

OVERVIEW



WILEY

Scholarly data mining: A systematic review of its applications

Amna Dridi | Mohamed Medhat Gaber | R. Muhammad Atif Azad | Jagdev Bhogal

School of Computing and Digital Technology, Birmingham City University, Birmingham, UK

Correspondence

Amna Dridi, School of Computing and Digital Technology, Birmingham City University, Millennium Point, Birmingham B47XG, UK.
Email: amna.dridi@mail.bcu.ac.uk

Abstract

During the last few decades, the widespread growth of scholarly networks and digital libraries has resulted in an explosion of publicly available scholarly data in various forms such as authors, papers, citations, conferences, and journals. This has created interest in the domain of big scholarly data analysis that analyses worldwide dissemination of scientific findings from different perspectives. Although the study of big scholarly data is relatively new, some studies have emerged on how to investigate scholarly data usage in different disciplines. These studies motivate investigating the scholarly data generated via academic technologies such as scholarly networks and digital libraries for building scalable approaches for retrieving, recommending, and analyzing the scholarly content. We have analyzed these studies following a systematic methodology, classifying them into different applications based on literature features and highlighting the machine learning techniques used for this purpose. We also discuss open challenges that remain unsolved to foster future research in the field of scholarly data mining.

This article is categorized under:

Algorithmic Development > Text Mining

Application Areas > Science and Technology

KEYWORDS

academic social network, citation analysis, conference analysis, document analysis, literature analysis, scholarly data mining, trend analysis

1 | INTRODUCTION

With the vast increase of research works undertaken in academia and industry and the widespread use of scholarly networks and digital libraries, we now have access to abundant academic resources. The volume of these resources has exceeded 114 million accessible on the web on 2014. The rate of the newly generated scholarly documents has accordingly overreached tens of thousands per day according to (Wu et al., 2014). As a consequence of this increasing volume of scholarly data, the extraction of useful knowledge and the understanding of the structure and dynamics of science are hampered. This has recently led to the emergence of *scholarly data mining* as an important research field, facing new challenges due to the typical nature of science, considering the complexity of the academic landscape and the 5V

feature (volume, variety, velocity, value, and veracity) of scholarly data (Kaisler, Armour, Espinosa, & Money, 2013; Xia, Wang, Bekele, & Liu, 2017).

Currently, the main problem to be faced by researchers and scholars is not simply obtaining any useful information from this accessible reservoir of data, but understanding the structure of the scholarly communication and track the dynamics of science, in order to provide better academic services for scholars and researchers. Nevertheless, this is not a trivial task since scholarly data is very different and usually includes some special features: complexity drawn from the fact that it involves various entities (papers, authors, and journals) and relationships among these entities; and veracity, which comes from author disambiguation and deduplication (Ferreira, Gonçalves, & Laender, 2012).

The extraction of useful knowledge from this amount of data is essential to provide support not only to scholars on their understanding of the rules and laws of science (Xia et al., 2017), but also to governments and institutions on several decision making processes such as policy making for fund disbursement, speculating upcoming research areas, and so on. In this regard, scholarly data mining gathers the most useful methods and techniques, such as machine learning (ML) techniques (Anderson, McFarland, & Jurafsky, 2012; Dey, Roy, Chakraborty, & Ghosh, 2017; Lu, Huang, Bu, & Cheng, 2018; Salatino, Osborne, & Motta, 2018; Sun, Kaur, Possamai, & Menczer, 2011; Taskin & Al, 2018; Zhao et al., 2018), to derive better academic sources to scholars and researchers. It is a process in which knowledge is extracted from the available scholarly data for different literature-based applications including *citation analysis* (Dey et al., 2017; Ding et al., 2014; Shi, Tong, Tang, & Lin, 2015; Taskin & Al, 2018), *document analysis* (Caragea, Bulgarov, & Mihalcea, 2015; Kim, Hansen, & Helps, 2018; Shardlow et al., 2018; Tuarob, Bhatia, Mitra, & Giles, 2016), *conference analysis* (Effendy, Jahja, & Yap, 2014; Effendy & Yap, 2016), *trend analysis* (An, Han, & Park, 2017; Dridi, Gaber, Azad, & Bhogal, 2019a, 2019b; Hou, Yang, & Chen, 2018; Rossetto, Bernardes, Borini, & Gattaz, 2018; Santa Soriano, Lorenzo Álvarez, & Torres Valdés, 2018; Zhang & Guan, 2017), and *literature analysis* (Dunne, Shneiderman, Gove, Klavans, & Dorr, 2012; Li, Council, Lee, & Giles, 2006; Liu, Huang, Yan, & Chen, 2015; Osborne, Motta, & Mulholland, 2013; Tan et al., 2016; Tang, 2016; Tang et al., 2008).

Although the study of big scholarly data is relatively new, some studies have emerged Xia et al. (2017) on how to investigate scholarly data usage in different disciplines. These studies motivate investigating the scholarly data generated via academic technologies such as scholarly networks and digital libraries for building scalable approaches for retrieving, recommending, and analyzing the scholarly content. Consequently, this has spawned five key applications that are mentioned above. Nevertheless, due to the increasing interest to scholarly data mining, it becomes essential to closely study the approaches for scholarly data analysis, categorize them based on the literature features or explore the techniques involved in mining scholarly data. In this regards, the aim of this article is to systematically review the most interesting research works published on the use of scholarly data mining. More than 90 research articles have been analyzed with special attention paid to the investigated literature features and the different analysis methods used. The final aim is therefore to provide the readers with a systematic revision about existing scholarly data mining applications; the involved techniques and the application areas. The study of the existing scholarly data mining was conducted from the content perspective. In other words, we have systematically investigated the content of works on big scholarly data and attempted to focus on the applications by semantically examining them. In contrast to (Khan, Liu, Shakil, & Alam, 2017; Xia et al., 2017) who have explored big scholarly data from big data perspective following a generic view, we have paid particular attention to the semantic interpretation of big scholarly data and emphasized their applications. This draws upon interesting insights into how scholarly data is learnt and used. It is hoped that these insights will contribute to a deeper understanding of scholarly data applications and provide an important opportunity to institutions and governments for decision making processes.

The article is structured as follows. In Section 2, we describe the systematic methodology followed. In Section 3, we analyze the research interests in scholarly data mining applications. In Sections 4 and 5, we analyze these applications following a literature-based analysis and we discuss the analysis methods used in these applications, while in Section 6, we show the areas of application of scholarly data mining. In Section 7, we discuss the open challenges in the area of scholarly data mining and present the limitations of the current review. Finally, in Section 8, we present the conclusions of this work.

2 | METHODOLOGY

In this review, we follow the systematic procedure proposed by Kitchenham (2004) to undertake a systematic review on scholarly data mining applications. We start by specifying the research questions being addressed and then we detect

relevant literature to augment our understanding of these questions, in a structured way. In this section, we present the details of the followed procedure.

2.1 | The need for a review

The importance of *scholarly data mining* has been raised for several key reasons. First, it is the availability of abundant academic resources. In addition to the scholarly documents such as papers, books, reports, and others, multiple associated data are available today including information about authors, citations, institutions, funds, and academic networks (Liu et al., 2018). Furthermore, there have been several initiatives by governments and organizations to digitize academic resources in order to meet the challenges of information explosion. As a matter of fact, the scholarly communication has been revolutionized drastically over the past two decades due to the unprecedented advancement in information and communication technology. This latter has brought a revolutionary form in archiving and accessing knowledge in the digitized form that used to be in the conventional print form. Second, due to this easily accessible reservoir of data, researchers and scholars are in an elevating need for a deeper understanding of the structure and dynamics of science. In other words, sophisticated techniques and tools are highly solicited to help researchers to learn better about knowledge production processes, curate insights from scholarly data, and speculate upcoming research topics. Third, the availability of this vast amount of data about scientists' collaborations, document sharing, and publications enables the evaluation of scientific impact of different entities including papers, authors, and journals. The measurement of this scientific impact is deemed vital for the governments and businesses for decision making processes such as funding allocation, research gap identification, university ranking determination, tenure, and recruitment decisions. Finally, going beyond the study of the scientific impact, scholarly data analysis also promotes the understanding of human social activities. It provides sociologists with valuable data to observe researcher interactions and community formation. It also allows countries to evaluate the impact of institutions or scientists to allocate resources. Overall, scholarly data mining contributes to the *Science of Science* (Fortunato et al., 2018; Light, Polley, & Börner, 2014) that advances our understanding of the structure and dynamics of science.

To our knowledge, there have not been enough attempts to closely study the approaches for scholarly data analysis, categorize them based on the literature features or explore the techniques involved in mining scholarly data. An effort in this direction could reveal new branches for future research in this area. In fact, there have been some recent reviews related to scholarly data but not deeply exploring the aforementioned specific issue. For instance, from one side, we found two reviews (Khan et al., 2017; Xia et al., 2017) that have treated scholarly data from a general perspective. They have studied the use of big data in scholarly ecosystems starting from scholarly data management and relevant technologies, passing through data analysis methods and finally looking into the research issues. From the other side, we found three reviews that were specific and have narrowed down the perspective. The first review (Ding et al., 2014) provides a comprehensive overview of citation analysis in terms of its theoretical foundations, methodical approaches, and example applications. The second review (Liu et al., 2018) addressed the issue of scholarly data visualization by focusing on the visualization tools and analytic systems. The third review (Bai et al., 2019) deals with the scientific recommendation problem as a sub-problem of scholarly data analysis. It provides a comprehensive review on the scholarly paper recommendation by reviewing the recommendation algorithms, introducing the evaluation methods of different recommender systems and highlighting the open issues in the paper recommendation systems.

These reviews are all recent which shows a great interest in the topic of scholarly data mining, not only for the opportunities it offers to the scientists and scholars to understand the unprecedented amount of scholarly data freely available, but also for institutions and governments that could take benefit from the proposed approaches for decision making processes.

2.2 | Review questions

In this review, we bring together the latest groundbreaking research on applications related to scholarly data mining and knowledge discovery from scientific data. For this matter, we provide an analysis of the used techniques and the application areas. Besides this, we also give an overview of the scope for future directions by addressing the challenges in the field. Specifically, we address the following research questions:

- Q1: What has been the interest in the topic since it started flourishing?
 Q2: Which applications scholarly data mining support?
 Q3: Which are the most commonly used techniques for mining scientific data?
 Q4: Which areas are associated with the greatest interest in the topic?
 Q5: What challenges that continue to exist in the field?

2.3 | Search process

We used two methods to obtain the articles reviewed in this work:

- *Database search* using queries related to the aim of this review. The following databases were used to gather articles combining scholarly data mining tools, techniques and applications: Web of Science¹, Google Scholar², DBLP³, ScienceDirect⁴, ACM Digital Library⁵, and IEEE Xplore Digital Library⁶. In each of these databases, we used search queries which search for terms—related to the topic of the review—in the title of publications such as “scholarly data mining,” “knowledge discovery,” “scientific data,” “trend analysis,” “scientific recommendation,” “citation analysis,” and “bibliometrics.”
- *Selection of related publication venues* that are known to publish in the area of scholarly data and related topics. Specifically, we found two venues. The first venue is Scientometrics journal⁷ which is a peer reviewed journal concerned with the quantitative aspects of the Science of Science and scientific research. The second venue is the Joint Conference on Digital Libraries (JCDL)⁸ that represents a major international forum focusing on digital libraries and associated issues.

As the search queries used were a mixed bag of generic and specific terms, we then refined the results by reading the titles and the abstracts within the articles, filtering out those that did not align with the scope of this review. Particularly, all papers had to be about applications of scholarly data mining and knowledge discovery from scientific data.

All the works studied in this review were published in peer reviewed journals or top conferences, which guarantees that they satisfy a certain standard of quality.

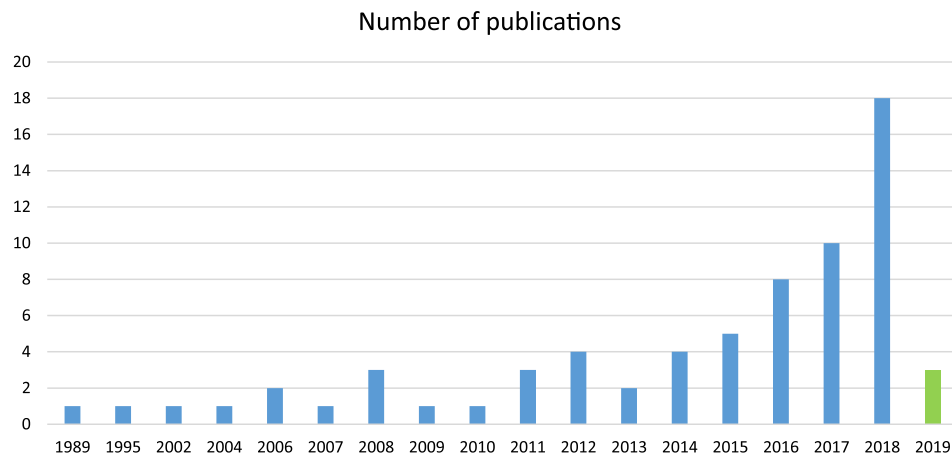
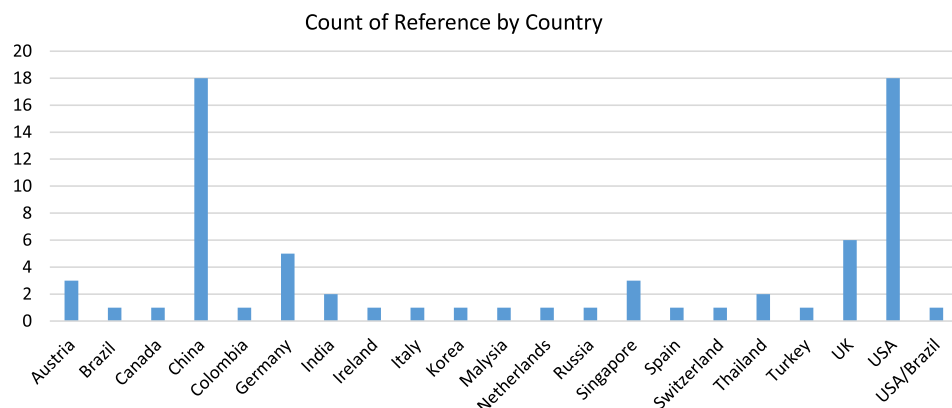
2.4 | Data extraction

After establishing the pool of articles related to the topic of this survey, we went through them extracting the following information:

- Year of publication; publication type (journal or conference); author's country; and number of citations.
- The application of scholarly data mining categorized into five categories based on the literature features: *citation analysis*, *document analysis*, *conference analysis*, *trend analysis*, or *literature analysis* that encloses all the previous features together.
- The technique used for the analysis ranging from *empirical studies* to *ML techniques*.
- The application area of the publication.

3 | RESEARCH INTEREST ANALYSIS

We have analyzed 95 publications, 64 of which are journal publications, while 31 are published in conferences. The increasing interest in the research of scholarly data mining becomes clear from the yearly increase in the number of publications related to this area mainly in the last 5 years starting from 2004 (see Figure 1). But, it is worth mentioning that the field traces its roots back in the late 1980's and mid 1990's through initial empirical studies principally on citation analysis. These early studies were focusing on the understanding of the citation behavior of scientific community. This may have contributed to the study of only citations as a literature feature of scientific publications during the first decade that preceded the emergence of the field. Also, this may justify why these analyses were mainly tackled in the area of Information Science (IS) rather than the area of Computer Science.

FIGURE 1 Number of publications by year**FIGURE 2** Number of publications by country**TABLE 1** Mean number of citations by year of the articles from 2011 to 2019

2011	2012	2013	2014	2015	2016	2017	2018	2019
29.33	93	61.5	66.75	13.6	15.87	13.1	7.4	3.33

It is also interesting to analyze the geographic distribution of publications. For this end, we took the country of publications' authors. Figure 2 shows the total number of publications corresponding to each country. What can be clearly seen in this figure is the dominance of both USA and China in publishing in the area, followed by the UK and Germany. Interestingly, this highly correlates with the report on the research outputs by country⁹, published by *Nature Index* in 2019. Also, it aligns with the global Research and Development (R&D) spending by country according to the latest statistics from the UNESCO Institute for Statistics¹⁰.

A further interest indicator is the number of citations of each work. We analyzed the mean number of citations per year in the last decade. The results are very sensitive to the low number of works per year that we have in the pool of studied articles. Table 1 illustrates the mean number of citations. It is somewhat intuitive that the mean number of citations of the papers published early in the time is significantly higher than the one of those published recently. However, it is worth noting that the mean number of citations is also relatively high during the last 2 years. This is insightful as it shows that the research area of big scholarly data is developing and becoming of interest nowadays.

4 | LITERATURE-BASED ANALYSIS

The ultimate goal of scholarly data mining is to understand the relational structure of science and provide the scholars with better academic services such as academic recommendation and literature organization. To this end, scholarly data mining involves various applications, which are mainly categorized based on the literature features; documents,

TABLE 2 Summary of references related to the applications of scholarly data mining

Applications	Sub-applications	References
Citation analysis	Citation and co-citation analysis	(Dey et al., 2017; Ding et al., 2014; Shi et al., 2015; Taskin & Al, 2018) (Jeong, Song, & Ding, 2014; Zhao et al., 2018) (Acuna, Allesina, & Kording, 2012; Caragea et al., 2015; Trujillo & Long, 2018) (An et al., 2017)
	Readership analysis	(Joseph & Thelwall Mike, 2017; Maflahi & Thelwall, 2018; Thelwall, 2018)
	Bibliometrics analysis	(Gleason, 1961; Godin, 2006; Lv et al., 2011; Martínez-Gómez, 2015; McBurney & Novak, 2002; Merediz-Solà & Bariviera, 2019; Mingers & Leydesdorff, 2015; Monroy & Diaz, 2018; Pilkington, 2004)
	Altmetrics analysis	(Bornmann & Haunschild, 2018; Nabout et al., 2018; Priem & Costello, 2010; Weller, Dröge, & Puschmann, 2011)
Document analysis	Authorship analysis	(Rexha, Kröll, Ziak, & Kern, 2018; Sun et al., 2011)
	Document structure analysis	(Boyack, Smith, & Klavans, 2018; Heffernan & Teufel, 2018; Lu et al., 2018; Weber & Gunawardena, 2011)
	Content analysis	(Caragea et al., 2015; Kim et al., 2018; Shardlow et al., 2018; Tuarob et al., 2016)
	Scientific recommendation	(Alam & Ismail, 2017; Zhao et al., 2018)
	Scientific text summarization	(Bhatia & Mitra, 2012; Jha, Abu-Jbara, & Radev, 2013; Mei & Zhai, 2008; Qazvinian & Radev, 2008)
Conference analysis		(Effendy et al., 2014; Effendy & Yap, 2016)
Trend analysis		(An et al., 2017; Hou et al., 2018; Rossetto et al., 2018; Santa Soriano et al., 2018; Zhang & Guan, 2017) (Dridi et al., 2019a, 2019b; Weismayer & Pezenka, 2017) (Effendy & Yap, 2017; Hoonlor, Szymanski, & Zaki, 2013)
Literature analysis		(Dunne et al., 2012; Li et al., 2006; Liu et al., 2015; Osborne et al., 2013; Tan et al., 2016; Tang, 2016; Tang et al., 2008)

citations, conferences, and trends. In this section, we examine these applications with respect to the aforementioned literature features. Table 2 summarizes these applications and categorizes the references with respect to the literature features they have investigated.

4.1 | Citation analysis

As citation counts remain the paramount measure of scientific impact, citations—as a literature feature—have presided the attention of researchers investigating the area of scholarly data mining. Therefore, citation analysis has been widely explored as an application of scholarly data mining.

Citation analysis has been extensively used to understand the scholarly communication through citation patterns (Ding et al., 2014). It does not only aim to assess the impact of research outputs, but it also aims to reveal the scholarly communication, map the landscape of scientific disciplines and track the knowledge transfer across domains.

In addition to the conventional study of citation and co-citation network to measure the impact of scientific papers, citation analysis involves different patterns that include *readership counts* (Joseph & Thelwall Mike, 2017; Maflahi & Thelwall, 2018; Thelwall, 2018), *bibliometrics* (Godin, 2006; Lv et al., 2011; Martínez-Gómez, 2015; McBurney & Novak, 2002; Monroy & Diaz, 2018; Pilkington, 2004) and even the metrics derived from social media—termed as *altmetrics* (Bornmann & Haunschild, 2018; Nabout et al., 2018; Priem & Costello, 2010; Weller et al., 2011).

4.1.1 | Citation and co-citation analysis

Citation and co-citation analysis is shaped around citation counting and citation relationships between documents/scholars that are cited together by other documents/scholars (Dey et al., 2017; Shi et al., 2015; Taskin & Al, 2018).

Citation analysis mainly studies the citation network, where nodes represent papers, authors, or journals, and edges represent the number of times each paper/author has been cited, co-authored, or co-cited (Ding et al., 2014). It, then,

measures the impact of published research quantitatively, and could be termed as *count-based citation analysis*. While citation count remains essential to measure the scientific impact, it fails to address the “how and why” questions of citation analysis. To fill this gap, *content-based citation analysis* (CCA; Ding et al., 2014) has been proposed as the next generation of citation analysis. It aims to study both syntactic and semantic levels of citations. The syntactic level considers the location of the reference in a citing article while the semantic level studies why a reference has been cited in a citing article. The content analysis of citations has included both manual approach and semi-automatic approach of natural language processing (NLP). The main goal of CCA is then to develop a code-book used to annotate citation contexts. NLP techniques have been used to extract the key concepts from citation contexts to understand the citing behavior. However, the identification of the best window size to extract the proper citation context and the detection of the correct citing paper sections are still an open challenge.

Co-citation analysis studies citation relationships in the co-citation network (Zhao & Ye, 2013). It measures the frequency with two papers/scholars are cited together by other papers/scholars. Typically, there are mainly two types of co-citation analysis methods namely *author co-citation analysis* Jeong et al. (2014) and *document co-citation analysis* (Trujillo & Long, 2018). The author co-citation analysis measures the similarity between co-cited authors by considering author's citation context. For this end, the citing sentences are extracted to obtain the topical relatedness between the cited authors instead of traditional author co-citation frequency. The citing sentence similarity is then measured by topical relatedness between two citing sentences. However, the document co-citation analysis enables to identify relevant literature and scholarly communities that may be left unnoticed in standard approaches to literature searching. Resulting networks help to identify gaps between published research areas. Document co-citation analysis is then proposed as a potential methodology to promote trans-disciplinary. In Trujillo and Long (2018), the authors have explored 229 source articles from the literature of systems thinking, extracted from the Web of Science Core Collection. After generating the document co-citation network, the authors have explored patterns in influential literature developed across different disciplines. For instance, they have demonstrated that community structure could be detected within the co-citation networks for systems thinking. Both of author and document co-citation analyses enable the identification of the intellectual structure of a research domain and the recognition of relevant scholarly communities.

Both citation and co-citation analyses have been used for different scholarly aims such as author names ambiguity (Sun et al., 2011), topic classification (Caragea et al., 2015), scientific success prediction (Acuna et al., 2012), identification of sleeping beauties (Dey et al., 2017), identification of dynamic knowledge flow patterns (An et al., 2017), and trend analysis Hou et al. (2018).

4.1.2 | Readership analysis

Readership analysis studies Mendeley¹¹ reader counts—that correspond to the number of readers of each article—and their evidence of early scholarly impact for published articles. Different works (Joseph & Thelwall Mike, 2017; Maflahi & Thelwall, 2018; Thelwall, 2018) have studied whether Mendeley reader counts reflect the scholarly impact of publications. While (Maflahi & Thelwall, 2018; Thelwall, 2018) have focused their studies on investigating the early scholarly impact of Mendeley reader counts for journal articles, (Joseph & Thelwall Mike, 2017) have studied whether this impact is equally true for conference papers. To do so, the authors have extracted Mendeley readership data and Scopus citation counts for both journal articles and conference papers published in 2011 in computer science and engineering. The authors have found Mendeley a moderate correlation between readership counts and citation counts for both journal articles and conference papers in Computer Science. However, the correlations were much lower between Mendeley readers and citation counts for conference papers than for journal articles in engineering. Therefore, there seem to be disciplinary differences in the usefulness of Mendeley readership counts as impact indicator for conference papers. Overall, all research works investigating Mendeley reader counts have found significant positive correlations between readership counts and citation counts, while Mendeley reader counts appear before citations. Readership analysis has been then proposed as a valuable early impact indicator for published research, addressing the issue of citations that take time to accumulate.

4.1.3 | Bibliometrics analysis

Bibliometrics analysis (Monroy & Diaz, 2018) focuses on the use of statistical analysis to examine scientific production patterns in a scientific field (Godin, 2006; McBurney & Novak, 2002). For instance, the authors in Monroy and

Diaz (2018) have applied time series tools to bibliometric data to conduct a comparative study of the dynamics of scientific production for several countries, in terms of papers published. They have compared the histories of scientific development of countries, aiming to understand the causes, and circumstances that led to dynamics of knowledge production. They have then identified the dynamical changes that affected global scientific production, and the instances where global production was influenced by social, political, and economic circumstances. On the other hand, bibliometrics have applied statistical analysis to assess relationships between authors, entities, journals, or countries, in addition to measuring the impact of research and linkage involving co-citations and keywords employed (Lv et al., 2011; Martínez-Gómez, 2015; Pilkington, 2004). In this context, the authors in Lv et al. (2011) have applied statistical analysis and knowledge visualization technology to study graphene literature from different subjects, authors, countries, and keywords distributed in several aspects of research topics. For this matter, the authors have collected and analyzed data from 1991 to 2010 from the Science Citation Index database, Conference Proceeding Citation Index database and Derwent Innovation Index database integrated by Thomson Reuters. Their bibliometric analysis has shown that the clusters distributed regularly in keywords of applied patents in recent 5 years due to the potential applications of graphene research gradually found.

Bibliometrics analysis is centrally, but not only, based on citation analysis. It also involves descriptive linguistics (Gleason, 1961), the development of thesauri, the evaluation of reader usage, and the analysis of associated keywords. All these bibliometrics patterns are used to identify research clusters, emerging topics, and leading scholars in bitcoin literature by analyzing 1,162 papers indexed in Web of Science (Merediz-Solà & Bariviera, 2019).

Bibliometrics are frequently used in the field of library and information science. A sub-field of it—that is concerned with the study of scientific publication—is called *scientometrics*, which is defined as “the study of the quantitative aspects of the process of science as a communication system” (Mingers & Leydesdorff, 2015).

4.1.4 | Altmetrics analysis

Altmetrics analysis (Bornmann & Haunschild, 2018; Nabout et al., 2018; Priem & Costello, 2010; Weller et al., 2011) supports the use of activities on on-line social media platforms as an early signal of research impact for scientific publications. Altmetrics seek new means of quantifying the impact of research outside the realm of research papers, such as online media and social network. This class of metrics includes mentions in the news, blogs, and on Twitter; article page-views and downloads; GitHub repository watchers.

Altmetrics have been considered as a measure of scientific dissemination and an early indicator of scientific influence and impact. For instance, they can point to interesting spikes in different types of attention. As a proof of evidence, some studies (Nabout et al., 2018) have shown that altmetrics are concordant with citation-based metrics. By way of illustration, (Nabout et al., 2018) have studied the correlation between traditional citation-based indicators and activities on online social media platforms in a dataset of 2,863 papers published in five ecological journals. Their result supported the use of activities on online social platforms as an early signal of research impact of ecological articles. However, this outcome is not totally supported by (Bornmann & Haunschild, 2018) who studied Twitter dataset to measure the impact of science and found that without considering the content of the tweets, simple counting can lead to wrong conclusions.

4.2 | Document analysis

A document represents an important literature feature in scholarly communication, which is defined by a set features itself including *author*, *content*, *structure*; and used for different scholarly usages such as *recommendation* and *summari-zation*. Document analysis includes then subsequent analyses detailed below.

4.2.1 | Authorship analysis

Authorship analysis has been treated differently in the literature. For instance, the authors in Rexha et al. (2018) have proposed to associate segments of text with their real authors using content-agnostic and stylometric features to solve the problem of authorship identification. For this purpose, two pilot studies have been conducted on a selected data from the free database created by the US National Library of Medicine—PubMed. Both studies aimed to understand

how humans can identify authorship among documents with high content similarity. The first study was a quantitative experiment involving crowd-sourcing, while the second was a qualitative one executed by the authors. Both experiments and observations contribute to automate the process of authorship identification as well as to distinguish specific features used by humans in their decision making process. In Sun et al. (2011), however, the authors have explored heuristic features based on citations and crowd-sourced topics to detect ambiguous author names in the context of social citation analysis systems such as *Scholarometer* system. Two classes of features were used. The first is a heuristic based on the percentage of citation accrued by the top name variations for an author, while the second feature class relies on crowd-sourced data to detect ambiguity at the topic level. The proposed approach succeeded to detect ambiguous author names in crowd-sourced scholarly data with an accuracy of 75%.

4.2.2 | Document structure analysis

Document structure analysis (Boyack et al., 2018; Heffernan & Teufel, 2018; Lu et al., 2018) studies the internal document structure by identifying the functional structure (further detailed at three levels: section header-based identification, section content-based identification, or paragraph-based identification [Lu et al., 2018]), identification of problems and solutions in a specific paper by making a binary decision about problem-hood and solution-hood of a given phrase in article Heffernan and Teufel (2018), or by studying research proposals and analyzing their discourse for clarity (Boyack et al., 2018). For the functional structure identification, the authors in Lu et al. (2018) have proposed a novel clustering algorithm to generate a domain-specific functional structure, applied to 300 research articles in computer science. The application of the proposed approach, in two tasks: academic search and keyword extraction, confirms that the identified structure obtains more relevant information and achieves better performance. However, for the identification of problems and solutions, the authors in (Heffernan & Teufel, 2018) have proposed an automatic classifier that makes a binary decision about problem-hood and solution-hood of a given phrase, that may or may not be a description of a scientific problem or a solution. The authors have defined a set of 15 features, including syntactic information (POS tags), document and word embeddings, and have applied several ML algorithms such as Naïve Bayes, Logistic Regression and Support Vector Machine, on a corpus of 2000 positive and negative examples of problems and solutions extracted from the 2016 ACL (Association of Computational Linguistics) anthology. The obtained results reveal the ability of the proposed classifier to distinguish problems from nonproblems with an accuracy of 82%, and solutions from nonsolutions with an accuracy of 79%. Regarding research proposal analysis, the authors in (Boyack et al., 2018) have used both citation and discourse analyses of 369R01 proposals submitted to the U.S. National Institutes of Health (NIH) by the University of Michigan Medical School, to discover possible predictors of proposal success. The analyses have focused on two issues: the Matthew effect in science—Merton's claim that eminent scientists have an inherent advantage in the competition for funds—and quality of writing or clarity. The obtained results suggested that a clearly articulated proposal is more likely to be funded than a proposal with lower quality of discourse.

Document structure analysis also includes the study of slide presentations (Weber & Gunawardena, 2011). It investigates the use of knowledge units—that represent scientific knowledge by combining elements of procedural, declarative, and structural knowledge—for the automated construction of slides. The knowledge units have been defined as the three paradigms of the research process *background*, *progress*, and *completed*.

4.2.3 | Content analysis

Document content analysis treats the document in two ways: a coarse-grained way by studying only keywords (Kim et al., 2018) and a fine-grained way by digging into the paper textual content (Caragea et al., 2015; Shardlow et al., 2018; Tuarob et al., 2016).

Keyword analysis (Kim et al., 2018) examines connections between keywords used to describe theses and dissertations in order to vividly picture similarities and differences among research domains. In this context, the authors in (Kim et al., 2018) have analyzed data from 29,435 dissertations and theses found in the ProQuest Theses and Dissertation database in the years 2009–2014. The obtained results identified interdisciplinary clusters, as well as the key differences in connections between the four computing disciplines in the database: computer science, computer engineering, information technology, and IS. However, textual content analysis involves different applications such as topic classification (Caragea et al., 2015); identification of research hypotheses (Shardlow et al., 2018); and extraction of algorithms (Tuarob et al., 2016).

For topic classification, the authors in (Caragea et al., 2015) have proposed a co-training approach that uses the text and citation information of a research article as two different views to identify the topic of an article. A subset sampled from the *CiteSeer*^x digital library, consisting of 3,186 labeled papers, has been used for topic classification with a co-training classifier. The obtained results showed that the proposed approach performs better than other semi-supervised and supervised methods. However, for the identification of research hypotheses, the authors in (Shardlow et al., 2018) have proposed a supervised method to extract new meta-knowledge dimensions that encode research hypotheses. A corpus of one thousand MEDLINE abstracts on the subject of transcription factors in human blood cells has been used, and a random forest classifier has been applied to achieve a better performance than previous efforts in detecting knowledge type, with a precision ranging from 86 to 100%. Regarding the extraction of algorithms, the authors in (Tuarob et al., 2016) have developed *AlgorithmSeer*, a system for extracting and searching for algorithms. To do so, hybrid ML approaches have been proposed to discover algorithm representations, and different techniques have been adopted to extract textual metadata for each algorithm. Finally, a demonstration version of *AlgorithmSeer* that is built on Solr/Lucene open source indexing and search system is presented and applied to over 200,000 algorithms extracted from over 2 million scholarly documents.

4.2.4 | Scientific recommendation

Scientific recommendation includes paper/topic recommendation (Alam & Ismail, 2017) and reviewer recommendation (Zhao et al., 2018). Scientific paper recommendation has been provided to assist scholars in finding relevant papers across the tremendous amount of academic information in the era of big scholarly data. In this context, (Kong, Mao, Wang, Liu, & Xu, 2018) have developed *VOPRec*, a scientific paper recommendation system based on vector representation learning of paper in citation networks. In fact, paper recommendation takes into account both text information of papers and structural identity with the citation network. Similarly, topic recommendation hinges upon bibliometric information of the literature to identify a suitable topic of current importance from a plethora of research topics. In this context, Alam and Ismail (2017) have developed *RTRS*—a recommender system for academic researchers—to assist both novice and experienced researchers in selecting research topics in their chosen field.

On the other hand, reviewer recommendation (Zhao et al., 2018) recommends suitable reviewers for a paper submitted for review. Reviewer recommendation is defined as a classification problem where both submissions and reviewers are described by some tags such as keywords and research interests and a Word Mover's Distance is used to measure the minimum distance between submissions and reviewers.

4.2.5 | Scientific text summarization

Scientific text summarization has been proposed to help scholars to know about the most influential content of the paper due to the vast growth of literature that makes it difficult for them to find high impact articles on unfamiliar topics (Mei & Zhai, 2008). Different approaches have been proposed to generate summaries of scientific articles. The authors in Jha et al. (2013) have presented a system that takes a topic query as input and generates a survey of the topic by first selecting a set of relevant documents, and then selecting relevant sentences from those documents. In Qazvinian and Radev (2008), on the other hand, the authors have proposed a citation summary network that uses a clustering approach where communities in the citation summary's lexical network are formed and sentences are extracted from separate clusters. Another summarizing problem has been tackled by Mei and Zhai (2008), which is summarizing the impact of a scientific publication. The authors have used language modeling methods—that incorporate features such as authority and proximity extracted from the citation context—to extract sentences that can represent the most influential content of the article. The scientific summarization includes also summarizing document-elements like tables, figures, and algorithms in scientific publications to augment search results and enable the retrieval of these document-elements (Bhatia & Mitra, 2012).

4.3 | Conference analysis

Conference analysis (Effendy et al., 2014; Effendy & Yap, 2016) studies conference categorization (Effendy & Yap, 2016) and relatedness between conferences (Effendy et al., 2014). A case-study approach was adopted by (Effendy et al., 2014) to assess the relatedness measures between conferences in computer science based on the computer science

bibliography DBLP¹². They have shown that the relatedness ranking produced can correlate well with the reputation ranking from CORE (Australian Computer Research and Education conference ranking)¹³.

Both studies help to understand the basis of conference reputation ratings, determine what conferences are related to an area and the classification of conferences into areas.

4.4 | Trend analysis

Trend analysis has received considerable interest in the past few years, because finding a research trend is a key to finding a niche in a particular field of interest, especially for those new to this field. The main goal of trend analysis is to reveal hidden trends within these vast resources, such as research trend evolution and community dynamics (Xia et al., 2017).

Different approaches in the literature dealt with trend analysis using different features such as citation counts, paper content especially keywords, or both of them. We can then categorize these approaches into three categories with respect to the features they have been using: *bibliometrics-based approaches* (An et al., 2017; Hou et al., 2018; Rossetto et al., 2018; Santa Soriano et al., 2018; Zhang & Guan, 2017) that are based on social network analysis, citation and co-citation analysis; *content-based approaches* (Dridi et al., 2019a, 2019b; Weismayer & Pezenka, 2017) that treat entities—essentially keywords—reflecting the paper content (Weismayer & Pezenka, 2017) or dig deeply into the paper content and study the associations between keywords (Dridi et al., 2019a, 2019b); and *hybrid approaches* (Effendy & Yap, 2017; Hoonlor et al., 2013) that combine both citation and content.

The bibliometrics-based approaches rely mainly on citation counts of published papers, and consequently find clues to topic evolution Taskin and Al (2018). For instance, the authors in (An et al., 2017) have considered both backward and forward citations to propose a hidden Markov model to identify temporal patterns of knowledge flows in business method patents. However, (Hou et al., 2018) have used a document co-citation analysis of a subsequent 7,574 articles published in 10 IS journals between 2009 and 2016, including 20,960 references, to study changes in the research topics in the IS domain. Similarly, Rossetto et al. (2018) have used citation and co-citation analysis to understand what are the main theoretical pillars that support the structure of innovation theories and fields. While citation counts may infer the importance of scientific work, they fail to delve into the paper content, which could lead to a more accurate computational history. For this reason, content-based approaches have emerged. For instance, some emerging works Anderson et al. (2012); Hall, Jurafsky, and Manning (2008); and Mortenson and Vidgen (2016) have proposed topic models to study the dynamics of research topics and accordingly the progress of science. While topic models try to extract semantics by capturing document level associations between words, they fail to detect pairwise associations between keywords. To overcome this problem, word embedding techniques have been proposed to conduct a fine-grained content analysis of scientific content. The first work was by He and Chen (2018) and aimed to track the semantic changes of scientific terms over time in the biomedical area. The second is the work proposed by (Dridi et al., 2019a, 2019b) that introduced a temporal word embedding approach for computational history applied to ML publications. The approaches detect the converging keywords that may result in trending keywords by computing the acceleration of similarities between keywords, their rankings and uprankings over successive timespans.

The hybrid approaches use both citation analysis and content analysis to detect research trends. For instance, Hoonlor et al. (2013) have analyzed data on grant proposals, ACM¹⁴ and IEEE¹⁵ publications using sequence mining, bursty word clustering. In like manner, Hou et al. (2018) tracked the evolution of research topics between 2009 and 2016 using the timeline knowledge map through Document-Citation Analysis of articles published in IS journals. They employed dual-map overlays of the IS literature to trace the evolution of the knowledge base of IS research based on scientometric indicators (H-index), citation analysis, and scientific collaboration. In the same context, Effendy and Yap (2017) obtained the computational history using the Microsoft Academic Graph (MAG)¹⁶ dataset. In addition to the citation-basic method, they used a content-based method by leveraging the hierarchical *FoS* (*Field of Study*) given by MAG for each paper to determine the level of interest in any particular research area or topic, and accordingly general publication trends, growth of research areas and the relationship among research areas in Computer Science.

4.5 | Literature analysis

Literature analysis encloses more than one literature feature (Dunne et al., 2012; Li et al., 2006; Liu et al., 2015; Osborne et al., 2013; Tan et al., 2016; Tang, 2016; Tang et al., 2008; Tao et al., 2017). It studies the key nodes of the academic social network such as papers, authors, citations, and corresponding relationships at the same time.

Generally, literature analysis has been shaped around the development of new tools and systems that support the exploration of scholarly data. This has been seen in the case of development of academic search systems that aim to comprehensively search and mine literature (Liu et al., 2015; Tan et al., 2016; Tang, 2016; Tang et al., 2008), such as *ArnetMiner* (Tang et al., 2008), *AMiner* (Tang, 2016) and *CiteSeerX* (Li et al., 2006). Another example of these tools is the study maps that have been built efficiently and thoroughly through topic analysis methods to dig into the underlying principles of a specific paper (Tao et al., 2017). Also, visualization has gained a great interest in literature analysis approaches as it helps to describe, analyze, simulate an academic social network and support community detection and collaboration networks. For instance, *Action Science Explorer* (ASE; Dunne et al., 2012) has been developed to show citation patterns and identify clusters; and *Rexplore* (Osborne et al., 2013) has integrated statistical analysis, semantic technologies, and visual analytics to provide effective support for exploring and making sense of scholarly data.

5 | SCHOLARLY DATA MINING METHODS

Scholarly data mining has been realized with different methods including statistical and empirical analysis, social network analysis, ML techniques, and NLP techniques. In the following, we briefly introduce scholarly data mining methods and we specify the applications they have been used for (see Table 3).

5.1 | Statistical and empirical analysis

Whereas statistics can broadly be defined as the discipline that deals with the collection; organization; analysis; interpretation and presentation of data, empirical analysis refers to the research that uses empirical evidence (Jan-Willem, 2014). Considering that using statistical methods in scientific studies is critical to determining the validity of empirical research, statistical methods, and empirical studies have been widely used together in scientific research. Defined as “research about research,” scholarly data mining has been particularly relying on statistical and empirical analysis mainly for citation analysis (Bornmann & Daniel, 2008; Bornmann & Haunschild, 2018; Cano, 1989; Godin, 2006; McBurney & Novak, 2002; Lv et al., 2011; Martínez-Gómez, 2015; Monroy & Diaz, 2018; Nabout et al., 2018; Pilkington, 2004; Priem & Costello, 2010; R. Shadish, Tolliver, Gray, & Gupta, 1995; Thelwall, 2018; Weller et al., 2011; Acuna et al., 2012). This is justified by the quantitative aspect provided by citation counts; they are measurable indicators of research impact. The quantitative aspect of citation counts has been used from different perspectives. The first perspective concerns the study of the scientific production. For instance, (Lv et al., 2011) have applied statistical analysis to evaluate global scientific production and developing trend of graphene research using the Science Citation Index, the Conference Proceeding Citation Index, and the Derwent Innovation Index database integrated by Thomson Reuters databases. Similarly, Martínez-Gómez (2015) have applied statistical and predictive analyses to 286 scientific works published between 1973 and 2013 in order to study the evolution of the research and the dissemination of knowledge. In the same context, (Acuna et al., 2012) have relied on statistics to track scientific careers and predict scientific success using h-index. They have used a dataset of 3,085 neuroscientists, 57 Drosophila and 151 evolutionary scientists to understand how science develops. However, (Monroy & Diaz, 2018) have used statistics to study the dynamics of scientific production of several countries in terms of papers published. They have analyzed Scopus database to identify dynamical changes that affected global scientific production such as social, political and economic circumstances. The second perspective concerns the study of the broad impact measurements of research beyond science, which is defined as *altmetrics* (Priem & Costello, 2010; Weller et al., 2011). In this context, (Nabout et al., 2018) have studied a dataset of 2,863 papers published in five ecological journals to study the correlation between traditional citation-based indicators and activities on online social media platforms such as Twitter and Mendeley. Similarly, (Bornmann & Haunschild, 2018) have studied Twitter data to measure the impact of science in order to fulfill the demands from governments and funding organizations. In addition to the statistical analysis, empirical and descriptive analyses have been used thoroughly (i) to study the origins of bibliometrics (Godin, 2006) and their purposes (McBurney & Novak, 2002); (ii) to study the citation behavior of scientists (Bornmann & Daniel, 2008; Cano, 1989) and explore the meanings of citations (R. Shadish et al., 1995); and (iii) to investigate the intellectual pillars of the technology management literature and explore differences in the research agendas of worldwide scholars (Pilkington, 2004).

Some other scholarly data mining applications have been realized with statistical and empirical studies, such as trend analysis (Kaempf, Tessenow, Kenett, & Kantelhardt, 2015) and literature analysis (Wu et al., 2014). For trend

TABLE 3 Summary of scholarly data mining methods and corresponding applications

		Citation analysis	Document analysis	Conference analysis	Trend analysis	Literature analysis
Statistical and empirical analysis		✓			✓	✓
Social network analysis		✓	✓	✓	✓	✓
Machine learning	Classification	✓	✓			
	Clustering	✓	✓		✓	
	HMM	✓				
	Ensemble learning		✓			
	Association rules		✓			
	Regression		✓			
	Deep learning		✓			
NLP	Topic models		✓		✓	
	Word embeddings		✓		✓	

analysis, the authors in Kaempf et al. (2015) followed a statistical analysis to measure a topic importance based on page-view time series of Wikipedia articles. They have studied the emergence and life cycle of the emerging Hadoop market. To do so, they have developed *ETOSHA*, an open source software framework for Wikipedia analysis. *ETOSHA* has been used to investigate the changes in the frequency of views of Wikipedia pages. These changes have been used as indicator of collective interests and social trends. More specifically, the statistical analysis follows both qualitative interpretation and quantitative measurement of the network properties of Wikipedia pages. This includes measuring the context sensitive relevance of Wikipedia topics with respect to local and global neighborhood. As a matter of fact, *ETOSHA* has initially relied on exploratory data analysis (Tukey, 1977), namely representation plots, to unveil existing implicit semantic relationships between Wikipedia pages to automatically discover the context the context neighborhoods. Then, based on these neighborhoods, *ETOSHA* has used relative relevance indexes including the time-dependant relevance index that identifies content relevance and public recognition of Wikipedia topics. Unlike Google search that fails to reveal how other keywords with strong relation influence trends, *ETOSHA* has leveraged context neighborhoods from Wikipedia page links to detect emerging trends. However, for literature analysis, descriptive statistics has been used to mine scholarly documents in a large-scale setting and provide scholarly applications, such as citation recommendation, expert recommendation, and collaborator discovery. For instance, (Wu et al., 2014) have built a scholarly big data platform based on CiteseerX system¹⁷ that integrates different services for scholarly data such as information extraction and user/log data analytics. The proposed platform is based on a virtual architecture using a private cloud with the design of the key modules, which included a focused crawler, a crawl-extraction-ingestion workflow, and distributed repositories and databases.

5.2 | Social network analysis

Due to the inherent social network generated from academic activities (such as citations, collaborations, and academic communications)—named academic social network (Mohamad, Lazim, & Rosle, 2018), social network analysis has been proposed to investigate the topologies and dynamics of this network (Xia et al., 2017). Social network analysis is mainly based on the graph theory (Deo, 1974) and aims to describe, analyze, and simulate an academic social network by representing, visualizing and detecting communities in a given network of main scientific entities such as researchers, papers, conferences, and citations. For instance, the citation network has been extensively studied to grasp the relationship among the scientific literatures (Rossetto et al., 2018). As well as, it has been used to detect the most influential nodes for graph summarization problem on citation networks (Shi et al., 2015), and to study the power-law link strength distribution in paper co-citation networks (Zhao & Ye, 2013). The paper network has been used to aid in the exploration of relationships among scientific documents for different purposes. For instance, Dunne et al. (2012) aimed to provide a summary, while identifying key papers, topics and research groups. For this end, they have

developed ASE (Action Science Explorer) and have tested it on a collection of 17,610 Computational Linguistics papers from the ACL Anthology Network. On the other hand, Li et al. (2006) have proposed CiteSeerX, which is a scientific literature library and search engine that automatically crawls and indexes scientific documents in the field of Computer and Information Science. In addition to the paper network that studies scientific papers externally based on their interconnections via citation network, topic network tends to study the papers internally by studying the topics they are discussing from different perspectives. For instance, Tao et al. (2017) have proposed a study map oriented method called RIDP (Reference Injection based Double-Damping Page Rank) that guides researchers to dig into the underlying principles of a specific paper. However, (Kim et al., 2018; Salatino, Osborne, & Motta, 2017) have studied the keywords associated with each paper in order to analyze the dynamics of research topics and visualize information on the growth and change in focus of research fields. The *academic network* has been studied by (Tang et al., 2008) to extract and mine academic social networks. They have provided search services for the academic network by extracting nearly half million research profiles.

Due to the strong relatedness between the aforementioned academic entities, social network analysis has been widely used to understand the large and heterogeneous networks formed by these entities and grasp the big picture of academic fields. Therefore, some works (Hoonlor et al., 2013; Osborne et al., 2013; Rossetto et al., 2018; Tan et al., 2016; Tang, 2016) have provided a systematic modeling approaches to gain a deep understanding of the large academic networks. For instance, (Tang, 2016) have developed *AMiner*—based on a large scholar dataset with more than 130,000,000 researchers' profiles and 100,000,000 papers from multiple publication databases—in order to study the heterogeneous networks formed by authors, papers they have published, and venues in which they were published. In the same context, Osborne et al. (2013) have developed *Rexplore*, which integrates statistical analysis, semantic technologies and visual analytics, to understand the dynamics of research areas, relate authors semantically, and perform fine-grained academic expert search along multiple dimensions.

It is worthy of note that the new learning paradigm *network representation learning* (Zhang, Yin, Zhu, & Zhang, 2018) has recently attracted some works in big scholarly data due to its ability to capture complex relationships across various disciplines such as citation networks. In this respect, Kong et al. (2018) have learned vector representation of papers with network embedding after bridging text information and structural identity with citation network, aiming to develop a robust scientific paper recommendation system. In other respects, (Liu et al., 2019) have proposed a novel model that relies on network representation learning to discover advisor-advisee relationships hidden behind scientific collaboration networks.

Social network analysis has been widely used in scholarly data mining applications including citation analysis (Rossetto et al., 2018; Shi et al., 2015; Zhao & Ye, 2013), literature analysis (Dunne et al., 2012; Li et al., 2006; Osborne et al., 2013; Tan et al., 2016; Tang et al., 2008; Tao et al., 2017), document analysis (Kim et al., 2018; Salatino et al., 2017), conference analysis (Effendy et al., 2014), and trend analysis (Hoonlor et al., 2013; Zhang & Guan, 2017).

5.3 | Machine learning techniques

Scholarly data mining involves different MLML techniques ranging from supervised approaches to unsupervised approaches, namely classification and clustering.

5.3.1 | Classification

In ML, classification refers to the task that requires the use of supervised learning algorithms to learn how to categorize a given set of data into classes (Alpaydin, 2010). Considering that the variety of scholarly data intrigues categorization, classification has been used for different scholarly applications including (i) CCA (Taskin & Al, 2018), where citations were divided into four main categories; citation meaning, citation purpose, citation shape, and citation array; (ii) early identification of sleeping beauties—scientific publications which do not get much cited for several years after being published, but then suddenly start getting cited heavily (Dey et al., 2017); (iii) paper reviewer recommendation (Zhao et al., 2018); (iv) identification of ambiguous author names (Sun et al., 2011); and (v) topic classification (Caragea et al., 2015).

Different classification techniques have been used. For instance, (a) Naïve Bayes Multinomial and Random Forest algorithms have been used for automatic citation sentence classification in a dataset of 423 peer-reviewed articles

associated with 12,881 references and 101,019 sentences, and have performed 90% success rate (Taskin & Al, 2018). (b) *Linear Support Vector Machine*, *Decision Tree*, and *KNN* have been used to classify papers as sleeping beauties or not (Dey et al., 2017). The classifiers have been applied to a dataset of more than 2 million papers published in the Computer Science domain and indexed by Microsoft Academic Search; and have achieved a precision of 73% in identifying sleeping beauties immediately after their year of publications. In a different task, *Support Vector Machine* and *Naïve Bayes Multinomial* have been used for topic classification of research papers (Caragea et al., 2015), applied to a subset sampled from the *CiteSeer*^{x18} digital library. (c) *Logistic Regression* algorithm has been proposed to detect ambiguous author names in crowd-sourced scholarly data extracted from *Scholarometer*¹⁹ (Sun et al., 2011). Two classes of features of a scholar's publications are supplied to the classifier: (i) name variations and citations and (ii) topic consistency; which helped to reach a 75% accuracy. In addition to the existing classification techniques, (Zhao et al., 2018) have proposed a novel classification method named. (d) *Word Mover's Distance Constructive Covering Algorithm (WMD-CCA)* to solve the reviewer recommendation problem as a classification issue. It has been applied to four public datasets and a synthetic dataset from *Baidu Scholar*²⁰ and has shown its effectiveness to solve the reviewer recommendation task as a classification issue and improve the recommendation accuracy.

5.3.2 | Clustering

Clustering, in ML, relies on unsupervised learning algorithms to divide data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups (Alpaydin, 2010). Due to the availability of unlabeled scholarly data, clustering has been used in different scholarly applications. As a matter of fact, clustering has been extensively used for document analysis (Anderson et al., 2012; Lu et al., 2018; Salatino et al., 2018). For instance, the authors in Anderson et al. (2012) and Salatino et al. (2018) have relied on clustering to group research topics. The former study has grouped topics into clusters based on how authors move through them, while the later has detected clusters of topics that exhibit dynamics correlated with the emergence of new research topics. On the other hand, the authors in Lu et al. (2018) have used clustering for document structure analysis; they have generated domain-specific structures based on high-frequency section headers in scientific documents of a domain.

Similarly to its utility in document analysis, clustering has been used in literature analysis to explore internal structure of papers and finding research topics (Liu et al., 2015). Furthermore, for trend analysis, clustering has been utilized to identify features of meta-knowledge (Zhang & Guan, 2017), and to investigate changes over time in the research landscape through clustering bursty keywords (Hoonlor et al., 2013).

Clustering has also served as a useful approach for citation analysis (Dunne et al., 2012; Hou et al., 2018). This has been shown in the case of clusters identification of citation patterns, which helps scholars by providing some forms of automated descriptions for interesting subsets of a document collection.

Different clustering techniques have been explored including *hierarchical clustering* (Anderson et al., 2012), *k-means* (Lu et al., 2018), and *advanced clique percolation method (ACPM)*, which is a novel clustering algorithm developed by Salatino et al. (2018) to detect clusters of topics in the evolutionary networks that exhibit an intensive activity in terms of pace of collaboration.

5.3.3 | Other ML techniques

Other than classification and clustering, different other ML techniques have been used for scholarly data mining. For instance, *Hidden Markov Model (HMM)*—which is defined as a statistical tool that models generative sequences that can be characterized by an underlying process generating an observable sequence (Baum & Petrie, 1966)—has been explored for citation analysis; In An et al. (2017), the authors identified dynamic patterns of knowledge flows driven by business method patents using HMM and patent citation data as an input. They have conducted a case study with the business method patents in 16 sub-classes related to secure transactions. Their analysis revealed that business method patents play increasingly important roles in advancement of business models. The proposed HMM based approach outperformed the existing research on knowledge flows that mainly focuses on static analysis while knowledge flows are intrinsically a dynamic phenomenon. Moreover, for document analysis, *ensemble learning*—which is a ML paradigm where multiple learners are trained to solve the same problem (Polikar, 2006)—and *association rules*—which is a rule-

based ML method, used to find correlations and co-occurrences between data sets (Piatetsky-Shapiro, 1991)—have been used by (Tuarob et al., 2016) to extract algorithm representations in a heterogeneous pool of scholarly documents. The proposed techniques discover pseudo-codes and algorithmic procedures, identify sections in scholarly documents, and use a heuristic that links different algorithm representations referring to the same algorithm together. The proposed techniques cover the limitations of the rule-based method proposed by Bhatia, Mitra, and Giles (2010) for pseudo-code detection, that assumes that each pseudo-code is accompanied by a caption. However, such an assumption is not usually true because of the wide variations in writing styles followed by different journals and authors. However, *regression*—which is defined as a set of statistical processes that attempt to determine the strength and character of the relationship between one dependent variable and a series of other variables—has been used by (Asooja, Bordea, Vulcu, & Buitelaar, 2016) to predict future keyword distribution in order to map scientific topic evolution over time. The prediction is based on historical data of 55k keywords extracted from LREC (Language Resources Evaluation Conference) conference proceedings from 2000 to 2014, and a time series dataset of topics and their popularity has been generated. Unlike existing approaches that simply map the evolution of scientific topics over years, the proposed approach automatically predicts keyword distribution. Consequently, it outperforms the methods based on topic modeling or clustering that require expert knowledge to manually label topics.

Deep learning—which is a form of ML based on artificial neural networks, which are capable to learn from unstructured and unlabeled data without human supervision—has been also used to analyze scientific literature. For instance, (Safder & Hassan, 2018) have designed a deep search system for algorithms from full-text scholarly big data. In contrast to traditional term frequency-inverse document frequency (TF-IDF) based approaches that use frequent terms as in bag of words models, the authors first generated a synopsis of the full-text document and then enriched it with sentences that classify as algorithm-specific metadata from full-text to improve the capabilities of algorithmic-specific searching tasks. These sentences were classified from deep learning based bi-directional long short term memory network model. The proposed model outperformed Support Vector Machine in classifying 37,000 algorithm-specific metadata sentences with 81% accuracy.

5.4 | Natural language processing techniques

Scholarly data mining involves scholarly text mining (Xia et al., 2017), which plays an important role in the analysis of document content. Thus, text mining and NLP techniques have been widely employed to analyze scientific publications.

Current research in scholarly text mining relies mainly on topical analysis. Indeed, *topic model*—which is defined as a statistical model for discovering the abstract topics that occur in a collection of documents (Blei, 2012)—, namely, *Latent Dirichlet Allocation* (Blei, Ng, & Jordan, 2003), has been extensively used either to assign topics to documents based on a given keyword set (document classification; Paul & Girju, 2009; Tang et al., 2008; Weismayer & Pezenka, 2017) or to detect groups of similar documents (document clustering; Anderson et al., 2012; Bakarov, Kutuzov, & Nikishina, 2018; Hall et al., 2008; Tang, 2016).

On the other hand, few recent works have explored *word embeddings* (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)—the newly discovered NLP technique that represents individual words as real-valued vectors in a predefined vector space—to analyze the content of scientific publications. For instance, the authors in He and Chen (2018) have proposed word embeddings to track the semantic changes of scientific terms over time in the biomedical area. Going beyond the existing studies on topic-level analysis, based on topic modeling techniques (Blei et al., 2003), that automatically detect research topics based on textual information and identify their novelty, the proposed approach investigates the impact of the novelty degree of research topics on the growth of scientific knowledge. In Tshitoyan et al. (2019), the authors have relied on word embeddings to capture latent knowledge from materials science literature and predict novel thermoelectric compositions. The authors have shown that—unlike supervised NLP (Friedman, Kra, Yu, Krauthammer, & Rzhetsky, 2001; Swain & Cole, 2016), which requires large hand-labeled datasets for training—word embeddings can be efficiently used to encode materials science knowledge present in the published literature as information-dense vector representations without human labeling or supervision. As a result, without any explicit insertion of chemical knowledge, these embeddings capture complex materials science concepts such as the underlying structure of the periodic table and structure–property relationships in materials. Lately, in Dridi et al. (2019a, 2019b), the authors have leveraged and learned word embeddings across time in order to study the change in pairwise similarities between scientific keywords over time. While topic models

intend to extract semantics by capturing document level associations among words (Anderson et al., 2012; Bakarov et al., 2018; Paul & Girju, 2009), they fail to detect pairwise associations of keywords. This is a considerable limitation since emerging topics often start first by an increasing closeness of keywords that may lead to a merge. Because of this, (Dridi et al., 2019a, 2019b) have used word embeddings to instantly detect converging keywords that may result in trending topics in the area of ML.

6 | PUBLICATIONS AREAS

Scholarly data mining has been applied to a wide range of disciplines ranging from neuroscience (Acuna et al., 2012) to literature studies (Martínez-Gómez, 2015).

Figure 3 presents distribution per application domains of scholarly data mining applications. What can be clearly seen from this figure is that computer & information science is the field with greatest number of publications. This observation is somehow obvious because the majority of scholars investigating scholarly data mining are coming from the area of computer science, where the investigation of their area of expertise is more convenient for interpretation and conclusion drawing. Different sub-areas of computer science have been studied such as artificial intelligence (Alam & Ismail, 2017; Dridi et al., 2019a, 2019b), computational linguistics (Anderson et al., 2012; Asooja et al., 2016; Bakarov et al., 2018; Hall et al., 2008; Paul & Girju, 2009), and big data (Kaempf et al., 2015). Besides, different scholarly data applications have been explored within the area of computer & information science such as document analysis (Alam & Ismail, 2017; Anderson et al., 2012; Salatino et al., 2017, 2018; Bakarov et al., 2018; Paul & Girju, 2009; Tuarob et al., 2016; Caragea et al., 2015), citation analysis (Taskin & Al, 2018; Weller et al., 2011), literature analysis (Dunne et al., 2012; Li et al., 2006; Osborne et al., 2013), conference analysis (Effendy et al., 2014; Effendy & Yap, 2016; Nuzzolese, Gentile, Presutti, & Gangemi, 2016; Tao et al., 2017), and trend analysis (Asooja et al., 2016; Dey et al., 2017; Dridi et al., 2019a, 2019b; Effendy & Yap, 2017; Kaempf et al., 2015).

Based on Figure 3, the second major part of studies has applied scholarly data mining to multidisciplinary area where more than one discipline has been studied (Boyack, van Eck, Colavizza, & Waltman, 2017; Godin, 2006; McBurney & Novak, 2002; Monroy & Diaz, 2018; Sun et al., 2011; Tang et al., 2008; Thelwall, 2018; Wu et al., 2014; Zhang & Guan, 2017).

Economy & Business area has also attracted the attention of scholars in scholarly data mining. They have investigated different aspects such as relations and economy (Santa Soriano et al., 2018); innovation and entrepreneurial ecosystem (Zhang & Guan, 2017); business (Rossetto et al., 2018) and business model innovation (An et al., 2017); and marketing and tourism (Weismayer & Pezenka, 2017).

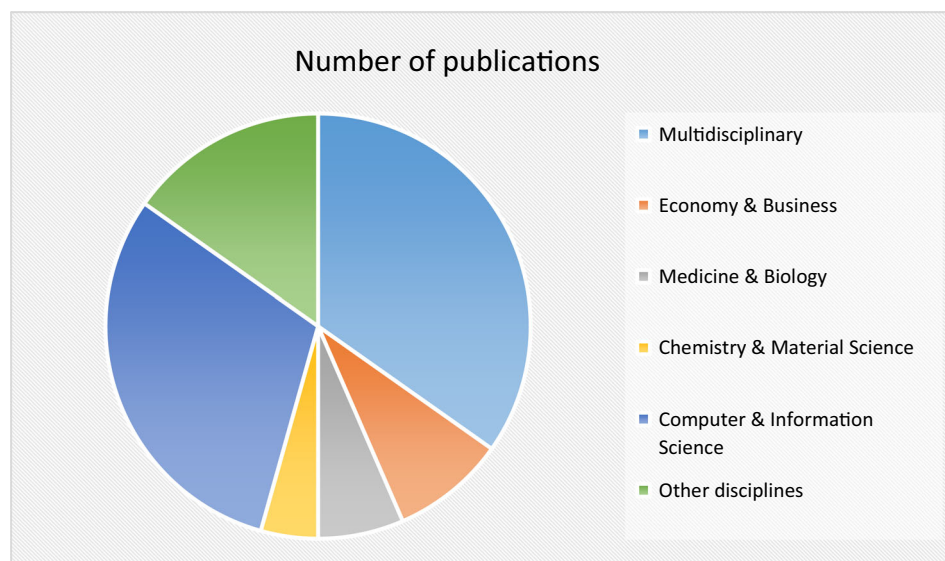


FIGURE 3 Distribution per application domains of scholarly data mining publications

The area of Medicine & Biology has been also studied through scholarly data mining. The existing studies (Liu et al., 2015; Rexha et al., 2018; Shardlow et al., 2018) have studied the document content of biomedical publications to extract knowledge from scientific literature and discover underlying interesting research topics. Similarly, the area of Chemistry & Material Science has seen the application of scholarly data mining for citation analysis (Lv et al., 2011) and trend analysis (Tshitoyan et al., 2019).

The other disciplines that have been involved in scholarly data analysis are as following: neuroscience (Acuna et al., 2012); ecology (Nabout et al., 2018), social science (Priem & Costello, 2010); education (Paul & Girju, 2009; Weber & Gunawardena, 2011); and translation and interpreting studies (Martínez-Gómez, 2015). The main scholarly data mining application applied to these area is citation analysis.

7 | DISCUSSION

The majority of the reviewed studies demonstrated that scholarly data mining can be effectively applied to a wide range of scholarly applications to learn about the structure and the dynamics of science. Our investigation suggested that scholarly data mining can be utilized to address different scholarly applications and provide better services to scholars and researchers such as academic recommendation, scientific text summarization and research trend prediction. This is significant because these services can potentially accelerate science and facilitate the identification of fundamental mechanisms responsible for scientific discovery (Fortunato et al., 2018). However, despite its notable advantages, scholarly data mining also brings challenges.

Collecting and processing scholarly data. Given the size of scholarly data and the complexity of its structure, collecting and processing it have become increasingly challenging. In fact, big scholarly data is characterized by the 5V feature (Xia et al., 2017). Veracity and variety make from scholarly data a complex system, where ambiguity is present and different entities are involved. The complexity of this system makes scholarly data management a challenging task.

Insufficiency of metrics to evaluate the research quality. Evaluating the research quality is an essential component of research assessment, and outcomes of such evaluations can help in institutional research strategies such as funding and recruitment. However, there are little standards to measure scientific performance objectively; metrics alone have been unable to achieve the task of predicting scientific impact and assessing research quality (Sahel, 2011). Improving existing research evaluation practices is, therefore, an urge.

Lack of gold standards. Some scholarly applications require gold standards to evaluate their outcomes such as trend analysis. However, there is no standards to use to perform comparative studies or to validate the obtained results. Most of existing studies on trend analysis have relied on descriptive analysis to present their studies, while the application of ML techniques requires standards to assess the quality of the proposed techniques, which makes this task challenging. In Dridi et al. (2019a, 2019b), the authors have attempted to build a gold standard relying on Google Trends hits. Their approach highlights the importance of promoting gold standards for the matter of trend analysis because this scholarly application represents an important direction toward knowledge discovery and the study of dynamics of science.

Limitations. There are some limitations in this systematic review that we wish to acknowledge. First, we included many scholarly data-related keywords in the search queries to cover as many related publications as possible. However, this process might miss some studies that failed to mention such terms. Second, we removed workshop articles before screening eligible publications aiming to keep high-quality research studies. However, this might miss some premature promising work. Third, we removed some studies that focused on bibliometrics analysis but neglected to use technical analysis techniques; they have used descriptive analysis. Finally, more details on trend analysis studies were excluded in order to keep the review of the applications as balanced as possible; which could be considered in future reviews about knowledge discovery and trend analysis.

8 | CONCLUSIONS

With the increasing availability of digital data on scientific publications coupled with the advances in computational technologies, exploring, studying, and analyzing vast amounts of scholarly data have become increasingly challenging. Scholarly data mining has thus offered opportunities to explore the structure and evolution of science. It helps, indeed, to curate and derive useful insights from scholarly data, which provides better scholarly services to scholars and

researchers. Moreover, scholarly data analysis can help governments and institutions in several decision making processes such as policy making for funding, identifying research gaps and speculating upcoming research areas. Since scholarly data has provided data scientists a fertile ground to explore, more investigations are needed to comprehensively study the topic of scholarly data mining.

To help with the investigation pursuit, we have provided a systematic review about scholarly data mining applications, performing a literature-based analysis, and description of more than 70 research papers; indicating the interest in the field from different perspectives such as the type of the techniques used and the discipline investigated. The value proposition of scholarly data mining is that with a deeper understanding of the structure of science, we can more effectively address scientific discovery problems and develop tools and policies that have the potential to accelerate science.

ACKNOWLEDGMENT

This work is supported by the Faculty of Computing, Engineering and Built Environment, Birmingham City University, through a Full Bursary Ph.D. Scholarship

CONFLICT OF INTEREST


The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Amna Dridi: Conceptualization, Data curation, Methodology, Writing-original draft. **Mohamed Medhat Gaber:** Methodology, Supervision, Writing-review & editing. **R. Muhammad Atif Azad:** Supervision. **Jagdev Bhogal:** Supervision.

ORCID

Amna Dridi  <https://orcid.org/0000-0002-0185-103X>

Mohamed Medhat Gaber  <https://orcid.org/0000-0003-0339-4474>

ENDNOTES

- ¹ <https://login.webofknowledge.com/>
- ² <https://scholar.google.com/>
- ³ <https://dblp.uni-trier.de/>
- ⁴ <https://www.sciencedirect.com/>
- ⁵ <https://dl.acm.org/>
- ⁶ <https://ieeexplore.ieee.org/Xplore/>
- ⁷ <https://link.springer.com/journal/11192>
- ⁸ <https://www.jcdl.org/>
- ⁹ https://www.natureindex.com/country-outputs/generate/All/global/All/n_article
- ¹⁰ <http://uis.unesco.org/en/news/new-uis-data-sdg-9-5-research-and-development>
- ¹¹ <https://www.mendeley.com/>
- ¹² <https://dblp.uni-trier.de/>
- ¹³ <https://www.core.edu.au/>
- ¹⁴ <https://dl.acm.org/>
- ¹⁵ <https://ieeexplore.ieee.org/>
- ¹⁶ <https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/>
- ¹⁷ <http://citeseerx.ist.psu.edu/>
- ¹⁸ <http://citeseerx.ist.psu.edu/index>
- ¹⁹ <https://scholarometer.indiana.edu/>
- ²⁰ <https://scolary.com/tools/baidu-scholar>

RELATED WIREs ARTICLES

[Emerging directions in predictive text mining](#)

REFERENCES

- Acuna, D., Allesina, S., & Kording, K. (2012). Future impact: Predicting scientific success. *Nature*, 489, 201–202.
- Alam, M. M., & Ismail, M. A. (2017). Rtrs: A recommender system for academic researchers. *Scientometrics*, 113, 1325–1348.
- Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed.). USA: The MIT Press.
- An, Y., Han, M., & Park, Y. (2017). Identifying dynamic knowledge flow patterns of business method patents with a hidden markov model. *Scientometrics*, 113, 783–802.
- Anderson, A., McFarland, D. and Jurafsky, D. (2012) *Towards A Computational History of the ACL: 1980-2008*. ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, Jeju Island, Korea. pp. 13–21.
- Asooja, K., Bordea, G., Vulcu, G. and Buitelaar, P. (2016) *Forecasting Emerging Trends from Scientific Literature*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Portorož, Slovenia. pp. 417–420.
- Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., & Xia, F. (2019). Scientific paper recommendation: A survey. *IEEE Access*, 7, 9324–9339.
- Bakarov, A., Kutuzov, A. and Nikishina, I. (2018). *Russian Computational Linguistics: Topical Structure in 2007-2017 Conference Papers*. Computational linguistics and intellectual technologies: Proceedings of the International Conference “Dialogue 2018.”, Moscow, Russia. pp. 1–13.
- Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematical Statistics*, 37, 1554–1563.
- Bhatia, S., & Mitra, P. (2012). Summarizing figures, tables, and algorithms in scientific publications to augment search results. *ACM Transactions on Information Systems*, 30(3), 13–24.
- Bhatia, S., Mitra, P. and Giles, C. L. (2010). *Finding Algorithms in Scientific Articles*. World Wide Web Conference, New York, NY, USA: Association for Computing Machinery. pp. 1061–1062.
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55, 77–84.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64, 45–80.
- Bornmann, L., & Haunschild, R. (2018). Allegation of scientific misconduct increases twitter attention. *Scientometrics*, 115, 1097–1100.
- Boyack, K. W., Smith, C., & Klavans, R. (2018). Toward predicting research proposal success. *Scientometrics*, 114, 449–461.
- Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2017). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, 12, 59–73.
- Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, 40, 284–290.
- Caragea, C., Bulgarov, F. and Mihalcea, R. (2015). *Co-training for Topic Classification of Scholarly Data*. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal. pp. 2357–2366.
- Deo, N. (1974). *Graph theory with applications to engineering and computer science (Prentice Hall series in automatic computation)*. Upper Saddle River, NJ: Prentice-Hall, Inc.
- Dey, R., Roy, A., Chakraborty, T., & Ghosh, S. (2017). Sleeping beauties in computer science: Characterization and early identification. *Scientometrics*, 113, 1645–1663.
- Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *The Journal of the Association for Information Science and Technology*, 65, 1820–1833.
- Dridi, A., Gaber, M. M., Azad, R. M. A. and Bhogal, J. (2019a) *Deephist: Towards a Deep Learning-based Computational History of Trends in the Nips*. International Joint Conference in Neural Networks, Budapest, Hungary. pp. 1–8.
- Dridi, A., Gaber, M. M., Azad, R. M. A., & Bhogal, J. (2019b). Leap2trend: A temporal word embedding approach for instant detection of emerging scientific trends. *IEEE Access*, 7, 1–1.
- Dunne, C., Shneiderman, B., Gove, R., Klavans, J., & Dorr, B. (2012). Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the Association for Information Science & Technology*, 63, 2351–2369.
- Effendy, S., Jahja, I. and Yap, R. H. (2014). *Relatedness Measures Between Conferences in Computer Science: A Preliminary Study Based on DBLP*. Proceedings of the 23rd International Conference on World Wide Web, WWW'14 Companion, Seoul, Korea. pp. 1215–1220.
- Effendy, S. and Yap, R. H. (2017). *Analysing Trends in Computer Science Research: A Preliminary Study Using The Microsoft Academic Graph*. Proceedings of the 26th International Conference on World Wide Web Companion, WWW'17 Companion, Perth, Australia. pp. 1245–1250.
- Effendy, S., & Yap, R. H. C. (2016). The problem of categorizing conferences in computer science. In N. Fuhr, L. Kovács, T. Risse, & W. Nejdl (Eds.), *Research and advanced technology for digital libraries* (pp. 447–450). Cham: Springer International Publishing.
- Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. (2012). A brief survey of automatic methods for author name disambiguation. *SIGMOD Record*, 41, 15–26.
- Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... Barabási, A.-L. (2018). Science of science. *Science*, 359, eaao0185.
- Friedman, C., Kra, P., Yu, H., Krauthammer, M., & Rzhetsky, A. (2001). GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17, S74–S82.
- Gleason, H. A. (1961). *An introduction to descriptive linguistics*. New York: Holt, Rinehart and Winston rev. ed. edn.
- Godin, B. (2006). On the origins of bibliometrics. *Scientometrics*, 68, 109–133.

- Hall, D., Jurafsky, D. and Manning, C. D. (2008) *Studying the History of Ideas Using Topic Models*. Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP'08. Honolulu, Hawaii: Association for Computational Linguistics. pp. 363–371.
- He, J., & Chen, C. (2018). Predictive effects of novelty measured by temporal embeddings on the growth of scientific literature. *Frontiers in Research Metrics and Analytics*, 3, 3.
- Heffernan, K., & Teufel, S. (2018). Identifying problems and solutions in scientific text. *Scientometrics*, 116, 1367–1382.
- Hoonlor, A., Szymanski, B. K., & Zaki, M. J. (2013). Trends in computer science research. *Communications of the ACM*, 56, 74–83.
- Hou, J., Yang, X., & Chen, C. (2018). Emerging trends and new developments in information science: A document co-citation analysis (2009–2016). *Scientometrics*, 115, 869–892.
- Jan-Willem, R. (2014) *Philosophy of statistics*, Stanford, CA, USA: Metaphysics Research Lab, Centre for the Study of Language and Information, Stanford University.
- Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics*, 8, 197–211.
- Jha, R., Abu-Jbara, A. and Radev, D. (2013) *A System for Summarizing Scientific Topics Starting From Keywords*. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (volume 2: Short papers), Sofia, Bulgaria. pp. 572–577.
- Joseph, A. K., & Thelwall Mike, K. K. (2017). Do mendeley reader counts reflect the scholarly impact of conference papers? An investigation of computer science and engineering. *Scientometrics*, 112, 573–581.
- Kaempf, M., Tessenow, E., Kenett, D., & Kantelhardt, J. (2015). The detection of emerging trends using wikipedia traffic data and context networks. *PLoS One*, 10, e0141892.
- Kaisler, S., Armour, F., Espinosa, J. A. and Money, W. (2013) *Big Data: Issues and Challenges Moving Forward*. 2013 46th Hawaii International Conference on System Sciences, Wailea, Hawaii, USA. pp. 995–1004.
- Khan, S., Liu, X., Shakil, K., & Alam, M. (2017). A survey on scholarly data: From big data perspective. *Information Processing & Management*, 53, 923–944.
- Kim, S., Hansen, D., & Helps, R. (2018). Computing research in the academy: Insights from theses and dissertations. *Scientometrics*, 114, 135–158.
- Kitchenham, B. (2004). *Procedures for performing systematic reviews* (p. 33). Keele, UK: Keele University.
- Kong, X., Mao, M., Wang, W., Liu, J., & Xu, B. (2018). Voprec: Vector representation learning of papers with text information and structural identity for recommendation. *IEEE Transactions on Emerging Topics in Computing*, 1, 1–1.
- Li, H., Councill, I., Lee, W. and Giles, C. (2006) *CiteSeerx: An Architecture and Web Service Design for An Academic Document Search Engine*. Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland. pp. 883–884.
- Light, R. P., Polley, D. E., & Börner, K. (2014). Open data and open code for big science of science studies. *Scientometrics*, 101, 1535–1551.
- Liu, J., Tang, T., Wang, W., Xu, B., Kong, X., & Xia, F. (2018). A survey of scholarly data visualization. *IEEE Access*, 6, 19205–19221.
- Liu, J., Xia, F., Wang, L., Xu, B., Kong, X., Tong, H., & King, I. (2019). Shifu2: A network representation learning based model for advisor-advisee relationship mining. *IEEE Transactions on Knowledge and Data Engineering*, 1–1.
- Liu, Y., Huang, Z., Yan, Y. and Chen, Y. (2015). *Science Navigation Map: An Interactive Data Mining Tool for Literature Analysis*. Proceedings of the 24th International Conference on World Wide Web, WWW'15 Companion, Florence, Italy. pp. 591–596.
- Lu, W., Huang, Y., Bu, Y., & Cheng, Q. (2018). Functional structure identification of scientific documents in computer science. *Scientometrics*, 115, 463–486.
- Lv, P. H., Wang, G.-F., Wan, Y., Liu, J., Liu, Q., & Ma, F.-C. (2011). Bibliometric trend analysis on global graphene research. *Scientometrics*, 88, 399–419.
- Maflahi, N., & Thelwall, M. (2018). How quickly do publications get read? The evolution of mendeley reader counts for new articles. *Journal of the Association for Information Science and Technology*, 69, 158–167.
- Martínez-Gómez, A. (2015). Bibliometrics as a tool to map uncharted territory: A study on non-professional interpreting. *Perspectives*, 23, 1–18.
- McBurney, M. K., & Novak, P. L. (2002) *What is Bibliometrics and Why Should You Care?* IEEE International Professional Communication Conference, Portland, OR, USA. pp. 108–114.
- Mei, Q. and Zhai, C. (2008) *Generating Impact-Based Summaries for Scientific Literature*. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, USA. pp. 816–824.
- Merediz-Solà, I., & Bariviera, A. F. (2019). A bibliometric analysis of bitcoin scientific production. *Research in International Business and Finance*, 50, 294–305.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013) *Distributed Representations of Words and Phrases and Their Compositionality*. Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5–8, 2013, Lake Tahoe, Nevada, United States. pp. 3111–3119.
- Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, 246, 1–19.
- Mohamad, M., Lazim, Y., & Rosle, S. (2018). Academic social network sites: Opportunities and challenges. *International Journal of Engineering & Technology*, 7, 133.
- Monroy, S. E., & Diaz, H. (2018). Time series-based bibliometric analysis of the dynamics of scientific production. *Scientometrics*, 115, 1139–1159.
- Mortenson, M. J., & Vidgen, R. (2016). A computational literature review of the technology acceptance model. *International Journal of Information Management*, 36, 1248–1259.
- Nabout, J. C., Teresa, F. B., Machado, K. B., do Prado, V. H. M., Bini, L. M., & Diniz-Filho, J. A. F. (2018). Do traditional scientometric indicators predict social media activity on scientific knowledge? An analysis of the ecological literature. *Scientometrics*, 115, 1007–1015.

- Nuzzolese, A. G., Gentile, A. L., Presutti, V. and Gangemi, A. (2016) *Conference Linked Data: The Scholarlydata Project*. The Semantic Web—ISWC 2016—15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II. pp. 150–158.
- Osborne, F., Motta, E., & Mulholland, P. (2013). Exploring scholarly data with rexplore. In H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, et al. (Eds.), *The semantic web—ISWC 2013* (pp. 460–477). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Paul, M. and Girju, R. (2009). *Topic Modeling of Research Fields: An Interdisciplinary Perspective*. International Conference Recent Advances in Natural Language Processing, RANLP, Borovets, Bulgaria. pp. 337–342.
- Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. In *Knowledge discovery in databases* (pp. 229–248). Cambridge: AAAI Press.
- Pilkington, A. (2004) *Defining Technology Management: A Citation/co-citation study*. 2004 IEEE International Engineering Management Conference (IEEE Cat. No.04CH37574), Singapore. Vol. 1, pp. 337–341.
- Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6, 21–45.
- Priem, J. and Costello, K. L. (2010) *How and Why Scholars Cite on Twitter*. Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem of ASIS&T'10, 75. Vol. 47, pp. 1–75. Pittsburgh, Pennsylvania: American Society for Information Science.
- Qazvinian, V. and Radev, D. R. (2008) *Scientific Paper Summarization Using Citation Summary Networks*. Proceedings of the 22nd International Conference on Computational Linguistics—Volume 1, COLING'08. pp. 689–696. Manchester, United Kingdom: Association for Computational Linguistics.
- Rexha, A., Kröll, M., Ziak, H., & Kern, R. (2018). Authorship identification of documents with high content similarity. *Scientometrics*, 115, 223–237.
- Rossetto, D. E., Bernardes, R. C., Borini, F. M., & Gattaz, C. C. (2018). Structure and evolution of innovation research in the last 60 years: Review and future trends in the field of business through the citations and co-citations analysis. *Scientometrics*, 115, 1329–1363.
- Safder, I. and Hassan, S.-U. (2018) *Ds4a: Deep Search System For Algorithms From Full-Text Scholarly Big Data*. 2018 IEEE International Conference on Data Mining Workshop (ICDMW), Singapore.
- Sahel, J.-A. (2011). Quality versus quantity: Assessing individual research performance. *Science Translational Medicine*, 3, 84cm13–84cm13.
- Salatino, A. A., Osborne, F. and Motta, E. (2018) *AUGUR: Forecasting the Emergence of New Research Topics*. Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA. pp. 303–312.
- Salatino, A. A., Osborne, F., & Motta, E. (2017). How are topics born? Understanding the research dynamics preceding the emergence of new areas. *PeerJ Computer Science*, 3, e119.
- Santa Soriano, A., Lorenzo Álvarez, C., & Torres Valdés, R. M. (2018). Bibliometric analysis to identify an emerging research area: Public relations intelligence—A challenge to strengthen technological observatories in the network society. *Scientometrics*, 115, 1591–1614.
- Shadish, R. W., Tolliver, D., Gray, M., & Gupta, S. K. S. (1995) Author judgements about works they cite: Three studies from psychology journals. *Social Studies of Science*, 25, 477–498.
- Shardlow, M., Batista-Navarro, R., Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2018). Identification of research hypotheses and new knowledge from scientific literature. *BMC Medical Informatics and Decision Making*, 18, 46.
- Shi, L., Tong, H., Tang, J., & Lin, C. (2015). Vegas: Visual influence graph summarization on citation networks. *IEEE Transactions on Knowledge and Data Engineering*, 27, 3417–3431.
- Sun, X., Kaur, J., Possamai, L. and Menczer, F. (2011) *Detecting Ambiguous Author Names in Crowdsourced Scholarly Data*. PASAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA. 568–571.
- Swain, M., & Cole, J. (2016). Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10), 1894–1904.
- Tan, Z., Liu, C., Mao, Y., Guo, Y., Shen, J. and Wang, X. (2016) *Acemap: A Novel Approach Towards Displaying Relationship Among Academic Literatures*. Proceedings of the 25th International Conference Companion on World Wide Web, WWW'16 Companion, Montreal, Canada. pp. 437–442.
- Tang, J. (2016) *Aminer: Toward Understanding Big Scholar Data*. Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, WSDM'16, San Francisco, California, USA. pp. 467–467.
- Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. and Su, Z. (2008) *Arnetminer: Extraction and Mining of Academic Social Networks*. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'08, Las Vegas, Nevada, USA. pp. 990–998.
- Tao, S., Wang, X., Huang, W., Chen, W., Wang, T. and Lei, K. (2017) *From Citation Network to Study Map: A Novel Model to Reorganize Academic Literatures*. Proceedings of the 26th International Conference on World Wide Web Companion, WWW'17 Companion, Perth, Australia. pp. 1225–1232.
- Taskin, Z., & Al, U. (2018). A content-based citation analysis study based on text categorization. *Scientometrics*, 114, 335–357.
- Thelwall, M. (2018). Differences between journals and years in the proportions of students, researchers and faculty registering mendeley articles. *Scientometrics*, 115, 717–729.
- Trujillo, C. M., & Long, T. M. (2018). Document co-citation analysis to enhance transdisciplinary research. *Science Advances*, 4, e1701130.
- Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., ... Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571, 95–98.

- Tuarob, S., Bhatia, S., Mitra, P., & Giles, C. L. (2016). Algorithmseer: A system for extracting and searching for algorithms in scholarly big data. *IEEE Transactions on Big Data*, 2, 3–17.
- Tukey, J. W. (1977). *Exploratory data analysis. Behavioral science: Quantitative methods*. Reading, MA: Addison-Wesley.
- Weber, R., & Gunawardena, S. (2011). Representing scientific knowledge. In *Cognition and exploratory learning in the digital age* (pp. 279–283). Rio de Janeiro, Brazil: Springer.
- Weismayer, C., & Pezenka, I. (2017). Identifying emerging research fields: A longitudinal latent semantic keyword analysis. *Scientometrics*, 113, 1757–1785.
- Weller, K., Dröge, E. and Puschmann, C. (2011) *Citation Analysis in Twitter*. Approaches for Defining and Measuring Information Flows Within Tweets During Scientific Conferences. In Sharp MSM2011, 1st Workshop on Making Sense of Microposts, Heraklion, Greece. pp. 1–12.
- Wu, Z., Wu, J., Khabsa, M., Williams, K., Chen, H., Huang, W., Tuarob, S., Choudhury, S. R., Ororbia, A., Mitra, P. and Giles, C. L. (2014). *Towards Building a Scholarly Big Data Platform: Challenges, Lessons and Opportunities*. IEEE/ACM Joint Conference on Digital Libraries, LONDON, United Kingdom, pp. 117–126.
- Xia, F., Wang, W., Bekele, T. M., & Liu, H. (2017). Big scholarly data: A survey. *IEEE Transactions on Big Data*, 3, 18–35.
- Zhang, C., & Guan, J. (2017). How to identify metaknowledge trends and features in a certain research field? Evidences from innovation and entrepreneurial ecosystem. *Scientometrics*, 113, 1177–1197.
- Zhang, D., Yin, J., Zhu, X. and Zhang, C. (2018) Network representation learning: A survey. *CoRR*, abs/1801.05852.
- Zhao, S., & Ye, F. (2013). Power-law link strength distribution in paper cocitation networks. *Journal of the American Society for Information Science and Technology*, 64, 1480–1489.
- Zhao, S., Zhang, D., Duan, Z., Chen, J., Zhang, Y.-P., & Tang, J. (2018). A novel classification method for paper-reviewer recommendation. *Scientometrics*, 115(1), 1–21.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Dridi A, Gaber MM, Azad RMA, Bhogal J. Scholarly data mining: A systematic review of its applications. *WIREs Data Mining Knowl Discov*. 2021;11:e1395. <https://doi.org/10.1002/widm.1395>