.: Unsupervised author disambiguation using Dempster–Shafer theory. Scientometrics **101**, 1955–1972 (2014)
51. Wang, J., et al.: A boosted-trees method for name disambiguation. Scientometrics **93**, 391–411 (2012)
52. Han, H., Giles, L., Zha, H., Li, C., Tsioutsiouliklis, K.: Two supervised learning approaches for name disambiguation in author citations. In: Proceedings of 4th Joint ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 296–305. ACM (2004)
53. Han, H., Zha, H., Giles, C.L.: Name disambiguation in author citations using a k-way spectral clustering method. In: Proceedings of 5th ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 334–343. ACM (2005)
54. Han, H., Xu, W., Zha, H., Giles, C. L.: A hierarchical Naive Bayes mixture model for name disambiguation in author citations. In: Proceedings of 20th ACM Symposium on Applied Computing (SAC), pp. 1065–1069. ACM (2005)
55. Bhattacharya, I., Getoor, L.: Relational clustering for multi-type entity resolution. In: Proceedings of 4th International Workshop on Multi-relational Mining (MRDM), pp. 3–12. ACM (2005)
56. Kang, I.-S., et al.: On co-authorship for author disambiguation. Inf. Process. Manag. **45**, 84–97 (2009)
57. Liu, W., et al.: Author name disambiguation for PubMed. J. Assoc. Inf. Sci. Technol. **65**, 765–781 (2014)
58. Qian, Y., Zheng, Q., Sakai, T., Ye, J., Liu, J.: Dynamic author name disambiguation for growing digital libraries. Inf. Retr. J. **18**, 379–412 (2015)
59. Strotmann, A., Zhao, D.: Author name disambiguation: What difference does it make in author-based citation analysis? J. Am. Soc. Inf. Sci. Technol. **63**, 1820–1833 (2012)
60. Jhawar, K., Sanyal, D.K., Chattopadhyay, S., Bhowmick, P.K., Das, P.P.: Author name disambiguation in PubMed using ensemble-based classification algorithms. In: Proceedings of 20th ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 469–470. ACM (2020)
61. Rehs, A.: A supervised machine learning approach to author disambiguation in the web of science. J. Informetr. **15**, 101166 (2021)
62. Kim, J., Kim, J., Owen-Smith, J.: Ethnicity-based name partitioning for author name disambiguation using supervised machine learning. J. Assoc. Inf. Sci. Technol. **72**, 979–994 (2021)
63. Mihaljević, H., Santamaría, L.: Disambiguation of author entities in ads using supervised learning and graph theory methods. Scientometrics **126**, 3893–3917 (2021)
64. Zhang, L., Ban, Z.: Author name disambiguation based on rule and graph model. In: Proceedings of 9th International Conference on Natural Language Processing and Chinese Computing (NLPCC), pp. 617–628. Springer (2020)
65. Kim, J., Owen-Smith, J.: ORCID-linked labeled data for evaluating author name disambiguation at scale. Scientometrics **126**, 2057–2083 (2021)
66. Boukhers, Z., Asundi, N.B.: Whois? Deep author name disambiguation using bibliographic data. In: Proceedings of 26th International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 201–215. Springer (2022)
67. Li, H., Cui, Y., Wang, T.: An effective approach for automatic author name disambiguation based on multiple strategies. In: Proceedings of 3rd International Conference on Computer Science and Software Engineering (CSSE), pp. 169–175. ACM (2020)
68. Rodrigues, N.D.S., Costa, A.R., Lemos, L.C., Ralha, C.G.: Multistrategic approach for author name disambiguation in bibliography repositories. In: Proceedings of 8th Annual International Conference on Information Management and Big Data (SIMBig), pp. 63–76. Springer (2021)
69. Waqas, H., Qadir, M.A.: Multilayer heuristics based clustering framework (MHCF) for author name disambiguation. Scientometrics **126**, 7637–7678 (2021)
70. D'Angelo, C.A., van Eck, N.J.: Collecting large-scale publication data at the level of individual researchers: a practical proposal for author name disambiguation. Scientometrics **123**, 883–907 (2020)
71. Zhang, Z., Yu, B., Liu, T., Wang, D.: Strong baselines for author name disambiguation with and without neural networks. In: Proceedings of 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp. 369–381. Springer (2020)
72. Mozafari, N.: A genetic-based approach for author name disambiguation problem. Iran. J. Inf. Process. Manag. **36**, 791–816 (2021)
73. Jinqi, Q., Luoyi, F., Xiaoying, G., Xinbing, W.: A network maximum flow based approach for author name disambiguation. J. Shanghai Jiaotong Univ. **54**, 111 (2020)
74. Zhou, Q., Chen, W., Wang, W., Xu, J., Zhao, L.: Multiple features driven author name disambiguation. In: Proceedings of IEEE International Conference on Web Services (ICWS), pp. 506–515. IEEE (2021)
75. Chuanming, Y., Yunci, Z., Aochen, L., Lu, A.: Author name disambiguation with network embedding. Data Anal. Knowl. Discov. **4**, 48–59 (2020)
76. Ma, Y., Wu, Y., Lu, C.: A graph-based author name disambiguation method and analysis via information theory. Entropy **22**, 416 (2020)
77. Ma, X., Wang, R., Zhang, Y., Jiang, C., Abbas, H.: A name disambiguation module for intelligent robotic consultant in industrial Internet of Things. Mech. Syst. Signal Process. **136**, 106413 (2020)
78. Wang, C., He, X., Zhou, A.: HEEL: exploratory entity linking for heterogeneous information networks. Knowl. Inf. Syst. **62**, 485–506 (2020)
79. Wang, H., et al.: Author name disambiguation on heterogeneous information network with adversarial representation learning. In: Proceedings of 34th AAAI Conference on Artificial Intelligence, pp. 238–245. AAAI Press (2020)
80. Pooja, K.M., Mondal, S., Chandra, J.: A graph combination with edge pruning-based approach for author name disambiguation. J. Assoc. Inf. Sci. Technol. **71**, 69–83 (2020)
81. Firdaus, et al.: Author identification in bibliographic data using deep neural networks. TELKOMNIKA Telecommun. Comput. Electron. Control **19**, 911–919 (2021)
82. Xiong, B., Bao, P., Wu, Y.: Learning semantic and relationship joint embedding for author name disambiguation. Neural Comput. Appl. **33**, 1987–1998 (2021)
83. Correia, A., et al.: AuthCrowd: author name disambiguation and entity matching using crowdsourcing. In: IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 150–155. IEEE (2021)
84. Zhang, Z., et al.: Author name disambiguation using multiple graph attention networks. In: Proceedings of International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2021)

85. Pooja, K.M., Mondal, S., Chandra, J.: Exploiting similarities across multiple dimensions for author name disambiguation. Scientometrics **126**, 7525–7560 (2021)
86. Zhang, L., Huang, Y., Yang, J., Lu, W.: Aggregating large-scale databases for PubMed author name disambiguation. J. Am. Med. Inf. Assoc. **28**, 1919–1927 (2021)
87. Firdaus., et al.: Author matching classification on a highly imbalanced bibliographic data using cost-sensitive deep neural network. In: Proceedings of International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), pp. 86–89. IEEE (2021)
88. Färber, M., Lamprecht, D.: The data set knowledge graph: creating a linked open data source for data sets. Quant. Sci. Stud. **2**, 1324–1355 (2021)
89. Backes, T., Dietze, S.: Lattice-based progressive author disambiguation. Inf. Syst. **109**, 102056 (2022)
90. Manzoor, A., Asghar, S., Amjad, T.: Toward a new paradigm for author name disambiguation. IEEE Access **10**, 76055–76068 (2022)
91. Färber, M., Ao, L.: The Microsoft Academic Knowledge Graph enhanced: author name disambiguation, publication classification, and embeddings. Quant. Sci. Stud. **3**, 51–98 (2022)
92. Qiping, D., Weijing, C., Ling, J., Yu'e, Z.: Author name disambiguation based on heterogeneous information network. Data Anal. Knowl. Discov. **6**, 60–68 (2022)
93. Santini, C., et al.: A knowledge graph embeddings based approach for author name disambiguation using literals. Scientometrics **127**, 4887–4912 (2022)
94. Waqas, H., Qadir, A.: Completing features for author name disambiguation (AND): an empirical analysis. Scientometrics **127**, 1039–1063 (2022)
95. Pooja, K.M., Mondal, S., Chandra, J.: Online author name disambiguation in evolving digital library. Neurocomputing **493**, 1–14 (2022)