On The Current State of Scholarly Retrieval Systems

Shah Khalid
School of Computer Science and Communication Engineering, Jiangsu University, China and
National University of Sciences and Technology (NUST), Islamabad, Pakistan
shahkhalid@ujs.edu.cn

Shah Khusro
Department of Computer Science,
University of Peshawar,
Peshawar, Pakistan
khusro@uop.edu.pk

Irfan Ullah
Department of Computer Science,
University of Peshawar,
Peshawar, Pakistan,
cs.irfan@uop.edu.pk

Godfrey Dawson-Amoah
School of Computer Science and Communication Engineering,
Jiangsu University, China
gdamoah@stmail.ujs.edu.cn

Abstract—**The enormous growth in the size of scholarly literature makes its retrieval challenging. To address this challenge, researchers and practitioners developed several solutions. These include indexing solutions e.g. ResearchGate, Directory of Open Access Journals (DOAJ), Digital Bibliography & Library Project (DBLP) etc., research paper repositories e.g. arXiv.org, Zenodo, etc., digital libraries, scholarly retrieval systems, e.g., Google Scholar, Microsoft Academic Search, Semantic Scholar etc., digital libraries, and publisher websites. Among these, the scholarly retrieval systems, the main focus of this article, employ efficient information retrieval techniques and other search tactics. However, they are still limited in meeting the user information needs to the fullest. This brief review paper is an attempt to identify the main reasons behind this failure by reporting the current state of scholarly retrieval systems. The findings of this study suggest that the existing scholarly retrieval systems should differentiate scholarly users from ordinary users and identify their needs. Citation network analysis should be made an essential part of the retrieval system to improve the search precision and accuracy. The paper also identifies several research challenges and opportunities that may lead to better scholarly retrieval systems.**

Keywords-information retrieval; scholarly search; scholarly users; citation networks

I. INTRODUCTION

A scholarly retrieval system is a sophisticated software that performs crawling, indexing, searching, and ranking to make scholarly data (research publications and related information including authors, publishers, citations, etc.), available to searchers. Several scholarly retrieval systems including Google Scholar, Microsoft Academic Search, CiteSeerX and Chinese Baidu Academic [1] are frequently used by modern-day online searchers. The retrieved scholarly documents include journal articles, conference proceedings, books, dissertations, technical reports, and patents. While some of these documents are freely accessible to all members of the public, access to others is limited only to subscribers. The academic web is growing, but there seems to be no definite agreement on its size. One estimation of the number of scholarly documents is 120 million, of which 25% is freely accessible [2]. Google Scholar has indexed nearly 160 million scholarly documents [3]. Microsoft Academic Search has indexed nearly 209.79 million [4]. The number of scholarly documents increases at an annual rate of over 1 million [5]. Such a huge collection of research publications is therefore challenging to process and find relevant papers effortlessly. Researchers are working on finding out a way for supporting scholarly search and making it more accessible. Their efforts resulted in several indexing solutions, publication repositories, digital libraries, research paper recommender systems, and scholarly retrieval systems. This paper aims to report on the current state of the scholarly retrieval system by identifying the commonalities and differences between web and scholarly users, surveying the search techniques of the available scholarly retrieval systems, and understanding the potential role citation networks analysis in retrieval relevant research publications.

II. THE CURRENT STATE SCHOLARLY RETRIEVAL SYSTEMS

Scholarly retrieval solutions take the user search query as input and check its relevance with publications using different ranking features [6-10]. As a complementary tool to the academic search, a research paper recommender system employs different filtering algorithms to find and recommend relevant papers based on users' implicit and explicit feedback as well as the content of these documents. In some cases, search and recommendation are employed in a search-recommendation hybrid manner, where keywords are first used to find an initial list of search results and then recommendations are applied to refine the search [11]. Without loss of generality, both architectures are highly related, and most of the techniques used for scholarly retrieval systems apply to scholarly recommender systems. Recommender systems are covered in several recent papers [11-13].

Corresponding author: Shah Khusro

A. The Structure of Scholarly Documents

Unlike general web, the unit of information to be retrieved by a scholarly search system is a research article that is retrieved based either on its content or some specific parts. A scholarly publication can be a journal article, conference paper, technical report, pre-print, thesis/dissertation, or a book. This paper considers research articles only, excluding theses/dissertations, technical reports, and books. A research paper has a well-defined structure and well-organized content to which writers are customarily constrained. Usually, the author follows the Author Guidelines or Instructions for Authors specifying the length, format, in-text citations, references, artwork, tables, etc., before the submission or after the manuscript is accepted for publication. The manuscript text is mainly unstructured [5], but it is sometimes considered semi-structured or even structured. Generally, research articles consist of a header, main-content, a bibliography, algorithms, tables, figures, mathematical equations and so on [14]. The header contains a title, authors, their emails and affiliations, abstract, and publication year, venue (journal, conference, etc.), volume and issue number, number of pages, etc. Figures and tables covey results and other structured information in a very symbolic and practical way. Algorithms are the step-by-step approach and effective way to present how a computational problem works. The mathematical computation is usually written in the form of equations. Bibliography (also known as references, incites, or notes) is a collection of cited publications listed at the end of the research article. They play a vital role in assessing the quality of the manuscript, helping the reader to learn more by accessing these links, and facilitates in creating citation networks. The extraction and usage of all essential components can enhance the ranking of scholarly retrieval systems [15-17]. By utilizing different tools such as OCR++ [18-20], Apache Tika [21], GROBID for header extraction [22], PDFFigures for table and figure extraction [23] and algorithm extraction [24], ParsCit for citations extraction [25], etc., documents can be parsed into different sections like title, abstract, body text, authors, venue, and references for optimizing retrieval. The metadata including title, author (name, email, affiliation), heading and section mapping, footnote, figures and table headings, URL, citation, and references can be extracted and processed in a usable format like XML or JSON [18]. The extraction and storage of figures can also play a vital role in the retrieval of relevant papers [26]. However, for an efficient scholarly retrieval system, it is essential to consider the structure and associated metadata of the scholarly documents in search, ranking, and recommendation [27].

B. Users of Scholarly Retrieval Systems

The scholarly users are different from the typical web searchers [28]. They have different search patterns: in general web search, the search activity is on a peak in the weekend and goes down during weekdays, but in scholarly search, the activity is on the peak during weekdays and (mostly) drops in weekends [29]. Academic searchers use scholarly retrieval platforms. The web searchers widely use the general web search engines. Table I summarizes the types of scholarly users and their requirements, based on [30].

TABLE I. TYPES OF SCHOLARLY USERS

User Category	Goal	Tasks
Readers or authors	Updating bibliography on a topic	• Find scholarly and recent publications. • Navigate through bibliography to find other relevant papers.
Reviewers	Freshness	• Check whether recent relevant papers are cited • Check if relevant papers are missing.
Editors	Scope	• Check the relevance of the submitted paper and of the papers cited in it. • Check if self-citations are present, count them and assess the paper.
Evaluators	Impact	• Check the number of citations of paper and author. • Check why a paper is cited (is it relevant or just for self-citation?). • Check citations evolution.
Event Organizers	Participants	• Find relevant paper regarding event. • Find authors of relevant papers.

1) Readers/Authors

Readers and authors usually search for and read scholarly documents to find novelties in literature either for learning or for developing new approaches to solve a problem. Because of the massive size of the academic web, it is unrealistic for a researcher to read every article related to the research subject [31]. In order not to overwhelm researchers with information overload, it is essential to provide a result set that includes the most relevant documents related to a given query.

2) Reviewers

Reviewers assess the quality of a submitted paper to ensure it meets the laid down standards. A critical aspect of a reviewer's job is to evaluate the citations used. For any given topic, there usually exist a set of core scholarly documents that need to be referenced in any new work because they establish the theoretical foundations of the topic. It is essential to ensure that all relevant information is made available to a reviewer during the evaluation process, which may aid in reviewing the submitted manuscript more efficiently.

3) Editors

The core mandate of editors is to evaluate the scope of a submitted scholarly document to ensure that it fits the platform (journal, conference proceedings, etc.) it is intended for. They also evaluate the number and quality of self-references in situations where they are used.

4) Evaluators

Evaluators belong to a category of users who usually carry out research aimed at determining the contributions of the author to the body of knowledge of a specific field of study.

5) Event Organizers

Event organizers are generally interested in getting to know the potential participants by using the citation network. The participants could include researchers working on insightful solutions in the given domain, students, authors of previously published relevant scholarly documents and any others who might have an interest in the event.

C. Approaches to Scholarly Search and Retrieval

The following sections discuss the approaches with which the scholarly search systems use to mitigate the problem of information overload for academic searchers.

1) Citation Graph-Based Approaches

The network of references forms the citation graph, in which the citing paper cites the cited ones. Several approaches practice citation graph [32-37] e.g. Sofia Search [37] produces a citation graph by starting from the initial set of papers and following the links of citing and cited papers until the desired number of candidate papers is found. It mimics a human in identifying candidate publications from the citation graph. From the citation graph of the seed papers, the approach generates a list of relevant papers. However, in the growth rate of research papers, the use of Sofia Search is limited. At first, it needs seed papers and a lower bound. Then, all the in-links and out-links are not equal and relevant [38-40]. Another representative technique of citation graph is bibliographic coupling that considers only out-links of a paper [41]. The similarity between P_1 and P_2 is computed as $\frac{|O_{P_1} \cap O_{P_2}|}{|O_{P_1} \cup O_{P_2}|}$ [42, 43], where O_{P_1} and O_{P_2} are the sets, having out-links of P_1 and P_2, respectively. The similarity between the two papers is equal to 0 when both sets O_{P_1} and O_{P_2} are empty. Bibliographic coupling has been practiced and worked well for classification of scholarly documents [43], plagiarism detection [44, 45], and similar legal judgments [46]. However, it is limited in retrieving relevant papers due to two reasons: a) bibliographic coupling misses some important papers not present in the out-cites, and b) it is unable to consider the in-cites of the papers. The citation context is used in [47] for retrieving relevant literature. However, extracting citation context is challenging due to the unavailability of full text and almost unable to reveal the main subject of the paper resourcefully [48, 49]. Several popular academic search engines including Google Scholar, PubMed and CiteSeer use the links between academic articles, provided by citation networks for documents ranking.

2) Content-Based Approaches

Content-based methods process textual content of the papers, which can be title, abstract, introduction, keywords and body of the articles. These methods weigh the article's influence by the frequency and position of the terms in the article [50]. Many techniques are based on the term weights to estimate the relevance of articles. The most widely used approach is the vector space model (VSM), which represents each article as a vector of term weights and the relevance is a measure in terms of some similarity measures such as cosine similarity between the query and document vectors. Many retrieval systems and applications practice VSM (e.g. [51]) even though, the cosine similarity does not perform well in many situations [52, 53]. Latent semantic analysis (LSA) improves the vector representation of a scholarly article by singular value decomposition (SVD) method [54, 55]. However, for the efficient retrieval of scholarly articles, LSA does not perform well comparatively [53]. Many scholarly retrieval systems prefer using BM25 which is among the best ranking techniques for scholarly retrieval [56]. However, to better meet the requirements of different scholarly users, researchers have also adopted hybrid approaches by combining content- and citation-based approaches, discussed below.

3) Hybrid Approaches

The hybrid approaches [43, 57-59] combine the best of the citation graph-based and content-based techniques to compute the relevance of documents to the search query. The proximity of citations is supportive in locating related articles [60, 61]. Two articles may be similar to each other if many articles in nearby locations cite them. However, all the papers are not publically available to locate the nearby locations and cannot guarantee the exact subject [48, 49]. The context passage around the citation indicates the main content of the cited paper [62], however, the cited paper can be focused on a different subject of context [62, 63]. Context passages are used for several other purposes in literature, like inter-article similarity estimation [64], disambiguation of named entities [65], topic-based retrieval [9, 47], identification of biomedical articles [50], and newspaper citations in scholarly search [66]. However, extracting context passage is challenging due to the unavailability of full-text and therefore inability to conclude the subject of the paper efficiently [48, 49]. Intuitively, many popular scholarly search engines like Semantic Scholar use hybrid approaches for ranking documents. Much research has been done on the effectiveness of academic search engines. Some authors use graph-based approaches for the effectiveness of the academic search [67-71] while bearing in mind that a citation graph is usually sparse and noisy[68]. The solution in [72] supports scholarly search using key-queries [73] and query covers [74] to enhance the effectiveness of the academic search. However, their approach takes a research article as input for key-phrase selection and weighting methods, which result in suboptimal ranking. Due to the massive expansion in research paper repositories, the scholarly search is a very hot and challenging domain for both researchers and developers. Although several approaches have been proposed in the literature to address the requirements of scholars, we are still away from an ideal academic search engine that meets the heterogonous needs of different categories of scholarly users with minimum effort. Further research is required to address the requirements of academic searchers.

D. Ranking Algorithms for Scholarly Search

There is no universal ranking algorithm that scholarly retrieval systems use to rank documents in response to user queries. In most cases, scholarly retrieval models are quadruples {D, Q, F, R (q, d)} [75]. D is the representation component that is usually searched in the collection set. Q is the logical view of the user need. F is a framework and reasoning component for modeling document representation, query, and their relationship. R (q, d) is a reasoning component to rank the document as per the query terms. Due to the advancements in IR, numerous technologies and techniques are used for enhancing scholarly retrieval systems. For instance, Semantic Scholar uses semantic technologies for accomplishing the task of locating relevant documents. AceMap [76] academic search system analyzes big scholarly document datasets using the "map" approach. Google Scholar and DBLP use text-based methods to navigate. These scholarly

retrieval systems use different ranking algorithms that place matching results in their order of relevance. Some systems let the user choose the ranking factor (publication date, number of citations, author or journal name and reputation, and relevance of the document based on some predefined designed criteria). The factor selected by the user is given more weight in determining the relevance of documents. Some other systems like Google Scholar do not allow users to intervene in the weighting factor of ranking.

In most scholarly retrieval systems, the relevance of a document is measured by considering different document elements. For instance, how repeatedly the search term is found in the document and in which field (i.e., title, abstract, body, etc.). Commonly, if the search term occurs more often in a document or a more important field of the document, it is considered more relevant. For example, the term in the title is weighted more heavily than its occurrence in the abstract and so on. The weight of each term in the document is assigned to the total ranking weight based on term position. Some of the document fields that may be weighted differently by scholarly search systems are shown in Table II. Due to the unavailability of data, the ranking algorithms and their attributes of all available scholarly retrieval systems was not considered. We slightly considered the ranking mechanism of Google Scholar, the most widely used scholarly search engine. It takes into account multiple factors such as relevance, citation count, author name, name of publisher etc. [77] when generating results to a user query. In assessing the relevance of a given document and query, Google Scholar gives higher weight to the title. The citation attribute also plays a vital role in the ranking, and therefore, the documents having relatively many citations are likely to be placed near the top of the result list. Author and journal or conference name can also affect the ranking of documents, i.e. a query having an author or journal/conference name is likely to be positioned in the top of the search results list. For example, most of the top results of a search for "information retrieval" are likely to be articles about various IR topics from the Information Retrieval Journal. Google Scholar also considers publication date and some other attributes in ranking [77]. Intuitively, without loss of generality as the research paper repository is growing rapidly, it is essential to consider all the desired components of the scholarly domain in ranking algorithms that shatter the metaphor of scholarly retrieval systems. These citation networks are discussed in detail below.

TABLE II. DIFFERENT FIELDS OF A SCHOLARLY DOCUMENT

Document text	**Document metadata**	**Electronic files metadata**
Title	Author names	Title
Abstract	Publication venue (journal, conference, etc.)	Author
Subheadings	Incites	Description
Body Text	Outcites	Size
Figures	Social tags	Filename
Tables	Social annotations	Date
Author keywords	Author reputation	No of terms

III. SCHOLARLY CITATION NETWORKS

In the scholarly domain, a citation network is a significant and critical part of scholarly retrieval models. A citation is a link from one scholarly document to another. When a document uses a text excerpt, an idea, a concept, a figure, an algorithm, etc. from another scholarly document, it usually refers to that document [78-80]. Citations are necessary because they help create links between publications and authors, give credit to authors, promote reusability and productivity, and provide a roadmap to discovery. Professional associations encourage scholars to replicate findings, results, improve research standards and give desired credit to scholars by citing their work when deemed relevant and related. In most cases the impact of an author, institution, journal or even a country concerning a particular field of study is measured in terms of citations count. For instance, a document with a lot of incites is considered more influential. The academic policymakers use citation networks via Google PageRank, and it's variants to quantify scholarly texts [81, 82]. Likewise, to credit authors, academic networks use the status of their citing authors to distinguish high-status authors in co-authorship networks [83, 84]. The citation networks also provide a technique to differentiate prestigious journals [85]. A journal is said to be prestigious if it has been cited by other prestigious journals and has numerous highly cited works. Institutions and countries are assessed using the same criteria [86]. Intuitively, citations are used in the retrieval models of many famous scholarly search engines such as Google Scholar in different ways like citation count, bibliographic coupling [41], and co-citation and citation context [47] in ranking results.

Citation networks form a complex graph. Consider a paper network where nodes are the scholarly documents and the edges are the citations between the papers, i.e. G (P,C), where P is the set of nodes (papers), and C is the set of edges (citations, i.e., in-cites and out-cites). It is a substantial complex graph that can have several sub-graphs including but not limited to paper graphs, author graphs, collaboration graphs and semantic graphs. For a collaboration graph, an edge (X, Y) exists if person X worked with person Y. In the case of semantic graphs, an edge (X,Y) exists if word X is associated with word Y. The insightful utilization of all the subgraphs in the scholarly network and their associated metadata can play a significant role in the efficiency of scholarly retrieval systems. For example, authors in [87] extracted the metadata from scholarly documents with the aim to create a knowledge-base of each scholarly article for efficient document retrieval. Citation networks can play a vital role in the systematic retrieval of scholarly literature. Much research has been carried out about how useful information from scholarly citation networks can be extracted and utilized for better information retrieval. CitNetExplorer [88] analyzes and visualizes citation networks to address citation-based scientific literature retrieval [89, 90]. The tool is helpful in finding full relevant papers about a specific topic for preparing a review article. Author in [91] extends the co-citation network by incorporating satellite documents. Co-citation is a relationship among two scholarly papers concurrently cited by a third scholarly document. When the co-citation linkage detects scholarly documents, it is conceivable to obtain more suitable search terms from the related document. Such terms may not have been included in the original seed document.

Despite their numerous benefits, the existing domain of citation networks considers all citations for a given document to be equally significant. This can lead to situations where inaccurate information is deemed to be relevant because several authors have cited it. For example, a paper titled "A vector space model for information retrieval" alleged to have been published in 1975 is considered the most commonly cited paper published by Gerard Salton even though it does not exist in reality [92]. The paper "Read before you cite!" suggests that authors read just 20% of the work they cite [93]. Other authors also concluded that 25% of the references are redundant, 40% are for aspiring only to minimum standards [94] and 62.7% address just definition, tools, etc. not attributed for a specific function [38-40]. All these show that to improve the quality of citation-based applications, citations should not be regarded as equally significant. To do this, several researchers have carried out research activities aimed at mitigating the challenges associated with citation networks [31], including:

- Scattering: There is no single authoritative place to keep a record of an entire academic citation network. Due to the distributed nature of the academic web, different search platforms have different citation metrics and analytics. This can pose a significant challenge when using citation networks to credit a paper, author, journal, institution, country, etc. [95].
- Uncertainty: The relationship between citations is not always available in repositories. For instance, in the ACM digital library, 18.5% of publications have no citing details while 55.6% lack any cited information [96]. Therefore different papers receive the same ranking score. Handling the erroneous and missing citations metadata (incites and out-cites) of scholarly documents is a massive challenge for academic retrieval systems [97].
- Restriction: As discussed earlier, the whole academic web is not freely available [2]. Since the citations used in a document are part of that document, the unavailability of such a document can be a profound challenge to accurate academic information retrieval.
- Integrating scholarly metrics and analytics: A citation network is a handy assessment tool for distinguishing different scholarly mark units [90, 95]. It is beneficial when determining the impact of papers, authors, co-authors, conferences, journals, institutions, projects, countries, etc., in a particular field of study. However, due to the challenges associated with citations, it may not always produce optimal results.
- Accessibility: Due to the growing rate of academic literature, it is challenging to locate relevant papers. More and more documents are published on a daily basis [97]. For instance, PloseOne alone published 30,000 documents with an average 85 documents per day in 2014 [97]. These publications inadvertently result in the addition of billions of nodes to the already existing citation networks. Web of Science, for example, accumulates about 1 billion citations per year. This makes the accessibility of citation network more challenging [97].
- Complex Graph: Citation network is a complex graph having some non-trivial features like instantaneous network evolution, complex nested topology, multiple nodes/edges and large-scale growing rate. These features make some of the algorithms needed to get optimal results inapplicable [98].

IV. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

The academic search is a fascinating research area. Several academic search engines exist today with Google Scholar being the dominant one. However, locating relevant documents is still challenging due to the high growth of the research papers repository. In this regard, much research is going on in the field of scholarly document ranking, retrieval, recommendation, and the proper exploitation of citation networks. We are still away from an efficient scholarly retrieval system. The reasons are many, including:

- Disambiguating authors: One can try Google Scholar while adding publications to his/her profile, where several articles are displayed being identified as possibly written by the user. For accurate disambiguation, email addresses, affiliations, city, country and field of expertise could be exploited together with the efficient classifier and machine learning algorithms.
- The limited use of semantic web technologies, especially ontologies and linked open data, makes popular scholarly retrieval systems limited, where reasoning and machine-understand-ability could bring fruitful results.
- Semantic web together with natural language processing could be employed in identifying and categorizing in-citations to differentiate between relevant papers and ones that were used for self-citations or improving/increasing the number of references in the bibliography section.
- User studies are required to understand the user interactions with scholarly retrieval systems to understand their information needs better so that more user-friendly solutions are produced. Given the high growth rate of the academic web, it is necessary to develop tools that realize and emphasize users' needs.
- The use of citation networks in scholarly retrieval and assessing the impact of scholarly works has achieved many fruitful results. However, efforts are required to exploit them to their fullest in building the scholarly reputation of the authors, research publications, and journals so that users could be able to judge the quality of a publication better. In this regard, the challenges mentioned above need the attention of researchers and practitioners.
- The performance of an academic search engine can be improved by in-depth insight into citation networks (i.e. paper network, author network, collaboration network and text network) and infer most influential citations. The relationship between citations, authors and publications can also be computed for each document to efficiently rank documents.

- The ranking algorithms of scholarly search engines are different, many factors of ranking documents are by nature ambiguous and confusing to formalize. Most of them are proprietary making it difficult to understand how they work. Therefore, detailed empirical studies are required, in order to understand their ranking techniques and devise solutions that are free, open-source, and which could be reproduced whenever required.

This review paper is an attempt to bring the attention of researchers and practitioners towards the endless possibilities in which a more efficient scholarly retrieval system could be developed. It emphasizes on mitigating the information overload that currently researchers, especially newcomers, face while trying to access the most intended and relevant papers. For a more efficient solution, it is essential first to understand user information needs, develop approaches in the light of these needs, and exploit citation networks and modern IR, machine learning, and semantic web technologies so that search engines could be able to better understand the content and provide access to the desired content timely and resourcefully.

REFERENCES

[1] Baidu Academic, available at: http://xueshu.baidu.com

[2] M. Khabsa, C. L. Giles, "The number of scholarly documents on the public web", PloS One, Vol. 9, No. 5, p. e93949, 2014

[3] E. Orduna-Malea, J. M. Ayllon, A. Martin-Martin, E. D. Lopez-Cozar, "About the size of Google Scholar: playing the numbers", available at: https://arxiv.org/abs/1407.6239, 2014

[4] Microsoft Academic, available at: https://academic.microsoft.com

[5] J. Wu, C. Liang, H. Yang, C. L. Giles, "CiteSeerX data: semanticizing scholarly papers", International Workshop on Semantic Big Data, San Francisco, USA, June 26 - July 1, 2016

[6] M. Liu, "Progress in documentation the complexities of citation practice: a review of citation studies", Journal of Documentation, Vol. 49, pp. 370-408, 1993

[7] D. Goldberg, D. Nichols, B. M. Oki, D. Terry, "Using collaborative filtering to weave an information tapestry", Communications of the ACM, Vol. 35, No. 12, pp. 61-70, 1992

[8] S. Bradshaw, "Reference Directed Indexing: Redeeming Relevance for Subject Search in Citation Indexes", in: International Conference on Theory and Practice of Digital Libraries, , pp. 499-510, Springer, 2003

[9] A. Ritchie, S. Teufel, S. Robertson, "Using Terms from Citations for IR: Some First Results", in: Advances in Information Retrieval, ECIR 2008, pp. 211-221, Springer, 2008

[10] A. Ritchie, Citation Context Analysis for Information Retrieval, University of Cambridge, 2009

[11] J. Beel, B. Gipp, S. Langer, C. Breitinger, "Research-paper recommender systems: a literature survey", International Journal on Digital Libraries, Vol. 17, No. 4, pp. 305-338, 2016

[12] K. Sugiyama, M. Y. Kan, "A comprehensive evaluation of scholarly paper recommendation using potential citation papers", International Journal on Digital Libraries, Vol. 16, No. 2, pp. 91-109, 2015

[13] C. He, D. Parra, K. Verbert, "Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities", Expert Systems with Applications, Vol. 56, pp. 9-27, 2016

[14] B. Sun, P. Mitra, C. Lee Giles, K. T. Mueller, "Identifying, indexing, and ranking chemical formulae and chemical names in digital documents", ACM Transactions on Information Systems (TOIS), Vol. 29, No. 2, p. 12, 2011

[15] S. Tuarob, S. Bhatia, P. Mitra, C. L. Giles, "AlgorithmSeer: A System for Extracting and Searching for Algorithms in Scholarly Big Data", IEEE Transactions on Big Data, Vol. 2, No. 1, pp. 3-17, 2016

[16] Y. Liu, K. Bai, P. Mitra, C. L. Giles, "TableSeer:automatic table metadata extraction and searching in digital libraries", 7th ACM/IEEE-CS Joint Conference on Digital libraries, Vancouver, British Columbia, Canada, June 17-22, 2007

[17] M. Khabsa, P. Treeratpituk, C. L. Giles, "AckSeer:a repository and search engine for automatically extracted acknowledgments from digital libraries", ACM/IEEE-CS Joint Conference on Digital Libraries, Washington, USA, June 10-14, 2012

[18] M. Singh, B. Barua, P. Palod, M. Garg, S. Satapathy, S. Bushi, K. Ayush, K. S. Rohith, T. Gamidi, P. Goyal, A. Mukherjee, "OCR++: A Robust Framework For Information Extraction from Scholarly Articles", 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, December 11-17, 2016

[19] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, E. A. Fox, "Automatic document metadata extraction using support vector machines", Joint Conference on Digital Libraries, Houston, USA, May 27-31, 2003

[20] M. Lipinski, K. Yao, C. Breitinger, J. Beel, B. Gipp, "Evaluation of header metadata extraction approaches and tools for scientific PDF documents", 13th ACM/IEEE-CS Joint Conference on Digital libraries, Indianapolis, USA, July 22-26, 2013

[21] Apache Tika, available at: https://tika.apache.org

[22] P. Lopez, "GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications", in: Research and Advanced Technology for Digital Libraries, pp. 473-474, Springer, 2009

[23] C. A. Clark, S. K. Divvala, "Looking Beyond Text: Extracting Figures, Tables and Captions from Computer Science Papers", in: AAAI Workshop: Scholarly Big Data, AAAI Publications, 2015

[24] S. Tuarob, S. Bhatia, P. Mitra, C. L. Giles, "Automatic detection of pseudocodes in scholarly documents using machine learning", 12th International Conference on Document Analysis and Recognition, Washington, USA, August 25-28, 2013

[25] I. G. Councill, C. L. Giles, M. Y. Kan, "ParsCit: an Open-source CRF Reference String Parsing Package", LREC, Vol. 8, pp. 661-667, 2008

[26] S. R. Choudhury, S. Wang, C. L. Giles, "Scalable algorithms for scholarly figure mining and semantics", International Workshop on Semantic Big Data, San Francisco , USA, June 26-July 1, 2016

[27] G. Veena, J. Mathew, J. Joseph, "A Survey on Search Systems for Extracting And Searching in Scholarly Big Data", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Special No. 14, pp. 98-103, 2016

[28] X. Li, M. D. Rijke, "Do Topic Shift and Query Reformulation Patterns Correlate in Academic Search?", in: Advances in Information Retrieval, Springer, 2017

[29] S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder, D. Grossman, "Temporal analysis of a very large topically categorized Web query log", Journal of the American Society for Information Science & Technology, Vol. 58, No. 2, pp. 166–178, 2007

[30] A. Di Iorio, R. Giannella, F. Poggi, S. Peroni, F. Vitali, "Exploring Scholarly Papers Through Citations", 2015 ACM Symposium on Document Engineering, Lausanne, Switzerland, September 8-11, 2015

[31] M. H. MacRoberts, B. R. MacRoberts, "Problems of citation analysis: A study of uncited and seldom ‐ cited influences", Journal of the American Society for Information Science and Technology, Vol. 61, No. 1, pp. 1-12, 2010

[32] X. Y. Liu, B. C. Chien, "Applying Citation Network Analysis on Recommendation of Research Paper Collection", 4th Multidisciplinary International Social Networks Conference, Bangkok, Thailand, July 17-19, 2017

[33] S. M. Mcnee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, J. Riedl, "On the recommending of citations for research papers", ACM Conference on Computer supported cooperative work, New Orleans, USA, November, 16-20, 2002

[34] A. Silvescu, A. Silvescu, P. Mitra, C. L. Giles, "Can't see the forest for the trees?: a citation recommendation system", ACM/IEEE-CS Joint Conference on Digital Libraries, Indianapolis, USA, July 22-26, 2013

[35] K. Sugiyama, M. Y. Kan, "Exploiting potential citation papers in scholarly paper recommendation", 13th ACM/IEEE-CS Joint Conference on Digital libraries, Indianapolis, USA, July 22-26, 2013

[36] Q. He, J. Pei, D. Kifer, P. Mitra, L. Giles, "Context-aware citation recommendation", International Conference on World Wide Web, Raleigh, USA, April, 2010

[37] B. Golshan, T. Lappas, E. Terzi, "SOFIA SEARCH: a tool for automating related-work search", ACM SIGMOD International Conference on Management of Data, Scottsdale, USA, May 20-24, 2012

[38] K. Toutanova, C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger", 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, Vol. 13, pp. 63-70, Hong Kong, October 7-8, 2000

[39] T. Chakraborty, R. Narayanam, "All fingers are not equal: Intensity of references in scientific articles", 2016 Conference on Empirical Methods in Natural Language Processing, Austin, USA, November 1-5, 2016

[40] S. Kumar, "Structure and dynamics of signed citation networks", 25th International Conference Companion on World Wide Web, Montreal, Canada, April 11-15, 2016

[41] M. M. Kessler, "Bibliographic coupling between scientific papers", American Documentation, Vol. 14, No. 1, pp. 10-25, 1963

[42] P. Calado, M. Cristo, E. Moura, N. Ziviani, B. Ribeiro-Neto, M. A. Concalves, "Combining link-based and content-based methods for web document classification", 12th International Conference on Information and Knowledge Management, New Orleans, USA, November 3-8, 2003

[43] T. Couto, M. Cristo, M. A. Goncalves, P. Calado, N. Ziviani, E. Moura, B. Ribeiro-Neto, "A comparative study of citations and links in document classification", 6th ACM/IEEE-CS Joint Conference on Digital Libraries, Chapel Hill, USA, June 11-15, 2006

[44] B. Gipp, Citation-based Plagiarism Detection: Detecting Disguised and Cross-language Plagiarism using Citation Pattern Analysis, Springer, 2014

[45] B. Gipp, N. Meuschke, "Citation pattern matching algorithms for citation-based plagiarism detection:greedy citation tiling, citation chunking and longest common citation sequence", International Symposium on Parallel Architectures, Algorithms, and Networks, Mountain View, USA, September, 19-22, 2011

[46] S. Kumar, P. K. Reddy, V. P. Reddy, A. Singh, "Similarity analysis of legal judgments", ACM Bangalore Conference, Bangalore, Karnataka, India, March 25-26, 2011

[47] S. Liu, C. Chen, K. Ding, B. Wang, K. Xu, Y. Lin, "Literature retrieval based on citation context", Scientometrics, Vol. 101, Vol. 2, pp. 1293-1307, 2014

[48] S. Teufel, "Argumentative Zoning for Improved Citation Indexing", in Computing Attitude and Affect in Text: Theory and Applications Vol. 20, pp. 159-169, Springer, 2006

[49] S. Mohammad, B. Dorr, M. Egan, A. Hassan, P. Muthukrishan, V. Qazvinian, D. Radev, D. Zajic, "Using Citations to Generate Surveys of Scientific Paradigms", Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Boulder, USA, May 31 - June 5, 2009

[50] R. L. Liu, "Retrieval of Scholarly Articles with Similar Core Contents", International Journal of Knowledge Content Development & Technology, Vol. 7, No. 3, pp. 5-27, 2017

[51] Apache Lucene, available at: http://lucene.apache.org

[52] J. S. Whissell, C. L. A. Clarke, "Effective measures for inter-document similarity", 22nd ACM International Conference on Information & Knowledge Management, San Francisco, USA, October 27 - November 1, 2013

[53] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, K. Borner, "Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches", Plos One, Vol. 6, No. 3, p. e18029, 2011

[54] P. Glenisson, F. Janssens, B. D. Moor, "Combining full text and bibliometric information in mapping scientific disciplines", Information Processing & Management An International Journal, Vol. 41, No. 6, pp. 1548-1572, 2005

[55] T. K. Landauer, D. Laham, M. Derr, "Colloquium Paper: Mapping Knowledge Domains: From paragraph to graph: Latent semantic analysis for information visualization", Proceedings of the National Academy of Sciences USA, Vol. 101, Suppl. 1, pp. 5214-5219, 2004

[56] S. E. Robertson, S. Walker, M. Beaulieu, P. Willett, "Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track", Nist Special Publication SP 500, pp. 253-264, 1999

[57] R. L. Liu, "Passage-Based Bibliographic Coupling: An Inter-Article Similarity Measure for Biomedical Articles", Plos One, Vol. 10, No. 10, p. e0142026, 2015

[58] K. W. Boyack, R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?", Journal of the American Society for Information Science & Technology, Vol. 61, No. 12, pp. 2389-2404, 2010

[59] F. Janssens, W. Glanzel, B. D. Moor, "A hybrid mapping of information science", Scientometrics, Vol. 75, No. 3, pp. 607-631, 2008

[60] B. Gipp, J. Beel, "Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis", 12th International Conference on Scientometrics & Informetrics, Rio de Janeiro, Brazil, July 14-17, 2009

[61] K. W. Boyack, H. Small, R. Klavans, "Improving the accuracy of co-citation clustering using full text", Journal of the American Society for Information Science & Technology, Vol. 64, No. 9, pp. 1759-1767, 2013

[62] X. Liu, J. Zhang, C. Guo, "Full-text citation analysis: A new method to enhance scholarly networks", Journal of the American Society for Information Science & Technology, Vol. 64, No. 9, pp. 1852-1863, 2013

[63] H. Small, "Interpreting maps of science using citation context sentiments: a preliminary investigation", Scientometrics, Vol. 87, No. 2, pp. 373-388, 2011

[64] B. Aljaber, N. Stokes, J. Bailey, J. Pei, "Document clustering of scientific texts using citation contexts", Information Retrieval, Vol. 13, No. 2, pp. 101-131, 2010

[65] P. I. Nakov, A. S. Schwartz, M. A. Hearst, "Citances: Citation sentences for semantic analysis of bioscience text", SIGIR 04 Workshop on Search & Discovery in Bioinformatics, Sheffield, UK, July 25-29, 2004

[66] M. A. J. Singh, D. S. Ravikumar, Newspaper Citation in Scholarly Publications: A Study on Financial Times Newspaper during 2001- 2010 as Reflected in Web of Science, Library Philosophy & Practice, University of Nebraska, 2018

[67] K. Sugiyama, M. Y. Kan, "Exploiting potential citation papers in scholarly paper recommendation", 13th ACM/IEEE-CS Joint Conference on Digital Libraries, Indianapolis, USA, July 22-26, 2013

[68] C. Caragea, A. Silvescu, P. Mitra, C. L. Giles, "Can't see the forest for the trees?: a citation recommendation system", 13th ACM/IEEE-CS Joint Conference on Digital Libraries, Indianapolis, USA, July 22-26, 2013

[69] C. Wang, D. M. Blei, "Collaborative topic modeling for recommending scientific articles", 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, August 21-24, 2011

[70] O. Kucuktunc, E. Saule, K. Kaya, U. V. Catalyurek, "TheAdvisor: a web service for academic recommendation", 13th ACM/IEEE-CS Joint Conference on Digital Libraries, Indianapolis, USA, July 22-26, 2013

[71] M. D. Ekstrand, P. Kannan, J. A. Stemper, J. T. Butler, J. A. Konstan, J. T. Riedl, "Automatically building research reading lists", 4th ACM Conference on Recommender Systems, Barcelona, Spain, September 25-30, 2010

[72] M. Hagen, A. Beyer, T. Gollub, K. Komlossy, B. Stein, "Supporting Scholarly Search with Keyqueries", 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016

[73] T. Gollub, M. Hagen, M. Michel, B. Stein, "From keywords to keyqueries: content descriptors for the web", 36th International ACM

SIGIR Conference on Research and Development in Information retrieval, Dublin, Ireland, July 28-August 1, 2013

[74] M. Hagen, B. Stein, "Candidate document retrieval for web-scale text reuse detection", International Symposium on String Processing and Information Retrieval, Pisa, Italy, October 17-21, 2011

[75] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley Longman Publishing, 1999

[76] Z. Tan, C. Liu, Y. Mao, Y. Guo, J. Shen, X. Wang, "AceMap: A Novel Approach towards Displaying Relationship among Academic Literatures", 25th International Conference Companion on World Wide Web, Montreal, Canada, April 11-15, 2016

[77] J. Beel, B. Gipp, E. Wilde, "Academic Search Engine Optimization (ASEO) Optimizing Scholarly Literature for Google Scholar & Co", Journal of Scholarly Publishing, Vol. 41, No. 2, pp. 176-190, 2009

[78] M. T. Luong, T. D. Nguyen, M. Y. Kan, "Logical structure recovery in scholarly articles with rich document features", in: Multimedia Storage and Retrieval Innovations for Digital Library Systems, pp. 270-292, IGI Global, 2012

[79] K. Siler, "Citation choice and innovation in science studies", Scientometrics, Vol. 95, No. 1, pp. 385-415, 2013

[80] C. L. Borgman, "Data, Data Citation, and Bibliometrics", Taiwan Data Curation and Citation Workshop, Taipei, Taiwan, December 5, 2016

[81] P. Chen, H. Xie, S. Maslov, S. Redner, "Finding scientific gems with Google's PageRank algorithm", Journal of Informetrics, Vol. 1, No. 1, pp. 8-15, 2007

[82] N. Ma, J. Guan, Y. Zhao, "Bringing PageRank to the citation analysis", Information Processing & Management, Vol. 44, No. 2, pp. 800-810, 2008

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

[84] F. Radicchi, S. Fortunato, B. Markines, A. Vespignani, "Diffusion of scientific credits and the ranking of scientists", Physical Review E, Vol. 80, No. 5, p. 056103, 2009

[85] E. C. Rosenthal, H. J. Weiss, "A data envelopment analysis approach for ranking journals", Omega, Vol. 70, pp. 135-147, 2016

[86] E. Yan, C. R. Sugimoto, "Institutional interactions: Exploring social, cognitive, and geographic relationships between institutions as demonstrated through citation networks", Journal of the American Society for Information Science and Technology, Vol. 62, No. 8, pp. 1498-1514, 2011

[87] J. Wu, J. Killian, H. Yang, K. Williams, S. R. Choudhury, S. Tuarob, C. Caragea, C. L. Giles, "Pdfmef: A multi-entity knowledge extraction framework for scholarly documents and semantic search", 8th International Conference on Knowledge Capture, Palisades, USA, October 7-10, 2015

[88] CiteNetExplorer, available at: http://www.citnetexplorer.nl

[89] N. J. Van Eck, L. Waltman, "Systematic Retrieval of Scientific Literature based on Citation Relations: Introducing the CitNetExplorer Tool", European Conference on Information Retrieval, Amsterdam, Netherlands, April 13-16, 2014

[90] N. J. van Eck and L. Waltman, "CitNetExplorer: A new software tool for analyzing and visualizing citation networks", Journal of Informetrics, Vol. 8, No. 4, pp. 802-823, 2014

[91] M. Eto, "Incorporating Satellite Documents into Co-citation Networks for Scientific Paper Searches", Joint Workshop on Bibliometric-Enhanced Information Retrieval and NLP for Digital Libraries, New Jersey, USA, June 19-23, 2016

[92] D. Dubin, "The most influential paper Gerard Salton never wrote", Library Trends, Vol. 52, No. 4, pp. 748-764, 2004

[93] M. V. Simkin, V. P. Roychowdhury, "Read before you cite!", Complex Systems, Vol. 14, pp. 269-274, 2003

[94] M. J. Moravcsik, P. Murugesan, "Some Results on the Function and Quality of Citations: Social Studies of Science", Social Studies of Science Vol. 3, No. 4, p. 538, 1988

[95] E. Yan, Y. Ding, "Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other", Journal of the American Society for Information Science and Technology, Vol. 63, No. 7, pp. 1313-1326, 2012

[96] Z. Jiang, X. Liu, "Recovering missing citations in a scholarly network: a 2-step citation analysis to estimate publication importance", 13th ACM/IEEE-CS Joint Conference on Digital libraries, Indianapolis, USA, July 22-26, 2013

[97] C. Chen, M. Song, "The Uncertainty of Science: Navigating Through the Unknown", in: Representing Scientific Knowledge, pp. 1-35, Springer, 2017

[98] H. Shakibian, N. M. Charkari, "Optimization problems in complex networks: Challenges and directions", 24th Iranian Conference on Electrical Engineering (ICEE), Shiraz, Iran, May 10-12, 2016