Scientometrics
https://doi.org/10.1007/s11192-020-03466-w

Finding rising stars in bibliometric networks

Ali Daud[1] · Min Song[2] · Malik Khizar Hayat[3] · Tehmina Amjad[3] · Rabeeh Ayaz Abbasi[4] · Hassan Dawood[5] · Anwar Ghani[3]

Received: 12 December 2019

Abstract
Finding rising stars (FRS) is a hot research topic investigated recently for diverse application domains. These days, people are more interested in finding people who will become experts shortly to fill junior positions than finding existing experts who can immediately fill senior positions. FRS can increase productivity wherever they join due to their vibrant and energetic behavior. In this paper, we assess the methods to find FRS. The existing methods are classified into ranking-, prediction-, clustering-, and analysis-based methods, and the pros and cons of these methods are discussed. Details of standard datasets and performance-evaluation measures are also provided for this growing area of research. We conclude by discussing open challenges and future directions in this prosperous area of research.

Keywords Finding rising stars (FRS) · Ranking · Prediction · Clustering · Analysis · Bibliometric networks

Introduction

Online databases, social networks, blogs, and social media are omnipresent in today's World Wide Web. Examples include Bibliometric Networks (BNs) (e.g. ACM, DBLP), social networks (e.g. Facebook, Foursquare, MySpace), blogs (e.g. WordPress, Joomla), and social media (e.g. Twitter). Intriguingly, in all aforementioned web sources, multiple types of links exist between various entities. For instance, in a BN that is most commonly investigated for FRS, possible types of entities include papers, authors, and publication venues, while links include author-paper, author-publication venue, paper–paper (citation) and author–author (coauthor). Such multiple interactions between different but related entities are rather dynamic and are continuously experiencing variations to their structure and attributes. Consequently, due to the evolution of a person's career over time, it is significant for decision makers and leaders to identify a high-potential candidate who is not an expert currently but is projected to be influential soon are called Rising Stars (Li et al. 2009). FRS is of paramount importance to appointing the right young people to available junior positions, to increase the productivity of an organization and to utilize the benefits of their

✉ Ali Daud
ali_msdb@hotmail.com

Extended author information available on the last page of the article

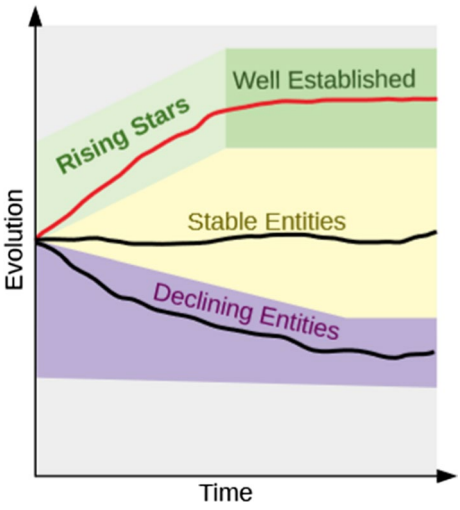

Fig. 1 The Evolution of a Person's Career over Time (G. Tsatsaronis et al., "How to become a group leader? or modeling author types based on graph mining," in Research and Advanced Technology for Digital Libraries, Springer 2011)

energetic and vibrant behavior. Figure 1 shows the typical evolution pattern for a person in different domains. Based on human intuition, the evolution of a person's career over time can usually have four notable patterns: (1) persons who start average but their performance declines over time due to less activity and will to make their career better, (2) people who take an average start but are capable of maintaining their pace over time with no extraordinary achievements, (3) entities who take an average start but rise quickly to become rising stars due to being very active and smart, leading to extraordinary achievements, and (4) entities who are well established due to their consistent remarkable achievements. There are many fields in which FRS would be beneficial, e.g. BNs, social networks including blogs and forums and production companies. A simple example is human resources (HR). In the HR sector, when appointing a new worker, an organization would like to forecast whether the applicant shows the talent of becoming an expert shortly, even if currently she has no conspicuous career achievements. In the business and marketing sector, companies used to analyze a new product using different business-oriented strategies before its official launch to determine whether it will attract a large number of customers shortly or not.

The significance of FRS has led to the proposals of several techniques over the last decade. The recent interests in FRS have been addressed in various application domains, like Bibliometric Networks (BNs) (Li and Tong 2015; Daud et al. 2015; Panagopoulos et al. 2017; Daud et al. 2013; Zhang et al.2016), Community Question Answer (CQA) networks (Le and Shah 2016), sports networks (Ahmad et al. 2017), and telecommunication networks (Daud et al. 2019).

Several efforts have been made in the literature to study the challenges involved in the ranking of entities and the identification of vital nodes in complex networks. A comprehensive review was conducted by Amjad et al. (2018a) to survey the methods for ranking of authors in academic social networks. The authors thoroughly studied and compared different ranking metrics including the iterative, link structure based, semantics based and temporal methods for ranking of authors. They emphasized that ranking methods use variable features for the purpose and all these features and their combinations produce variable results and influence the results in their way. These features and their combinations can be used in different scenarios depending upon the requirement of the problem. These

methods generate changing results with a change of variables. Benchmarking standards are not present for the performance evaluation of these methods due to the unavailability of the ground truth.

A detailed study was conducted by Liao et al. (2017) to survey the static and time-aware ranking algorithms and their applications in evolving networks. They studied the influence of network progression on well-established invariable algorithms. They also examined the impact of the temporal dimension in tasks like network traffic prediction, and identification of vital nodes. They identified that there is a need for extensive performance estimation of static and temporal network ranking methods and is an open challenge for future researchers. They concluded that overlooking the time variable can cause less optimized or even false results.

A wide-ranging survey was conducted by Lu et al., regarding the detection of vital nodes within a complex network. They found that identifying a generic index that is capable of measuring a node's importance in all possible situations is not practical. They studied the methods for the recognition of specific imperative nodes. Based on structural information they introduced the centrality indices and ranked these nodes with iterative methods. They also studied the importance of a node by considering the influence of the removal of one node or a group of nodes from the network. They also studied the detection of a set of fundamental nodes along with the discovery of individual critical nodes. They found that the performance of an algorithm depends on the objective functions under consideration. Identifying the dominant node or the most dominant set of nodes within a specified time instead of a steady-state is an open research challenge.

Recently, a study was conducted by Zhou et al. (2019) to find fast influencers in a complex network for real-world applications like viral marketing and online information spreading. The study discriminated between the usual nodes and fast influencers who can spread a message in a short time. Brohi and Lehnertz studied the vital edges instead of vital nodes with the help of centrality measures (Bröhl and Lehnertz 2019). They studied to find the edges in a network that are important among other pairs of vertices.

To the best of our knowledge, however, no dedicated and comprehensive study of FRS yet exists, despite the mounting significance of a fresh research area because of the increasing interest and availability of data.

It is important to note that in several application domains, the data are naturally represented as a network, with clearly a defined set of nodes and edges. However, in some domains, it is not obvious how to represent data as a network. It is likely to depend on the nature of the research question being addressed. In this study, we aim to bridge the gap between the increasing need for FRS and the lack of existence of a comprehensive assessment of FRS methods. In all, the major contributions made in this study are summarized as follows:

- Provides a chronological study of existing methods that can provide a roadmap to extract useful insights for the research community
- Provides a classification of methods based on particular application domains and identifies key features to study the semantics of the problem.
- Provides Insights about standard FRS datasets.
- Provides useful recommendations and directions for future research.

The rest of this paper is organized as follows. Section 2 provides a taxonomy of methods, which each taxonomic group is further sub-categorized based on different methods. Section 3 reviews some applications of FRS methods in networks other than bibliometric

networks. Section 4 provides data sets used in different methods and highlights how FRS methods are evaluated. In Sect. 5, we recommend future research directions, and Sect. 6 concludes this study.

Finding rising stars in bibliometric networks

In this section, we discuss different application domains wherein the problem of FRS has been addressed. Next, we discuss the pros and cons of existing methods. Lastly, we walk through the historical paradigm of these methods (see Table 1) and provide a comprehensive summary of methods bibliometric networks (BNs), Within the BNs, we have made four subcategories based on the data mining functionalities used for FRS. Careful analysis of the historical paradigm of networks and methods provide us with several interesting trends.

BNs are built by using online publication attributes along with coauthor and citation associations among them. The evolution and growing usage of online databases have forced BNs to cope with substantially rich information and interesting and valuable knowledge. For instance, BNs, as shown in Fig. 2, contain information on *papers*, *authors*, *venues* (conferences or journals), *titles/terms,* and *citations*. Convincingly, FRS problem in these networks may prove to discern interesting results as compared to author ranking (Amjad et al. 2015), expert finding(Daud et al. 2010), author name disambiguation (Wu et al. 2014), author interest finding (Daud 2012), research collaboration recommendation (Guns and Rousseau 2014), citation content analysis (Zhang et al. 2013), citation count prediction (Yan et al. 2011), for the research community.

For instance, when universities recruit young faculty members, they aim to hire those candidates who may prove to be Rising Stars in the future. Methods of FRS in BNs are sub-categorized into four categories: (1) Ranking Methods, (2) Prediction Methods, (3) Clustering Methods, and (4) Analysis Methods. Methods in each of these categories are discussed below.

Ranking methods

Ranking is one of the most popular approaches to FRS; proposed studies adopting rankings are summarized in Table 2. PubRank (Li et al. 2009), for example, is based on information derived from the node's out-links and is fundamentally different from the algorithms which use in-links, particularly the state-of-the-art PageRank (Page et al. 1999). In PubRank, the authors used three major features. First, the degree of mutual influence between researchers in the citation network is modeled using a novel link weighting strategy in which a novice researcher may prove to be a Rising star in the future by leveraging and collaborating with her seniors. Second, the authors used the past performance of a researcher, which is measured by investigating the quality of initial publications by a novice researcher following her publications at top-level venues. Finally, the authors took into account the chronological changes in the networks. The research area of evolution pattern for a researcher reflects her collaboration strength; a rapid rise in this measure may push one to turn into a star. PubRank has two major limitations. Firstly, the Rising Stars do not reflect the author's contribution-oriented mutual influence. However, the author's order indicates the contribution of each author, as the first author is generally believed to be the main contributor in terms of work. The novice author who co-authors with eminent authors as a first author has

Table 1 Historical Paradigm of Methods in Bibliometric Networks

Year/task	Bibliometric networks			
	Ranking	Prediction	Clustering	Analysis
2009	Li et al. (2009)			
2013	Daud et al. (2013)			
2015		Li and H. Tong (2015, Daud et al. (2015), and Dong et al. (2015)		
2016	Zhang et al. (2016a, b, Wijegunawardana and Mohan (2016))	Li et al. (2016)		
2017	Daud et al. (2017)	Ning et al. (2017a, b) and Zia et al. (2017)	Panagopoulos et al. (2017)	Amjad et al. (2017)
2018		Amjad et al. (2018)		Ding et al. (2018)
2019		Nie et al. (2019) and Bin-Obaidellah and Al-Fagih (2019)	Zhu et al. (2019)	
2020				Daud et al. (2020)

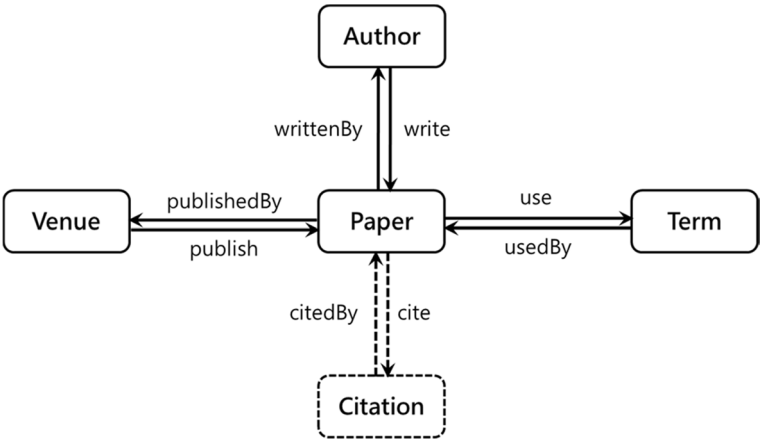

Fig. 2 Schema for BNs

a greater chance of being a star than a novice collaborator who is listed as a second or later author. Lastly, using static ranking is not practical, owing to node's (publication venues) rapid evolution pattern in BNs.

PubRank was later improved by introducing author's "contribution-oriented mutual influence" and "dynamic publication venue" scores in StarRank (Daud et al. 2013). The co-author order is taken into account in this method by calculating a distinct weight for each coauthor based on the order of appearance (in case the authors were listed alphabetically in the paper then this is not applicable as these days most papers lists authors based on contribution). For instance, if an author L has collaborated on one paper with K as the third author, then the contribution is calculated as 1/3. Further, the dynamic publication venue is another significant factor in FRS. The entropy of venues is calculated using words appearing in the title of papers to represent a dynamic publication venue score: the lower the entropy, the higher the venue quality, although there are few exceptions, e.g. Nature, Science and PLOS One like journals, which publish on diverse topics and still have high impact factors. In other words, high-quality venues show less disorder in certain publication trends. Entropy can be calculated as a result of two situations. One possibility is an inconsistent or incomplete list of available online venues and others can be a delay in updating of venues. StarRank outperformed PubRank by exterminating its limitations, although it has its limitations. StarRank has also ignored citations information just like PubRank and did not provide detailed experimental evidence of entropy of title as equivalent to the citations or rank of the venues.

Collaboration is undeniably a fundamental feature of a researcher's academic career. Zhang et al. (2016) proposed method of CocaRank based on collaboration caliber for FRS in BNs. In contrast to previous methods, the authors modeled BNs as heterogeneous networks. Researcher's ability to collaborate with their peers is computed as Collaboration Caliber (Coca) of the researcher. The Coca score for each researcher is calculated using entropy-based statistics. Subsequently, the PageRank (Page et al. 1999) score for each paper in a citation network is calculated to generate HITS (Kleinberg 1999) scores of authors and journals in author-paper and author-journal networks, respectively. Finally, Coca and HITS scores are merged to compute the ultimate score for each researcher using the CocaRank method. Author-paper networks are explored in

Table 2 Summary of ranking-based rising stars methods in BNs

Model	Year	Main idea	Data Set	Findings	Limitations
PubRank (Li et al. 2009)	2009	Used mutual influence and publication venues	DBLP	Collaboration with senior authors can impact the standing of the researchers significantly.	Lack of coauthor's citation-based influence, author contribution-based influence and dynamic venue worthv (Daud et al. 2013).
StarRank (Daud et al. 2013)	2013	Used author-weighted contribution and publication venues	DBLP	Impact of collaboration with senior Authors as well as incorporation of dynamic publication venues can significantly identify the rising stars.	Lack of coauthor' citations-based influence, venues-based influence (Zhang et al. 2016; Daud et al. 2017) lack of social features
CocaRank (Zhang et al. 2016)	2016	Used collaboration caliber	APS	When caliber based collaboration is used for finding Rising Stars, there is a rapid increase in citation count in initial 3-5 years, decrease in 5-7 years, and flat afterwards.	Overlooked the diverse types of nodes and edges in heterogenous academic networks (Wijegunawardana and Mohan 2016). Social features are also ignored.
IIRL (Zhang et al. 2016)	2016	Author, co-author (social), content, venue, and temporal features-based ranking	AMiner	Temporal features are the best indicators for Rising Stars prediction, Venue features are less important	Low prediction accuracy and performance. Ignores multiobjective optimization (Wijegunawardana and Mohan 2016). Lacks academic social networks, such as ResearchGate or Mendeley.
MOO (Wijegunawardana and Mohan 2016)	2016	Multiobjective optimization and rank aggregation methods applied to coauthor network, citation network, and venue scores	AMiner/ IS Stack Exchange	Distinct Rising Stars detection can be done by key player identification and rank aggregation	Weighted mutual influence is ignored (Daud et al. 2017). Academic social networks such as ResearchGate or Mendeley.

Table 2 (continued)

Model	Year	Main idea	Data Set	Findings	Limitations
ScholarRank (Zhang et al. 2016)	2016	Considers both statistical indicators and influence calculation methods	APS	Enhanced features and diverse sub-networks were found to be more top ranked Rising Stars	Ignores multiobjective optimization (Wijegunawardana and Mohan 2016). Still, multiple factors overlooked pertinent to researcher's weighted influence, sociality and paper download times.
WMIRank (Daud et al. 2017)	2017	Weighted mutual influence of coauthors, order of appearance, and venues	AMiner	Mutual influence of coauthors, author contributions, papers, venues provides better Rising Stars	Lacks academic social networks such as ResearchGate or Mendeley. Ignores multiobjective optimization (Wijegunawardana and Mohan 2016).

(Zhang et al. 2016) with a combination of citation and paper-journal networks. Zhang et al. found rapid-rising young researchers and proposed a versatile feature-based Impact Increment Ranking Learning (IIRL) algorithm. They found that venue-based features proved to be less useful than other feature classes, such as author, coauthor (social), content, and temporal features. They further analyzed the results and showed that prediction models follow similar patterns for different research topics (Zhang et al. 2016).

In addition to preceding studies, the problem of FRS is investigated using diverse available data sources to investigate how the significance of a beginner researcher advances over time (Wijegunawardana and Mohan 2016). The heterogeneous data include AMiner and Information Security Exchange forum. They employed three measures: (1) coauthor network PageRank slope, (2) Citation network PageRank slope and (3) publication venue score slope. Scores are combined by using multi-objective optimization and rank-aggregation methods (Wijegunawardana and Mohan 2016).

Similar to CocaRank (Zhang et al. 2016), Zhang et al. also proposed another ranking-based method of FRS called ScholarRank (Zhang et al. 2016). The difference between the two techniques is the networks used to demonstrate the efficiency of algorithms. In the ScholarRank method, citation, paper-journal, and paper-author networks are used to measure the mutual reinforcement process in BNs. The mutual influence as measured in StarRank (Daud et al. 2013) and mutual reinforcement process among authors in BNs are used in calculating the author's ScholarRank score. The feature of mutual influence is used in a number of studies where it has provided significant improvement in FRS in BNs (Li et al. 2009).

Daud et al. (2017) proposed a method called Weighted Mutual Influence Rank (WMIRank), in which weighted mutual influence in coauthor networks is used to find Rising Stars. The feature space is three-dimensional, exploring the citation influence of coauthors, the order of appearance, and publication venues. StarRank also explored the order of appearance while measuring the author's contribution-oriented feature. The greater the rank of an author on a paper, the higher the contribution score of that author. Instead, WMIRank used the number of citations with weighted influence for measuring author's contribution.

Gopavarapu et al. presented a method of finding the rising stars based on PubRank in 2019 (Gopavarapu et al.2019). They presented a method using mutual influence based features including on author's contribution, co-author's citations, the order of appearance of author name and venue of publications. The authors claim that work is based on PubRank but their work is more inspired or a close copy of features used by Daud et al. (2013) in StarRank (2013). Gopavarapu et al. have used the features of order of author's name and publication venue in a way very similar to work of Daud et al. even the values in example calculations and equations are copied from Daud et al.'s work (for reference see values in Table 1 and Eq. 5 of Daud et al. which are unethically copied by Gopavarapu et al. on pages 442 and 443). Gopavarapu et al. have not provided dataset details and have only mentioned that data of their college faculty was used for experimentation and college name is not mentioned as well. Performance evaluation was not performed for the measuring goodness of the so-called proposed method.

Summarizing the above discussion, it can be seen that intuitively, collaboration with well-known researchers can lead to a rising future career. A current low-profile researcher may be a rising star in the future if she collaborates with renowned researchers in her preliminary publications.

Prediction methods

Supervised machine-learning techniques are applied to a wide range of application domains and are well-accepted for FRS (Table 3). BNs that comprise a broad range of features such as collaboration (Zhang et al. 2016; Dong et al. 2015) and paper count benefit from prediction abilities of supervised machine-learning techniques.

In Zhang et al. (2016), the authors presented collaboration caliber-based ranking of Rising Stars. However, in Dong et al. (2015), the authors used collaboration signature with predictive case studies. They defined the collaboration signature based on the distribution of collaboration strengths with which a researcher collaborates in her academic network. It is accomplished by four measures: "sociability", "dependence", "diversity", and "self-collaboration". They experimentally demonstrated that collaboration signatures help in identifying Rising Stars in the early stages of their careers.

Prediction of Rising Stars enables employers to pinpoint persuasive researchers in coauthor networks. Classification and prediction models are quite capable of handling the Rising Stars prediction problem. Daud et al. (2015) and Amjad et al. (2018) investigated the problem in BNs using the discriminative and generative models. From each category, two models are selected: Maximum Entropy Markov Model (McCallum et al. 2000) (MEMM), Classification and Regression Tree (CART) (Loh 2011) as discriminative, Bayes Network (BN) (Friedman et al. 1997) and Naive Bayes (NB) (Zhang 2004; Rish 2001) as generative models for classification task. Author-, coauthor-, and venue-based feature categories are explored and venue category is found to be most effective. MEMM performed better for predicting an average number of citations and CART performed better for predicting an average relative increase in citations.

The mechanisms that leverage high-impact scientific work have many significant implications including recruitment search, personal career development, and authority of research resources. Despite much progress in prediction models several key algorithmic challenges for long-term scientific impact prediction still need to be investigated. The iBall-a family of algorithms to foresee the enduring scientific impact at an initial stage was proposed. iBall can be generalized to both classification and regression models. It handles several aspects that leverage scholarly impact: "scholarly feature design", "non-linearity", "domain heterogeneity", and "dynamics". FRS is formulated as an optimization problem and scalable algorithms are provided to solve the problem. The validity of algorithms is performed on the American Physical Society publication dataset (Li and Tong 2015).

Forecasting research impact always has significant implications in research resources allocation, finding potential collaborators, tracking research frontiers, and so on. To address this problem, Li et al. (2016) proposed a predictive model called iPath with improved prediction consistency and enhanced parameter smoothness for predictive model parameters for FRS. Extensive empirical results prove that the iPath algorithm is more effective than iBall.

Publication venues play an important role in rise or fall of certain bibliometric entities such as authors, co-authors, and citations of papers. For instance, a paper published at a significant venue can increase the citations for the respective author—hence, an increase in the significance of the author. Consequently, predicting rising venues is persuasive from the citations point of view. Zia et al. (2017) addressed the problem of rising venue prediction in citations networks by proposing five prediction features including citation count, publications counts, cited to, and cited by at venue level. Also,

Table 3 Summary of prediction-based rising stars methods in BNs

Model	Year	Main idea	Data Set	Findings	Limitations
Collaboration signature (Dong et al. 2015)	2015	Explored future scientific impact correlation with collaboration signatures as a distribution of collaboration strengths in her academic ego network	AMiner	Similar level authors retain similar collaboration, signatures, collaboration signatures are strong indicator of scientific impact	Collaboration signatures are only validated in computer science publications, causality among collaboration signature development and scientific growth is not investigated (Dong et al. 2015). Impact forecasting for point prediction (Li and Tong 2015)and impact pathway is not considered (Li et al. 2016)
Feature-based Rising Stars(Daud et al. 2015)	2015	Author, coauthor, and venue classed features are explored	AMiner, DBLP	Among three featured classes, venue is found the most effective. CART and MEMM perform better with respect to total citations and an average relative rise in total number of citations, respectively.	Academic social networks such as ResearchGate or Mendeley is not considered. Impact forecasting for point prediction (Li and Tong 2015) and impact pathway is not considered (Li et al. 2016).
iBall family of algorithms (Li and Tong 2015)	2015	A model to handle the "non-linearity", "scholarly feature design", "domain-heterogeneity" and "dynamics" for predicting the persistent scientific impact in the initial years of researcher	APS	Citation record of a research entity (e.g., author, venue, paper) in initial three years is a stronger indicator of its enduring impact	Impact forecasting is for point prediction only, and is not directly applicable to forecast the impact pathway (Li et al. 2016).
iPath (Li et al. 2016)	2016	A model to handle impact forecasting for both point prediction and impact pathway	AMiner	Prediction consistency and parameter smoothness is flexible for handling linear and non-linear models, domain heterogeneity and dynamics	Academic social networks such as ResearchGate or Mendeley inclusion related parameters are not explored in proposed model
Rising Venues (Zia et al. 2017)	2017	ML-based features to predict rising venues in citations networks	DBLP	Citation count, publications count, cited to, and cited by are the effective features for rising venue prediction	More Bibliometric features to be explored

Table 3 (continued)

Model	Year	Main idea	Data Set	Findings	Limitations
StarRank (Ning et al. 2422)	2017	Citation-based influential researcher's prediction	APS	StarRank's performance is better with SVM	Apply StarRank and alike models to other dataset domains
Social Gene (Ning et al. 2017)	2017	Factors analysis method predict rising star in scholarly networks	APS	Social Gene performed better than PubRank	Explore the commotion between rank of rising researchers and the social genes and also internal relation among social genes
Academic Rising Star Prediction (Nie et al. 2019)	2019	Non-iterative hierarchical weighted evaluation model	AMiner	Features based on quality of citing paper and co-authors are used to predict rising stars.	Inclusion of sentiment of citations in the training process can further improve the model and its results.
Prediction of rising stars (Bin-Obaidellah and Al-Fagih 2019)	2019	Use of scientometric indicators and machine learning indicators	WoS	Scientometric indicators when used as features in classification task and multiple linear regression can give good results of prediction.	Comparative analysis among the sued scientometric indicators was not provided which in fact can show that which indicatiors are better for the prediction.

the authors employed four ML algorithms including Bayesian Network, Support Vector Machine, Multilayer Perceptron, and Random Forest, whereas Support Vector Machine appeared with higher accuracy.

Evaluating a researcher's influence has been a real problem for higher education and research institutions. Ning et al. (2017) neural network-based model StarRank to predict the influential researchers. The assessment of a researcher's worth is primarily based on the papers she published. However, about PubRank (Li et al. 2009), the authors mentioned that newly published papers receive fewer citations as compared to the earlier published papers. It could eventually diminish the excellence of a rising researcher. Also, a less-cited paper has no noteworthy influence on the researcher's excellence. To address the problem, the authors also considered the number of papers cited in a certain paper to measure the influence of that paper on the researcher. Ultimately, the rising researcher is predicted based on the assumption that an excellent paper will certainly cite other excellent papers and probably be cited by other excellent papers. Therefore, regardless of the paper publishing time, a high quality paper will have a high influence on the researcher.

Finding rising academic supervisor/researcher is important with respect to the research institutions as well as students. Most of the times institutions have to deal with the fresh graduates to hire—it necessitates the need to forecast rising researchers from the institution's and as well as student's point of view. Ning et al. (2017) proposed an operational method, Social Gene, to predict rising stars in academic networks. In order to address rising researcher prediction problem, the authors used inner characteristics for academic entities termed as social genes rather than using bibliometric information. Overall, Social Gene method used 14 initial features comprising the information for co-author and publication. Analytical Hierarchy Process (AHP) is used as an operational research tool for decision support—which separates the complex problem into several hierarchies. The subjective assessment and objective method are united in order to evaluate the weight of factors (features).

A machine learning based method was presented by Nie et al. for prediction of academic rising stars (Nie et al. 2019). They proposed a hierarchical non-iterative weighted model for the said task. The model uses the features based on citing paper's quality and impact of collaborators. The increment of score of young authors was used to label the authors as rising stars. Multiple features were extracted and studied with the help of multiple classifiers for better results. The results show that author's venu based features were one of the best predictors for the academic rising stars and the author based features were of small relevance.

Another method for the prediction of academic rising stars was presented by Omar Bin-Obaidellah and Ashraf E. Al-Fagih et al. used the scientometric indicators for the said task (Bin-Obaidellah and Al-Fagih 2019). They applied machine learning methods (Multiple linear regression, KNN and SVM) on data collected from web of science ranging from 2001 to 2015. The papers of only two affiliations were collected, including King Fahd University of Petroleum & Minerals- KFUPM, and Indian Institute of Technology- IIT. The authors concluded that their scientometric indicators can be used for the prediction task but they have not provided a comparative analysis that out of eight indicators, which indicators were more suitable for the said task and which ones have low impact.

Clustering methods

Unsupervised machine-learning techniques are also applied to find Rising Stars and have proven their effectiveness in vast application areas. Pertaining to state-of-the-art complex BNs, it is usually tough to infer performance indicators that are relevant to dynamics and

the career evolution of a researcher over time. Extending the area of Rising Stars finding techniques unsupervised machine learning is also used.

An initial effort was made to explore productivity, impact, and collaborations as a triad of key performance indicators for FRS in BNs using unsupervised learning (clustering) (Panagopoulos et al. 2017). The researchers are grouped into an automatically learned optimal number of clusters using clustering validity metrics—namely, the Davies Bouldin Index (DBI). Results demonstrate that the authors who were able to perform consistently for all key performance indicators are the Rising Stars (Table 4).

An unsupervised method for finding the rising stars in different technological fields was presented in 2019 that involves the mining of patent information from co-inventor network (Zhu et al. 2019). Considering technology performance, sociability and innovation caliber, the authors designed multiple features and assigned weights to all features using entropy. K-Means algorithm was used to cluster the inventors according to their profiles. The study empirically analyses three types of stars including tech-oriented Rising technology Stars (RTS), social-oriented RTSs and innovation-oriented RTSs. All-round RTSs were also discovered as entities that have potential in at least three of the mentioned categories.

Analysis method

Being successful is not the same for each individual as it fits into the semantics of one's vision of what victory means. Every single researcher needs to define the type of values, abilities, aspirations, incentives, or goals she wants to achieve. For a successful career in academia, young researchers yearn for an entry in one of the world's top institutions and to work with the pioneers of their respective fields.

To identify those researchers who are most likely to succeed, Amjad et al. (2017) conduct an extensive analysis to measure that how productive an author can be in her initial phase of career and hereafter (Table 5). They used the following core parameters for their analysis: publications, citations, H-index (Hirsch et al. 2005), sociability, first and last publication, and longevity. They completed the analysis using two subsets of publications—one for identification and the other for verification. AMiner data are whittled down by applying filters including a senior researcher filter, which eliminates those with higher H-index values and publication counts; a junior researcher filter, which weeds out the novice researchers with lower publication counts; and an author collaboration filter, which removes authors who have collaborated with a renowned researcher and is likely may become a rising star.

The findings of (Amjad et al. 2017) in BNs reflect the variety of motives feasible for solving the FRS problem. The junior researcher can rely on well-known senior collaborators to advance their careers. In contrast, if one fails to collaborate with senior researchers initially, she may have the opportunity to collaborate with senior researchers later. This could also be the reason for becoming a rising star as senior researchers are attracted towards junior researchers owing to their valuable work.

Evaluating rising stars in academia significantly help for tasks such as resource allocation, decision support, research funding, and other real-world problems. Ding et al. (2018) mined the core factors using a decision tree—influencing the future impact of authors in the academic social network. The authors used American Physical Society (APS) dataset for the experimentation which contains the papers and citation information. The authors used PubRank as the benchmark method, however, it doesn't count the publication quality for the authors. As an alternative, they used PubRank as the paper quality and called

Table 4 Summary of clustering-based rising stars method in BNs

Model	Year	Main idea	Data Set	Findings	Limitations
Feature-based clustering method for FRS (Panagopoulos et al. 2017)	2017	Detecting groups of authors based on predefined quantity, impact, and collaboration-based features	Scopus	Continuous improvement in key performance indicators is a must for initial years of career. Strong collaborations play key roles in scientific longevity of researchers	Limited period for power graphs, usage of limited clustering approaches such as multi-clustering to group different types of objects collectively
Identifying rising technology stars in co-inventor networks (Zhu et al. 2019)	2019	Finding the technology rising stars according to their potential in technology performance, sociability and innovation caliber.	Derwent Innovation Index patent database	innovation caliber is less important than the other two traits. Feature 'potential' is most important for tech-oriented and social-oriented RT Stars	More features, indicators and variables need to be explored for RT stars.

Table 5 Summary of Analysis-based Rising Stars Method in BNs

Model	Year	Main idea	Data Set	Findings	Limitations
Features-based Analysis of Rising Stars (Amjad et al. 2017)	2017	Features-based analysis with intuition that collaboration with senior researchers as a major indicator of being a rising star	AMiner	Initial senior collaboration can polish a junior researcher's future career; lacking senior collaboration initially but achieving such collaboration, later on, can do the same trick in most cases	Effect of causality was not considered (Amjad et al. 2017), statistical significance testing of the hypothesis was not performed
Factor-based analysis and evaluation of Rising Stars (Ding et al. 2018)	2018	Inner factor-based analysis to predict the future impact of authors	APS	Citation proved to be the influential factor to gauge the future impact of authors	The datasets which include more bibliometric information other than citation should be considered
Falsely Predicted Rising Stars (Daud et al. 2020)	2020	Analyze the reasons behind the falsely predicted rising star to reduce the false positive rate of these methods	DBLP	Five major reasons were identified related to falsely predicted rising stars	Further improvement is required to improve to FRS methods so that anticipated false-positive cases can be identified at an 3early stages.

in extend-PubRank. The experimental results show 80% accuracy in terms of citations for rising star prediction.

The finding rising star methods can be applied in many fields like bibliometric networks, social networks like the forums, blogs, and depending upon the nature of the field in which they are applied, the accuracy of these methods can be very critical. In Daud et al. (2020), Daud et al. have analyzed the reasons behind the falsely predicted rising stars. They analyzed the case of bibliometric networks and studied the individual cases who were identified as rising stars but were unfortunately not able to show performance accordingly. The DBLP dataset was studied and authors came up with five major reasons behind the falsely predicted rising stars.

The methods discussed in Sect. 2 are focusing on the problem of finding the rising stars in bibliometric networks. Most of these methods fall into the category of ranking methods. The ranking based FRS methods mainly are experimented for only one academic social network and lacks in comparison of results from more than one network. Most of these methods also ignore multi-objective optimization (Dong et al. 2015). Some of these methods overlooked the diverse types of nodes and edges in heterogeneous academic networks or ignore the effect of weighted mutual influence of co-authorship or co-citations. The prediction based methods suffer the main problem that collaboration signature development and scientific growth is not investigated (Dong et al. 2015). Impact forecasting for point prediction (Li and Tong 2015) and impact pathway is also not considered (Li et al. 2016). Apart from that, these methods are also mainly tested for only one network. Very little work was done for FRS suing the clustering based and analysis based ranking methods. These domains need the attention of new researchers to get more significant findings.

Application of FRS methods in other domains

From the literature surveyed, we can see that most of the work done in the field of FRS in BNs. However, although very little, we found the applications of FRS methods in networks other than BNs. In this section, we provide the details of application FRS methods from these miscellaneous domains including Community Question Answering (CQA) networks, Sports Networks, and Telecommunication Networks.

CQA comprise an important component of social networks. CQA is built by exploring the question–answering forums entitie's attributes and associations between entities, such as users, posts, and topics. In CQA networks, there is relatively a small ratio of users who have higher ratings, which they can achieve by providing quality posts, and eventually, earn an illustrious reputation in the community. Such users attract much traffic to the forums and are also crucial for site development in terms of hits or clicks. To identify those individuals who are new to these forums and possibly will attain a top rating in the future is a challenging task. As in contrast to BNs, CQA networks contain dearth information on collaborators, peers, and relationships, such as "friendships" on Facebook.

Preliminary effort (Le and Shah 2016) has been made to attempt to find Rising Stars in CQA networks (Table 6). A three-step model is proposed wherein, first, features are extracted, then training and testing sets are separated out, and finally, classification algorithms—namely, Logistic-regression (log-reg) (Loh 2011)), Support Vector Machine (SVM) (Cristianini and Shawe-Taylor 2000), Decision Trees (DT) (Quinlan 1999), Random Forest (RF) (Breiman et al. 2001), and Adaptive Boosting (AdaBoost) (Freund et al. 1999)—are executed to detect Rising Stars. The feature space is categorized into four

Table 6 Summary of prediction-based rising stars method in CQA networks

Model	Year	Main Idea	Data Set	Findings	Limitations
Feature-based CQA Rising Stars Prediction Method (Le and Shah 2016)	2016	Argued that unlike other domains CQA lacks collaborator and friend information and proposed four feature categories: "personal features", "community features", "temporal features", and "consistent features".	Stack Overflow	Rising Stars can be identified after short observations, 85% accuracy of forecasting Rising Stars. Decision trees result in low accuracy whereas RandomForest and AdaBoost result in high accuracy	Feature's applicability on other domains is questionable. Difference between focused and non-focused CQA sites is not studied (Le and Shah 2016)
Identifying Potential Answerers in Community Question-Answering (Le and Shah 2018)	2018	By maintaining the personal profiles of the users, identifying the potential answerers can increase the chance that questions will get answered on a CQA forum	Yahoo Answers and Stack Overflow	A short set of 1000 users was identified in both datasets from whom at least one will answer the question with a probability of more than 50%. Incorporating user's interest can further increase this probability	The method was not tested on unanswered questions. The questions that receive at least one answer were considered for experimentation (Le and Shah 2018).

sub-categories: “personal features” (participated topics, number of posts, average length of posts, question–answer ratio), “community features” (average score, average favorite marked, average comments posts received), “temporal features” (average delay between the posts, delay among two recent posts, post’s trend), and “consistent features” (standard gap between the posts, standard scores between posts, standard time gap between the post in the second half, and activity (less/more) on posts).

Not all features are of equal significance. To convey the feature’s significance level, RF is used to evaluate the importance of each. “Number of posts”, the “average gap between the posts”, and the “average score of the posts” are found to be most substantial among the unabridged feature set. Feature significance is also measured using four subcategories. The personal features are found to be more accurate. Furthermore, a total of 3.4 million users with 21.2 million posts for 6 years are explored from Stack Overflow CQA forum by (Le and Shah 2016).

Another very significant aspect of CQA was raised by Le and Shah in (2016). They suggested that CQA sites can enhance the experience of a user by recognizing the most probable answerers and directing the most appropriate questions to them. Finding potential answerers increases the chance that a question is answered or answered more quickly. To handle the problem they used prediction to identify the potential answerers based on question content and user profiles. A record of past activities of users is maintained and a score is computed using question and user profiles when a new question is posted. The proposed method predicted a group of 1000 users out of which at least one user will answer the question with a probability greater than 50% from datasets of Yahoo Answers and Stack Overflow.

Sports networks are built by exploring the sports entitie’s attributes and associations between players, teams, and places where the matches are played. FRS in the sports domain is necessary to select the best players for a team. The preliminary effort is made by Ahmad et al. (2017) to predict Rising Stars in the game of cricket, the second-most prevalent game around the globe after football (Table 7). Given a cricketer from two defined categories (batsman and bowler), the authors predicted whether a cricketer is a RS by using rich feature space. The predominant features for batsman consist of co-batsman runs, co-batsman average, and co-batsman strike rate, whereas predominate features for a bowler include co-bowler economy, co-bowler average, and co-bowler strike rate. The co-batsman category attained 88% f-measure accuracy, thus declared as the most prominent feature for the prediction of Rising Stars. As far as the generative (BN and NB) and discriminative (SVM and CART) models are concerned, NB is dominant over all of the rest, with an accuracy of 94.5%. However, in general, the generative models are proved better than discriminative models in foreseeing the Rising Stars.

Telecom networks are built by exploring the telecom entitie’s attributes and associations between entities, such as business manager, senior business manager, and regional business manager.

FRS in the telecom domain is necessary to maximize business profit. Businesses can benefit from the information in their systems, but as written, you’re saying that the growing need is what provides the insights and recommendations. Maximum profit generation with minimum investment is the key goal of customer relationship management. Trend analysis, statistical functions, projection, and prediction are effective tools to stimulate business ahead of time for in-time decisions making and planning. Specifically, the decision to place the best people at best positions is a key role provided by human resource departments. The boom of the telecom industry has made the hiring of rising business managers the backbone of the business, as a business’s success depends on fixed line operators that are

Table 7 Summary of prediction-based rising stars method in sports networks

Model	Year	Main Idea	Data Set	Findings	Limitations
Feature-based Cricket Rising Stars Prediction Method (Ahmad et al. 2017)	2017	Using co-players, teams, and opposite team attributes in Cricket Sports ODI Format for finding rising batsman and bowlers	ESPN Cricinfo—Statusguru	Co-batsman category overwhelmed the others in batting, team category overtook in bowling, NB generative model is dominant overall among generative and discriminative models	Social media discussions about players are not considered, while they are found useful recently for players ratings

based on the productivity of business managers. Every business manager focuses on how the business can be enhanced with respect to a number of viable features.

To address the problem of Rising Star's prediction in the telecom sector, Daud et al., made an effort to predict rising business (telecom) managers via investigating two key variables: average revenue (AR) and an average relative increase in revenue (ARIR) (Daud et al. 2019) (Table 8). Diverse feature space based on co-business manager, senior business manager, and regional general manager is explored using supervised machine learning models—namely, Naïve Bayes (NB), Bayes Network (BN), Neural Network (NN), and Support Vector Machine (SVM). Co-business manager category attained 80% f-measure accuracy thus declared the most prominent feature for Rising Stars prediction. Among generative (BN and NB) and discriminative (SVM and NN), NB dominated.

Datasets and performance evaluation

Datasets

Most tasks are highly dependent on dataset's availability and usage. The datasets used from different real-world application domains for FRS are discussed in this section. Table 9 gives an overview of features mainly included in datasets calculated from bibliometric networks and Table 10 gives an overview of features included by other generic networks, apart from bibliometric networks, for the FRS problem.

DBLP

DBLP[1] is an online database containing metadata of major publications related to computer science. This project started as a small investigational web server and evolved into a popular open-data repository of computing-related publications. The mission of this data source is to provide bibliometric meta-data of research artifacts. As of June 2017, DBLP indexed around 3.8 million papers authored by more than 1.7 million researchers. Specifically, DBLP indexed more than 31,000 conference or workshop proceedings, 32,000 journal volumes, and 23,000 monographs. All data are freely available to download.[2]

AMiner

AMiner[3] is a free online service used to index and search BNs. It intends to carry out data mining tasks, web search, and indexing through the online publication database. It was created as a research project especially relevant for BN analysis, expert finding, and influence extraction. All the papers appearing in it are peer-reviewed. Since 2006, more than 3 million publications by about 1,300,000 researchers have been placed in this repository.

[1] http://dblp.uni-trier.de/.
[2] http://dblp.uni-trier.de/xml/.
[3] http://aminer.org.

Table 8 Summary of Prediction-based Rising Stars Method in Telecom Networks

Model	Year	Main Idea	Data Set	Findings	Limitations
Feature-based Rising Business Manager Prediction Method (Daud et al. 2019)	2019	Prediction of Rising Business Managers using Co-Business Manager, Senior Business Manager and Regional Business Manager feature categories	PSTN, Broadband (PTCL)	Co-business manager features are more useful, NB is dominant overall	Geographical, educational, and personal family features are not considered which have significant impact on someone's managerial abilities

Table 9 Main features included in datasets of academic networks

	Features Included
AMiner	Title, Authors, Venue, Year, Number of citations, References, Abstract
DBLP	Title, Authors, Pages, Year, Volume, Number, DOI, URL
APS	DOI, Journal, Volume, Issue, First page and last page, Title, Authors, Affiliations, Publication history, Table of contents heading, Article type, Copyright information.
Scopus	Document type, Abstract, Keywords and index terms, Cited references, Affiliation data, Author profiles, and ORCID integration.
SNAP	Collaboration network and ground-truth communities extracted from DBLP dataset

Table 10 Main features included in datasets of miscellaneous networks

	Features Included
Stack Overflow	Unique question id, User id of questioner, Score of the question, Time of the question, a comma-separated list of the tags, Number of views of a question, Number of answers for a question, Unique answer id, User id of answerer, Score of the answer, Time of the answer
ESPN Cricinfo	Data for Test Matches, ODIs, T20Is including Bating, Bowling, Fielding, All-round, Partnership, Team, ICC Rankings, Umpire and Referee
Public Switched Telephone Network	In time provisioning, In time rectification, Registered fault, Repeated fault, Repeated fault, Net additions, Disconnection rate, Denied order, Broadband connections, Customer retention, Customer win back, Customer segmentation, Exchange capacity, Collection ratio
KONECT	Code, Category, Data source, Vertex type, Edge type, Format, Edge weights, Size, Volume, Average degree, Maximum degree, Size of largest connected component, Wedge count, Claw count, Triangle count, Square count, Power law exponent, Gini coefficient, Relative edge distribution entropy, Assortativity, Clustering coefficient, Diameter, Mean shortest path length, and spectral norm

American physical society (APS)

The American Physical Society[4] archives detailed information on publications from 12 physical journals, including authors, publication's DOI, date of publication, venues, author's affiliations, and citations. It also includes the publication's application towards theory, experiments, education, and application of physics in the development of computer technology.

Scopus

Scopus[5] is a bibliometric database encompassing citations and abstracts for academic journal articles. It incorporates almost 22,000 journals from more than 5000 publishers. Of these, 20,000 are peer-reviewed journals in social sciences, medical, scientific, and

[4] https://journals.aps.org/archive/.
[5] https://www.scopus.com/.

technical domains. It is maintained and owned by the publisher Elsevier. It is only available through subscription. Scopus also covers patent databases.

Stack overflow

The Stack Overflow[6] is a focused CQA site covering programming-related questions. It is distinct from other CQA sites in that all the questions are pertinent to the programming. Users of Stack Overflow participate in various activities like questioning, answering, and voting a post-up or down. Users can achieve a good reputation by posting highly up-voted and quality answers to the questions.

ESPN cricinfo

The ESPN Cricinfo[7] is a reliable website that collates data for all cricket matches played since 1779. It contain information on all player categories including batsman, bowler, wicket-keeper, and all-rounders. Additionally, it contains data on different cricket formats, including one-day international matches, twenty-twenty international matches, and tests. One can also find domestic-level data for numerous cricket playing nations. The database also includes news and stories about matches and players.

Public switched telephone network

The data are acquired from Pakistan Telecommunication Company Limited (PTCL) for a maximum period of one year on special request for research-oriented purposes. It comprises information on the business manager (BM), Senior Business Manager (SBM), and Regional General Manager (RGM). It also deals with the two major telecom products: Public Switched Telephone Network (PSTN) and Broadband.

Stanford network analysis platform (SNAP)

SNAP[8] is a network analysis and graph mining library which is general purpose. It provides simple to use, high-level procedures for study and exploitation of large networks. It can process substantial networks with hundreds of millions of nodes and billions of edges and is written in C++ with some modules in python. It can competently handle the large graphs, calculates structural properties, generates regular and random graphs, and supports attributes on nodes and edges (Leskovec et al. 2016).

Koblenz network collection (KONECT)

KONECT[9] is a collection of large network datasets of several types including weighted, unweighted, signed, directed, undirected, bipartite and rating networks to execute research

[6] https://archive.org/details/stackexchange.
[7] http://stats.espncricinfo.com/ci/engine/stats/index.html.
[8] http://snap.stanford.edu/.
[9] http://konect.uni-koblenz.de/.

in network sciences and other similar disciplines. It is initiated and organized by the Institute of Web Science and Technologies at the University of Koblenz–Landau.

Performance evaluation

The performance of proposed methods can be measured in different ways according to their functionality. Both quantitative and qualitative methods are usef to evaluate perforamce of FRS methods. Quantitatively, a number of methods were used to evaluate the performance of methods that classify Rising Stars (Daud et al. 2015; Zhang et al. 2016; Ahmad et al. 2017; Zhang et al. 2013) by labelling top ranked authors based on citations as rising stars and remaing as not rising stars. Typically, for binary classification problems, standard measures—namely, precision, recall, and F1-score—are used for measuring performance by Daud et al. (2015) and Panagopoulos et al. (2017), Daud et al. 2013, and Zhang et al. 2016; Le and Shah 2016; Ahmad et al. 2017). E.g. if there are total 500 authors then they are ranked as per number of total citations they had and top 100 are labelled as rising stars and bottom 100 are labelled as not rising stars. One can try diffent number such as top 50 as rising stars and bottom 50 as not rising stars.

The Spearman Correlation Coefficient (SCC) and Pearson Correlation Coefficient (PCC) are two noteworthy performance evaluation measures for ranking-based methods (Zhang et al. 2016; Dong et al. 2015). E.g. the correlation between number of citations based ranks and authors ranked as rising stars on the basis of proposed methods is calculated to see the overlaps. The more the overlaps the better the proposed methods are.

These methods are also evaluated using qualitative performace measures, such as, do authors have brilliant google scholar profiles, their top cited paper at least has 100 citations, are they member of prestigious publication groups e.g. ACM, IEEE, AAAS, SIAM, etc., do they have remarkable achievements e.g. IBM outstanding technical achievement awards, IBM Canada research impact of the year award or IBM outstanding innovation award or best student paper award (Daud et al. 2017).

Challenges and future directions

The mounting needs of FRS revealed a number of challenges and future research directions. We present the most significant directions in this section.

Falsely predicted rising stars

Common attributes used to find Rising Stars are mainly based on the co-author's worth and publishing in top-tier venues at the initial stages of one's career. Sometimes, scholars are lucky enough to have an expert supervisor or co-author form high reputed universities of developed countries who enables them to publish in a top–tier publication venue. Some of these scholars who belong to developing countries where advanced research labs are not available. These scholars have to move back to their countries after completing their PhDs, hence their performance degrades due to reduced research facilities in their home countries. In some cases, young researchers accept an industry job where there is less emphasis on publishing, thus making their performance to slow down. This would lead them to be falsely predicted as Rising Stars, and this issue needs a detailed investigation. Plentiful different attributes, such as collaboration, incoming and outgoing citations, and mutual

influence of collective attributes (Daud et al. 2017), could help in drawing and enhancing the trustworthiness of found Rising Stars.

Use of multiple data sources

Traditional databases, which have officially recorded data and social networks, have data generated by internet users that can be merged to find Rising Stars in an improved way. For example, in BNs FRS can be performed on the basis of traditional BNs (Scopus, DBLP, CiteSeer*X*) and bibliometric networks (ResearchGate.net, Mendeley.com, Academia.edu, scores or LinkedIn skill endorsements). Similarly, in a sports network, Rising Stars can be found by using records of matches played and discussions about the players on social media.

Impact of audience

The impact of citing papers (audience) could be a significant factor in foreseeing Rising Stars in BNs. It can be assumed that, once a paper is cited by an influential audience, it is likely to become influential itself. However, determining audience worth will result in a multi-influence distribution measurement that would enlarge the complexity of the problem.

Assorted application domains

Several RS-finding methods have proposed models that are based on BNs. However, plentiful different features, such as collaboration signatures (Dong et al. 2015), applying machine learning techniques (Daud et al. 2015; Amjad et al. 2018), and WMIRank (Daud et al. 2017). They can be explored in terms of versatile application domains, including sports other than cricket, different community question-answer forums by offering incentives to the site users and hiring business managers in different industrial organizations, such as Telecom Industry, Marketing, and Advertisement Industry, etc. Consequently, the scope of the problem can be enhanced and made applicable to different real-world problems.

Unearthing influential factors

The rising star prediction depends on a vast range of factors. Persuasively, there exist a number of attributes to approach and rectify the FRS problem in different domains. The significant factors could include: enhancing prediction accuracy (Zhang et al. 2016), RS centrality measure (Wijegunawardana and Mohan 2016), social relations of scholars, paper download times (Zhang et al. 2016), and considering varied types of nodes and links for notching essential indicators for ranking Rising Stars (Zhang et al. 2016). Additionally, optimized usage of clustering methods is also an interesting area that needs further exploration (Panagopoulos et al. 2017).

Long-term prediction impact models

Forecasting potential ability by understanding the dynamic mechanism of an entity is an area that has been extensively researched recently. Forecasting long-term scientific impact

models (Li and Tong 2015) and modeling impact pathway prediction problem different factors, such as prediction consistency and parameter smoothness (Li et al. 2016), needs more exploration into the applications of real-world large datasets. The impact pathway prediction problem for handling both linear and non-linear models of different classes also needs to be explored for FRS in different domains, as so far, they are explored for BNs only.

Conclusions

In this study, detailed insights about FRS are provided. One can see that so far most of the work is done in the bibliometric networks and there remains a lot of room to research and study other domains such as blogging, question-answering networks, sports networks, telecom networks, and e-commerce. Mostly co-entity category features proved effective for FRS in different application domains. The dynamic nature of FRS presents a number of unique research issues in different domains. The snapshot of the work done in this area shows open challenges and research opportunities, from developing new models like (Li and Tong 2015) to their applications (Le and Shah 2016; Ahmad et al. 2017). We believe that the methods presented will be useful for the researchers in this area now and in the future.

Acknowledgements This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF- 2019S1A5C2A03083499).

References

Ahmad, H., Daud, A., Wang, L., Hong, H., Dawood, H., & Yang, Y. (2017). Prediction of rising stars in the game of cricket. *IEEE Access, 5,* 4104–4124.

Amjad, T., Daud, A., & Aljohani, N. R. (2018a). Ranking authors in academic social networks: a survey. *Library HiTech, 36*(1), 97–128.

Amjad, T., Daud, A., Che, D., & Akram, A. (2015). MuICE: mutual influence and citation exclusivity author rank. *Information Processing and Management, 52*(3), 374–386.

Amjad, T., Daud, A., Khan, S., Abbasi, R. A., & Imran, F. (2018). Prediction of rising stars from pakistani research communities. In *2018 14th International Conference on Emerging Technologies (ICET)*, pp. 1–6.

Amjad, T., et al. (2017). Standing on the shoulders of giants. *Journal of Informetrics, 11*(1), 307–323.

Bin-Obaidellah, O., & Al-Fagih, A. E. (2019). Scientometric Indicators and Machine Learning-Based Models for Predicting Rising Stars in Academia. In: *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, Sarawak, Malaysia, Malaysia, 2019, pp. 1–7.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Bröhl, T., & Lehnertz, K. (2019). Centrality-based identification of important edges in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 29*(3), 33115.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Daud, A. (2012). Using time topic modeling for semantics-based dynamic research interest finding. *Knowledge Based System, 26,* 154–163.

Daud, A. et al. (2017). Finding rising stars in co-author networks via weighted mutual influence. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 33–41.

Daud, A., Abbasi, R., & Muhammad, F. (2013). Finding rising stars in social networks. In *International Conference on Database Systems for Advanced Applications* (pp. 13–24). Berlin: Springer

Daud, A., Ahmad, M., Malik, M. S. I., & Che, D. (2015). Using machine learning techniques for rising star prediction in co-author network. *Scientometrics, 102*(2), 1687–1711.

Daud, A., Amjad, T., Khaliq, T., & Dawood, H. (2020). All that Glitters is not Gold: Falsely Predicted Rising Stars. *Researchpedia Journal of Computing*. (Accepted)

Daud, A., Li, J., Zhou, L., & Muhammad, F. (2010). Temporal expert finding through generalized time topic modeling. *Knowledge Based System, 23*(6), 615–625.

Daud, A., ul Islam, N., Hayat, M. K., Abbasi, R. A., & Dawood, H. (2019). Prediction of rising business managers in telecommunication networks. *Telemat. Inform*.

Ding, F., Liu, Y., Chen, X., & Chen, F. (2018). Rising star evaluation in heterogeneous social network. *IEEE Access, 6,* 29436–29443.

Dong, Y., Johnson, R. A., Yang, Y., & Chawla, N. V. (2015). Collaboration signatures reveal scientific impact. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pp. 480–487.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *The Japanese Society for Artificial Intelligence, 14*(771–780), 1612.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*(2–3), 131–163.

Gopavarapu, A. R., Sowmya, K. S., Abhishek, B. S., & Babu, P. V. (2019). Finding rising stars in co-author networks

Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics, 101*(2), 1461–1473.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America, 102*(46), 16569–16572.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM JACM, 46*(5), 604–632.

Le, L. T., & Shah, C. (2016). Retrieving rising stars in focused community question-answering. In *Asian Conference on Intelligent Information and Database Systems*, pp. 25–36.

Le, L. T., & Shah, C. (2018). Retrieving people: Identifying potential answerers in Community Question-Answering. *Journal of the Association for Information Science and Technology, 69*(10), 1246–1258.

Leskovec, J., & Sosič, R. (2016). Snap general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology, 8*(1), 1.

Li, X.-L., Foo, C. S., Tew, K. L., & Ng, S.-K. (2009). Searching for rising stars in bibliography networks. In *International conference on Database Systems for Advanced Applications*, pp. 288–292.

Li, L., & Tong, H. (2015). The child is father of the man: Foresee the success at the early stage. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 655–664.

Li, L., Tong, H., Tang, J., & Fan, W. (2016). ipath: Forecasting the pathway to impact. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 468–476.

Liao, H., Mariani, M. S., Medo, M., Zhang, Y.-C., & Zhou, M.-Y. (2017). Ranking in evolving complex networks. *Physics Reports, 689,* 1–54.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 14–23.

Lü, L., Chen, D., Ren, X.-L., Zhang, Q.-M., Zhang, Y.-C., & Zhou, T. (2016). Vital nodes identification in complex networks. *Physics Reports, 650,* 1–63.

McCallum, A., Freitag, D., & Pereira, F. C. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. *Icml, 17,* 591–598.

Nie, Y., Zhu, Y., Lin, Q., et al. (2019). Academic rising star prediction via scholar's evaluation model and machine learning techniques. *Scientometrics, 120,* 461–476.

Ning, Z., Liu, Y., & Kong, X. (2017). Social gene—A new method to find rising stars. In *2017 International Symposium on Networks, Computers and Communications (ISNCC)*, pp. 1–6

Ning, Z., Liu, Y., Zhang, J., & Wang, X. (2017b). Rising star forecasting based on social network analysis. *IEEE Access, 5,* 24229–24238.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web.

Panagopoulos, G., Tsatsaronis, G., & Varlamis, I. (2017). Detecting rising stars in dynamic collaborative networks. *Journal of Informetrics, 11*(1), 198–222.

Quinlan, J. R. (1999). Simplifying decision trees. *International Journal of Human-Computer Studies, 51*(2), 497–510.

Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, Vol. 3, pp. 41–46.

Tsatsaronis, G., et al. (2011). How to become a group leader? or modeling author types based on graph mining. In S. Gradmann, F. Borri, C. Meghini, & H. Schuldt (Eds.), *Research and advanced technology for digital libraries* (pp. 15–26). Berlin: Springer.

Wijegunawardana, P., Mehrotra K., & Mohan, C. (2016). Finding rising stars in heterogeneous social networks. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 614–618.

Wu, H., Li, B., Pei, Y., & He, J. (2014). Unsupervised author disambiguation using Dempster-Shafer theory. *Scientometrics, 101*(3), 1955–1972.

Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011). Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1247–1252.

Zhang, H. (2004). The optimality of naive Bayes. *AA, 1*(2), 3.

Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology, 64*(7), 1490–1503.

Zhang, C., Liu, C., Yu, L., Zhang, Z.-K., & Zhou, T. (2016). Identifying the academic rising stars. *ArXiv Prepr. ArXiv160605752*.

Zhang, J. et al. (2016). Cocarank: A collaboration caliber-based method for finding academic rising stars. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 395–400.

Zhang, J., Ning, Z., Bai, X., Wang, W., Yu, S., & Xia, F. (2016). Who are the rising stars in academia?. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pp. 211–212.

Zhou, F., Lü, L., & Mariani, M. S. (2019). Fast influencers in complex networks. *Communications in Nonlinear Science and Numerical Simulation, 74,* 69–83.

Zhu, L., Zhu, D., Wang, X., Cunningham, S. W., & Wang, Z. (2019). An integrated solution for detecting rising technology stars in co-inventor networks. *Scientometrics, 121,* 137–172.

Zia, M. A., Zhang, Z., Li, G., Ahmad, H., & Su, S. (2017). Prediction of rising venues in citation networks. *Journal of Advanced Computational Intelligence and Intelligent Informatics, 21*(4), 650–658.

Affiliations

Ali Daud[1] · Min Song[2] · Malik Khizar Hayat[3] · Tehmina Amjad[3] · Rabeeh Ayaz Abbasi[4] · Hassan Dawood[5] · Anwar Ghani[3]

Min Song
min.song@yonsei.ac.kr

Malik Khizar Hayat
khizar.mscs741@iiu.edu.pk

Tehmina Amjad
tehminaamjad@iiu.edu.pk

Rabeeh Ayaz Abbasi
rabbasi@qau.edu.pk

Hassan Dawood
hassan.dawood@uettaxila.edu.pk

Anwar Ghani
anwar.ghani@iiu.edu.pk

[1] Department of Computer Science and Artificial Intelligence, University of Jeddah, Jeddah, Saudi Arabia

[2] Department of Library and Information Science, Yonsei University, Seoul, Korea

[3] Department of Computer Science and Software Engineering, IIU, Islamabad, Pakistan

[4] Department of Computer Science, QAU, Islamabad, Pakistan

[5] Department of Software Engineering, University of Engineering and Technology, Texila, Pakistan