Ranking authors in academic social networks: a survey

Tehmina Amjad and Ali Daud
International Islamic University, Islamabad, Pakistan, and
Naif Radi Aljohani
Faculty of Computing and Information Technology, The University of King Abdulaziz, Jeddah, Saudi Arabia

Received 6 May 2017
Revised 12 September 2017
Accepted 9 November 2017

Abstract

Purpose – This study reviews the methods found in the literature for the ranking of authors, identifies the pros and cons of these methods, discusses and compares these methods. The purpose of this paper is to study is to find the challenges and future directions of ranking of academic objects, especially authors, for future researchers.

Design/methodology/approach – This study reviews the methods found in the literature for the ranking of authors, classifies them into subcategories by studying and analyzing their way of achieving the objectives, discusses and compares them. The data sets used in the literature and the evaluation measures applicable in the domain are also presented.

Findings – The survey identifies the challenges involved in the field of ranking of authors and future directions.

Originality/value – To the best of the knowledge, this is the first survey that studies the author ranking problem in detail and classifies them according to their key functionalities, features and way of achieving the objective according to the requirement of the problem.

Keywords Academic social networks, Author ranking, Expert finding, Learning-based ranking, Link analysis, Text similarity ranking

Paper type Literature review

1. Introduction

With the emergence of social network, the world has become a very small place where people are connected to each other via satellite channels, wireless communications 3G/4G networks and many more. We can define a social network as a network within which individuals and/or organizations are arranged as nodes (called actors) and are largely interconnected via edges signifying various relationships, for example, co-authorship, citations, references, recommendation, friendship, likes and dislikes, etc. Representative social networks that are very popular include Twitter, Facebook, Flickr, Instagram, YouTube, etc. Social networks are often represented as graph structures to facilitate mining and analysis of the networks.

Academic social networks (ASNs) are a subclass of social networks with scientific researchers as the main actors who collaborate in a research and appear as co-authors of publications. Such networks are now materialized on the internet and are well supported by various social networking service platforms. Many online publication repositories, such as Citeseer[1] and DBLP[2] are good examples of materialized ASNs on the Internet. They are frequently used for various mining tasks such as author ranking and expert recommendation. With production of a large number of scientific articles, finding relevant information has become a problem in recent years. With an exponential increase in the size of the data, growing computational powers and economical storage mechanism, the problem of finding relevant information has gathered attention of the researchers. With an increase in the size of scholarly data, ranking in ASNs has also become an integral

The work is supported by Higher Education Commission (HEC), Pakistan startup research grant under Interim Placement of Fresh PhDs program 2011, and the Indigenous Ph.D. Fellowship Program.

Library Hi Tech

0737-8831
DOI 10.1108/LHT-05-2017-0090

part of these networks. These methods are required for expert finding, research grant recommendations, finding relevant reviewers and members for editorial panels of journals, workshops and conferences, faculty promotions and relevant tasks in ASNs. There are some intrinsic problems that are involved in ranking of ASNs. These include the dependability of the results of ranking with the attributes used as ranking criteria. Review of these ranking criteria like the number of publications, the number of citations, the citation date, the context of citations, the prestige of authors of citing article topic sensitivity, temporal dimension and so forth would throw light on the role of these ranking criteria.

Various ranking methods have been proposed to quantify the scientific output and quality of researchers/authors. Most of the research articles are co-authored by multiple researchers. Using generic ranking models do not necessarily generate satisfactory results, as these methods tend to treat all authors equally, whereas each author may have contributed to a co-authored work highly differently. Instead of simply counting and grouping the publications and citations of researchers, more appropriate and sophisticated methods for author ranking are expected to produce far better results for decision making and are thus much needed. Apart from ranking methods, several other areas are well investigated for web databases, such as, research collaboration (Guns and Rousseau, 2014), citation content analysis (Zhang *et al.*, 2013), research community mining (Daud *et al.*, 2009a; Daud and Muhammad, 2012), citation recommendation (Daud *et al.*, 2009).

Gupta *et al.* (2013) had published their results on a comparative study of various link analysis methods. They discussed the pros and cons of several generic ranking algorithms including citation count, PageRank, and Hyperlink-Induced Topic Search (HITS). Jiang *et al.* (2013) published another comparative study on link analysis methods. They studied link analysis methods and compared citation counting and summation of paper ranks. Different from these two prior surveys, in our work we studied the author ranking methods in a more thorough and detailed way – we came up with a classification structure for the existing ranking methods and we incorporated more methods into our investigation: probabilistic and learning-based methods. We expect our output (as summarized in this article) to be more thorough and helpful to researchers who want to hold a quick grasp of the status quo of the research in this topic area.

In this study, we classify a wide range of existing author ranking methods into three main types based on their functionalities: these are link analysis, text similarity ranking and learning-based methods. Each category is further divided into more specific subcategories. The classification criteria utilized are inferred by thoroughly studying and analyzing the available literature and contribute to the paper.

Currently, no benchmarking standard particularly designed for author ranking is available. This makes the comparison process to be qualitative most of the time, instead of quantitative. In this survey, we shall also discuss the limitations of the existing methods, and in doing so, we put more emphasis on addressing future directions and inspiring new ideas and methods for solving the current problems.

The article is structured as follows: following a general instruction (Section 1), in Section 2 we present some basic concepts which form a convenient basis for the subsequent discussion; in Section 3, which reflects our key contribution, we bring up a classification scheme for current author ranking methods by which, we put existing ranking methods into three main types (or categories) and numerous subtypes (or subcategories); in Section 4, we briefly address the available data sets and evaluation metrics proposed for author ranking; in Section 5, we point out future directions, and, finally, in Section 6, we conclude the paper.

2. Basic concepts

In this section, we introduce the basic concepts related to ASNs for the convenience of our subsequent discussion.

2.1 Author and co-author

An author is an entity who can claim intellectual contribution in the accomplishment of the research described in a scholarly article. The scholarly article can be published in the form of a paper or a book and contribution can be based on author's study, analysis and/or experimentation. An academic author is usually a researcher conducting a study in a particular academic discipline. When multiple researchers collaborate together and produce a joint output in the form of a publication they are said to be co-authors of that publication.

2.2 References and citations

When an author cites or refers to an existing work in his/her paper, the cited work is called a reference. (From now on, we may simply use the terms, paper or article, to imply any form of publication produced by researcher). Because of the outgoing nature, the references appeared in a paper are also called out-links. The references of a paper are cited in the text of the paper and are listed at the end of the paper with details such as author names, paper title, publishing venue and date, and page numbers, etc. The list of the references of a paper is also known as the bibliography of the paper. Citation of a paper "A" occurs when the authors of paper "B" mention a reference of paper "A" in paper "B". In the context of ASNs, we call each citation that a paper received (i.e. it is mentioned in another paper) an incoming reference of the paper, which thus is also termed as an in-link.

2.3 Co-authorship networks, author-citation networks and paper-citation networks

The ASNs are usually represented as graphs in which nodes stand for authors or papers, and edges represent a certain relationship between the nodes such as authorship (between an author node and a paper node) and co-authorship (between multiple authors with regard to a common paper), author-citation (between two author nodes via their papers), and *paper-citation* (between two paper nodes). Accordingly, we differentiate three types of graphs (or networks): co-authorship graph which represents a co-authorship network highlighting the collaboration relationship among authors, author author-citation graph which represents a citation network highlighting the citation relationship happened between authors, and paper-citation graph which represents a citation network where papers directly refer to each other. Evidently, citation relationship is a weaker relationship as compared to co-authors relationship, since authors who cite each other's work may not actually know each other, while authors who collaborate on a common publication must (usually) have already known each other. As examples, graph G1 in Figure 1 illustrates co-authors relationship between authors (where, A1 and A2 are co-authors of paper P1; P2 is solely authored by A2; P3 is co-authored by A1 and A3; and P4 is co-authored by A2 and A3); graph G2 is an author-citation graph (where, A1 cites A2 and A3; A2 cites A4; A3 cites A2; and A4 cities A1 and A3); and graph G3 is a paper-citation graph (where, P1 cites P2 and P4; P2 cites P4; P3 cites P2; and P4 cites P3).

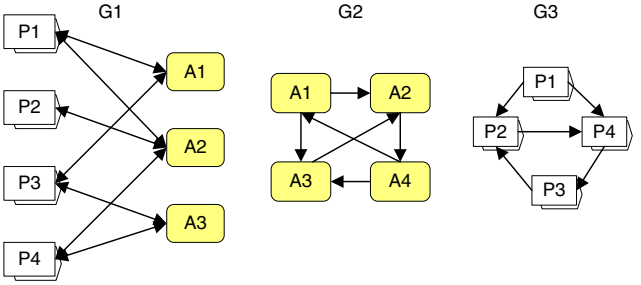

Figure 1.
Examples of co-authorship graph (G1), author-citation graph (G2), and paper-citation graph (G3)

2.4 Author ranking and expert finding

Author ranking is to computationally decide the ranks of authors with respect to their research output and performance as compared to other authors. The criteria of this ranking may include variables such as number of publications, number of citations, ranks (or impact factors (IFs) of publication venues, etc. Author ranking is used to discover recommended experts in a particular academic discipline. Therefore, the terms, author ranking and expert finding, have been used alternately in the literature as well as in this survey.

2.5 Evaluation measures

Evaluation of ranked results is challenging as the ground truth is unavailable. This phenomenon makes the task of performance measuring and evaluation hard and tricky. The familiar, traditional measures of recall and precision are not straightforwardly applicable to author ranking. Researchers need to manually identify a set of relevant truth values that can be used for this purpose. Thus, varied data sets have been picked by various researchers for evaluating their work. This makes the comparison of the ranking methods a challenging task. Common data sets and standard benchmarks for author ranking are very much in need.

2.6 Author name disambiguation

Resolving the name ambiguity in bibliographic databases is called author name disambiguation. A name can either shared by multiple authors or multiple variant names of a single author can also create ambiguity. Techniques are required to disambiguate the authors from one and another before finding their rankings so that correct ranks can be assigned to them.

3. Classification of author ranking methods

In this study, we have classified the methods according to their functionalities, i.e. the way or method of solving a problem. The key objective of all the methods is same, i.e. the ranking of authors and it is achieved by different approaches and by considering different features as weights. Based on an extensive and in-depth review of the related literature, we discovered that almost all the author ranking methods we reviewed are centered on the key functionality of each respective method. Therefore, these methods are best classified according to their key functionalities and our survey of these methods is best set forth according to the categories of these methods.

Figure 2 depicts the classification hierarchy of the major methods that we have included in this survey. At the top level, these author ranking methods are classified into three main categories: link analysis methods, text similarity methods, and learning-based methods. The first main category, link analysis methods, includes the methods that calculate the rank of an academic object, specifically authors, by taking into account the linkage structure of a relevant graph. These methods are further divided into two subclasses: iterative and bibliometric methods. The iterative methods follow a number of iterations to calculate the ranks, while the bibliometric methods are based on some sort of calculation involving the bibliometric citations. The second main category, text similarity methods, unlike the first category, finds some similar text from relevant text data and utilizes the data for the calculation of rank. The third main category, learning-based methods, applies the machine learning approach and classification rules to compute the ranks of academic objects. Table I, from a historical view, shows representative author ranking methods as they are related to the three main categories we introduced above.

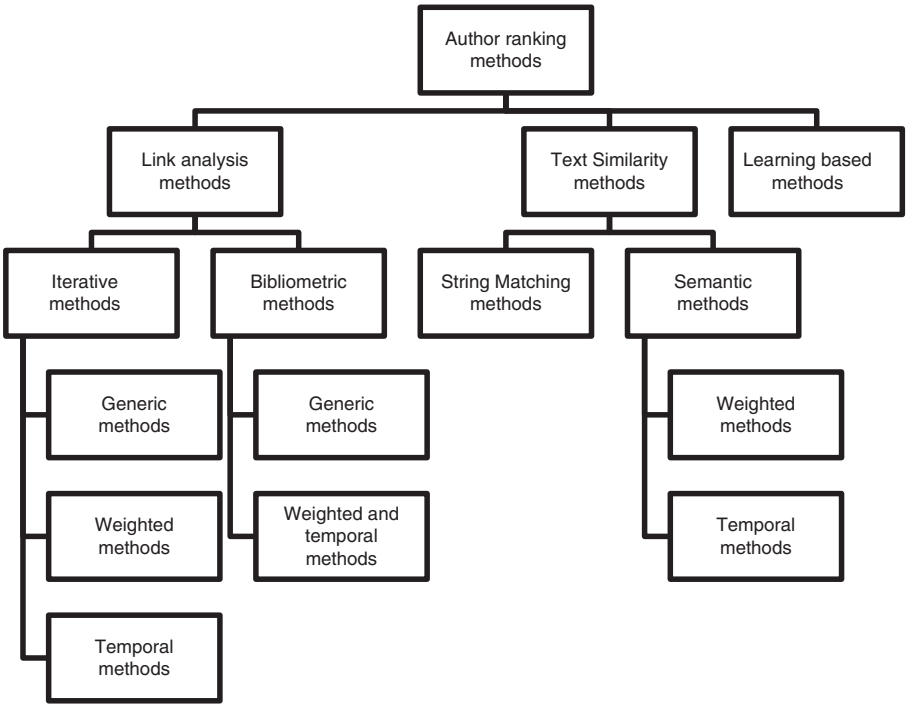

Figure 2.
Classification of ranking methods

3.1 Link analysis methods

Link analysis methods model the ASNs as graphs in which the nodes represent actors and the edges characterize interactions between these actors. These methods are further classified into subcategories: iterative and bibliometric methods. We address each subcategory details below.

3.1.1 Iterative link analysis methods. In iterative link analysis methods, a set of instructions is computed repeatedly until a stopping criterion is reached or until algorithm converges. A good example of this subcategory is the PageRank algorithm (Page *et al.*, 1999). Its basic idea is that a page is considered to have more significance if many other important pages points towards this page. This means that the rank of a page is distributed among other pages linked to it. Thus, the rank of a page is computed in an iterative manner. Iterative link analysis methods, particularly for author ranking, can be further differentiated as generic, weighted, and temporal methods.

3.1.1.1 Generic methods. This type of author ranking methods utilizes the link structure of the relevant nodes while computing the rank of the authors. According to the basic concepts of PageRank (Brin and Page, 1998; Page *et al.*, 1999), a page is termed essential if the pages linked to it are important. It was pointed out that there is an important linkage between PageRank and citation analysis.

There are few limitations of PageRank, which are relatively new pages, although important or of high quality, but as having fewer links to it, may be unfairly ranked lower; the initial distribution of rank is equal for all the pages without any differentiation of good or bad quality; the rankings provided are not content-based as they only analyze the link structure; and it is easy to tamper because people can make fake pages pointing to a page in order to increase its rank.

Table I.
Historical paradigms of author ranking methods from 1999 to 2017.

	Link analysis		Text similarity			
Year/type	Bibliometric	Iterative	String matching based	Semantics based	Temporal	Learning based
1999		HITS (Kleinberg, 1999), PageRank (Page *et al.*, 1999)				
2005	h-index (Hirsch, 2005)	Author-Rank (Liu *et al.*, 2005)				
2006	g-index (Egghe, 2006)		Formal models for expert finding (Balog *et al.*, 2006)			
2007	R and AR indices (Jin *et al.*, 2007), H-core (Burrell, 2007a), m-quotient (Burrell, 2007b)					
2008	Kth rank (Sekercioglu, 2008), successive g-index (Tol, 2008), weighted h-index (Egghe and Rousseau, 2008)	PR for bibliographic networks (Fiala *et al.*, 2008), expertise search with temporal dimension (Li and Tang, 2008)		Mixture model (Zhang *et al.*, 2008), simultaneous modeling of papers, authors and venues (Tang *et al.*, 2008)		
2009	Weighted h-index (Zhang, 2009)	Weighted PR (Ding *et al.*, 2009), diffusion of scientific credits (Radicchi *et al.*, 2009)			Semantics and temporal maven search (Daud *et al.*, 2009b)	Expert2Bole (Yang *et al.*, 2009)
2010	Consistent ranking of authors and journals (Bouyssou and Marchant, 2010)		Dependencies in expert finding (Yang and Zhang, 2010), integration of multiple features for expert finding (Zhu *et al.*, 2010)		Expert finding with topic modeling (Daud *et al.*, 2010)	Discriminative models (Fang *et al.*, 2010)
2011		WPR (Yan and Ding, 2011), WPR (Ding, 2011a),				
Link analysis algorithm		Topic-based PR (Ding, 2011b), Rare Rank (Wei *et al.*, 2011)		Learning-based method (Moreira *et al.*, 2011)		

(*continued*)

	Link analysis		Text similarity			
Year/type	Bibliometric	Iterative	String matching based	Semantics based	Temporal	Learning based
(Gollapalli *et al.*, 2011)						
2012		Ranking of publications and authors (Gupta *et al.*, 2013)	Venue based ranking (Daud and Hussain, 2012)		Temporal Author Topic Model (Daud, 2012), time aware PageRank (Fiala, 2012)	
2013	Co-author core based indices (Ausloos, 2013)			TWFG (Lin *et al.*, 2013), Author-Document-Topic graph model (Gollapalli *et al.*, 2013)		
2015		Mutual influence based ranking (Amjad, Daud, Che, Akram, 2015)		THR (Amjad, Ding, Daud, Xu, Malic, 2015)		
2016		Mutual influence based ranking (Amjad, Daud, Akram, Muhammed, 2016)				
2017	Domain-Specific Index (Amjad and Daud, 2017)					

Table I.

HITS is another popular method that utilizes the network and link structure for ranking the web pages (Kleinberg, 1999). Jon Kleinberg contributed with the important concepts of Hub and Authority in his design of HITS. Hubs are the pages that serve as substantial catalogues, linking up the pages with actual information. Authorities are pages that actually contain the required data, and are pointed to by (many) hubs. Hub and authority score is calculated for all pages. Generally, a reliable hub should provide links of many authorities and a reliable authority is the one that is pointed by many hubs. It is not uncommon that many pages can simultaneously act as a hub and an authority. Unlike PageRank, HITS is query-dependent as page scores are calculated at query time (considering page contents). This is both an advantage and a disadvantage because HITS requires a neighborhood graph to be built at query time. It also makes HITS vulnerable to spamming. People may tamper HITS by adding links to and from their pages, and influence the hub and authority scores of their pages. The computational complexity of PageRank and HITS is a challenging issue as the two algorithms are both based on eigenvector. PageRank and HITS are general ranking methods, for the ranking of generic entities; and they provide the baselines for more domain-specific ranking methods. In the following, we address the ranking methods that are more specific to author ranking.

Most of the author ranking techniques consider authors or papers of authors, instead of general web pages, as nodes in their graph structures. These graphs typically model co-authorship networks or citation networks, as we introduced in Section 2.

Fiala *et al.* (2008) presented a variation of the original PageRank algorithm based on co-author graphs, incorporating the number of citations as well. They tested the performance of an algorithm by applying it on data from DBLP digital library. The ranking results were compared to the victors of the ACM E.F. Codd Innovations Award. Results show that the proposed algorithm works well as compared to the original PageRank algorithm. The major limitation of this technique is that it initially distributes the rank equally among all authors, like the original PageRank method.

Ranking methods have also been exercised for expert finding in a certain domain. Gollapalli *et al.* (2011) proposed an application of PageRank for expert finding. Expert finding models normally use the documents as confirmation of expertise while ranking authors. Unlike most of the models, Gollapalli *et al.* also used other sources of evidence like association of a document with its venue, and number of citations. They presented a link analysis based model for integrating multiple sources of evidence in the PageRank algorithm to rank experts. The claim that the proposed method is applicable for other academic objects like venues is not empirically validated.

Some iterative methods were proposed that focus on finding the rising stars from the research communities (Daud *et al.*, 2013, 2017; Li *et al.*, 2009). The rising star means the researchers who are starting their career and have a potential to rise in their near future. A detailed analysis and study was conducted by Amjad *et al.* (2017) to find the correlation between success and co-authorship of a junior researcher with a senior researcher.

3.1.1.2 Weighted methods. Various extensions of PageRank have been presented in the literature, specifically for ranking of authors. The parameters, e.g., number of publications or number of citations have been added to the original PageRank as weights.

Liu *et al.* (2005) applied different measures of centrality like degree, closeness, betweeness and PageRank to co-authorship networks extracted from digital libraries (DL). An Author-Rank algorithm (a weighted version of PageRank) was proposed for finding the rank of an author in an undirectional co-authorship network considering the frequency of collaboration. They studied author centrality within a co-authorship network to calculate the status of an author. The co-author network was derived from the "Advances in Digital Libraries (ADL)", "DL" and "Joint Conference on Digital Libraries (JCDL)" conferences

from 1994 to 2004. Weighted and directed network was chosen to represent the collaboration relationship among entities. A range of centrality measures was applied to study this network, which is an alternative centrality metric to study the properties of similar networks. In this study, authors tried to eliminate the major limitation of the PageRank algorithm, i.e. the initial uneven distribution of rank to all nodes. Their proposed initial distribution depends upon co-authorship link weights which were calculated from the co-authorship frequency of all authors. The results show that Author Rank and PageRank show similar performance. However, both are better than other measures like degree, closeness and centrality. For validation of their results, they compared them with previous program committee members of same DL. The comparisons assume that the program committee members are esteemed researchers.

Radicchi *et al.* (2009) presented a subjective form of PageRank method for the positioning of authors. A weighted-directed co-citation network was used for this study. The ranking was performed by taking into account the diffusion of credits exchanged by authors. The main focus of study was to distinguish the citations, and to give more importance to the references coming from prominent authors as compared to less prominent ones. The study also involves the non-local nature of the diffusion process in which any author can affect the credit of any other far away author. The results articulate that the weighted indicators acquired by the proposed method correlate with the scientific accomplishments calculated by some prestigious prices from the domain of physics, including Wolf prize, Nobel prize, Dirac medal, Boltzmann medal and Planck medal.

Ding *et al.* (2009) presented an application of the basic PageRank algorithm for finding author's impact. They considered a co-citation network as a test bed, and proposed a weighted PageRank algorithm. An author co-citation network was constructed by extracting the data of 108 authors who have received more than 200 citations. They studied the correlation between different damping factors like 0.85, 0.75, 0.65, 0.55, 0.35, 0.25 and 0.15. The results have been compared with h-index, centrality measures and citation ranking. The findings are as follows:

- citation ranking is highly correlated with PageRank;
- citation rank and PageRank do not show correlation with results of centrality measure; and
- h-index and centrality measure are correlated.

Later, Yan and Ding (2011) presented an application of the basic PageRank algorithm to find author's impact. The major difference from the study of Ding *et al.* (2009) is that this work considers the undirected co-authorship network instead of co-citation network for the experimentation. First, they evaluated the correlation of PageRank with citation ranking. Second, a range of different damping factors (0.15-0.85 with increment of 0.1) was tested to evaluate their effect on ranking results. Third, they proposed a weighted PageRank algorithm named Author-Rank that uses total number of citations as weighting criteria. Finally, they evaluated the performance of the proposed Author Rank with h-index, citation and PC members of ISSI conferences. The results prove that the proposed method performs better.

Ding (2011a) studied the application of weighted PageRank for the author-citation network. The main idea behind the approach is to find out that how different weighted PageRank algorithms can be used to find the popularity and esteem of a scholar. The network model under consideration was a directed weighted graph and the weighting criteria were the citation count and publication count. Nodes with the higher weights have a higher chance of being visited by a random surfer. Different damping factors are used:

- 0.15 to represent an equal opportunity of getting a citation;

- 0.5 to represent that the scholarly articles normally take a short path of length equals to 2; and
- 0.85 to represent a network topology.

They compared the PageRank and weighted PageRank algorithms with popularity rank and prestige rank presented by Ding and Cronin (2011) for highly cited authors. The results show that "popularity-rank" and "prestige-rank" holds correlation with PageRank and Weighted-PageRank. However, in case of finding a prize winner, the prestige rank performs better than all other measures used.

If we see the formula of PageRank algorithm, we can divide it into two parts. The first part is static and involves the damping factor and total number of nodes. The second part is actually the dynamic and iterative part. In articles of Ding (Ding, 2011a; Ding *et al.*, 2009; Yan and Ding, 2011), the modification is done in the first part of the formula, and weights are added to it. This leads to the fact that major limitations of PageRank are still there, because it divides the initial rank equally to all its nodes, whether they are significant or not.

Hong and Baccelli (2012) presented an addition to PageRank family. They performed the simultaneous ranking of papers and authors. For this purpose, they used the bipartite graphs as there are nodes of two types. They presented an extension of page rank algorithm: PR-G for a global graph for both authors and papers, PR-A for the author's graph, and PR-P for papers. They simulated different scenarios to compare these extensions with existing methods like h-index and citation count. They demonstrated that better qualitative results can be attained by incorporating author paper graphs.

Two weighted versions of PageRank, MuICE and MINCC were presented by Amjad, Daud, Che, Akram (2015) and Amjad, Daud, Akram, Muhammed (2016), respectively, which includes influence of co-authors on an author. They argued that it is not only the progress of an author what is important of his/her ranking but the influence of the co-authors on the author is very significant, especially when co-authors are senior. In MuICE, the conjoint effect of co-authors on each other was considered in terms of the total number of papers, the total number of received citations and the publications as a first author. Apart from that, the presence of increasing number of exclusive authors in citing papers was also considered. The results show that ranks of authors are not only determined by their own publications and citations but they were influenced by the progress of their co-authors significantly. MINCC considers the mutual influence among authors according to their number of publications. Apart from that, they also incorporated the normalized weight of citations according to position of author's name in the paper. Table II provides a summary of generic and weighted iterative methods to provide a quick overview for the readers.

3.1.1.3 Temporal methods. Time weighted ranking involves the consideration of temporal dimension while ranking of authors. Most of the users require up to date information while searching answers for a query. The temporal dimension can be very important in many ways. With time information, one can estimate the experience or seniority of an author. The authors can also change their field of interest with time. For example, author "A" is an expert of Data Mining and he publishes many papers in his field till 2006. Later on, he changes his interest and starts focusing on social networks. In 2012, he can be called an expert of Data Mining, if time factor is ignored while ranking. On the other hand, if we involve the time factor, there may be many other authors who worked on latest trends in Data Mining while author "A" was focusing on social networks. Naturally, the experts ranked according to a specific topic in a given time cannot remain same forever.

Unfortunately, the temporal dimension has gained attention of the researchers mostly for ranking of general entities and not for authors specifically. Different methods for ranking of general entities are presented as Fresh information retrieval (Sato *et al.*, 2003), Timed PageRank (Yu *et al.*, 2004), temporal ranking for search engine results (Jatowt *et al.*, 2005)

Table II.
Summary of link analysis based generic and weighted iterative methods

Reference No.	(Liu *et al.*, 2005)	(Fiala *et al.*, 2008)	(Ding *et al.*, 2009)	(Radicchi *et al.*, 2009)	(Yan and Ding, 2011)	(Yan and Ding, 2011)	(Gollapalli *et al.*, 2011)	(Hong and Baccelli, 2012)	(Amjad, Daud, Che, Akram, 2015)
Study	State of the DL after 10 years of activity	Modification of PageRank algorithm	Effect of different damping factors	Diffusion of credits exchanged by authors	Effect of different damping factors	Measure popularity and prestige of an author	Expert finding	Ranking of papers and authors	Ranking of authors
Proposed	Author Rank for weighted directional network	Bibliographic PageRank	A weighted PageRank	Science Author Rank Algorithm (SARA)	A weighted PageRank	A weighted PageRank	A modified PageRank	Extension of PageRank	A weighted and modified version of PageRank
Network type	Weighted and directed co-authorship network	Co-authorship network	Co-citation network	Weighted and directed co-citation network	Co-authorship network	Weighted and directed co-citation network	Typed-graphs	Bipartite paper author graph	Co-authorship network
Weights	Co-authorship frequency	Number of citations	Co-citation frequency	Number of authors and their number of references	Citation counts	Number of citations and number of publications	Edge-type weights	Graph has nodes of two types, papers and authors	Number of papers, citations and first author publications
Comparison with	degree, closeness and betweenness centrality metrics	ACM E.F. Codd Innovations awards	Citation ranking, centrality measure, h-index	Citation Count, Balanced Citation Count	h-index, citation ranking, PC membership of ISSI conferences	PageRank, prestige rank, popularity rank	Probabilistic model	h-index and citations count	Citation count, Author statistics from aminer.org
Data set	ACM,-IEEE-and joint-ACM/IEEE digital-library-conferences	DBLP	108 most cited authors from 1970 to 2008 in field of IR	Publications from 1893-2006 in physical review	Taken from data set of [41]	Collected from Web of Science 1956-2008	ArnetMiner, UvT Expert Collection	Simulations on synthetic data	ArnetMiner

(*continued*)

Table II.

Reference No.	(Liu *et al.*, 2005)	(Fiala *et al.*, 2008)	(Ding *et al.*, 2009)	(Radicchi *et al.*, 2009)	(Yan and Ding, 2011)	(Yan and Ding, 2011)	(Gollapalli *et al.*, 2011)	(Hong and Baccelli, 2012)	(Amjad, Daud, Che, Akram, 2015)
Finding/ strengths	Author Rank correlates with PageRank	Proposed method outperforms PageRank	Proposed method correlates with PageRank	Proposed method shows improvement over Citation count and Balanced citation count	Proposed method outperforms PageRank	Proposed method correlated with popularity rank and prestige rank	Proposed model is flexible and can be used for other academic entities	Qualitative gain by considering authors and papers graph	Influence of co-authors of authors on their standing
Limitations	Main limitation of PageRank (all nodes are assigned same weight initially) exists	Penalize cited author for co-authorship with citing author	Main limitation of PageRank exists	Main limitation of PageRank exists	Main limitation of PageRank exists	Main limitation of PageRank exists	Main limitation of PageRank exists, all nodes are treated uniformly	Main limitation of PageRank exists, all nodes are treated uniformly	Main limitation of PageRank exists

and Time-Weighted PageRank (Manaskasemsak *et al.*, 2011). While Yu *et al.* (2005) added temporal dimension in traditional PageRank algorithm specifically for ranking research publications, they considered the reputation based factors for ranking and included publication date and dates of citations as a temporal factor with the help of the linear regression method.

Very little work was found that specifically addressed the problem of author ranking. Li and Tang (2008) applied the temporal aspects for the problem of expert finding. They combined the social network within the random walk model and modeled the time information by using the forward and backward propagation process. They represented the academic objects as nodes and relationships as links and modeled the time varying information. The basic idea was to divide the whole heterogeneous network into time slices (Gs) where G represents the graph and s represents the number of time windows. Though time dimension was considered, assigning all in-links to all authors of a paper and no consideration of the semantics are drawbacks of their work.

A time aware PageRank algorithm was presented by Fiala (2012). The proposed method modifies the PageRank algorithm in such a way that citations between different authors can be weighted depending upon the information extracted from the co-authorship graph. They also emphasized upon publication date and citation date. The weights to the citations are assigned on the basis when two authors have collaborated with each other. This information is extracted from the co-authorship network.

3.1.2 Bibliometric link analysis methods. Bibliometric methods are used to analyze data from the citation analysis to determine the impact of authors, journals/conferences and publications. These methods can be useful in measuring the output of a scientific research. They can also be helpful in giving an idea of how the researchers work and collaborate. These methods have a longer history than link analysis methods. In fact, the link analysis methods were inspired by bibliometric methods.

The bibliometric methods are normally quantitative and non-iterative in nature, which count the number of citations that have been made for a scientific work or paper. For example, IF is used to find the quality of a journal through an arithmetic mean of the number of citations to articles published in a journal. The journals that achieve high IF are considered more significant and prestigious as compared to the journals that have low IF. IF can only be used to compare the journals only within a field and index authors indirectly through their published papers in a journal. H-index (Hirsch, 2005) is a state of the art which can be used directly to index authors and other academic objects. The idea is that the author with a higher average number of citations has high h-index which is considered better. Bibliometric link analysis methods can be further categorized into three types: generic, weighted and temporal methods.

3.1.2.1 Generic methods. Hirsch (2005) proposed the h-index to calculate the research output and significance of worked performed by a researcher. Due to its wide range of application, it has become an important tool. The limitation of h-index is its insensitivity to one or several exceptionally highly cited papers. It considers that a highly cited paper is important, but it does not consider the actual number of citations. Also, if a paper is once selected in a top group and its h-index is calculated, any subsequent citations are not considered at all even if they double or triple.

Egghe (2006) proposed the g-index, a variant of h-index to remove the above-mentioned limitations of the h-index while keeping its benefits. Unlike h-index, g-index escalates when the number of received citations increases. To calculate the g-index, all articles of an author are organized in a decreasing order of number of citations, just like h-index. According to h-index, the papers on the rank 1..h have at least h citations. So overall they have at least h2 citations. But it is not necessary that h is the highest rank, whereas in case of g-index,

the first g papers should have a minimum number of g2 citations together. It is shown that the g-index has a greater discriminatory power than that of h-index (Tol, 2008).

F-index was proposed by Katsaros *et al.* (2009) which is against the idea of totally removing the effect of self-citations. It studies the significance of an article not only by its received citations, but also by the presence of unique authors. An increasing number of exclusive authors as citing authors of a paper represents wider penetration of the work.

Previous methods did not consider the paper and the journal's importance simultaneously. Bouyssou and Marchant (2010) argued that a uniform method is required for the ranking of journals as well as authors because the quality of journal and quality of author both are related to the work published in journals. The perception behind the idea is that the researchers with high prestige publish in highly ranked journals, hence both entities are interrelated. It must be noted that h-index and its variants emphasize more on quantity than quality, because they focus on the number of citations of a paper, hence more focus on visibility of an author. Another point of debate is the chance of inconsistency that may arise due to self-citation. Most of these indices consider the papers as if they have a single author, neglecting the impact of co-authors, as normally the papers are co-authored.

Ausloos (2013) presented a scientometric method to measure the impact of an author focusing the co-author's core. Instead of directly counting the number of citations, they measured the performance of a researcher in a scientific network. They focused on the role of co-authors by measuring their co-authored papers and their received citations. In the proposed solution, they showed that a co-author C has J papers co-authored with one or numerous collaborators. By arranging all co-authors of that researcher according to the number of their co-authored papers and giving them rank r, starting from $r = 1$ to most creative one, we can get a relationship of J inversely proportional to r. Based on Hirsch core they proposed "co-author core," and introduced ma and aa indices. Domain-specific index (DSI) was presented by Amjad and Daud (2017) which was based on h-index but unlike h-index it assigns the weight of citations to the authors considering the interest of that author in that specific topic. Authors can be interested in more than one topic and can have a different level of expertise in all their fields. DSI was capable of finding their distinct domain-specific index in all their fields of interests. Table III provides a summary of generic bibliometric author ranking methods to give the reader a quick overview.

3.1.2.2 Weighted and/or temporal bibliometric methods. Burrell (2007b) stated that the number of years is an important factor in author ranking and proposed the m-quotient. To calculate m-quotient, the h-index of the authors is divided by the total years of his academic career. Though weighting factor in terms of years is considered, all the authors in a paper are given the same credit for the contribution. Consequently, Sekercioglu (2008) presented the kth-rank index to measure the contribution of co-authors in a paper instead of giving an equal contribution to all authors in a paper having multiple authors. According to the kth-rank index, every co-author contributes 1/k of the first author. The h_w index (a citation-based weighted h-index) was proposed by Egghe and Rousseau (2008) which considers the performance changes. Egghe (2008) also presented fractional h and g-indices. The study includes h-index and g-index of authors, when authorship of the cited articles is counted in a fractional way. According to Egghe, there are two possible methods for this purpose. One is to count the citations of these papers in a fractional way and the other is to rank papers in a fractional way to give credit to an author.

The temporal dimension is also considered by the methods of bibliometric category. Jin *et al.* (2007) presented R and AR indices, which complement the h-index. The R-index measures the number of citations in h-core, while AR also considers the age of publications. This creates an index that can actually increase and decrease with time. It is calculated by

Ref No.	(Hirsch, 2005)	(Egghe, 2006)	(Katsaros *et al.*, 2009)	(Bouyssou and Marchant, 2010)	(Ausloos, 2013)	(Amjad and Daud, 2017)
Year	2005	2006	2009	2010	2013	2017
Proposed	h-index	g-index	f-index	Consistent ranking of authors and journals	Co-author core	Domain-specific index
Data set	Highly cited researchers from 1983-2002	Citation data of two authors: L. Egghe and H. Small	(www.cs.ucla.edu/~palsberg/h-number.html	Used mathematical proofs	Publication list of a research group, SUPRATECS	Arnetminer
Finding/ strengths	Consideration of author own paper and citations is intuitive as compared to considering the impact factor in which he/she publishes	Giving author credit for his/her highly cited papers is important	Self-citations, impact can be considered through number of authors in citing papers instead of ad-hoc weight assigned to self-citations	Indexing of authors also depends upon the journals in which they publish	Finding value of an author in co-author's core	Ranks authors according to their topic specific citations instead of whole citations
Comparison with	Impact factor	h-index	h-index	h-index	h-index	Ad-hoc h-index
Limitations	Considers all citations, can be manipulated by self citations, citations and publication dates are not considered. Topic insensitive	Citations and publication dates are not considered. Topic insensitive	Citations and publication dates are not considered. Topic insensitive	Citations and publication dates are not considered. Topic insensitive	Citations and publication dates are not considered. Topic insensitive	Citations and publication dates are not considered.

Table III.
Summary of link analysis based bibliometric methods

obtaining the square root of the total number of citations present in the Hirsch core to calculate the index.

Burrell (2007a) presented another temporal method for measuring author's research output. The h-index identifies the most constructive core of the researchers output according to their received citations. Burrell called this core as h-core, and studied the size of h-core. They also studied the A-index and emphasized upon time dependent and dynamic nature of these measures. The interrelationships of these measures are also discussed and it is found that A-index is a linear function of h-index and time. It is also established that the h-core has an approximate square-law relationship with h-index, A-index and time. Table IV provides a summary of weighted and temporal bibliometric author ranking methods to give the reader a quick overview. Two somewhat similar methods named Consistent Annual Citations-Index (Daud and Muhammad, 2014) and Variation-Index (Daud, Saleem Yasir, Muhammad, 2013) were proposed that considered how consistently a paper of a researcher is able to get attention of other researchers over a period of time (in years) in a scientific community.

3.2 Text similarity methods

Text similarity methods identify the similar text in query terms and the document set to rank authors. These methods are categorized into string matching and and semantics-based methods.

3.2.1 String matching based methods. String matching methods simply match the query words with the publication set of authors without considering semantics. Some string matching methods also involve the use of language models (LMs) or probabilistic models. The term LMs are used interchangeably with the term Probabilistic models. It was first introduced by Ponte and Croft (1998). A LM or probabilistic model finds the probability of generating a query q given document d: $p(q|d)$. The documents which have a larger probability value are ranked higher.

String matching methods were used for the problem of expert finding and author ranking. Balog *et al.* (2006) addressed the following question "what is the probability of a candidate ca being an expert given the query topic q?" They presented two general strategies for expert finding within a document collection by using generative probabilistic models. The first method modeled the authors by using the documents they are directly associated with. The second model founnd the documents related to a topic first and then the authors associated with them. Both models ranked the candidates based on the likelihood that the applicant is an expert with respect to a query topic, but the models differed in the way they do it. In the first model, they created a textual depiction of the knowledge of individuals with respect to the documents with which they are related. From this representation, they evaluated the probability of the query topic to rank the candidates. The second model ranked the documents according to the query and found likelihood of the candidate to be an expert by taking into account the set of related documents.

Yang and Zhang (2010) utilized the associations that may exist among the query terms for expert finding by saying that only modeling dependencies among the authors and query terms is not enough. They proposed a method based on language modeling by fusing two kinds of dependencies within an integrated framework. Zhu *et al.* (2010) presented a language modeling method for expert finding that incorporates multiple features. The idea behind this scheme is that multiple document features can affect the expert finding. Document features can include several levels of relationships between specialists and a query topic. Results show that process of expert finding can be improved to achieve better results if these document features are also incorporated.

Daud and Hussain (2012) argued that the existing LMs perform text-based matching of query terms with the documents of candidate experts and do not consider the venue

Table IV.
Summary of weighted and temporal bibliometric author ranking methods

Ref No. Year	(Burrell, 2007b) 2007	(Jin *et al.*, 2007) 2007	(Burrell, 2007a) 2007	(Sekercioglu, 2008) 2008	(Egghe and Rousseau, 2008) 2008	(Egghe, 2008) 2008
Proposed	m-quotient	R and AR indices	Time-dependent study of h-core and A-index	kth-rank index	Weighted h-index	Fractional h and g-indices
Data set	Synthetic example[a]	Multiple data sets like Web of Science, World Science Series, Price Awardees	Synthetic example	Synthetic example	Publications in 2000 of five small European countries	Synthetic example
Findings/ strengths	Consideration of authors research activity years is important to make junior and senior authors comparable	An index that increases and decreases with time	Time-dependent behavior of A-index and size of the h-core	Giving same contribution to all authors of a paper is intuitively not right as different authors have different contribution	Effect of continuous and discrete setting for rank of scientists	Fractional way is adopted to study status of the cited articles
Comparison with	h-index	h-index, A-index, g-index	h-index, A-index	h-index	h-index, R and A indices	Non fractional h and g-indices
Limitations	Topic insensitive, author's position not considered	Topic insensitive	Topic insensitive	Topic insensitive, time of publication not considered	Topic insensitive	Topic insensitive, author's position not considered

Note: [a]Synthetic examples: method not tested on real world data set

of publication before ranking them. They presented a language modeling method that takes into account the publication venue while ranking the experts in a given field named as Influence language modeling for expert finding. The results show that papers which are published in a high level venue are more appreciated than the papers which are published in a low level venue and should contribute while ranking authors. Usmani and Daud (2017) presented a unified method for ranking of authors with consideration of publication and venue of publication at the same time. Table V provides the summary of text similarity methods used for the author ranking problem.

3.2.2 Semantics-based methods. Semantics-based methods are employed to capture semantics between author publications and query terms. Probabilistic Semantic Analysis (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) are state-of-the-art models used for semantically ranking authors. Semantic ranking methods consider the polysemy/synonymy of words while ranking the authors in contrast to text similarity based methods which follow exact word matching and ignore semantics.

Semantic web is an extension of standard web with machine readable metadata. Semantic search aims to improve the accuracy of the results retrieved by the search engine by understanding the intention of the searcher. While ranking semantically, a search engine needs to involve the meaning of a web page or document to find its relevancy to a given query. Swoogle (Ding *et al.*, 2004) and XSEarch (Cohen *et al.*, 2003) are the example of semantic ranking.

Semantic ranking has become very important with the growth of available data. Consider a scenario where the documents are semantically related to the field of expertise of an author. If we are using common text similarity matching techniques like TF-IDF to find an author related to a query, we will not be able to find the most relevant results. Therefore, involving semantics is necessary for such scenarios. Generally, the semantic ranking

Ref No. Year	(Balog *et al.*, 2006) 2006	(Yang and Zhang, 2010) 2010	(Zhu *et al.*, 2010) 2010	(Daud and Hussain, 2012) 2012
Study	Expert Finding	Dependencies between query terms in expert finding	Integration of multiple document features for expert finding	Expert finding with respect to publication venue
Proposed	Generative probabilistic models	Language modeling based method	Language modeling based method	Influence language models
Data set	TREC Enterprise	TREC Enterprise	TREC Enterprise	DBLP
Findings/ strength	Model that traces documents on topic, and then finds the related experts	Usefulness of the window-based model and the combined framework. Dependencies within the query were exploited for expert finding	Formal methods for accommodating numerous document attributes for expert finding	Influence language modeling performs better than existing language modeling methods
Evaluation measures	MAP, Precision, MRR	MAP	MAP, precision	Citations count taken from Google Scholar
Limitations	Based on intra-document frequencies, dependencies between the query terms were ignored	texts that do not match the window-n restriction are removed, so classical language model is also combined. multiple document features were ignored	Relationship of query terms with multiple document features was not considered	Dependencies or relationships between query terms and multiple document features were not considered

Table V.
Summary of text similarity methods for author ranking

methods do not directly model the relevance of a query and document. Rather, they use a latent semantic layer to model relevance between a query and document. That is why the authors whose support documents are associated with the same theme layer are ranked higher, even if they do not contain the query terms.

3.2.2.1 Weighted methods. Various researchers tried to include semantics in a ranking method. A mixture model for-expert finding was presented by Zhang *et al.* (2008). They showed the importance of semantics by assuming that there is a hidden "semantic" layer $\Theta = \{\theta_1, \theta_2, [\ldots], \theta_k\}$ between query q and document d_j. Each hidden theme θ_m is semantically related to several queries and support documents. For example, in scenarios where we try to search an author related to a query by exact matching a person who belongs to semantically related words of that query, will not be retrieved through common TF-IDF matching techniques. This means we need to introduce semantics in such cases. They attempted to identify the semantic knowledge that relates the query term and support documents. They used Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) to present a mixture model to identify the semantic knowledge. Prior to that, simple LMs have been used for the purpose, but with the use of PLSA the semantics are involved to capture the hidden themes. In this way, a specialist whose support documents are related to the semantics of a query is ranked higher even if the query terms are not present in the documents. They tested the proposed method on ArnetMiner and found better results.

Tang *et al.* (2008) presented another topic modeling method for modeling not only authors but also the papers and publication venues. The main idea is to model authors, papers and publication venues together by using a probabilistic topic model. They assumed that modeling all these objects separately would produce unsatisfactory results, therefore they proposed a simultaneous modeling. For this purpose, three different variations of Author-Conference-Topic (ACT) model are proposed to achieve better performance. The ACT models are used with random walk models. This is a very general approach, and different combinations of topic model and random model can be applied in different ways. It is also applicable in fields of blog search and social search.

Ding (2011b) presented a topic-based algorithm-for-ranking of authors. The main idea behind this approach is to enhance the semantics of ranking authors by topic-dependent ranks based on the mixture of a topic model and a weighted PageRank algorithm. They combined the LDA (Blei *et al.*, 2003) with PageRank for author co-citation networks. LDA was used to capture the topic-wise features of nodes by assuming a hidden structure for a set of topics that link the words and documents. They used the ACT model to calculate topic distribution of publications containing titles and authors. The corresponding PageRank scores and topic distributions of each five topics for the top 100 highly cited authors were selected.

Wei *et al.* (2011) presented Rare Rank algorithm, which is an extension of the PageRank algorithm. It semantically ranks the documents by modeling the behavior of a researcher rather than a random surfer. The main idea behind the model is the presence of a knowledge base which contains a terminological topic ontology and academic objects including researchers, journals and/or conferences and publications. This simulates the environment in which a researcher is searching for some required documents and his behavior is based upon his rational thinking, rather than randomness. To generate the results, the link information, for example, citations and the content information are combined. The damping factor is changed to 0.95 to model the random walk factor that is reduced. First of all, the knowledge base is represented as a directed graph. Then the domain topic ontology is applied to the graph to enhance its semantics. The similarity measure is calculated by using LDA (Blei *et al.*, 2003). The transition probability matrix is constructed in a manner that reflects behavior of a researcher. Thus, the derived results naturally contain both the relevance and the quality. Lin *et al.* (2013) presented a topic sensitive approach for expert finding.

They presented a combination of content-based methods and link structure based methods in such a way that they can jointly consider all the personal information about an author as well as network information. The data set is taken from the Web of Science ranging from publication from 2001 to 2008. They presented a topical weighted factor based graph model (TWFG), and compared it with existing methods like topic-based PageRank presented in Ding (2011b) and topic model and citation count. As an evaluation measure, they considered the convergence rate of SIGIR PC and NDCG, and found better results than existing methods. Gollapalli *et al.* (2013) presented Author-Document-Topic graph-based models for expert finding. They considered a tripartite graph structure that contains authors, documents and topics as nodes. They presented two models, which can respond to name-based and topic-based queries. First method is an extension of PageRank for the graphs which have multiple type of edges. The second method is based on weighted, undirected, tripartite graph having authors, documents and topics for finding content-based similarity via document-topic edges. Two data sets are used for experimentation: a subset from ArnetMiner and Citeseer and UvT collection. The comparison of the proposed methods is performed with okapi BM25. Results show that new methods are good enough to provide a unified framework to rank authors in response of two types of queries: the name-based queries and topic-based queries. Topic-based Heterogeneous rank was presented by Amjad, Ding, Daud, Xu, Malic (2015) in which authors, papers and venues were ranked simultaneously considering the effect of all on each other. Apart from simultaneous ranking of these academic entities, the main limitation of PageRank, i.e. assigning the same rank to all nodes initially was also addressed. ACT model was used to assign initial ranks based on topic-wise probability distribution of all entities. The data set was retrieved from Web of Science. Topic-based Heterogeneous Rank can combine information about publications, authors and journals/conferences to realistically rank academic entities in a heterogeneous environment along with impact of their topics. Amjad, Bibi, Shaikh, Daud (2016) presented a method for ranking of authors by assigning topic-based weights of received citations to multi-authored papers. Table VI provides a summary of semantic topic-based methods for author ranking. The methods discussed in next subsection are also based on semantics but are described under a different category because along with semantics, they also incorporate the temporal dimension.

3.2.2.2 Temporal methods. Some of the semantics-based methods have also included the temporal dimension for ranking of authors. Hence, the methods discussed in this section not only consider the semantics, but also include the time dimension. A comprehensive topic modeling method for maven search is presented by Daud *et al.* (2009) to identify a person with a given expertise. They presented "semantics and temporal information based maven search (STMS)" method to find out the hidden topics between the authors, venues and time simultaneously. According to the STMS method, within a venue c, every author from set of K authors is represented as a multinomial distribution θ_r over topics and every topic is a multinomial distribution Φ_z over words and multinomial distribution Ψ_z representing year of venue for that topic. They also elaborated the inference making process for topics and authors of new venues and how author correlations can be discovered. They also explained the bad effect of sparseness of data in the information retrieval process.

Daud *et al.* (2009c, 2010) presented temporal expert finding methods with the help of topic modeling. These two models were based on ACT model proposed by Tang *et al.* (2008). Daud *et al.* (2009c, 2010) not only considered the semantics, but also involved the temporal aspect. This ensures that users can find out the experts specifically interested in a field in a given time period. Generic semantics-based models are not capable of finding similar topics in different years. As highly dynamic data keep on changing with time, reflecting the ups and downs in trends is important. In these models, the conference influence and time information are simultaneously modeled.

Table VI.
Summary of Semantic Topic-based ranking methods

Reference No. Year	(Zhang *et al.*, 2008) 2008	(Tang *et al.*, 2008) 2008	(Ding, 2011b) 2011	(Wei *et al.*, 2011) 2011	(Lin *et al.*, 2013) 2013	(Gollapalli *et al.*, 2013) 2013	(Amjad, Ding, Daud, Xu, Malic, 2015) 2015
Study	Use of Probabilistic Latent Semantic Analysis to find hidden semantics between term and documents	Simultaneous modeling papers, authors, and publication venues	A combination-of-topic-model-weighted PageRank	Model the behavior of a rational researcher, rather than a random surfer	A grouping of subject based and link structure based methods	Ranking of experts	Simultaneous modeling papers, authors, and publication venues, with respect to topic
Proposed	A mixture of composite, and hybrid language model	Unified topic modeling method and its assimilation with random walk	Topic-based PageRank for author co-citation network	Algorithm-for-ranking entities-in-semantic search-applications	A topical weighted factor graph model (TWFG)	Graph-based model for expertise retrieval	Method for topic-based ranking in a heterogeneous graph
Comparison with	Language models	Language models	PageRank	PageRank	Topic model, citation count, and topic-based PageRank (Ding, 2011b)	BL (Okapi) and BL(Prob)	Non-topic-based ranking
Data set	ArnetMiner	ArnetMiner	15,367 papers, 350,750 citations and 25,762 authors from 1956 to 2008	IRIS2 publication-ontology and-knowledge base ACM-SW	Web of science from 2001 to 2008	ArnetMiner, citeseer, and UvT expert collection	Information retrieval based data from Web of science
Evaluation measure	Precision, MAP	Precision, MAP	Principle Component Analysis	Precision and NDCG	Coverage rate of SIGIR PC and NDCG	Precision, Recall and MAP	Qualitative analysis
Finding/ strengths	Use of a latent theme layer between a query and documents to find semantically related experts of a topic	probabilistic topic model for authors, papers and publication venues can be used for effective ranking	PageRank was modified and topic-based ranking of authors was derived	Rare Rank algorithm facilitates the rational researcher, instead of a random surfer	TWFG has better results when coverage rate of SIGIR PC is considered	Unified framework ranks experts for names and topic queries	Topic-based ranking effectively identifies academic entities. Basic limitation of PageRank was addressed
Limitations	Link structure was not considered. Time dimension, which is important with topic was not considered	Results dependent on parameter tuning. Time dimension not considered	Basic limitation of PageRank exists. Time dimension not considered	Basic limitation of PageRank exists. Time dimension not considered	Results dependent on parameter tuning. Time dimension not considered	Basic limitation of PageRank exists. Time dimension not considered	Time dimension not considered

Strategies to model the year-by-year interest of researchers or static researcher interest were presented in the study of Mimno and McCallum (2007), Rosen-Zvi *et al.* (2010), Steyvers *et al.* (2004), while, topic over time approaches as in the study of Wang and McCallum (2006) models the evolution of topics, but they ignore the researcher's interest. Daud (2012) proposed a novel that is Temporal Author Topic by combining the static researcher's interest Author Topic with Topic over Time. This approach models the researcher's interest with respect to the topics by modeling all years simultaneously.

3.3 Learning-based methods

Learning-based methods are flavors of supervised or semi-supervised methods of machine learning which build a ranking model automatically with the help of training data by optimizing the feature set (Fang *et al.*, 2010; Moreira *et al.*, 2011). The use of machine learning methods were adopted in information retrieval to construct retrieval formulas which are capable of finding the query relevance with the documents. Very small number of learning-based methods for ranking of authors has been found in literature. In learning-based methods, our focus is on predicting whether an author is an expert in a given field or not. In fact, the foundations of classification problem are involved in learning-based methods.

Yang *et al.* (2009) proposed an expert finding system which used the learning-based methods for finding a function to rank the candidate experts. They divided the whole method into three parts. First is the data preparation in which they gathered data from structured (DBLP) as well as non-structured (web pages) sources and created an academic network database. The second part is the expert finding phase in which a supervised learning algorithm Ranking SVM (Herbrich *et al.*, 1999) is used to rank candidates. An already labeled training data $L = \{(x_i, y_i)\}_{i=1}^{l}$ and unlabeled test data $S = \{x_i'\}_{i=1}^{u}$ was used. Ranking SVM aims to learn a ranking function $f \in F$ which can predict the relative order of instances: $x_i > x_j \leftrightarrow f(xi) > f(xj)$. Third part is the Bole Search in which they find best supervisors of a specific field. Their methods discover a latent space while learning the ranking function. Some of the features of authors that they used are "the year he/she published his/her first paper, the number of papers of an expert, the number of papers in last two years, the number of papers in last five years, the number of citations of all his/her papers, the number of papers cited more than 5 times, the number of papers cited more than ten times, PageRank score etc." They are using the features based on LMs and they obtained better results than results of language modeling approaches.

The retrieval system of discriminative probabilistic models tries to estimate the probability that a given document is evaluated to be relevant or irrelevant with respect to a user query. Fang *et al.* (2010) argued that discriminative models can give better performance than the generative models like statistical language modeling. They presented a discriminative learning framework for expert finding. This framework combines document evidence and document candidate relations within a unified model. Some of the features that they used are LM, PageRank, URL length, anchor text, title, exact name match, name match, last name match, etc. The main benefit that we can attain from this method is the capability to incorporate variable document evidence and document candidate association attributes.

Moreira *et al.* (2011) explored the benefits of using learning-based methods for expert finding problem. They actually performed experiments on existing state-of-the-art methods by applying different combinations of features. They combined multiple indicators of expertise, which are derived from the textual contents, from the graph structure and information provided by researcher's profile. The state-of-the-art methods used are SVMrank (Joachims, 2006) and SVMmap (Yue *et al.*, 2007).

The presented method is a supervised learning-based method which works in two steps:

(1) training: learning of a ranking function that can sort experts; and

(2) testing: determine the similarity between an expert and a new query by application of learned ranking function.

The DBLP data set has been used for experiments, covering the data of both, the journals and the conferences. To train and validate the ranking model, a set of queries is also required and for relevance judgment ArnetMiner is used. The features used in this method are the textual relevance, the profile features and the graph features. The results show that the proposed learning-based method is better than the conventional learning-based methods.

Some examples of the features used by different learning-based methods are PageRank (as document authority information), in-degree, URL length (Zhu *et al.*, 2010), graph-based expert authority (Chen *et al.*, 2006), internal structure of the document to show the association of expert with contents of the document (Balog and De Rijke, 2008a, b), non-local evidence (Balog and De Rijke, 2008a, b) etc. The machine learning techniques were exploited by Daud *et al.* (2015) to find the rising stars from the research community as well. Table VII provides a summary of learning-based methods for author ranking.

3.4 Comparison between methodologies

From the discussion above, one can see that there is a wide range of ranking methods available in the literature. Different types have evolved with the emergence of different ranking scenarios and conditions. With time improvements have been made in existing methods making more recent methods more powerful than the previous ones in almost all categories. Link analysis methods are mostly applied to the situations where we are interested in evaluating relationships among the nodes. It is helpful in the analysis of all types of networks, information retrieval, and knowledge discovery. These methods can also be applied in an unsupervised way. Bibliometric methods analyze the data quantitatively, whereas link analysis methods can analyze quantitatively as well as qualitatively. Semantics-based methods are applied when we are interested in finding people related to a

Ref No.	(Yang *et al.*, 2009)	(Fang *et al.*, 2010)	(Moreira *et al.*, 2011)
Year	2009	2010	2011
Proposed	Method to learn function for expert ranking	Discriminative learning framework for expert search	A learning-based method
Data set	DBLP and web pages of authors	TREC corpora	DBLP through ArnetMiner project
Learning algorithm	Rank-SVM	discriminative learning framework	SVMrank (Joachims, 2006), SVMmap (Yue *et al.*, 2007)
Evaluation measure	Precision, MAP	Precision, MAP, MRR, R-Prec	Precision, MAP
Comparison with	RSVM, language model, expert finding	Document model (Balog *et al.*, 2006)	Expert finding (Yang *et al.*, 2009)
Findings/ strengths	Use of learning-to-rank tools for learning an function for ranking of authors	Proposed method integrates various textual features and document candidate relationships into a unified way using discriminative methods	Textual similarity between documents and queries, graph structure with the citation patterns for the community of experts, and profile information about the experts was used with learning-to-rank models
Limitations	Could have more personalized or customized by adding more features	Query-dependent ranking not considered	Query-dependent ranking not considered

Table VII. Summary of learning-based methods to rank authors

specific topic. The aim is to capture the meaning hidden in the search string. Temporal methods are applied when aim is to capture the relevant documents with respect to a time frame. A combination of semantics and temporal ranking methods can be helpful in finding the experts in a given field at a specified time. Learning-based methods are applied when training data are available, and semi-supervised and supervised methods can be applied to generate the ranking model.

4. Data sets and performance measures

In this section, we will introduce the data sets used for experimentation of author ranking and the evaluation measures used in all categories.

4.1 Data sets

The well-known bibliographic databases like DBLP, OPD, and TREC have been widely utilized by the researchers. DBLP is the most widely used database for purposes of citation analysis and expert search (Daud *et al.*, 2009b; Daud and Hussain, 2012; Fiala *et al.*, 2008; Liu *et al.*, 2005; Moreira *et al.*, 2011). Its basic reason, perhaps, is that the citation records in DBLP are represented in a well-structured format, i.e., XML. It is a very large and comprehensive data set that covers the journal and conference publications in the field of computer science. Researchers can extract subsets of this data set according to the requirement of their domain of study. Researchers have used its different statistical parameters according to the requirement of their study for evaluation of their proposed methods. In the study of Daud *et al.* (2009b), the researchers have used the data only from the year 2003 to 2007, including 112,317 authors and 62,563 publications. In the study of Liu *et al.* (2005), the researchers have extracted the data of ACM-DL (1995-2000), IEEEADL (1994-2000), and JCDL (2001-2003) only, including all long and short papers, posters, demonstrations, and organizers of workshops. Their extracted data set contains 1,567 authors, 759 publications, and 3,401 co-authorship relationship pairs.

IRIS2 publication ontology and knowledge base ACM-SW were used by Wei *et al.* (2011). E-Society Project and Open Directory Project were used by Manaskasemsak *et al.* (2011). In total, 43 most recurrent queries from the query log of ArnetMiner were collected and divided into two subsets and two experiments were carried out (Tang *et al.*, 2008). ArnetMiner is also used by Gollapalli *et al.* (2011). The data set used for expert finding task of TREC enterprise was used by Balog *et al.* (2006), Fang *et al.* (2010), Yang and Zhang (2010). TREC Enterprise tracks from 2005 to 2008 were used by Fang *et al.* (2010), in which for the years 2005-2006, the document collection was crawled from the World Wide Web Consortium (W3C) and for the years 2007-2008, the document collection was crawled from the website of Commonwealth Scientific and Industrial Research Organization (CSIRO). In Balog *et al.* (2006), the researchers have used the data set of the 2005 edition of the TREC Enterprise track with document collection used is the W3C corpus, which is a heterogeneous document repository containing a mixture of different document types crawled from the W3C website. In Yang and Zhang (2010), the researchers have conducted the experiments on the data set of TREC 2007 enterprise. Their data collection was crawled from publicly available pages of Australia's national science agency CSIRO which includes 370,715 web documents.

4.2 Performance measures

The basic issue faced by the researchers is how to measure the performance of the proposed methodology with standard/huge databases, because truth values are not available. In this field of ranking as ground truth values are not available, researchers cannot directly apply quantitative measures; instead first they have to prepare some records that can be matched with results.

The quantitative measures like precision at k (P@k) and Mean Average Precision (MAP) are used by Balog *et al.* (2006); Fang *et al.* (2010); Moreira *et al.* (2011); Salton *et al.* (1975); Tang *et al.* (2008); Wei *et al.* (2011); Zhang *et al.* (2008). Precision is the fraction of retrieved documents that are relevant to the search. The formula is as follows:

$$\text{Precision} = \frac{\{\text{Relevant documents}\} \cap \{\text{retireved documents}\}}{\{\text{retrieved documents}\}} \quad (1)$$

Precision can be used when a user wants to see only the first k retrieved results. As truth values for the ranking of authors are not available, so applying the quantitative measures, like precision, requires some values to be compared with the results. For example, Moreira *et al.* (2011) have applied 13 query topics from the computer science domain on Arenetminer data set and compared the precision of their retrieved results with those values. Gollapalli *et al.*, 2011; Tang *et al.*, 2008; Zhang *et al.*, 2008) have used the method of pooled relevance judgment, as ground truth values are not available. They first collected the top 30 results from the three data sets (Libra, Rexa, and ArnerMiner) in a single list. Then, one faculty member and two graduate students provided human judgments to finally make one complete list with which they compared the results of their proposed methods using P@k and MAP. Wei *et al.* (2011) used the human judgment of relevance and quality to evaluate the produced rankings, and compared the retrieved results by using P@k and NDCG. Discounted Cumulative Gain (DCG) uses a categorized similarity scale of documents from the result set to estimate the significance of a document based on its rank in the result list. The cumulative gain at each position should be normalized across queries to get Normalized Discounted Cumulative Gain (NDCG), as a comparison of results of one query with another cannot be done consistently by using DCG alone. NDCG is a measure of ordering accuracies when there are numerous stages of relevance judgment. Given a query and a ranking, NDCG computed at top k documents is given as follows:

$$\text{NDCG}k(\tau) = \text{Zk} \sum_{\text{i}=1}^{\text{k}} \frac{2r(i) - 1}{\log\ 2\ (\text{i}+1)} \quad (2)$$

where τ is the ranking, r(i) is the gain value and Zk is the normalization factor. NDCG is also used by Manaskasemsak *et al.* (2011).

Few other quantitative measures like degree, closeness and betweenness centrality measures have been used by Liu *et al.* (2005). The total number of edges that are adjacent to the node defines the degree centrality of a node. When talking about authors, degree centrality represents how many connections tie one author to his/her immediate neighbors in the network. Closeness centrality is determined by finding shortest path distances of a node to all other nodes in the network. A central author therefore has many short connections to other authors in the network. Betweenness centrality measures the number of times a node acts as a bridge along the shortest path between two other nodes. In terms of authors, we can say that how often a particular author is found on the shortest path between any pair of authors in the network. Liu *et al.*, (2005) extracted the names of all JCDL, ADL and DL program committee members from the conference web sites or printed proceedings. Then, they matched the top 50 results retrieved by degree, closeness, betweenness, PageRank, and AuthorRank one by one against each JCDL committee member to identify matches for the purpose of evaluation.

The Spearman correlation coefficient is another important metric used for evaluation in the study of Liu *et al.* (2005). It is used to measure the strength of correlation between two variables. They calculated the Spearman correlation coefficient between degree centrality, PageRank, and AuthorRank and found that PageRank and AuthorRank are more closely

related than degree centrality. IT is also used in the study of Yan and Ding (2011) to find the correlation between Popularity Rank and Prestige Rank. In the study of Yan and Ding (2011), Spearman correlation coefficient is used to plot values of PageRank with different damping factors. In the study of Fiala *et al.* (2008), it is used to plot the correlation between different ranking schemes.

Instead of comparing the results with some sort of ground truth values, OSim and KSim are two measures which can be used to compare the results of two ranking algorithms. OSim is a similarity measure which indicates the degree of overlap between the top k URLs of two rankings $\tau 1$,$\tau 2$. The formula is as under:

$$\text{OSim}k(\tau 1, \tau 2) = \frac{|R1 \cap R2|}{k} \tag{3}$$

where, k is the total number of queries, $R1$ is the ranking of query 1 and $R2$ is the ranking of query 2. KSim is a variant of Kendall's distance measure. It is the number of pairwise swaps necessary to align two lists of ranking. It indicates the degree to which the relative orderings of the top k URLs of two rankings are in agreement. The formula is as under:

$$\text{KSim}k(\tau 1,\ \tau 2) = \frac{\left|\left\{(u, v) : \tau' 1,\ \tau' 2 \text{ agree on order of} (u, v) \text{and } u \neq v\right\}\right|}{|U| \times (|U| - 1)} \tag{4}$$

where, $\tau 1$ and $\tau 2$ are lists of URLs and (u, v) is a pair of distinct nodes. OSim and KSim are used by Manaskasemsak *et al.* (2011).

5. Future directions and research challenges

In this section, we will discuss research issues, open challenges and future directions for author ranking.

5.1 Future directions

From the literature survey, we found that in ranking of academic objects, there is a lot of room for improvement. For example, ranking of authors with respect to their contribution in a publication (Sekercioglu, 2008), change in the field of interest over time (Li and Tang, 2008), number of unique authors who cite a publication (Katsaros *et al.*, 2009), considering semantics (Daud *et al.*, 2009b) and feature set optimization using learning-based methods (Fang *et al.*, 2010; Moreira *et al.*, 2011) is very important. Quantification of the author's contribution in a publication is necessary because usually all the authors do not contribute equally, e.g. first author is usually the main contributor of that work (Sekercioglu, 2008). While ranking an author, the change of field of interest is an important consideration (Daud, 2012) as the authors who are experts in a field at a given time may not be the expert in the same field in a later period due to change of interest (Daud *et al.*, 2010). Using maximum features and their optimization is also very important for ranking authors (Fang *et al.*, 2010) as previously a few features were used in linkage analysis and probabilistic methods.

The ranking is mostly done for authors and simultaneous ranking of all objects like authors, publications, and venues need to be done in future. The foremost requirement that we identified from this study is the need of a benchmarking standard in the field of ranking as there are no standard rules which make it challenging to evaluate and compare different types of ranking methods. This research gap was also been highlighted by Jiang *et al.* (2013). Along with that, a very comprehensive experimental comparison of different types of ranking methods is an open challenge.

5.2 Research challenges

We categorize the challenges of ranking of academic objects into five types.

The first challenge arises in the field of finding the actual contribution of an author in a co-authored paper. One can solve the problem of author ranking in such a way that contribution of all co-authors is considered (Hirsch, 2010). In this way, justice will be done to the authors of papers that have multiple authors. A very important point that needs attention in the future is that the weight of citations is given equally to all the co-authors. More investigative works are required in order to provide the actual weight of citations to all authors.

The second challenge arises with the application of probabilistic methods for ranking of academic objects. Probabilistic ranking methods need to be explored for different combinations of features of language modeling. For example, the query expansion and its effect on different features like PageRank, URL length, in-degree, out-degree, etc. can be discovered.

The third challenge comes from the need of finding temporal trends in documents and other entities together with considering the contents of documents. The temporal aspects need to be addressed in more detail as time factor can be a powerful weight that can reflect the changes in the field of interest with respect to time (Fiala *et al.*, 2008). The effect of time period whose data are selected is also required to be tested in the future. Study of impact of time has also been highlighted by Jiang *et al.* (2013).

The fourth challenge is associated with the semantic ranking of academic objects. The semantic ranking of authors can be helpful for the people who are interested in finding researchers in their own field of interest. Topic modeling methods can also be utilized to find out the semantic associations between any two given nodes in a network. In the field of semantics, many hidden themes can be associated with a given query which needs to be found out automatically.

The fifth and final challenge is associated with learning-based methods. Learning-based methods can also be presented in a generalized form rather than a ranking of authors only. A hybrid of discriminative and generative models can be more effective to get the best of most methods.

6. Conclusions

The academic objects like author, conferences, journals and papers are the main entities to be ranked. In this study, we have studied the methods for ranking of authors and we have classified them according to their way of calculations. There are methods which can rank any one, two or three of the mentioned academic objects but more generic solutions are missing which can be applied to all those academic objects. We observed that most of the work done for the ranking of the authors is in category of link analysis methods. The link analysis based methods use several features like the number of publications, the number of citations, the position of authors, the influence of co-authors, the effect of topic, and different combinations of these features. All the features and their combinations have their own effect on the ranking results and can be used in different scenarios. These methods can give variable results under different situations. For example, if the objective is to find topic-based ranking of authors, a generic method that only uses the number of publications or citations cannot give required results. Similarly, if question is about finding the influence of a senior author on a junior, a topic-based method is not applicable. We conclude that all ranking features have very strong influence on the results of a methods and hence selection of a ranking method must be done keeping in view the required results and objectives.

The temporal methods can provide meaningful results while also incorporating the time of publication and received citations for ranking. Temporal models with topic sensitivity can give significant results in scenarios when objective is to find topic-based

experts in a certain time range. The learning-to-rank methods get less attention, as compared to link analysis and bibliometric methods. The strengths of supervised learning methods can be further explored to find ranking functions for expert finding. We also conclude that hybrid models are required in future that can consider the dependencies between query terms that can incorporate multiple document and ranking features, and consider the link structure of the network as well. These hybrid methods must be dynamic enough to allow the user to select different features under different scenarios to fit the requirements of a certain scenario. Such hybrid models can possibly give required results under variable situations.

All the methods found in literature use different data sets, are based on different type of networks, use different metrics, parameters or weighing criteria for ranking and have different performance evaluation methods. The main reason behind this scenario is unavailability of standards, hence making the task very complicated and challenging.

Notes

1. *citeseerx.ist.psu.edu/*
2. www.informatik.uni-trier.de/~ley/db/

References

Amjad, T. and Daud, A. (2017), "Indexing of authors according to their domain of expertise", *Malaysian Journal of Library & Information Science*, Vol. 22 No. 1, pp. 69-82.

Amjad, T., Bibi, S., Shaikh, M.A. and Daud, A. (2016), "Author productivity indexing via topic sensitive weighted citations", *Science International*, Vol. 28 No. 4, pp. 4135-4139.

Amjad, T., Daud, A., Akram, A. and Muhammed, F. (2016), "Impact of mutual influence while ranking authors in a co-authorship network", *Kuwait Journal of Science*, Vol. 43 No. 3, pp. 101-109.

Amjad, T., Daud, A., Che, D. and Akram, A. (2015), "MuICE: mutual influence and citation exclusivity author rank", *Information Processing & Management*, Vol. 52 No. 3, pp. 374-386.

Amjad, T., Ding, Y., Daud, A., Xu, J. and Malic, V. (2015), "Topic-based heterogeneous rank", *Scientometrics*, Vol. 104 No. 1, pp. 313-334.

Amjad, T., Ding, Y., Xu, J., Zhang, C., Daud, A., Tang, J. and Song, M. (2017), "Standing on the shoulders of giants", *Journal of Informetrics*, Vol. 11 No. 1, pp. 307-323.

Ausloos, M. (2013), "A scientometrics law about co-authors and their ranking: the co-author core", *Scientometrics*, Vol. 95 No. 3, pp. 895-909.

Balog, K. and De Rijke, M. (2008a), "Associating people and documents", Advances in Information Retrieval, Vol, 4956, Lecture Notes in Computer Science Advances in Information Retrieval, ECIR, Springer, Berlin and Heidelberg, pp. 296-308.

Balog, K. and De Rijke, M. (2008b), "Non-local evidence for expert finding", *Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, CA, 26-30 October*, pp. 489-498.

Balog, K., Azzopardi, L. and De Rijke, M. (2006), "Formal models for expert finding in enterprise corpora", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, 6-11 August*, pp. 43-50.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *The Journal of Machine Learning Research*, Vol. 3 No. 1, pp. 993-1022.

Bouyssou, D. and Marchant, T. (2010), "Consistent bibliometric rankings of authors and of journals", *Journal of Informetrics*, Vol. 4 No. 3, pp. 365-378.

Brin, S. and Page, L. (1998), "The anatomy of a large-scale hypertextual web search engine", *Computer Networks and ISDN Systems*, Vol. 30 No. 1, pp. 107-117.

Burrell, Q.L. (2007a), "On the h-index, the size of the Hirsch core and Jin's A-index", *Journal of Informetrics*, Vol. 1 No. 2, pp. 170-177.

Burrell, Q.L. (2007b), "Hirsch's h-index: a stochastic model", *Journal of Informetrics*, Vol. 1 No. 1, pp. 16-25.

Chen, H., Shen, H., Xiong, J., Tan, S. and Cheng, X. (2006), "Social network structure behind the mailing lists: ICT-IIIS at TREC 2006 expert finding track", *Proceedings of the Text REtrieval Conference, (TREC '05)*, *Gaithersburg, MD*.

Cohen, S., Mamou, J., Kanza, Y. and Sagiv, Y. (2003), "XSEarch: a semantic search engine for XML", *VLDB Endowment, Proceedings of the 29th International Conference on Very Large Data Bases*, Vol. 29, pp. 45-56.

Daud, A. (2012), "Using time topic modeling for semantics-based dynamic research interest finding", *Knowledge-Based Systems*, Vol. 26 No. 1, pp. 154-163.

Daud, A. and Hussain, S. (2012), "Publication venue-based language modeling for expert finding", *Proceedings of International Conference on Future Communication and Computer Technology (ICFCCT 2012)*, *19-20 May*.

Daud, A. and Muhammad, F. (2012), "Group topic modeling for academic knowledge discovery", *Applied Intelligence*, Vol. 36 No. 4, pp. 870-886.

Daud, A. and Muhammad, F. (2014), "Consistent annual citations based researcher index", *Collnet Journal of Scientometrics and Information Management*, Vol. 8 No. 2, pp. 209-216.

Daud, A., Abbasi, R. and Muhammad, F. (2013), "Finding rising stars in social networks", Database Systems for Advanced Applications, Vol. 7825, DASFAA, Lecture Notes in Computer Science, Springer, Berlin and Heidelberg, pp. 13-24.

Daud, A., Saleem Yasir, S.M. and Muhammad, F. (2013), "V-index an index based on consistent researcher productivity", *IEEE 16th International, Multi Topic Conference, Lahore, 19-20 December*, pp. 61-65, doi: 10.1109/INMIC.2013.6731325.

Daud, A., Shaikh, A.M.A.R. and Rajpar, A.H. (2009), "Scientific reference mining using semantic information through topic modeling", *Mehran University Research Journal of Engineering and Technology*, Vol. 28 No. 2, pp. 253-262.

Daud, A., Ahmad, M., Malik, M.S.I. and Che, D. (2015), "Using machine learning techniques for rising star prediction in co-author network", *Scientometrics*, Vol. 102 No. 2, pp. 1687-1711.

Daud, A., Li, J., Zhou, L. and Muhammad, F. (2009a), "Conference mining via generalized topic modeling", *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Vol. 5781, Machine Learning and Knowledge Discovery in Databases. ECML PKDD, Lecture Notes in Computer Science, Springer, Berlin and Heidelberg, pp. 244-259.

Daud, A., Li, J., Zhou, L. and Muhammad, F. (2009b), "A generalized topic modeling approach for Maven search", *Advances in Data and Web Management, Proceedings of International Asia-Pacific Web Conference and Web-Age Information Management (APWEB-WAIM)*, *Springer, Suzhou*, *2-4 April*, pp. 138-149.

Daud, A., Li, J., Zhou, L. and Muhammad, F. (2009c), "Exploiting temporal authors interests via temporal-author-topic modeling", *5th International Conference on Advanced Data Mining and Applications, Springer*, *Beijing*, *17-19 August*, pp. 435-443.

Daud, A., Li, J., Zhou, L. and Muhammad, F. (2010), "Temporal expert finding through generalized time topic modeling", *Knowledge-Based Systems*, Vol. 23 No. 6, pp. 615-625.

Daud, A., Aljohani, N.R., Abbasi, R.A., Rafique, Z., Amjad, T., Dawood, H. and Alyoubi, K.H. (2017), "Finding rising stars in co-author networks via weighted mutual influence", *Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee*, *Perth*, *3-7 April*, pp. 33-41.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V. and Sachs, J. (2004), "Swoogle: a search and metadata engine for the Semantic Web", *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, *Washington, DC*, *8-13 November*, pp. 652-659.

Ding, Y. (2011a), "Applying weighted pagerank to author citation networks", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 2, pp. 236-245.

Ding, Y. (2011b), "Topic-based pagerank on author cocitation networks", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 3, pp. 449-466.

Ding, Y. and Cronin, B. (2011), "Popular and/or prestigious? Measures of scholarly esteem", *Information Processing & Management*, Vol. 47 No. 1, pp. 80-96.

Ding, Y., Yan, E., Frazho, A. and Caverlee, J. (2009), "Pagerank for ranking authors in co-citation networks", *Journal of the American Society for Information Science and Technology*, Vol. 60 No. 11, pp. 2229-2243.

Egghe, L. (2006), "An improvement of the h-index: the g-index", *ISSI Newsletter*, Vol. 2 No. 1, pp. 8-9.

Egghe, L. (2008), "Mathematical theory of the h-and g-index in case of fractional counting of authorship", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 10, pp. 1608-1616.

Egghe, L. and Rousseau, R. (2008), "An h-index weighted by citation impact", *Information Processing & Management*, Vol. 44 No. 2, pp. 770-780.

Fang, Y., Si, L. and Mathur, A.P. (2010), "Discriminative models of integrating document evidence and document-candidate associations for expert search", *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, 19-23 July*, pp. 683-690.

Fiala, D. (2012), "Time-aware pagerank for bibliographic networks", *Journal of Informetrics*, Vol. 6 No. 3, pp. 370-388.

Fiala, D., Rousselot, F. and Ježek, K. (2008), "Pagerank for bibliographic networks", *Scientometrics*, Vol. 76 No. 1, pp. 135-158.

Gollapalli, S.D., Mitra, P. and Giles, C.L. (2011), "Ranking authors in digital libraries", *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, Ottawa, 13-17 June*, pp. 251-254.

Gollapalli, S.D., Mitra, P. and Giles, C.L. (2013), "Ranking experts using author-document-topic graphs", *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, Indianapolis, IN, 22-26 July*, pp. 87-96.

Guns, R. and Rousseau, R. (2014), "Recommending research collaborations using link prediction and random forest classifiers", *Scientometrics*, Vol. 101 No. 2, pp. 1461-1473.

Gupta, S., Duhan, N. and Bansal, P. (2013), "A comparative study of page ranking algorithms for online digital libraries", *International Journal of Scientific and Engineering Research*, Vol. 4 No. 4, pp. 1225-1233.

Herbrich, R., Graepel, T. and Obermayer, K. (1999), "Large margin rank boundaries for ordinal regression", *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, pp. 115-132.

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*, Vol. 102 No. 46, pp. 16569-16572.

Hirsch, J.E. (2010), "An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship", *Scientometrics*, Vol. 85 No. 3, pp. 741-754.

Hofmann, T. (1999), "Probabilistic latent semantic indexing", *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, 15-19 August*, pp. 50-57.

Hong, D. and Baccelli, F. (2012), "On a joint research publications and authors ranking", p. 9, available at: http://hal.inria.fr/hal-00666405

Jatowt, A., Kawai, Y. and Tanaka, K. (2005), "Temporal ranking of search engine results", *Web Information Systems Engineering – WISE, 6th International Conference on Web Information Systems Engineering*, Springer, New York, NY, pp. 43-52.

Jiang, X., Sun, X. and Zhuge, H. (2013), "Graph-based algorithms for ranking researchers: not all swans are white!", *Scientometrics*, Vol. 96 No. 5, pp. 743-759.

Jin, B., Liang, L., Rousseau, R. and Egghe, L. (2007), "The R-and AR-indices: complementing the h-index", *Chinese Science Bulletin*, Vol. 52 No. 6, pp. 855-863.

Joachims, T. (2006), "Training linear SVMs in linear time", *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, 20-23 August*, pp. 217-226.

Katsaros, D., Akritidis, L. and Bozanis, P. (2009), "The f index: quantifying the impact of coterminal citations on scientists' ranking", *Journal of the American Society for Information Science and Technology*, Vol. 60 No. 5, pp. 1051-1056.

Kleinberg, J.M. (1999), "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, Vol. 46 No. 5, pp. 604-632.

Li, X.-L., Foo, C.S., Tew, K.L. and Ng, S.-K. (2009), "Searching for rising stars in bibliography networks", Database Systems for Advanced Applications, Vol. 5463, DASFAA, Lecture Notes in Computer Science, Springer, Berlin and Heidelberg, pp. 288-292.

Li, Y. and Tang, J. (2008), "Expertise search in a time-varying social network", *IEEE, Ninth International Conference on Web-Age Information Management, Zhangjiajie Hunan, 20-22 July*, pp. 293-300.

Lin, L., Xu, Z., Ding, Y. and Liu, X. (2013), "Finding topic-level experts in scholarly networks", *Scientometrics*, Vol. 97 No. 3, pp. 797-819.

Liu, X., Bollen, J., Nelson, M.L. and Van de Sompel, H. (2005), "Co-authorship networks in the digital library research community", *Information Processing & Management*, Vol. 41 No. 6, pp. 1462-1480.

Manaskasemsak, B., Rungsawang, A. and Yamana, H. (2011), "Time-weighted web authoritative ranking", *Information Retrieval*, Vol. 14 No. 2, pp. 133-157.

Mimno, D. and McCallum, A. (2007), "Expertise modeling for matching papers with reviewers", *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, CA, 12-15 August*, pp. 500-509.

Moreira, C., Calado, P. and Martins, B. (2011), "Learning to rank for expert search in digital libraries of academic publications", Progress in Artificial Intelligence, Lecture Notes in Computer Science, Springer, Berlin and Heidelberg, pp. 431-445.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999), "The pagerank citation ranking: bringing order to the web technical report", Stanford Digital Library Technologies Project.

Ponte, J.M. and Croft, W.B. (1998), "A language modeling approach to information retrieval", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, 24-28 August*, pp. 275-281.

Radicchi, F., Fortunato, S., Markines, B. and Vespignani, A. (2009), "Diffusion of scientific credits and the ranking of scientists", *Physical Review E*, Vol. 80 No. 5, p. 56103.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P. and Steyvers, M. (2010), "Learning author-topic models from text corpora", *ACM Transactions on Information Systems (TOIS)*, Vol. 28 No. 1, p. 4.

Salton, G., Wong, A. and Yang, C.-S. (1975), "A vector space model for automatic indexing", *Communications of the ACM*, Vol. 18 No. 11, pp. 613-620.

Sato, N., Uehara, M. and Sakai, Y. (2003), "Temporal ranking for fresh information retrieval", Association for Computational Linguistics, AsianIR '03 Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages, Vol. 11, Sapporo, 7 July, pp. 116-123.

Sekercioglu, C.H. (2008), "Quantifying coauthor contributions", *Science*, Vol. 322, p. 371.

Steyvers, M., Smyth, P., Rosen-Zvi, M. and Griffiths, T. (2004), "Probabilistic author-topic models for information discovery", *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, 22-25 August*, pp. 306-315.

Tang, J., Jin, R. and Zhang, J. (2008), "A topic modeling approach and its integration into the random walk framework for academic search", *Proceedings of IEEE International Conference on Data Mining*, *Pisa*, *15-19 December*, pp. 1055-1060.

Tol, R.S. (2008), "A rational, successive g-index applied to economics departments in Ireland", *Journal of Informetrics*, Vol. 2 No. 2, pp. 149-155.

Usmani, A. and Daud, A. (2017), "Unified author ranking based on integrated publication and venue rank", *International Arab Journal of Information Technology*, Vol. 14 No. 1, pp. 14-20.

Wang, X. and McCallum, A. (2006), "Topics over time: a non-markov continuous-time model of topical trends", *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Philadelphia, PA*, *20-23 August*, pp. 424-433.

Wei, W., Barnaghi, P. and Bargiela, A. (2011), "Rational research model for ranking semantic entities", *Information Sciences*, Vol. 181 No. 13, pp. 2823-2840.

Yan, E. and Ding, Y. (2011), "Discovering author impact: a pagerank perspective", *Information Processing & Management*, Vol. 47 No. 1, pp. 125-134.

Yang, L. and Zhang, W. (2010), "A study of the dependencies in expert finding", *Third International Conference on Knowledge Discovery and Data Mining, IEEE*, *Phuket, 9-10 January*, pp. 355-358.

Yang, Z., Tang, J., Wang, B., Guo, J., Li, J. and Chen, S. (2009), "Expert2bole: from expert finding to bole search", *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD'09)*, pp. 1-4.

Yu, P.S., Li, X. and Liu, B. (2004), "On the temporal dimension of search", *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, *New York, NY*, *19-21 May*, pp. 448-449.

Yu, P.S., Li, X. and Liu, B. (2005), "Adding the temporal dimension to search-a case study in publication search", *IEEE, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, *19-22 September*, pp. 543-549.

Yue, Y., Finley, T., Radlinski, F. and Joachims, T. (2007), "A support vector method for optimizing average precision", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *Amsterdam*, *23-27 July*, pp. 271-278.

Zhang, C.-T. (2009), "A proposal for calculating weighted citations based on author rank", *EMBO Reports*, Vol. 10 No. 5, pp. 416-417.

Zhang, G., Ding, Y. and 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*, Vol. 64 No. 7, pp. 1490-1503.

Zhang, J., Tang, J., Liu, L. and Li, J. (2008), "A mixture model for expert finding", Advances in Knowledge Discovery and Data Mining, Vol. 5012, PAKDD Lecture Notes in Computer Science, Springer, Berlin and Heidelberg, pp. 466-478.

Zhu, J., Huang, X., Song, D. and Rüger, S. (2010), "Integrating multiple document features in language models for expert finding", *Knowledge and Information Systems*, Vol. 23 No. 1, pp. 29-54.

Corresponding author
Tehmina Amjad can be contacted at: tehminaamjad@iiu.edu.pk