information

MDPI

Article

An Overview on Evaluating and Predicting Scholarly Article Impact

Xiaomei Bai [1,2], Hui Liu [1,*], Fuli Zhang [3], Zhaolong Ning [1], Xiangjie Kong [1], Ivan Lee [4] and Feng Xia [1]

1 School of Software, Dalian University of Technology, Dalian 116620, China; xiaomeibai@outlook.com (X.B.); Zhaolongning@dlut.edu.cn (Z.N.); xjkong@ieee.org (X.K.); f.xia@ieee.org (F.X.)
2 Computing Center, Anshan Normal University, Anshan 114007, China
3 Library, Anshan Normal University, Anshan 114007, China; zfuli@outlook.com
4 School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, SA 5095, Australia; ivan.lee@unisa.edu.au
* Correspondence: liuhui1126@dlut.edu.cn; Tel.: +86-411-6227-4391

Academic Editor: David Bawden
Received: 23 May 2017; Accepted: 23 June 2017; Published: 25 June 2017

Abstract: Scholarly article impact reflects the significance of academic output recognised by academic peers, and it often plays a crucial role in assessing the scientific achievements of researchers, teams, institutions and countries. It is also used for addressing various needs in the academic and scientific arena, such as recruitment decisions, promotions, and funding allocations. This article provides a comprehensive review of recent progresses related to article impact assessment and prediction. The review starts by sharing some insight into the article impact research and outlines current research status. Some core methods and recent progress are presented to outline how article impact metrics and prediction have evolved to consider integrating multiple networks. Key techniques, including statistical analysis, machine learning, data mining and network science, are discussed. In particular, we highlight important applications of each technique in article impact research. Subsequently, we discuss the open issues and challenges of article impact research. At the same time, this review points out some important research directions, including article impact evaluation by considering Conflict of Interest, time and location information, various distributions of scholarly entities, and rising stars.

Keywords: scholarly big data; article impact; machine learning; data mining

1. Introduction

Scholarly impact acts as one of the strongest currencies in the academia, and it is frequently measured in terms of citations of research articles. Citations indicate the impact of scholars, articles, journals, institutions, and other scholarly entities [1]. The influence of an article is often quantified as an index, which represents its contributions for improving research finding by other scholars [2].

Researching the impact of scientific articles mainly focuses on two interrelated questions: how to assess the past impact of an article, and how to accurately predict its future impact? The study of article impact is important for evaluating the impact of individual scientists, journals, teams, institutions, and even for countries. It is also crucial for addressing the following fundamental problems, such as rewards, funding allocation, promotion, and recruitment decisions. Evaluating and predicting article impact have attracted great attention in the academic and scientific arena over the past decades. The changes occur from one dimension to multiple dimensions, from unstructured metrics to structured metrics (Figure 1). Citations [3] are a popular indicator to measure article impact. However, it only focuses on the perspective of single dimension. Altmetrics [4] provide information on downloads, views, shares, and citations to assess article impact from a multidimensional perspective. PageRank has

been introduced to evaluate article impact [5], which can be viewed as a milestone in impact research. It has shown a structured method to quantify article impact. Meanwhile, in order to objectively evaluate article impact and accurately predict its future impact, machine learning and data mining techniques play crucial roles, such as mining the important characters of scholarly networks and optimizing the performance of algorithms [6].

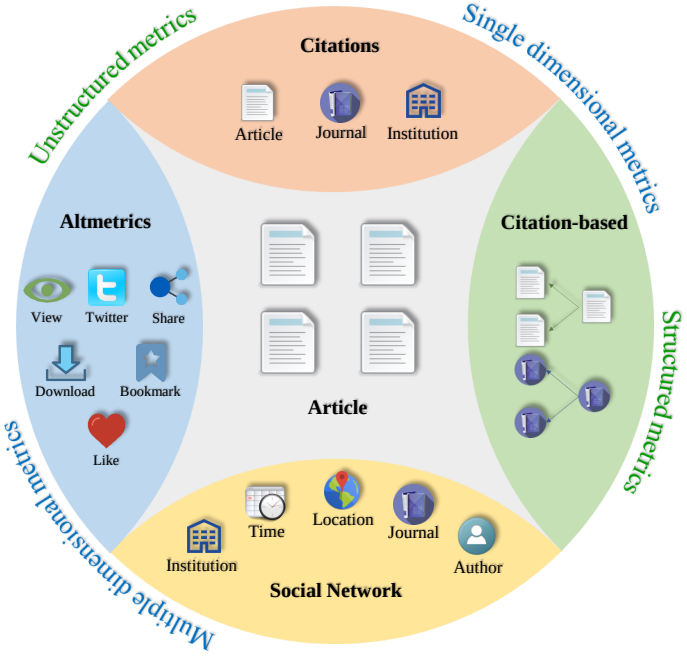

Figure 1. Methods of evaluating and predicting article impact.

What drives the rapid development in evaluating and predicting article impact? The past decade has witnessed the rapid growth in the ability of network platforms to gather and transport a large number of academic data, i.e., a phenomenon usually referred to as "Big Scholarly Data" (see Figure 2). Different networks with various scholarly entities and their relationships can be observed from Figure 2. Scholars can collect such data to solve the problems of scholarly impact evaluation. They can obtain useful insights from such datasets by leveraging statistical analysis, machine learning, data mining, and network science techniques. The academic data with exponential growth become essential to develop the scholarly impact metrics. The metrics combine the statistical and computational considerations. However, one problem cannot be ignored. That is these datasets are personalized. SCOPUS contains abstracts and citations of journal papers. Web of Science offers online scientific citations by Thomson Reuters. PubMed includes more than 23 million citations for biomedical literature. CiteULike allows users to search and share scholarly papers. Mendeley can not only be used to manage references, but also it is an academic social platform. Digital Bibliography & Library Project (DBLP) shows publications of journals and conferences, not including citation information. Microsoft Academic Graph (MAG) includes heterogeneous information with publication records, authors, institutions, journals, conferences, fields of study and citation relationships. In these raw data, the most prominent problems are loss and incompletion of data, which probably will result in poor performance of evaluation and prediction to some extent. Data cleaning and supplement are necessary for accurately capturing the evaluative and predictive results. Besides, these data sets can be jointly investigated to complement one another. For example, DBLP does not include citation information, but it has an effective mechanism to process name disambiguation. Integrating DBLP dataset and the citation information of SCOPUS can meet the needs for some scholarly analysis.

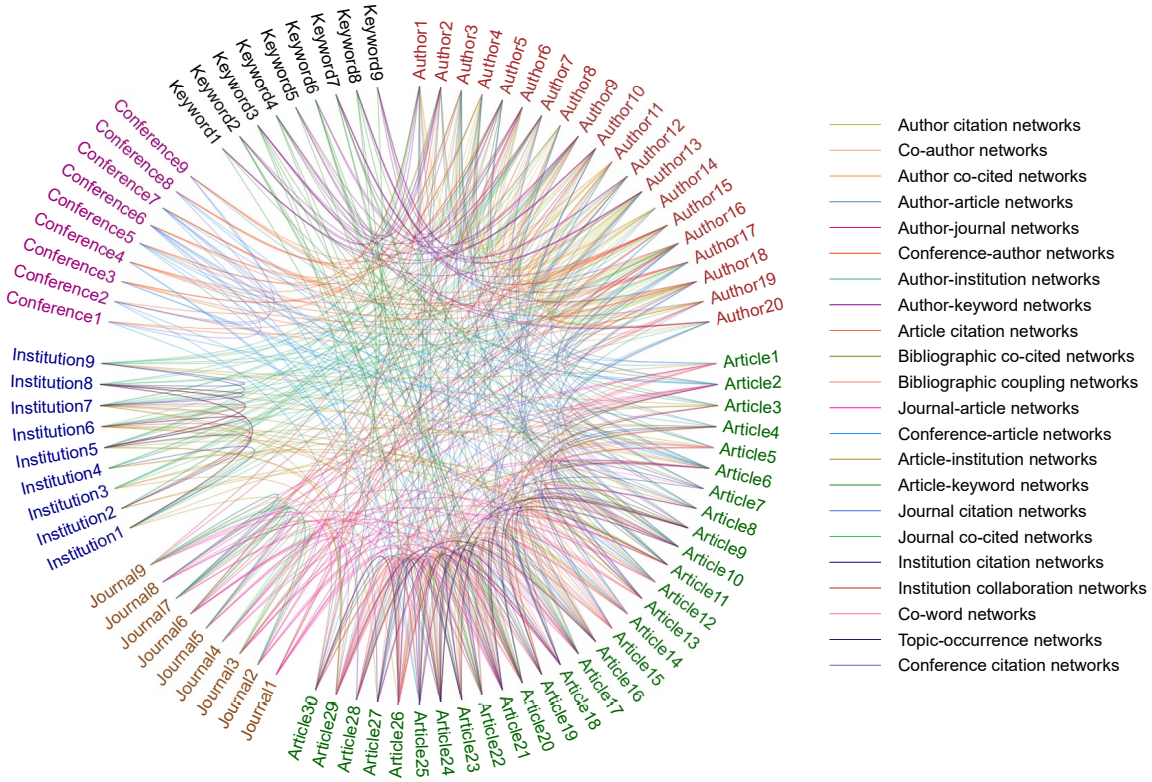

Figure 2. Characterizing scholarly networks.

2. Key Techniques

In this section, we discuss four crucial techniques for evaluating and predicting article impact including statistical analysis, machine learning, data mining and network science techniques.

2.1. Statistical Methods

Statistical methods cover the process of collecting, dealing with, analysing and explaining data. Researchers can gain science knowledge from data through statistics. Statistical analysis is mainly interested in analyzing and understanding data, including regression models, variable selection, principal component analysis, factor analysis, cluster analysis, canonical correlation analysis, time series analysis, probability and density estimation, and so on [7,8]. Regression models contain single variable regression and multiple variables regression. Statistical techniques are usually used in most fields of nature and social science, such as finance, medical treatment, industry etc. In researching scholarly impact fields the benefits of statistical techniques are as follows:

- Pre-process data;
- Optimize parameters, for instance, multiple variables linear regression;
- Select features and improve models for scholarly evaluation and prediction. For example, use the massive existing statistics to estimate a probability density function;
- Analyse scholarly data to obtain statistical data, and then use statistical model to predict the trends of impact, top scholars, top articles, etc.

To explore the relationships between citations and citation distance, statistical analysis such as grouping and clustering may be applied. When we use group analysis technique, an appropriate segmentation point of citation distance is crucial. Usually, the selection of segmentation point depends on the experimental data. An advantage of group analysis technique is relatively easy to deal with data. However, due to the compulsory group, the disadvantage of group analysis is obvious.

Clustering analysis technique remedies the drawback of compulsory group such as Density-based Spatial Clustering of Applications with Noise algorithm can be used to analyze the relationships between citations and citation distance based on the density of institutions.

In short, statistical analysis can be used to pre-process data, capture intermediate results or gain final results in evaluating and predicting research of scholarly impact. For example, a multivariate linear regression was used to estimate the parameters of three algorithms for evaluating the impact of papers [9]. Based on principal component analysis, a factor analysis was used to explore the main components in bibliometric and altmetric indicators [10]. Especially when scholars predict citations of a paper or a scholar's H-index, they usually give an estimative range instead of a specific value.

2.2. Machine Learning

Machine learning is one of the most rapidly developing techniques, and it can help computers to address the problems learnt through experience [6]. Machine learning mainly includes three major paradigms: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is widely applied in spam classifying of e-mails, face identifying, and medical diagnosis fields. It aims to generate predictions according to its mapping functions. Relying on different mapping functions, learning algorithms are divided into neural networks [11], support vector machines (SVM) [12], decision trees [13], logistic regression [14], and decision forests [15]. The mapping functions are driven by different kinds of application needs. Unsupervised leaning focuses on direct inference of predictions without the help of the training sample of previous solved cases [16]. The purpose of reinforcement learning is to learn a mapping function by desponding on intermediate between supervised and unsupervised leaning in training data. Reinforce learning has been successfully applied in human-level control [17].

In recent years, one prominent progress in supervised learning involves deep neural networks. Deep learning [18] has played an important role in computer vision, speech recognition, natural language translation, and collaborative filter. Deep learning algorithms can be used to discover useful representations of the input data without the requirement of labelled training data. The development of machine learning is closely related to other research fields progress. As machine learning theory develops, we will see the benefits it brings us. Machine learning contributes to scholarly impact research as follows:

- Design effective algorithms fitting to various scholarly sources of data;
- Predict future trends such as articles impact and scholars' impact in future;
- Conduct scholarly recommendation such as recommending collaborators, the articles with top impact in various research fields.

Currently, some researchers have leveraged machine learning techniques to successfully predict scholarly impact, including articles, scholars, institutions and even countries. The commonly used methods for predicting scholarly impact include neural networks, SVM, Markov [19], XGboost [20], etc. In term of the performance of predicting the scholarly impact, neural networks model is better than Markov model. Neural networks model can be used to deal with large amount of data compared to Markov model and SVM model. SVM model can not only be directly used to regress, but also be used to classify. The SVM model is more suitable to deal with a small amount of scholarly data. Otherwise, mixing SVM model and neural networks model can obtain better performance of prediction compared to the independent SVM or the single neural networks model. The predictive power of XGboost is better than Markov, neural networks, and SVM. However, a disadvantage of XGboost is that it needs to adjust a large of number of parameters. In the future, we believe that machine learning can provide more support to resolve emerging issues about scholarly impact, and also can provide models for understanding learning in scholarly impact, biological evolution, neural systems and other research fields.

2.3. Data Mining

Data mining is used to discover knowledge hidden in a large amount of data, including spatial data mining, temporal data mining, sequence data mining and intention data mining [21]. Data mining has important applications in finance, telecommunication, science, and engineering fields. In recent years, we have witnessed a rapid expansion in the ability to collect data from various sensors and online media platforms in different formats. For instance, a large source of data is going to be generated from online platforms like Facebook, Twitter, and Google. Big data drives scholars to continually explore useful patterns for better services. Meanwhile, it gives a challenge for big data mining. Data mining can benefit scholarly impact research as follows:

- Mining heterogeneous academic networks, such as article-author networks, author-journal networks, author-institution networks, etc.;
- Exploring the complex relationships among academic entities, including the relationships of papers, authors, journals, conferences, teams, institutions and countries;
- Seeking automatically patterns in scholarly data to predict future trends and improve predicting performance;
- Mining large data streams for effective scholarly recommendations;
- Cleaning scholarly data to gain valuable information;
- Integrating diverse kinds of scholarly data.

In brief, data mining can solve the scholarly evaluation and predication problems by analyzing data in database. Discovering meaningful patterns in scholarly networks will lead to some advantages. Useful patterns allow scholars to predicate scholarly impact based on new data. For example, in order to predict the impact of an article, we may first train the data of previous years by applying machine learning techniques, like neural network, Markov and SVM models. In addition, we predict future impact of an institution on testing datasets. How to express a pattern is important. The expressions of a pattern can be presented in two ways: transparent box and black box. The former's construction discloses the structure of the pattern by explaining something about scholarly data, while the latter's construction is inexplicable. Data mining also involves learning for finding structural patterns in scholarly data. It helps to explain data before making predictions. In data mining, machine learning is applied in many research fields. It is used to capture the explicit knowledge structures which are important to preform well on new data.

2.4. Network Science

Network science can help to understand the structure of networks, development and weaknesses. Despite apparent diversities, a lot of networks generate, evolve, and are driven by some basic laws and mechanisms. For instance, degree distribution has been proved to be the power law; small world property is an important principle in many networks. Two important organizing principles of the evolution of networks were introduced, i.e. preferential attachment and fitness [22].

Well-known scholarly network structures are complex, including homogeneous and heterogeneous networks [23], directed and undirected networks [24]. Homogeneous citation networks contain article-article networks, author-author networks, journal-journal networks, and word-word networks, etc. Figure 3 shows the citation relationships of article-article networks generated by random extracted 486 articles and their references from APS dataset, 561 edges in total. Each circle represents an article and the links represent citation relationships. Blue, yellow and green represent nodes with small, medium and large degrees. Heterogeneous citation networks include article-author networks, author-journal networks, article-journal networks, etc. In particular, co-author networks and co-word networks are also important homogeneous networks in scholarly impact studies. Citation network is a representative directed network, showing a link from a citing paper to a cited paper, while co-author network is a undirected network. The important indices of nodes in undirect networks include degree centrality, betweens centrality, closeness centrality, k-shell, k-core, and eigenvector centrality.

In directed networks, two representative algorithms, i.e., PageRank [25] and HITS [26] algorithms, are commonly used to calculate the importance degree of nodes.

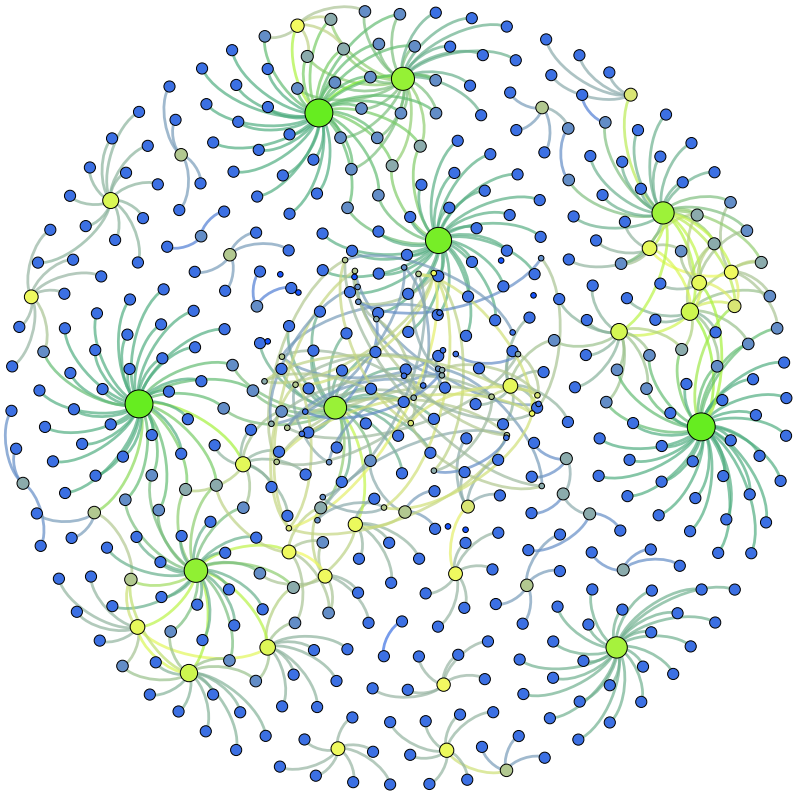

Figure 3. Characterizing citation relationships of article-article networks. The degrees of nodes range from small (blue) to large (green). The larger the degree of a node is, the more references the article has.

In diversified scholarly networks, network analysis plays a key role mainly in the following three aspects. First, network analysis helps identifying key nodes in scholarly networks. These key nodes are a series of scholarly entities, including top influential articles, top influential authors, top influential journals, top influential teams, top influential institutions, co-authors with super tie, academic rising star [27,28], serendipity in scientific collaboration [29], Sleeping Beauties in science [30], etc. We also need to study the difference of the important degree of various nodes in unweighted and weighted networks. Most scholarly networks are weighted [31,32], but we cannot always obtain appropriate weights. However, an appropriate weight is the key for quantifying scholarly impact. For example, impact of an article is no longer a simple citation count. The importance degree of each article in citation networks should consider the authors' authorities of citing articles and published journal's prestige of the article through analyzing the citation networks. Citation-based structured measurements have provided new perspective for evaluating scholarly impact. Second, network analysis helps to explore the most important structure features, such as what features determine scholars' success [33], success of an article [34], and success of teams [35]. Third, network analysis helps quantifying the relationships among scholarly entities, including articles, authors, journals, conferences, institutions, teams and countries. For example, previous researchers have quantified the relationships of co-authors in scientific community. It indicates scientific collaboration with weak, strong, and super ties from longitudinal perspective [36]. All in all, network structured analysis provides a solution to quantifying the scholarly impact.

In the next section, we will introduce article impact metrics and prediction, mainly including two aspects: core methods and their recent research progress.

3. Article Impact Metrics

Figure 4 provides a framework for evaluation of article impact to build and test a set of scholarly data models, including data collection, data pre-processing, data analysis, features selection, algorithms design, optimizing algorithms and evaluation of algorithms. Datasets refer to original datasets like DBLP, APS and MAG. According to targets of evaluation, we can use the original datasets or complement the original datasets by crawling necessary data from scholarly websites like SCOPUS. Researching the various relationships including citation, co-author, co-cited, etc. is beneficial to assess the impact of scholarly impact. For example, identifying different citation relationships provides an objective evaluation method [9]. In the assessment framework, the assessment method is the most central part. Currently, there are several types of assessment methods: citations, Altmetrics and citations-based structured metrics. The validity of the verification method is an essential part. Common evaluation methods include Spearman's correlation coefficient, recommendation intensity and so on. Impact metrics can briefly be divided into two categories: unstructured metrics (or statistical metrics) and structured metrics according to the way of measurement.

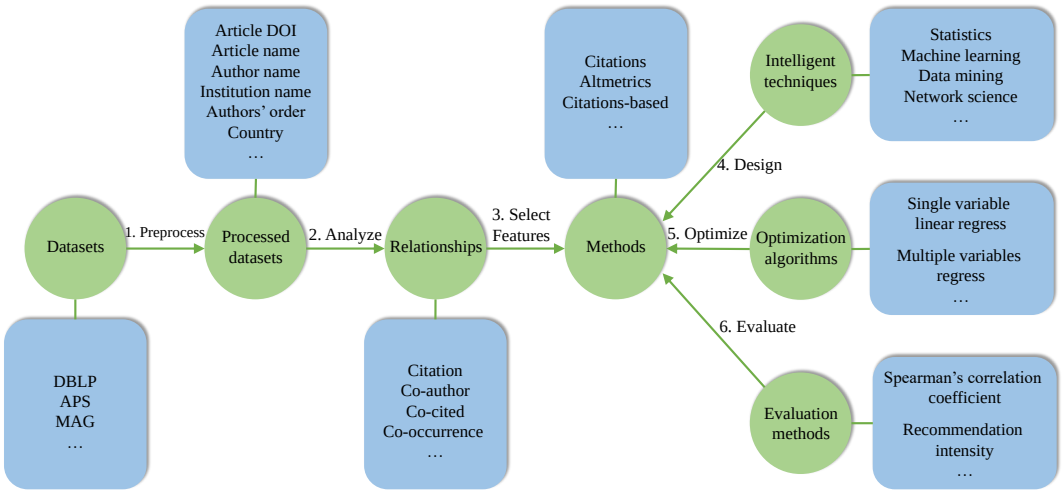

Figure 4. Frameworks of evaluating article impact.

Citations as statistical method are perhaps the oldest and most widely used metric for article impact evaluation. Citations as measuring metric are always under dispute. From the perspective of objective evaluation, can original citations truly characterize the quality of article? The answer is obviously no. The biggest obstacles are self-citation and mandatory citation [37], which have increased the difficulty of objectively measuring article impact. How to accurately identify a variety of self-citation and mandatory citation is challenging. Meanwhile, negative citation has attracted scholars' attention [38]. However, scientific researchers do not stay at distinguishing the citation patterns. The scholarly publications are undergoing the changing from traditional prints to online platforms. The change generates some open issues. Meanwhile, it presents an opportunity to characterize article impact from multidimensional perspective.

Altmetrics [39] emerge at the historic moment and obtains much attention in academic community. Altmetrics are the study of measuring the scholarly impact based on activities in social media platforms, and go beyond citations [40]. Altmetrics present various quantitative values including citations, downloads, mentions, tweets, shares, views, discussions, saves and bookmarks from statistical perspective. Altmetrics scores (mentioned in blogs) can be used to identify highly cited articles. At the same time, Altmetrics can complement and improve evaluation of article impact with new insights [10]. Although broadening the evaluation methods for measuring scholarly impact, Altmetrics lack the authority and credibility as metrics. It is partly because Almetrics are easy to be gamed by malicious scholars [41].

Citations-based structured methods have made some progress. One significant measurement of impact metrics in recent years involves homogeneous and heterogeneous networks [42], including citation networks, co-author networks, co-citation networks, article-author networks, article-journal networks, author-journal networks. The diversity of scholarly networks can satisfy the diverse needs of applications with different scholarly structures capturing different kinds of scholarly characters. One thing can be certain: citations-based structural metrics can generate a truer measure of the importance of an article than citations alone. Previous researchers have contributed to the structural metrics for evaluating article impact [5,43–48]. These assessment methods mostly are based on PageRank algorithm and HITS algorithm. PageRank algorithm provides a fast and objective ranking way to rank the nodes in network. In a citation network, papers with higher PageRank scores have more chances to be visited. PageRank is more suitable for homogeneous networks. In scholarly networks, HITS algorithm distinguishes the scholarly entities as authorities and hubs based on the local structure, and calculates their scores in a mutual reinforcing way. HITS algorithm can also be applied to heterogeneous networks like paper-author network and paper-journal network, in which the authors and journals are regarded as hub nodes, and the papers are regarded as authority nodes. It is worth mentioning that S-index metric measured article impact through influence propagation in heterogenous citation networks [49]. Meanwhile, Neil Shah et al. suggested a good impact metric should consider the following six aspects: volume sensitivity, prestige sensitivity, robustness, extensibility, temporality, interpretability and computability. Exploiting network structure characters may provide an opportunity to develop a refined and objective metric for measuring the scholarly impact.

While co-citation analysis can be utilised to associate the relevance across different disciplines and to identify the bridging nodes [50], it should be noted that citation-based metrics are biased by diverse domain sizes and citation activities [51]. Domain variation may hamper a fair evaluation for scholarly impact, such as scholarly papers in some disciplines are cited much more or much less compared to others [52]. Two important reasons cause the above results. One is uneven number of cited papers each article in different domains, the other is unbalanced cross-discipline citations. Although scholarly papers can be cited by different domains, Schneider et al. [53] suggested relative citation pattern within disciplines should be considered for the evaluation of scholarly impact.

4. Article Impact Prediction

Prediction of future impact is an emerging area, researching on the "science of science". Impact prediction is more important compared to impact evaluation. Impact prediction can directly allocate funds, scientific awards, and other decisions. Figure 5 provides a flowchart of a computational model for predicting article impact. The left column (Input) is the input data, capturing publication, citation, downloads, reviews, and other information. The center column (Model) describes model learning and testing. The right column (Output) provides a few specific examples which the model can predict.

Specially, article impact prediction has attracted a lot of attention in recent years. Predicting an article impact mainly focuses on predicting citations or citation distributions through network science, data mining and machine learning techniques (see Table 1). Early citations of an article played a critical role for predicting its long-run citation [54]. They showed that university ranking with cumulative citations can be easily predicted by early received citations across the economics discipline at a university. Cao et al. [55] presented a Gaussian mixture model to predict future citations of papers based on short-term citation activities. Peter et al. [56] constructed a keyword-term network to predict the numbers of citations in the future by analyzing the recursive centrality measures, indicating document centrality has higher predictive ability for the future citations of papers. Based on quantile regression, Stegehuis et al. [57] proposed a model to predict the probability distribution for future citations of an article, and considered two key features: early citations and journal impact factor. Yu et al. [58] leveraged four categories of features, including articles, authors, citations, and journals to

predict future citations of an article based on stepwise regression analysis. Based on co-authorship networks, a Machine Learning Classifier was developed to predict whether a publication would get high citations [59]. Based on Random forest classifier, they showed a supervised classification model, in which multidimensional feature vectors were considered to predict the future citations of a paper. Wang et al. [3] constructed a generative model for predicting long-term impact of an article by using three key factors: preferential attachment, citation trend, and fitness. In short, previous researchers are mostly based on early citations for predicting the impact of paper. They mainly focus on the autocorrelation of historical data in citation network. However, a common drawback of these predictive methods is that they are dependent too much on historical citations. Exploring the fundamental characteristics of citations yielded may be able to find a novel predictive method, ignoring the early citations. In recent years, with the development of social media, social media activities are used to reflect the underlying impact of an article. For example, Tweets can predict whether an article can be cited frequently when an article was published for 3 days [60]. Based on a heterogeneous scholarly network, Mohan et al. [61] predicted academic impact by integrating the bibliometric data with the social data like weblogs and mainstream news, indicating that graph-based measure can reasonably predict the impact of early stage researchers.

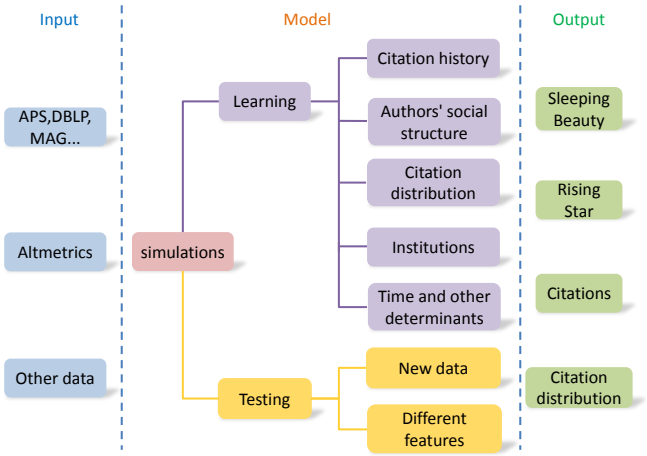

Figure 5. Flowchart of predicting article impact.

Table 1. Several representative methods for predicting article impact.

Features	Prediction Goal	Main Techniques
early citations, Journal Impact Factor	quantile of citations distribution	quantile regression
authors characteristics, institutional factors, features of article organization, research approach	citations	multivariate analysis
Social dimension: co-authorship networks	citations	random forest classifier
year, page count, author count, author name, journal, abstract length, title length, special issue, etc.	long-term citations	random forest
Altmetrics: tweeter	citations	correlation analysis, linear regression analysis

There is an increasing interesting in identifying Sleeping Beauties in science. Sleeping Beauty in scientific community refers to that the value of an article can be recognized only after years of publication [30]. Ke et al. suggested a common mechanism using a parameter-free method to identify Sleeping Beauties on large-scale datasets.

5. Open Issues and Challenges on Article Impact Metrics

Despite pioneers have obtained success, article impact remains a young field with many open issues. In previous researches, many different datasets are usually used to quantify scholarly article impact. These granulitic and inconsistent data have been applied in various scholarly researches. Sharing datasets are necessary and valuable for objectively evaluating article impact and generating new metrics. Unified and consistent scholarly datasets are an open issue. Citation-based structured metrics are relatively new and have got less attention. Researchers consider that the important degrees of citation structures is newly shaped by PageRank and HITS algorithms introduced in scholarly networks. In addition, social dimensioned assessment and citation distributions have been less explored. Thus, multidimensional metrics for quantifying article impact are an open issue. Altmetrics have been considered for complementing article-level metrics. Pioneered researchers have made some progress. Altmetrics for evaluating scholarly article impact is still an open issue. In this open issue, locating the reasonable and available benchmarks is an urgent need to be solved.

5.1. Unified and Consistent Scholarly Datasets

With the rapid emergence of a large number of social platforms, scholarly datasets present hitherto unknown event in academia. Even though these datasets possess personalized characters, they have the problems of missing data, repeated data, data uncertainty phenomena. Evaluation metrics based on these inconsistent datasets can bring some problems. For example, reproducing scientific findings in previous researches can be realized. Therefore, unified and consistent scholarly datasets should be ascertained and shared by scientific researchers in academia for impact metrics.

5.2. Multidimensional Metrics

In previous researches, citation-based structured metrics mainly consider the dimensions of authors, journals, articles and time. Each author's importance in citation networks is usually ignored. An article generative impact is regarded as the same no matter who cites it. In fact, citing authors' impact in citation networks should be investigated for objectively quantifying article impact. Copying the same citations from other articles is a frequently observed practice in academic publications [62]. Therefore, an article may get more citations through frequency-dependent copying if it is cited by experienced scholars. The article impact can be influenced by many factors such as authors' social relationships, citation distributions of authors, journals, institutions and countries. In particular, identifying anomalous citation patterns and weakening citation strength are critical for objectively measuring article impact [9]. Although analysing Conflict of Interest (COI) relationships between authors has given a solution to identify anomalous citations. We need to mine COI relationships for more objective assessment in a further step. These problems have not been addressed. Therefore, future impact metrics need to explore the importance in citation networks, authors' social relationships, various citation distributions, etc.

5.3. Altmetrics

Altmetrics are recent article-level metrics [63]. Altmetrics are usually considered as the complement beyond citations. Altmetrics have some merits for evaluating. However, Altmetrics are only based on web usage statistics [64]. They are more easily manipulated by factitiously downloading, sharing, commenting, etc. What can be done to guarantee the credibility of data on social media for evaluating article impact? What can be measured by Altmetrics? How to select sources of data for Altmetrics? What relationships exist between Altmetrics and citations? Using data analysis techniques

to explore Almetrics indicators in depth provides a possible solution to validating Altmetrics. There are many explored opportunities in article impact researches.

5.4. Benchmarks

Available and credible benchmarks are key to measuring article impact. Despite past decades witnessed important progress, it is difficult to verify the performance of article impact metrics. Without right datasets and standards, developed metrics are not contextually robust and cannot be understood [65]. Therefore, how to select benchmarks based on unified and consistent scholarly datasets with the aim of objectively quantifying impact is an important open issue.

6. Open Issues and Challenges on Article Impact Prediction

Despite our research has summarized article impact prediction so far, a great number of further issues and challenges call for our attention to predict impact accurately. In this section, we point out some potential issues except for unified scholarly datasets and benchmarks.

6.1. Sleeping Beauty

Despite of the previously analyzed Sleeping Beauties phenomena, various issues remain to be addressed in the corresponding researches. How to identify Sleeping Beauties in science? How to predict impact of Sleeping Beauties? Whether the trending topics are related to Sleep Beauties? Whether the trending topics have contributed to predict Sleep Beauties? Whether the correlations between Sleep Beauties and different journals, between Sleep Beauties and institutions can influence the impact of Sleeping Beauties? Therefore, more efforts are needed to explore these critical scientific problems.

6.2. Multidimensional Prediction

Despite pioneered researchers have obtained success from multidimensional perspective in predicting article impact, a full integration of multidimensional datasets needs to be explored in a further step. Characterizing the breadth and the depth of an article impact is unfortunately only from one single perspective. For example, previous researches generally focused on early citations to predict impact of an article [54]. However, little attention has been paid to location information such as institutions and countries, social relationships and citation distributions for predicting impact. Therefore, future research needs to predict article impact from multiple dimensions.

6.3. Rising Star

Predicting the fast-rising citations for an article in the future provides valuable guidance to the academia. It can help the academia to find out popular topics or new topics, advanced techniques, significant findings, etc. Meanwhile, a direct benefit is to avoid wasting time in the ocean of scholarly data for researchers. What are the features contributed to enhance an article impact? Finding these features is beneficial to predict rising star in articles.

7. Conclusions

This article presents a detailed overview of evaluating and predicting article impact. It discusses the open issues and challenges that need to be solved in a further step. At first, we have given a simple introduction about article impact research. Next, we have elaborated on core methods and recent progress. Then, we have introduced some key techniques, and some opportunities can be seen by leveraging statistics, machine learning, data mining and network science techniques. Finally, we have presented open research issues regarding the assessment and prediction of article impact, and pointed out potential research directions.

Author Contributions: X.B. conceived the study and wrote the manuscript; F.X. supervised the design and the development of the proposed study; F.Z. contributed statistical analysis work; X.B., H.L., F.Z., Z.N., X.K., I.L. and F.X. revised the manuscript; and all authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aguinis, H.; Suárez-González, I.; Lannelongue, G.; Joo, H. Scholarly impact revisited. *Acad. Manag. Perspect.* **2012**, *26*, 105–132.
2. Gargouri, Y.; Hajjem, C.; Larivière, V.; Gingras, Y.; Carr, L.; Brody, T.; Harnad, S. Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS ONE* **2010**, *5*, e13636.
3. Wang, D.; Song, C.; Barabási, A.L. Quantifying long-term scientific impact. *Science* **2013**, *342*, 127–132.
4. Piwowar, H. Altmetrics: Value all research products. *Nature* **2013**, *493*, 159.
5. Chen, P.; Xie, H.; Maslov, S.; Redner, S. Finding scientific gems with Google's PageRank algorithm. *J. Informetr.* **2007**, *1*, 8–15.
6. Jordan, M.; Mitchell, T. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260.
7. Johnson, R.A.; Wichern, D.W. *Applied Multivariate Statistical Analysis*; Prentice Hall: Upper Saddle River, NJ, UAS, 2002; Volume 5.
8. Di Ciaccio, A.; Coli, M.; Ibanez, J.M.A. *Advanced Statistical Methods for the Analysis of Large Data-Sets*; Springer: Berlin/Heidelberg, Germany, 2012.
9. Bai, X.; Xia, F.; Lee, I.; Zhang, J.; Ning, Z. Identifying anomalous citations for objective evaluation of scholarly article impact. *PLoS ONE* **2016**, *11*, e0162364.
10. Costas, R.; Zahedi, Z.; Wouters, P. Do 'altmetrics' correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 2003–2019.
11. Rojas, R. *Neural Networks: A Systematic Introduction*; Springer: Berlin/Heidelberg, Germany, 2013.
12. Hearst, M.A.; Dumais, S.T.; Osman, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18–28.
13. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106.
14. Hosmer, D.W., Jr.; Lemeshow, S. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2004.
15. Ho, T.K. Random decision forests. In Proceedings of the Third International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282.
16. Hastie, T.; Tibshirani, R.; Friedman, J. Unsupervised learning. In *The Elements of Statistical Learning*; Springer: New York, NY, USA, 2009; pp. 485–585.
17. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533.
18. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.
19. Jarrow, R.A.; Lando, D.; Turnbull, S.M. A Markov model for the term structure of credit risk spreads. *Rev. Financ. Stud.* **1997**, *10*, 481–523.
20. Chen, T.; He, T. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
21. Bhise, R.; Thorat, S.; Supekar, A. Importance of data mining in higher education system. *IOSR J. Hum. Soc. Sci.* **2013**, *6*, 18–21.
22. Barabási, A.L. Network science. *Philos. Trans. R. Soc. Lond. A Math. Phys. Eng. Sci.* **2013**, *371*, 20120375.
23. Yan, E.; Ding, Y. Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *J. Am. Soc. Inf. Sci. Technol.* **2012**, *63*, 1313–1326.
24. West, J.D.; Jensen, M.C.; Dandrea, R.J.; Gordon, G.J.; Bergstrom, C.T. Author-level Eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *J. Am. Soc. Inf. Sci. Technol.* **2013**, *64*, 787–801.
25. Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web*; Technical Report; Stanford InfoLab: Stanford, CA, USA, 1999.

26. Kleinberg, J.M. Authoritative sources in a hyperlinked environment. *J. ACM* **1999**, *46*, 604–632.
27. Zhang, C.; Liu, C.; Yu, L.; Zhang, Z.K.; Zhou, T. Identifying the Academic Rising Stars. *arXiv* **2016**, arXiv:1606.05752.
28. Zhang, J.; Ning, Z.; Bai, X.; Wang, W.; Yu, S.; Xia, F. Who are the Rising Stars in Academia? In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, Newark, NJ, USA, 19–23 June 2016; pp. 211–212.
29. Sugiyama, K.; Kan, M.Y. Serendipitous recommendation for scholarly papers considering relations among researchers. In Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, Ottawa, ON, Canada, 13–17 June 2011; pp. 307–310.
30. Ke, Q.; Ferrara, E.; Radicchi, F.; Flammini, A. Defining and identifying Sleeping Beauties in science. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 7426–7431.
31. Bai, X.; Zhang, J.; Cui, H.; Ning, Z.; Xia, F. PNCOIRank: Evaluating the Impact of Scholarly Articles with Positive and Negative Citations. In Proceedings of the 25th International Conference Companion on World Wide Web, Montréal, QC, Canada, 11–15 April 2016; pp. 9–10.
32. Zhu, X.; Turney, P.; Lemire, D.; Vellino, A. Measuring academic influence: Not all citations are equal. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 408–427.
33. Sutherland, K.A. Constructions of success in academia: An early career perspective. *Stud. High. Educ.* **2015**, *42*, 743–759.
34. Letchford, A.; Moat, H.S.; Preis, T. The advantage of short paper titles. *R. Soc. Open Sci.* **2015**, *2*, 150266.
35. Anicich, E.M.; Swaab, R.I.; Galinsky, A.D. Hierarchical cultural values predict success and mortality in high-stakes teams. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 1338–1343.
36. Petersen, A.M. Quantifying the impact of weak, strong, and super ties in scientific careers. *Adv. Short Pap. Titles* **2015**, *112*, E4671–E4680.
37. Esfe, M.H.; Wongwises, S.; Asadi, A.; Karimipour, A.; Akbari, M. Mandatory and self-citation; types, reasons, their benefits and disadvantages. *Sci. Eng. Ethics* **2015**, *21*, 1581–1585.
38. Catalini, C.; Lacetera, N.; Oettl, A. The incidence and role of negative citations in science. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 13823–13826.
39. Priem, J. Altmetrics. In *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*; MIT Press: Cambridge, MA, USA, 2014; pp. 263–288.
40. Kwok, R. Research impact: Altmetrics make their mark. *Nature* **2013**, *500*, 491–493.
41. Cheung, M.K. Altmetrics: Too soon for use in assessment. *Nature* **2013**, *494*, 176.
42. Yan, E.; Ding, Y. Measuring scholarly impact in heterogeneous networks. *Proc. Am. Soc. Inf. Sci. Technol.* **2010**, *47*, 1–7.
43. Wang, Y.; Tong, Y.; Zeng, M. Ranking Scientific Articles by Exploiting Citations, Authors, Journals, and Time Information. In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, WA, USA, 14–18 July 2013.
44. Walker, D.; Xie, H.; Yan, K.K.; Maslov, S. Ranking scientific publications using a model of network traffic. *J. Stat. Mech. Theory Exp.* **2007**, *2007*, P06010.
45. Sayyadi, H.; Getoor, L. FutureRank: Ranking Scientific Articles by Predicting their Future PageRank. In Proceedings of the SIAM International Conference on Data Mining (SDM 2009), Sparks, NV, USA, 30 April–2 May 2009; pp. 533–544.
46. Zhou, Y.B.; Lü, L.; Li, M. Quantifying the influence of scientists and their publications: Distinguishing between prestige and popularity. *New J. Phys.* **2012**, *14*, 033033.
47. Wang, S.; Xie, S.; Zhang, X.; Li, Z.; Yu, P.S.; Shu, X. Future influence ranking of scientific literature. In Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, PA, USA, 24–26 April 2014.
48. Liu, Z.; Huang, H.; Wei, X.; Mao, X. Tri-Rank: An Authority Ranking Framework in Heterogeneous Academic Networks by Mutual Reinforce. In Proceedings of the IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI), Limassol, Cyprus, 10–12 November 2014; pp. 493–500.
49. Shah, N.; Song, Y. S-index: Towards better metrics for quantifying research impact. *arXiv* **2015**, arXiv:1507.03650.
50. Small, H. Maps of science as interdisciplinary discourse: Co-citation contexts and the role of analogy. *Scientometrics* **2010**, *83*, 835–849.

51. Kaur, J.; Radicchi, F.; Menczer, F. Universality of scholarly impact metrics. *J. Informetr.* **2013**, *7*, 924–932.
52. Radicchi, F.; Fortunato, S.; Castellano, C. Universality of citation distributions: Toward an objective measure of scientific impact. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 17268–17272.
53. Schneider, M.; Kane, C.M.; Rainwater, J.; Guerrero, L.; Tong, G.; Desai, S.R.; Trochim, W. Feasibility of common bibliometrics in evaluating translational science. *J. Clin. Transl. Sci.* **2017**, *1*, 45–52.
54. Bruns, S.B.; Stern, D.I. Research assessment using early citation information. *Scientometrics* **2015**, *108*, 917–935.
55. Cao, X.; Chen, Y.; Liu, K.J.R. A data analytic approach to quantifying scientific impact. *J. Informetr.* **2016**, *10*, 471–484.
56. Klimek, P.; Jovanovic, A.S.; Egloff, R.; Schneider, R. Successful fish go with the flow: Citation impact prediction based on centrality measures for term-document networks. *Scientometrics* **2016**, *107*, 1265–1282.
57. Stegehuis, C.; Litvak, N.; Waltman, L. Predicting the long-term citation impact of recent publications. *J. Informetr.* **2015**, *9*, 642–657.
58. Yu, T.; Yu, G.; Li, P.Y.; Wang, L. Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics* **2014**, *101*, 1233–1252.
59. Sarigöl, E.; Pfitzner, R.; Scholtes, I.; Garas, A.; Schweitzer, F. Predicting scientific success based on coauthorship networks. *EPJ Data Sci.* **2014**, *3*, doi:10.1140/epjds/s13688-014-0009-x.
60. Eysenbach, G. Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *J. Med. Internet Res.* **2011**, *13*, E123.
61. Timilsina, M.; Davis, B.; Taylor, M.; Hayes, C. Towards predicting academic impact from mainstream news and weblogs: A heterogeneous graph based approach. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; pp. 1388–1389.
62. Simkin, M.; Roychowdhury, V. Read Before You Cite! *Complex Syst.* **2003**, *14*, 269–274.
63. Thelwall, M. Data Science Altmetrics. *J. Data Inf. Sci.* **2016**, *1*, 7–12.
64. Barbaro, A.; Gentili, D.; Rebuffi, C. Altmetrics as new indicators of scientific impact. *J. Eur. Assoc. Health Inf. Libr.* **2014**, *10*, 3–6.
65. Wilsdon, J. We need a measured approach to metrics. *Nature* **2015**, *523*, 129.