Knowledge Extraction, Growth Study and Impact Prediction for Scientific Documents

Mayank Singh

Knowledge Extraction, Growth Study and Impact Prediction for Scientific Documents

Thesis submitted to the
Indian Institute of Technology, Kharagpur
for award of the degree

of

Doctor of Philosophy

by

Mayank Singh

Under the supervison of

Prof. Pawan Goyal
and
Prof. Animesh Mukherjee

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

February 2019

APPROVAL OF THE VIVA-VOCE BOARD

Date: _____

Certified that the thesis entitled **"Knowledge Extraction, Growth Study and Impact Prediction for Scientific Documents"** submitted by **Mayank Singh** to the Indian Institute of Technology, Kharagpur, for the award of the degree Doctor of Philosophy has been accepted by the external examiners and that the student has successfully defended the thesis in the viva-voce examination held today.

(Member of the DSC) (Member of the DSC) (Member of the DSC)

(Supervisor) (Joint Supervisor)

(External Examiner) (Chairman)

Certificate

This is to certify that the thesis entitled **"Knowledge Extraction, Growth Study and Impact Prediction for Scientific Documents"** submitted by **Mayank Singh (13CS92R02)** to the Indian Institute of Technology Kharagpur, is a record of bonafide research work carried under my supervision and is worthy of consideration for the award of the Doctor of Philosophy of the Institute. To the best of my knowledge, the results embodied in this thesis have not been submitted to any other University or Institute for the award of any other Degree or Diploma.

Date:
Place: IIT Kharagpur

Dr. Pawan Goyal
Assistant Professor
CSE, IIT Kharagpur

Dr. Animesh Mukherjee
Associate Professor
CSE, IIT Kharagpur

Declaration

I certify that

a. The work contained in this thesis is original and has been done by myself under the general supervision of my supervisors.

b. The work has not been submitted to any other Institute for any degree or diploma.

c. I have followed the guidelines provided by the Institute in writing the thesis.

d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

e. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

f. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Date: _____

Place: IIT Kharagpur Mayank Singh

Acknowledgements

I would like to express my sincere gratitude to Prof. Pawan Goyal and Prof. Animesh Mukherjee for their continual support, motivation and providing me the opportunity to carry out this research work. I also thank Prof. Soumen Chakrabarti (IIT Bombay) for his active collaboration and meticulous suggestions. I would also like to express thank to Dr. Lipika Dey and Dr. Arindam Pal (TCS Innovation labs) for their collaboration in conducting patent text studies and introducing me to the amazing field of intellectual property rights. It was a pleasure to work with Dr. Constantine Dovrolis (Georgia Tech, Atlanta) on Hourglass effect in citation networks. I would also like to thank Dr. Tridib Mukherjee (Xerox Research Center, India) for motivating and guiding me to work on interesting social problems using mobile technology. Special thanks to all current members and alumni of Complex Network Research Group ($CNeRG$) for their constant support and for the confidence that they have placed on me. I am very grateful to all Professors in $CNeRG$, especially Prof. Niloy Ganguly, for productive suggestions on presentation and writing skills, and wonderful reading group discussions. A special thanks to Dr. Tanmoy Chakraborty (IIIT Delhi) for developing a research acumen during my initial research days.

I am also very grateful to many people for helping me throughout my stay in IIT Kharagpur. In particular, I am very grateful to Amrith Krishna, Sandipan Sikdar, Rohit Verma, Satadal Sengupta, Suman Kalyan Maity, Barnopriyo Barua, Priyank Palod and Siva Prakasam for immense help and coordination. I thank Rajdeep Sarkar for excellent collaborative efforts and many fruitful discussions. I owe this work to my wife Ginni Singh for her constant love, encouragement and moral support, and my parents Anil Kumar Singh and Kiran Singh, for their continuous encouragement and blessings.

Finally, I thank $CNeRG$ for financial help on many occasions. I am grateful to Tata Consultancy Service Limited (TCS) for financially supporting my entire Ph.D. tenure. I am also greatly thankful to Google, Flipkart, SIGIR, KDD, and IARCS for supporting my conference travels at several instances.

Author's Biography

Mayank Singh has received his B.Tech. degree in 2012 from Computer Science and Engineering department of Indian Institute of Technology Jodhpur. He has been pursuing Ph.D. at the department of Computer Science and Engineering, IIT Kharagpur, since December 2013. Currently, he is TCS research fellow. His research interest includes Data Mining with applications in Digital Libraries, Natural Language and Processing, and Complex Networks.

Publications from the Thesis

1. **Mayank Singh**, Rajdeep Sarkar, Atharva Vyas, Pawan Goyal, Animesh Mukherjee, and Soumen Chakrabarti. "Automated Early Leaderboard Generation From Comparative Tables". In Proceedings of the 41st European Conference on Information Retrieval, 2019.

2. **Mayank Singh**, Pradeep Dogga, Sohan Patro, Dhiraj Barnwal, Ritam Dutt, Rajarshi Haldar, Pawan Goyal, and Animesh Mukherjee. "CL Scholar: The ACL Anthology Knowledge Graph Miner". In Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2018.

3. **Mayank Singh**, Rajdeep Sarkar, Pawan Goyal, Animesh Mukherjee, and Soumen Chakrabarti. "Relay-Linking Models for Prominence and Obsolescence in Evolving Networks". In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1077–1086. ACM, 2017.

4. **Mayank Singh**, Ajay Jaiswal, Priya Shree, Arindam Pal, Animesh Mukherjee, and Pawan Goyal. "Understanding the Impact of Early Citers on Long-Term Scientific Impact". In Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries, pp. 1-10. ACM, 2017.

5. **Mayank Singh**, Abhishek Niranjan, Divyansh Gupta, Nikhil Angad Bakshi, Animesh Mukherjee, and Pawan Goyal. "Citation Sentence Reuse Behavior of Scientists: A Case Study on Massive Bibliographic Text Dataset of Computer Science". In Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries, pp. 277-280. IEEE Press, 2017.

6. **Mayank Singh**, Soham Dan, Sanyam Agarwal, Pawan Goyal, and Animesh Mukherjee. "AppTechMiner: Mining Applications and Techniques from Scientific Articles". In Proceedings of the 6th International Workshop on Mining Scientific Publications, pp. 1–8. ACM, 2017.

7. **Mayank Singh**, Barnopriyo Barua, Priyank Palod, Manvi Garg, Sidhartha Satapathy, Samuel Bushi, Kumar Ayush, Krishna Sai Rohith, Tulasi Gamidi, Pawan Goyal, and Animesh Mukherjee "OCR++: A Robust Framework for Information Extraction from Scholarly Articles". In the 26th International Conference on Computational Linguistics (COLING), pp. 3390–3400. 2016.

8. **Mayank Singh**, Tanmoy Chakraborty, Animesh Mukherjee, and Pawan Goyal. "Is this conference a top-tier? ConfAssist: An Assistive Conflict Resolution Framework for Conference Categorization". Journal of Informetrics (JOI) 10(4), pp. 1005-1022. 2016.

9. **Mayank Singh**, Vikas Patidar, Suhansanu Kumar, Tanmoy Chakraborty, Animesh Mukherjee, and Pawan Goyal. "The Role of Citation Context in Predicting Long-Term Citation Profiles: An Experimental Study Based on a Massive Bibliographic Text Dataset". In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM), pp. 1271-1280. ACM, 2015.

10. **Mayank Singh**, Tanmoy Chakraborty, Animesh Mukherjee, and Pawan Goyal. "ConfAssist: A Conflict Resolution Framework for Assisting the Categorization of Computer Science Conferences". In Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), pp. 257-258. ACM, 2015.

11. **Mayank Singh**, Soumajit Pramanik, and Tanmoy Chakraborty. "PubIndia: A Framework for Analyzing Indian Research Publications in Computer Science". D-Lib Magazine 21, no. 11/12, 2015.

Abstract

Understanding scholarly articles is a key ingredient of impressive research recipe. Scholarly articles keep the scientific community up to date with the current research and development results and ideas. With the tremendous advancement in Internet infrastructure, we are witnessing an ongoing explosion in scholarly information that is generated. In this thesis, we attempt to introduce, study and solve some of the challenges emanating from scholarly volume overload. In particular, we look into three different dimensions: (i) metadata, structure, bibliography and experimental performance extraction from scholarly articles, (ii) designing network-assisted aging growth models for evolving citation networks with novel proposal of temporal summaries, and (iii) leveraging textual and network information to design long-term scientific impact prediction frameworks. While the first objective is related to the curation of scientific data, the second one pertains to its growth and the third one demonstrates real-world application.

Curation: We develop two open-source knowledge extraction frameworks for scholarly articles. The first framework, *OCR++*, is a hybrid framework to extract textual information such as metadata, structure, and bibliography from PDF research articles. OCR++ employs a variety of Conditional Random Field (CRF) models and hand-written rules specially crafted for handling various tasks. The second framework mines experimental performance from papers embedded within comparative tables to construct performance tournament graphs that encode information about performance comparisons between scientific papers. We also present a bunch of different ways to aggregate the tournament edges and a bunch of ways to score and rank papers on the basis of this incomplete and noisy information. We show that our scheme of ranking brings forth the state-of-the-art papers at the top of the ranklist unlike well-established academic search systems like Google Scholar and Semantic Scholar that mostly place highly cited papers at the top of the ranklist. Also, the system is useful in automatically discovering and maintaining leaderboards in the form of partial orders between papers.

Growth: We next present the first plausible network-driven set of models for obsolescence in the context of research paper citations, based on a natural notion of *relay-linking*. In fact, we observe that such a relaying process indeed exists by conducting a challenging micro-scale experiment on real data. We propose several measurements on citation network that constitute temporal signatures summarizing the coexistence of entrenchment and obsolescence. Our proposed network influenced models of aging mimic temporal signatures of real networks better than state-of-the-art aging models.

Application: Finally, as an application of curated scientific knowledge, we improve upon scientific impact prediction. We present empirical evidences of high correlation between (i) two textual features extracted from the citation contexts and (ii) three different characteristic properties of early citers with long-term citation counts of the paper. We append these features along with various other features available at the time of publication to improve the prediction accuracy of state-of-the-art baselines with high margin.

Keywords: scholarly knowledge, information extraction, citation network, performance graph, network growth, long-term scientific impact.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Scientific articles are the principal medium of communication that keep the scientific community up to date with the current research and development results. They comprise publications that report theoretical, empirical, and application works in natural and social sciences and engineering. Scientific knowledge is continuously growing with new research that builds its foundation on earlier research. Current research works may improve, corroborate, or refute existing knowledge in a specific domain. Scientific articles comprise patent articles, academic research articles, white papers, books, etc. The number of English-language scientific articles is estimated to be above 114 million by the beginning of the year 2013 [94]. Jinha *et al.* [89] showed that scientific volume is increasing with an annual rate of $\sim 3\%$. This overwhelming volume of articles is a monumental source of scientific knowledge and know-how. In order to maintain this vast resource of knowledge, we should continuously focus on the development of scalable archival systems, academic library systems, etc.

With tremendous growth in Internet infrastructure, several online scholarly systems exists that not only substantially contribute in article curation ($ArXiv$, $NCBI$, etc.), search and recommendation ($Google\ Scholar$, $Microsoft\ Academic\ Search$, $Semantic\ Scholar$, $Aminer$, etc.), but also in social tasks like community building ($ResearchGate$), researcher identification ($ORCHID$), code/data sharing ($GitHub$), knowledge sharing ($Stack\ Overflow$), etc. These systems are not only immensely helpful to the research community in searching through the vast volume of data, but also play a critical role in community building. This ever-growing popularity of academic systems has led to several seminal works on optimization, archival, scalability, retrieval, electronic publishing, and privacy.

The current thesis focuses on a better understanding of academic research

articles. The thesis introduces several interesting challenges in this area and tries to provide elegant solutions to some of them. The academic research articles are selected due to their easy availability. However, at several places, the thesis claims to achieve similar performance on non-scholarly datasets as well. The next section describes some of the intriguing challenges that we intend to address in the current thesis.

1.1 Major challenges

The applications of data science to the scientific corpus is in its preliminary stages. Current academic systems leverage rich semi-structured data present within research articles. However, exponential growth in research volume has led to the intractable problems of scientific data accessibility [58], reproducibility [12], text reuse and obfuscation [40, 118], and impact measurement [168] that hampers our end goal of efficient research understanding. Some of the fundamental challenges are summarized below:

- **Inefficiency and insufficiency of current scholarly extraction systems:** Majority of scientific articles exist in Portable Document Format (PDF) [89]. Processing of PDF scientific articles remains a major bottleneck in automatic metadata extraction. Processing inaccuracies exist mainly due to inherent OCR technology limitations. In addition, lack of a standard extraction methodology to support diverse formatting styles adopted by different publishing venues leads to insufficiency in scholarly extraction systems. Machine learning research in this area has been dramatically limited by the lack of large-scale annotated corpora. Also, comparatively less attention has been paid to document structure analysis than metadata and bibliography extraction from scientific articles.
- **Unexplored performance comparisons for ranking research articles:** Performance comparison against competitive methods is a fundamental aspect of reporting research. However, the ordering of competing techniques in a leaderboard depends on a large number of factors, including the task being solved, the data set(s) used, the experimental conditions, and the choice of performance metrics. Current state-of-the-art scholarly information extraction systems like $GROBID$ [115], $ParseCit$ [48], etc., fail to extract rich semantic performance information present in tables and charts. As a consequence, present academic search systems utilize relevance, recency, and popularity based metrics like citation count, velocity, etc., instead of performance-based metrics

for ranking research articles. This inherent discrimination against reported performance results in inadequacy of comparative leaderboards. A primary difficulty in the development of such system is the unavailability of relevant data. Even if such a dataset is available, developing meaningful data mining techniques to extract from the noisy scientific text would be equally challenging.

- **Incomplete understanding of citation growth dynamics:** Scientific research volume is growing exponentially with more and more research articles entering into the scholarly system citing existing research papers. This fast-evolving scientific network (hereafter *'citation network'*) throws several challenges and opportunities. Therefore, understanding and modeling growth factors is of prime importance for the entire research community. Citation network growth modeling has received much attention in the last three decades [8, 55]. Despite this interest in general network modeling, not many efforts have been put forward to specifically understand several physical phenomena underlying the scientific network. Few examples of such physical phenomena are aging [72, 186], prominence [62], topic shifts [138], etc. Some of the highly celebrated models like preferential attachment [8] and link copying models [100], while enabling elegant analysis, only capture rich-gets-richer effects, not aging and decline. Recent aging models are complex and heavily parameterized; most involve estimating 1–3 parameters per paper [168]. These parameters are intrinsic: they explain the decline in terms of events in the past of the same paper and do not explain, using the network, where the citation might go instead. Also, traditional characterization of linking dynamics is insufficient to judge the faithfulness of models. Collecting real evidence to show the existence of link diversion in the scientific network is an impractical task due to its inherent structure. Also, developing a modeling framework in absence of such real evidence presents a significant difficulty. Therefore, an immediate task would be to determine whether this phenomenon could be captured in a real citation network.

- **Unexplored early information to determine future scientific impact:** Success of a research work is estimated by its scientific impact. Quantifying scientific impact through citation counts or metrics has received much attention in the last two decades [136, 182]. However, prediction of future citation counts is an extremely challenging task because of the nature and dynamics of citations [61, 70]. For example, in contrast to the popular perception of unique citation profile, Chakraborty *et al.* [31] showed the existence of six different citation profiles. Re-

cent advancement in the prediction of future citation counts has led to the development of complex mathematical and machine learning based models. The existing supervised models have employed several paper, venue, and author-centric features that can be obtained at the publication time [31, 182]. Despite this enormous interest, the characteristics of early information generated immediately after publication have not been dealt with in-depth. Information available within 1–2 years after publication includes initial popularity trends measured by citation counts, textual patterns of incoming citation contexts, popularity and productivity of early citing authors, etc. Availability of rich time-stamped scholarly data is a major challenge. Also, curation of citation contexts from article's full text presents a challenging IR task.

With these challenges at hand, the thesis aims to solve different problems in scholarly science particularly concentrating on the following - (i) information extraction from scholarly articles, (ii) mining performance comparisons to rank scholarly articles, (iii) modeling scientific growth through relay-linking phenomenon, and (iv) estimating the long-term scientific impact. While the first two objectives are related to the curation of scientific data, the third one pertains to its growth, and the fourth one demonstrates real-world application.

1.2 Objectives

The thesis addresses four different issues which contribute to the four different chapters –

1. **Knowledge extraction from scholarly articles:** Here we intend to perform robust information extraction from research articles. In specific, we aim to develop a robust scholarly information extraction framework that automatically processes PDF scientific articles in a scalable fashion. We believe that a key strategy to tackle problems (described in the previous section) is to analyze research articles from different publishers to identify generic patterns and rules, necessary for various information extraction tasks.

2. **Mining performance comparisons to rank scholarly articles:** We plan to develop a novel bibliometric system that robustly mines experimental performance information reported in scientific articles. Performance information is created by comparing multiple competitive methods against several evaluation metrics. We plan to extract performance information from comparative tables. We also aim to propose a novel

performance tournament graph with papers as nodes, where edges encode noisy performance comparison information extracted from papers. In the face of noisy extractions, we plan to develop several approaches to rank papers, identify the best of them, and show that commercial academic search systems fail miserably at automatic leaderboard discovery and at finding papers listed in widely-used lists of state-of-the-art papers in several areas of Computer Science.

3. **Modeling scientific growth through relay-linking phenomenon:** Our next objective is to study the underlying behavior of link formation in scholarly citation networks. We aim to develop a network growth model based on surprising inversion or undoing of triangle completion, where an old node diverts a citation to a younger follower in its immediate vicinity. We argue that, in contrast to traditional growth models, the proposed model will provide better fitness against the real citation network.

4. **Estimating long-term scientific impact:** As our final objective we intend to leverage information from curated scientific data to develop predictive systems for long-term popularity prediction. Citation count of a publication is the most commonly accepted metric by the research community to evaluate the impact and quality of a research article. We aim to augment information available at publication time with early information available soon after publication. We plan to utilize both textual and network information generated within 1–2 years after publication to estimate future citation counts. Specifically, we plan to study early indicators of long-term popularity.

1.3 Knowledge extraction from scholarly articles

Obtaining structured data from documents is necessary to support retrieval tasks. Various scholarly organizations and companies deploy information extraction tools in their production environments. Through a comprehensive literature survey, we find comparatively less research in document structure analysis than metadata and bibliography extraction from scientific documents. We propose *OCR++*, an open-source framework designed for a variety of information extraction tasks from scholarly articles including metadata (title, author names, affiliation, and e-mail), structure (section headings and body text, table and figure headings, URLs and footnotes) and bibliography (citation instances and references). We analyze a diverse set of scientific articles written in the

English language to understand generic writing patterns and formulate rules to develop this hybrid framework. Extensive evaluations show that the proposed framework outperforms the existing state-of-the-art tools with a huge margin in structural information extraction along with improved performance in metadata and bibliography extraction tasks, both in terms of accuracy and processing time.

1.4 Mining performance comparisons to rank scholarly articles

A leaderboard is a tabular presentation of performance scores of the best competing techniques that address a specific scientific problem. Manually maintained leaderboards take time to emerge, which induces a latency in performance discovery and meaningful comparison. This can delay dissemination of best practices to non-experts and practitioners. Regarding papers as proxies for techniques, we present a new system to automatically discover and maintain leaderboards in the form of partial orders between papers, based on performance reported therein. In principle, a leaderboard depends on the task, data set, other experimental settings, and the choice of performance metrics. Often there are also tradeoffs between different metrics. Thus, leaderboard discovery is not just a matter of accurately extracting performance numbers and comparing them. In fact, the levels of noise and uncertainty around performance comparisons are so large that reliable traditional extraction is infeasible. We mitigate these challenges by using relatively cleaner, structured parts of the papers, e.g., performance tables. We propose a novel performance improvement graph with papers as nodes, where edges encode noisy performance comparison information extracted from tables. Every individual performance edge is extracted from a table with citations to other papers. These extractions resemble (noisy) outcomes of 'matches' in an incomplete tournament. We propose several approaches to rank papers from these noisy 'match' outcomes. We show that our ranking scheme can reproduce various manually curated leaderboards very well. Using widely-used lists of state-of-the-art papers in 27 areas of Computer Science, we demonstrate that our system produces very reliable rankings. We also show that commercial scholarly search systems cannot be used for leaderboard discovery, because of their emphasis on citations, which favors classic papers over recent performance breakthroughs.

1.5 Modeling scientific growth through relay-linking phenomenon

Understanding scientific literature growth has received much attention over the last three decades. Citation network evolving over time by adding new papers and citation links show the fascinating interplay between prominence and obsolescence. With rapidly growing publication repositories, understanding the networked process of obsolescence is equally important for an emerging field. We propose several measurements on citation network that constitute a *temporal signature* summarizing the coexistence between entrenchment and obsolescence.

Our study led to a family of network growth models, roughly speaking: to add a citation in a new paper, choose an existing paper p_0, but if it is too old, walk back along a citation link to p_1 and (optionally) repeat the process. We call this hypothesized process *triad uncompletion* and the associated generative model *relay-linking*. Triad uncompletion is reverse of extremely popular triangle completion process where if links (u, v) and (v, w) are present, we add a new link (u, w)). In sharp contrast to existing work, we avoid modeling aging as governed by network-exogenous rules or distributions (whose complexity scales with the number of nodes). Even though, our models have only two global parameters shared over all nodes, the proposed relay-linking models mimic temporal signatures of real networks better than state-of-the-art aging models. We also conduct an interesting study to show that temporal signatures for various research communities can yield further insights into their comparative dynamics.

1.6 Estimating long-term scientific impact

Highly-cited works remain as one of the most important criteria for various organizations (e.g., companies, universities and governments) to identify the best talents, especially at their initial stages. An early estimate would help in identification of promising articles that could accelerate research and dissemination of new knowledge. The existing works have used various venue and author-centric features, along with the citation information from the initial years for the task of citation prediction. We argue that the features extracted from the citation contexts can be extremely helpful for the future impact prediction. A citation context is, in principle, a set of sentences where a paper is referred to. We show that even using some very simplistic features extracted from the citation context can boost the performance of a citation prediction system significantly.

Next, we study the influence of early citing authors whom we call *early citers* (EC) on the long-term scientific impact of a paper. This study proposes several interesting properties of EC which leads to a brand new paradigm in citation behavior analysis. Using a massive computer science bibliographic dataset we identify two distinct categories of EC – we call those authors who have high overall publication/citation count in the dataset as *influential* and the rest of the authors as *non-influential*. We investigate three characteristic properties of EC and present an extensive analysis of how each category correlates with citation count in terms of these properties. In contrast to popular perception, we find that influential EC negatively affects future citation counts possibly owing to *attention stealing*. A detailed inspection of author collaboration network reveals that this stealing effect is more profound if an EC is nearer, 1,2-hop neighbor, to the authors of the paper being investigated.

1.7 Contributions

This thesis is primarily aimed at understanding different aspects of scholarly information. Albeit there is a multitude of work in this topic with contributions from the domains of Computer Science, Information Science, and Physics, plenty of fundamental questions still remain unanswered. Here we aim to answer some of them. One of the primary objectives of the current thesis is to develop publicly available prototype systems along with the proposed idea that not only provide a testbed for validation of ideas but also act as indicators of acceptability of new ideas. We summarize (we elaborate in the forthcoming chapters) below the primary contributions of the thesis.

1.7.1 Knowledge extraction from scholarly articles

Although significant efforts have been put toward efficient extraction and curation of scientific data, most of the works are domain specific and error-prone. We show development of *OCR++*, an open-source framework designed for a variety of information extraction tasks from PDF scholarly articles including metadata (title, author names, affiliation and e-mail), structure (section headings and body text, table and figure headings, URLs and footnotes), and bibliography (citation instances and references). Some of the interesting contributions are as follows:

- We analyze a diverse set of scientific articles written in the English language to understand generic writing patterns and formulate a huge set of hand-written rules.

- Hand-written rules are combined with a variety of Conditional Random Field (CRF) models to develop a first-of-its-kind hybrid information extraction framework.

- Extensive evaluations show that the proposed framework outperforms the existing state-of-the-art tools with a huge margin in structural information extraction along with improved performance in metadata and bibliography extraction tasks, both in terms of accuracy and processing time.

- The current version of *OCR++* is deployed at `http://www.cnergres.iitkgp.ac.in/OCR++/home/` along with the entire source code publicly available.

1.7.2 Mining performance comparisons to rank scholarly articles

We also present development of a new bibliometric system that robustly mines performance comparison information reported in comparative tables. Some of the main contributions are as follows:

- We identify table citations as a rich source of performance comparisons and present development of information extraction system to extract labeled performance tournament graphs from tables.

- We show that the proposed extraction mechanism is extremely open-domain and style-agnostic while demonstrating that performance comparisons can still be inferred from the extractions.

- We also propose several reasonable edge aggregation strategies to simplify and featurize the performance improvement graph, in preparation for ranking papers.

- We report an extensive, first-of-its-kind experiment that demonstrates severe limitations of current academic search systems in retrieving performance leaderboards.

- We adapt two widely-used tournament solvers and find that they are better than some simple ranking baselines. However, we can further improve on tournament solvers using simple variations of PageRank on a graph suitably derived from the tournament.

1.7.3 Modeling scientific growth through relay-linking phenomenon

The rate at which papers in evolving citation networks acquire links shows complex temporal dynamics. We propose a new temporal sketch of an evolving citation network and introduce several new characterizations of a network's temporal dynamics. Then we propose a new family of frugal aging models with no per-node parameters and only two global parameters. Our work reveals certain interesting results:

- We reconcile physical phenomenon of obsolescence vs. entrenchment in citation network. We propose two temporal signatures to summarize the coexistence between entrenchment and obsolescence.
- We discuss insufficiency in intrinsic obsolescence based models. Several previous growth models hypothesize that aging papers lose probability of getting cited, but none of these models *use the graph structure* to predict where these citations are likely to be redistributed.
- We show micro-scale circumstantial evidence to prove our hypothesis that at a given point in time, an old popular paper p_0 begins to lose citations in favor of a relatively young paper p_1 that cites p_0.
- Our model is based on a surprising inversion or undoing of triangle completion, where an old node relays a citation to a younger follower in its immediate vicinity. Our proposed new family of frugal aging models requires no per-node parameters. Instead we show that only two global parameters are sufficient.
- We propose three metrics (*distance*, *turnover*, and *divergence*) to measure the closeness between real temporal signatures against the simulated network. We minimize an optimization function using grid search approach to obtain an optimal set of global parameters.
- As an interesting application, we show that estimated *turnover* values negatively correlate with impact factor (IF10) for the four conference subsets we choose.

1.7.4 Estimating long-term scientific impact

The success of academic entities like research papers, authors, publication venues, organizations, etc. is estimated by their scientific impact. Although significant efforts have been put forward to understand the scientific impact,

most of the works are field-specific and highly debatable. The current thesis studies popularity dynamics of research articles.

1. We conduct an experimental study to understand the role of citation context in predicting long-term citation profiles. We show that features gathered from the citation contexts of the research papers can be very relevant for long-term scientific impact ($LTSI$) prediction. The other contributions of the work are as follows:

    - We create a massive dataset consisting of more than 26 million citation contexts for nearly 1.5 million research papers in the computer science domain, crawled from Microsoft Academic Search (MAS).
    - We extract two features from the citation contexts – average countX (number of times a paper is cited within the same article, averaged over all the citing papers) and average citeWords (number of words within the citation context, averaged over all the citing papers).
    - We then append these features along with various other features available at the time of publication in an earlier framework based on stratified learning [31] improving the prediction accuracy by 8-10% on average over the best performing baseline.

2. In a subsequent work, we aim to better understand the complex nature of the *early citers* (EC) and study their influence on long-term scientific impact. EC represents the set of authors who cite an article early after its publication (within 1–2 years).

    - We identify two important categories of EC – we call those authors that have high publication/citation count in the data as *influential* and the rest of the authors as *non-influential*.
    - We analyze three different characteristic properties of EC.
    - We empirically show that early citations might not be always beneficial; in particular early citations from influential EC negatively correlates with the $LTSI$ of a paper.
    - We build a citation prediction model incorporating the EC features; the prediction model by far outperforms the baseline predictions.

1.8 Organization of the thesis

The rest of the thesis is organized into six chapters.

Chapter 2 presents a detailed literature survey on different aspects associated with scholarly information which includes curation (crawling, processing, search, and ranking), growth, and application to real-world systems (long-term scientific impact estimation).

Chapter 3 is dedicated to our first objective of knowledge extraction from scholarly articles. In specific we propose framework for extracting metadata, structure, and bibliographical information from PDF research articles.

Chapter 4 is dedicated to our second objective of mining performance comparisons to rank scholarly articles. We propose a framework for extracting performance information embedded inside comparative tables from research articles. We further show that performance curation leads to improved visibility of the state-of-the-art articles and automatic leaderboard discovery.

Chapter 5 deals with our third objective of studying the growth of scientific volume. More specifically, we propose a relay-linking based growth model motivated by the fact that an older paper loses citations to a newer paper due to aging effect. We further propose temporal signatures to show the interplay of entrenchment and obsolesce together and to measure fitness between simulated and real data.

Chapter 6 presents our final objective of leveraging scientific information to improve the scientific impact prediction systems. In particular, we aim to perform the specific task of predicting long-term citation counts of a paper by utilizing early information available within 1–2 years after publication.

Chapter 7 concludes the thesis by summarizing the contributions and pointing to future directions that have been opened up from this thesis.

CHAPTER 2

Related Work

Effective scientific literature understanding plays a critical role toward research community's common goal of *"March for Science"*. However, unprecedented growth in research volume has led to several problems in accessibility [58, 76, 77, 103, 177], archival [44, 66, 146, 152], knowledge extraction [4, 59, 153], search and ranking [68, 92, 104, 150, 174], recommendation [21, 132, 162, 167], reproducibility [12, 17, 133], text reuse and obfuscation [40, 64, 85, 118], and summarization [122, 141, 142, 163]. In this chapter, we discuss relevant literature related to the objectives of this thesis. In specific, we look into three separate directions. The first part deals with knowledge extraction from scholarly articles in general, followed by detailed description of the existing academic extraction systems. We then move into second part where we look into the existing network modelling strategies for evolving graphs in general, and citation networks in particular. Finally we look into studies on long-term scientific impact prediction.

2.1 Knowledge extraction from scientific articles

In last two decades, the scientific community has witnessed substantial growth in availability of scientific articles with the adoption of the Open Access publishing model. As a consequence, new methodologies and automated tools to ease the extraction, semantic representation and browsing of information from papers are necessary [145]. Obtaining structured data from scientific research articles is necessary to support scholarly retrieval tasks [15]. Majority of scholarly extraction systems are 'textual' or 'text-based'. Among textual extraction

systems, we find comparatively less research in document structure analysis than metadata and bibliography extraction from scientific documents.

We discuss several state-of-the-art systems that utilize textual information present in scholarly articles. Majority of scientific documents exist in PDF format due to ease in portability and printing [89]. Recent advancements in digital libraries, particularly, preprint servers like $arXiv$[1] have massively contributed to the availability of LaTeX source code repositories. Recently, similar submission policies have been introduced by ACL[2], Elsevier[3], PLOS[4], etc. Next, we describe some state-of-the-art textual scholarly knowledge extraction systems –

2.1.1 ParsCit

ParsCit [48] is an open-source reference string parsing package. ParsCit utilizes trained Conditional Random Fields (CRF) [102] model used to label the token sequences in the reference string. They utilize a heuristic model to identify reference strings (citation contexts) from a plain text file and map them to parsed references. The package comes with utilities to run it as a web service or as a standalone utility.

2.1.2 GeneRation Of BIbliographic Data (GROBID)

The most popular scholarly extraction system is GROBID [115]. GROBID is a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured TEI-encoded documents with a particular focus on technical and scientific publications. Similar to ParsCit, it also utilizes Conditional Random Fields (CRF) as the learning algorithm. It supports header extraction of bibliographical information (e.g., title, abstract, authors, affiliations, keywords, etc.), references extraction and parsing, parsing of references in isolation, extraction of patent and non-patent references in patent publications, parsing of names, in particular author names in header, and author names in references, parsing of affiliation and address blocks, parsing of dates, and full text extraction from PDF articles, including a model for the the overall document segmentation and a model for the structuring of the text body. GROBID is currently deployed in production at ResearchGate[5], HAL Re-

[1] https://arxiv.org
[2] http://acl2014.org/CallforPapers.htm
[3] https://www.elsevier.com/authors/author-schemas/latex-instructions
[4] http://journals.plos.org/plosone/s/latex
[5] researchgate.net

search Archive, the European Patent Office[6], INIST[7], CERN[8], etc. GROBID currently supports batch processing, a comprehensive RESTful API, a JAVA API, a relatively generic evaluation framework (precision, recall, etc.), and the semi-automatic generation of training data.

2.1.3 PDFX

PDFX [43] is a rule-based system designed to reconstruct the logical structure of scholarly articles in PDF form, regardless of their formatting style. The system's output is an XML document that describes the input article's logical structure in terms of title, sections, tables, references, etc. and also links it to geometrical typesetting markers in the original PDF, such as paragraph and column breaks. The key aspect of PDFX is that the rule set used relies on relative parameters derived from font and layout specifics of each article, rather than on a template-matching paradigm.

2.1.4 SVMHeaderParse

SVMHeaderParse [75] is a Support Vector Machine (SVM) [45] classification-based method for metadata extraction from the header part of research papers. Authors show that it outperforms other machine learning methods on the same task. The method first classifies each line of the header into one or more of 15 classes. The classes are title, author, affiliation, address, note, email, date, abstract, introduction, phone, keyword, web, degree, pubnum and page. An iterative convergence procedure is then used to improve the line classification by using the predicted class labels of its neighbor lines in the previous round. Further metadata extraction is done by seeking the best chunk boundaries of each line. SVMHeaderParse is deployed by $CiteSeer^{x}$[9] for header extraction.

2.1.5 Other systems

In addition to above state-of-the-art scholarly extraction systems, several other systems exist like CERMINE [164], FLUX-CIM [46], PDFMEF [178], Xtract [16], PDFMeat[10], Mendeley Desktop[11], etc. Lipinski *et al.* [110] conducted a comparison study and observed that GROBID performed best, followed

[6] https://www.epo.org/
[7] www.inist.fr/
[8] https://home.cern/
[9] http://citeseerx.ist.psu.edu/
[10] https://github.com/mankoff/pdfmeat
[11] http://www.mendeley.com/

by Mendeley Desktop. SciPlore's Xtract, PDFMeat, and SVMHeaderParse also delivered good results depending on the metadata type to be extracted.

2.1.6 Extraction algorithms

Researchers follow several different approaches for individual extraction tasks. The approaches based on image processing segment document image into several text blocks. Further, each segmented block is classified into a predefined set of logical blocks using machine learning algorithms. Gobbledoc [124] used X-Y tree data structure that converts the two-dimensional page segmentation problem into a series of one-dimensional string-parsing problem. Dengel *et al.* [53] employed the concept language of the GTree for logical labeling. Similar work [57] presented a hybrid approach to segment an image by means of a top-down technique and then bottom-up approach to form complex layout component.

Similarly, current state-of-the-art systems use support vector machine (SVM) [75] or conditional random field (CRF) [48, 102, 116] based machine learning models for information extraction. A study by [71] compares ParsCit (a CRF based system) and Mendeley Desktop (a SVM based system). They observed that SVMs provide reasonable performance in solving the challenge of metadata extraction than CRF based approach. One reason for the more reliable SVM results may lie in the bad splitting of sequences extracted from PDFs. Since every line break marks the beginning of a new sequence, sequences belonging to the same metadata field but ranging over several lines are broken apart. This happens especially for longer titles. While the SVM context model can recover from such bad splits, CRFs cannot. CRFs do not consider labeling information from previous sequences and hence may more easily fail in finding bad splits. Rule-based approaches [16, 43] are also utilized to extract metadata information.

2.2 Mining performance comparisons to rank scholarly articles

Majority of performance comparison information is present in comparative tables and charts. We, therefore, discuss several state-of-the-art systems that extract information embedded in tables and charts. We also discuss works that leverage citation graphs for search and recommendation of scholarly information.

2.2.1 Automatic chart detection and extraction

Various studies have addressed the problem of detection and extraction of information from scientific charts. Mitra *et al.* [121] developed a machine-learning-based approach for automatic categorization of figures embedded in scientific articles. Chen *et al.* [36] described a method to automatically annotate each text in the statistical chart with semantic roles like axis labels, caption, etc. However, the method does not extract original data values from the chart. In [37], data values are extracted into XML format from black and white or grayscale charts. Savva *et al.* [148] proposed a method to automatically infer original chart data values upon the manual specification of chart regions in the textual area. The limitation of the manual specification was removed in [83]. However, they handle only single-series bar charts. Extraction mechanism proposed in [5] is fully automated using image processing and text recognition techniques combined with various heuristics derived from the graphical properties of bar charts. It improves on [83] by overcoming the need of manual efforts. Al-Zaidy *et al.* [7] utilizes machine learning methods for automatic information extraction from bar charts. Cliche *et al.* [41] proposed deep learning methods to extract data values from scatter plots.

2.2.2 Automatic table detection and extraction

Table detection, extraction, and annotation have been important research problems for years. To handle these issues, different approaches are designed for different types of documents. Table extraction from HTML documents is a well-studied problem. Several studies have attempted to address these challenges by utilizing pre-defined layout approaches [84, 154], heuristics-based approaches [96, 97, 125, 140], and statistical approaches [125, 135, 172], as well as a mixture of both heuristic and statistical approaches [173]. The pre-defined table layout based algorithms use a set of strict, pre-defined table layout information to detect tables. For a given type of image, it is usually able to have a satisfactory detection performance. However, its extension ability is very limited. The heuristics-based algorithms use a set of rules or pre-defined syntax rules of table grammar to derive decisions. The complex heuristics are usually based on the local analysis. It sometimes has an even more complicated post-processing part. As for statistical or optimization based algorithms, they either do not need parameters or need free parameters which are used in the process are obtained by an off-line training process. Many statistical systems utilize supervised machine learning algorithms like Decision trees [125, 172], Hidden Markov models (HMM) [10], Conditional Random Field (CRF) [135], SVM [172], or Neural Network [126] for table detection and extraction.

Although considerable research has been done to extract tables from HTML or image documents, extracting un-tagged tables (e.g. in PDF format) in digital libraries is difficult. Tools that are specifically dedicated to table extraction from PDFs include Tabula[12], TableSeer [111], Pdf-table-extract[13], Pdf2table [184], and PDF-TREX [128]. Tabula utilizes state-of-the-art tool Apache PDFBox to generate immediate XML character-level annotated data. XML files are further processed using a set of heuristics and computer vision algorithms to generate the final output. However, Tabula only works on text-based PDFs. TableSeer crawls digital libraries, detects tables from documents, extracts tables metadata, indexes, and ranks tables, and provides a user-friendly search interface. Pdf-table-extract utilizes heuristic-based line and cell finding algorithms. It outputs JSON, XML, and CSV lists of cell locations, shapes, and contents, and CSV and HTML versions of the tables. Pdf2table is based on a heuristic approach that performs table recognition, in which information organized in the tabular structure is recognized, and table decomposition in which recognized elements are assigned to a table model. It also provides an additional interface to manually edit parts of the extracted output. The approaches of Pdf2table, TableSeer, and PDF-TREX suffer from several extraction inaccuracies. These systems utilize *Table Lines* for detection and extraction. However, this may lead to under- or over-segmentation because it depends on predefined thresholds, spanning cells, and it would be difficult to distinguish between tables if a page contains more than one table [123].

2.2.3 Leveraging citation graphs in academic systems

Derek J De Solla Price [51] laid the foundation stone of citation graphs in bibliometry. The majority of academic search systems utilize citation graphs for search and recommendation [14, 80, 160, 161]. They are also used to characterize popularity dynamics [13, 156, 168, 171, 180], topic evolution [79, 90], and community detection [119]. Commercial scholarly search systems like Google Scholar (GS)[14], Microsoft Academic Search (MAS)[15], Semantic Scholar (SS)[16], AMiner (AM)[17] etc., incorporate relevance, age, citation trajectory, citation velocity, and impact factor for ranking papers. These systems utilize scholarly extraction tools (described in Section 2.1) for obtaining citation information.

[12] https://github.com/tabulapdf/tabula
[13] https://github.com/ashima/pdf-table-extract
[14] https://scholar.google.co.in/
[15] https://academic.research.microsoft.com/
[16] https://semanticscholar.org/
[17] https://aminer.org/

The limited early visibility of state-of-the-art papers, along with diverse tasks, data sets and baseline approaches, leads to several ill-effects in comparison studies. Kharazmi *et al.* [95] present evidence that comparison with a strong baseline is more informative than with multiple weaker baselines. They found that the Information Retrieval community continues to test against weaker baselines. A similar study [52] shows that comparison with multiple weak baselines led to "excessive optimism" in the progress made in an area of research. Thus, automatic leaderboard generation is an interesting research challenge. Recent works [78] have focused on automatic synthesis matrix (review matrix) generation from multiple scientific documents. To the best of our knowledge, no existing academic systems factor in *comparative experimental performance* reported in papers.

2.3 Modelling scholarly network growth

In his classic papers, Price [51, 139] presents evidences of obsolescence in bibliography network. Recently, Parolo *et al.* [131] presented evidence that it is indeed becoming "increasingly difficult for researchers to keep track of all the publications relevant to their work", which can lead to reinventions, redundancies, and missed opportunities to connect ideas. Based on analysis of citation data, they propose a pattern of a paper's citation counts per year, which peaks within a few years and then the typical paper fades into obscurity. Such works have seen considerable press following, with headlines[18] ranging from the tongue-in-cheek "Study shows there are too many studies" to the more alarmist "Science is 'in decay' because there are too many studies".

On the other hand, Verstak *et al.* [166] claim that fear of evanescence is misplaced, and that older papers account for an increasing fraction of citations as time passes. In a related vein, when PageRank began to be used for ranking in Web search, there was a concern that older pages have an inherent — and potentially unfair — advantage over emerging pages of high quality, because they have had more time to acquire hyperlink citations. In fact, algorithms have been proposed to compensate for this effect [38, 130]. (In that domain, clickthrough also provides valuable support for recency to combat historic popularity.)

So where does reality lie between entrenchment and obsolescence? Chakraborty *et al.* [33] present a nuanced analysis that naturally clusters papers into the ephemeral and the enduring. This gives hope that not all creativity is lost in the sands of time, but neither do older papers capture all our attention.

[18] http://www.independent.co.uk/news/science/there-are-too-many-studies-new-study-finds-10101130.html

Others [168, 171] model aging as intrinsic to a paper, reducing the probability of citing it as it ages, but do not prescribe where the diverted citations end up.

In an interesting work on explaining aging by attention stealing, Waumans *et al.* [175] present several shreds of evidence of attention-stealing from parent paper by child paper. They show that the arXiv article titled "Notes on D-Branes" [137] published in the year 1996 started losing its citations in the very next year (1997). The reason for attention stealing is attributed to four papers that cite [137] and go further on the same topic. In another example, the paper titled "Theory of Bose-Einstein condensation in trapped gases" [49] from the American Physical Society dataset[19] suffers from a similar stealing effect. This paper starts losing attention to its three child papers six years after publication. In all the three cases, the title clearly indicates the scientific content continuity in the child paper.

Albert and Barabasi's remarkable scale-free model (preferential attachment or PA) [8] "explained" power law degrees, but failed to simulate many other natural properties, such as bipartite communities. The "copying model" [100] gave a better power law fit and explained bipartite communities. Recent works [106, 168, 171] have shown that classical network growth models do not capture aging. Dorogovtsev *et al.* [55] empirically showed that power law aging function better fits real citation networks. A similar study by Hajra *et al.* [72] reconfirms the previous claim. Additionally, they show the existence of two exponents and a possibility of a crossover from one to the other. Universally, the crossover value was roughly close to ten years after publication. Recently, Wang *et al.* [171] modelled aging using an exponential decay function. They proposed that the probability of citing paper p at time t is proportional to the product $k_p(t)e^{-\lambda(t-b_p)}$, where $k_p(t)$ is the number of citations p has at time t, b_p is its birth epoch, and λ is a global decay parameter.

A more sophisticated model by Wang *et al.* [168] involves three model parameters η_p, μ_p, σ_p per paper. In effect, this model is just a reparameterization to achieve *data collapse* [20] — collapsing apparently diverse citation trajectories into one standard function of age. This thesis hypothesizes that the reason is that aging papers lose probability of getting cited, but none of the aging models *use the graph structure* to predict where these citations are likely to be redistributed. This limitation also applies to Hawkes processes [11, 60].

Another interesting approach in modeling complex networks is through generative models. Majority of these models learn from data by leveraging *Bayesian non-parametrics*. Schmidt *et al.* [151] derived an infinite mixture model as an infinite limit of a finite parametric model, inferring the model parameters by Markov chain Monte Carlo, and checking the model's fit and predictive perfor-

[19] http://journals.aps.org/datasets

mance. Similar works [127, 35] present application of Bayesian non-parametrics in collaborative filtering, link prediction, and graph and network analysis. Practical application also requires a choice of priors that are tractable— though the consequences of which remain unclear, e.g., for consistency and practical performance [86].

2.4 Long-term scientific impact

The success of a research work is estimated by its scientific impact. Quantifying scientific impact through citation counts or metrics [18, 56, 67, 81] has received much attention in the last two decades. Although these approaches are quite popular, they appear to be highly debatable [82, 101]. Additionally, they fail to take into account the future accomplishments of a researcher/article. In recent years, several researchers have investigated the problem of future citation count prediction [31, 136, 168, 182, 183, 185]. While some works propose complex mathematical models [120, 158, 168, 170, 171, 179] incorporating ageing assumptions, majority of the works focused on supervised machine learning models [32, 63, 182, 183]. Recently, Didegah *et al.* [54] presented an overview of the literature on predicting scientific impact at the later time period. Majority of the past works have proposed a set of features and used a supervised learning model to predict the citation count at a later time point. Next, we discuss several discriminating features utilized in supervised learning models –

2.4.1 Discriminating features

Many works use only the information available at the time of publication to predict future citation count, while other works also use information available from the initial years after publication.

Information available at the time of publication

Several predictive frameworks [32, 63, 182, 183] leverage information available at publication stage of a scientific article. This information can be categorized into three types (i) paper-centric, (ii) author-centric, and (iii) venue centric. Paper-centric information includes team-size, pages [114], number of articles for the first author [63], number of citations for the first author [63], number of affiliations [63], the journal impact factor [63], title, abstract, content novelty, reference count, and several diversity measures like reference diversity, keyword diversity, and topic diversity. Author-centric information includes author's h-index, productivity (measured by publication count), popularity (measured by

citation count) [29], past influence (measured by citation count of author's most popular work), sociality (measured by PageRank value in co-authorship network), and versatility (measured by author's topic distribution). Venue centric information includes venue's popularity (measured by citation count), venue centrality (measured by PageRank value in the venue-venue citation network), and past influence [182, 183].

Information generated after publication

In addition to information available at publication time, few works have leveraged information generated within 1–2 years after publication. There are few recent works [23, 170] that present an empirical analysis of the correlation between short-term and long-term citation counts. To the best of our knowledge, most of the post-publication information is limited to early popularity measured by incoming citation counts within 1–2 years after publication. Stern [159] reports that shortly after the appearance of a publication the combined use of early citations and impact factors yields a better prediction of the long-term scientific impact of the publication than the use of early citations only. Chakraborty *et al.* [32] showed that citation counts accumulated within the first year after publication as a feature can significantly improve the prediction accuracy. Brody *et al.* [25] used "number of times downloaded" data within the first 6 months after publication.

2.4.2 Several citation trajectories

A common underlying implicit assumption among research community is that the citation trajectories of all published papers have similar characteristics. However, an analysis of 463,348 papers from Physical Review (PR) corpus observed that such an assumption is flawed and suggests high heterogeneity in the citation histories [169]. In a similar attempt, Chakraborty *et al.* [31] showed the existence of six different patterns of citation profiles of research papers based on the number and position of peaks in the citation profile. They proposed two-stage prediction model, which maps a query paper into one of the six categories in the first stage, and then in the second stage, a regression module is run only on the subpopulation corresponding to that category to predict the future citation count of the query paper.

2.4.3 Predictive frameworks

Predictive frameworks are classified into three broad categories — (i) machine learning models, (ii) mathematical models, and (iii) graph-based predictive

models:

Machine learning models

Among machine learning (ML) based prediction models, majority of the works have utilized support vector regression (SVR) [31, 183], classification and regression tree (CART) [27, 183] and linear and multiple regression models [98, 114]. Within ML models, we categorize works into three types based on the temporal availability of features – (a) features available at the time of publication [27, 63, 98, 113, 182], (b) features available after publication [25], and (c) combination of (a) and (b) [31]. Callaham *et al.* [27] used features like journal impact factor, research design, the number of subjects, rated subjectivity for scientific quality, news-worthiness, etc. Further, they train decision trees to predict citation counts of 204 publications from emergency medicine specialty meeting. Livne *et al.* [113] used five group of features – authors, institutions, venue, references network, and content similarity to train an SVR model. Similarly, Kulkarni *et al.* [98] also used information present at the publication time. They trained linear regression to predict citation count for five year ahead window using 328 medical articles. Yan *et al.* [182] introduced features covering venue prestige, content novelty and diversity, and authors' influence and activity. Another work used data generated after the publication to predict citation count [25]. In this study, the downloaded data within the first six months after publication was used as a predictive feature. Chakraborty *et al.* [31] proposed a two-stage prediction model by adopting stratified learning approach, whereby, in the first stage, the model maps a query paper into one of the six citation profiles (described in Section 2.4.2), and then in the second stage a regression module is run only on the subpopulation corresponding to that category to predict the future citation count of the query paper. The prediction model consumes information present at the publication time as well as citation information generated within the first two years after publication.

Mathematical models

The use of early citations to predict $LTSI$ has been studied in various papers using mathematical models. Wang *et al.* [170] and Mingers *et al.* [120] proposed models that described how publications accumulate citations over the time. Stegehuis *et al.* [158] employed two predictor models (journal impact factor and early paper citations) to predict a probability distribution for the future citation count of a publication. They only considered accumulated citations within one year after publication. This is in contrast to the approach proposed by Wang *et al.* [168] where they allow predictions to be made fairly

soon after the appearance of a publication. They propose three fundamental citation driving mechanisms – (a) preferential attachment, (b) aging and novelty, and (c) importance of a discovery. The importance of discovery depends on so many intangible and subjective dimensions that it is impossible to objectively quantify them. Therefore, they introduced a fitness parameter (η) as a collective measure to capture the community's response to a work. The fitness parameter for each publication is sampled from a distribution ($\rho(\eta)$). Their proposed model collapses the citation histories of papers from different journals and disciplines into a single curve indicating that all papers tend to follow the same universal temporal pattern. More recent work by Xiao *et al.* [179] explored paper-specific covariates and a point process model to account for the aging effect and triggering role of recent citations. Liangyue *et al.* [108] propose a joint predictive model to forecast the long-term scientific impact problem, formulated as a regularized optimization problem. Their work addresses four key algorithmic challenges, including the scholarly feature design, the non-linearity, the domain-heterogeneity, and the dynamics. Further, they propose a fast online update algorithm to adopt joint predictive model efficiently over time. They observe that citation history is a strong indicator of long-term impact and using additional contextual or content features brings little marginal benefits in terms of prediction performance.

Graph-based predictive models

Pobiedina and Ichise [136] introduce a new feature GERscore (Graph Evolution Rule score), based on frequent graph pattern mining techniques, for citation prediction. Yu *et al.* [185] propose a new data structure namely discriminative term buckets to capture both document similarity and potential citation relation. They also propose metapath based feature space to interpret structural information in citation prediction. Along with these novel ideas, they present an extensive analysis of differences between citation prediction problem and the related work, e.g., traditional link prediction solution.

In this thesis, we utilize and construct several time-stamped scholarly datasets and extraction tools to address the above issues that remained unattended so far in the literature. The observations that we make are very unique and the conclusions that we draw, thereby, are significantly novel adding huge value to the rich digital library literature.

CHAPTER 3

Knowledge extraction from scholarly articles

This chapter is devoted to our first objective - knowledge extraction from scholarly articles. More specifically, we develop a framework *OCR++* that performs information extraction from PDF research articles.

3.1 Introduction

Obtaining structured data from documents is necessary to support retrieval tasks [15]. Various scholarly organizations and companies deploy information extraction tools in their production environments. Google scholar[1], Microsoft academic search[2], Researchgate[3], CiteULike[4], etc., provide academic search engine facilities. European publication server (EPO)[5], ResearchGate and Mendeley[6] use *GROBID* [115] for header extraction and analysis. A similar utility named *SVMHeaderParse* is deployed by $CiteSeerX$[7] for header extraction.
Through a comprehensive literature survey, we find comparatively less research in document structure analysis than metadata and bibliography extraction from scientific documents. The main challenges lie in the inherent errors in OCR processing and diverse formatting styles adopted by different publishing venues.

[1] http://scholar.google.com
[2] http://academic.research.microsoft.com
[3] https://www.researchgate.net
[4] http://www.citeulike.org/
[5] https://data.epo.org/publication-server
[6] https://www.mendeley.com
[7] http://citeseerx.ist.psu.edu/

We believe that a key strategy to tackle this problem is to analyze research articles from different publishers to identify generic patterns and rules, specific to various information extraction tasks. We introduce *OCR++*, a hybrid framework to extract textual information such as (i) metadata – title, author names, affiliation, and e-mail, (ii) structure – section headings and body text, table and figure headings, URLs and footnotes, and (iii) bibliography – citation instances and references from scholarly articles. The framework employs a variety of Conditional Random Field (CRF) models and hand-written rules specially crafted for handling various tasks. Our framework produces comparative results in metadata extraction tasks. However, it significantly outperforms state-of-the-art systems in structural information extraction tasks. On average, we record an accuracy improvement of 50% and a processing time improvement of 52%. We claim that our hybrid approach leads to higher performance than complex machine learning models based systems. We also present two novel use cases including extraction of public dataset links available in the proceedings of the NLP conferences archived in the ACL anthology.
In Section 3.2, we present the overall framework with detailed description of several extraction modules. Section 3.3 presents detailed evaluation against state-of-the-art scholarly extraction tool *GROBID*. Section 3.4 presents two interesting usecases. In Section 3.5, we present the current deployment information. Section 3.6 summarizes the current chapter.

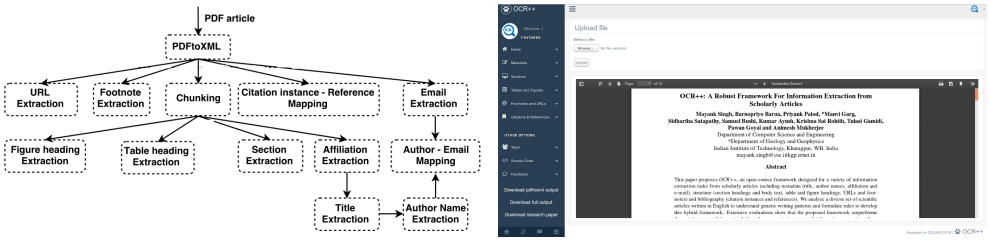
3.2 Framework overview

OCR++ is an extraction framework for scholarly articles, completely written in Python (Figure 3.2). The framework takes a PDF article as input, (1) converts the PDF file to an XML format, (2) processes the XML file to extract useful information, and (3) exports output in structured TEI-encoded[8] documents. We use open source tool *pdf2xml*[9] to convert PDF files into rich XML files. Each token in the PDF file is annotated with rich metadata, namely, x and y co-ordinates, font size, font weight, font style, etc. (Figure 3.1).
Figure 3.2(a) describes the sub-task dependencies in *OCR++*. The web interface of the tool is shown in Figure 3.2(b). We leverage the rich information present in the XML files to perform extraction tasks. We use several NLP features in the current extraction framework. Some of these include part-of-speech (POS) tags, token length, orthographic case information, etc. Although each extraction task described below is performed using machine learning models

[8]http://www.tei-c.org/index.xml

[9]URL: http://sourceforge.net/projects/pdf2xml/. We employ version 1.2.7 developed for 64 bit Linux systems.

Figure 3.1: Screenshot of *pdf2xml* tool output.



(a) Sub-tasks dependencies in *OCR++*.

(b) Screenshot of OCR++ web interface.

Figure 3.2: OCR++ framework overview and user interface.

as well as handwritten rules/heuristics, we only include the better performing scheme in our framework. Next, we describe each extraction task in detail.

3.2.1 Chunking

As a first step, we segment XML text into *chunks* by measuring distance from neighboring text and differentiating from the surrounding text properties such as font-size and bold-text.

3.2.2 Title extraction

We train a CRF model to label the token sequences using 6300 training instances. Features are constructed based on generic characteristics of formatting styles. The token level feature set includes boldness, relative position in the paper, relative position in the first chunk, relative size, the case of the first character, boldness + relative font size overall, the case of the first character in the present and the next token, and the case of first character in present and the previous token.

3.2.3 Author name extraction

In this sub-task, we use the same set of features as described in the title extraction sub-task to train the CRF model along with a heuristic that the tokens eligible for the author names are either present in the first section or within 120 tokens after the title. Different author names are distinguished using heuristics, such as the difference in y-coordinates, tab separation, etc. Further, false positives are removed using heuristics such as length of consecutive author name tokens, symbol or digit in token, and POS tag. The first word among consecutive tokens is considered as the first name, the last word as the last name, and all the remaining words are treated together as the middle name.

3.2.4 Author email extraction

An email consists of a username, a sub-domain name, and a domain name. In case of scholarly articles, usually, the usernames are written inside brackets separated by commas and the bracket is succeeded by the sub-domain and domain name. On manual analysis of a set of scholarly articles, we find four different writing patterns, author1@cse.domain.com, {author1, author2, author3}@cse.domain.com, [author4, author5, author6]@cse.domain.com and [author7@cse, author7@ee].domain.com. Here, 'cse' and 'ee' represent sub-domain instances. Based on these observations, we construct handwritten rules to extract emails.

3.2.5 Author affiliation extraction

We use handwritten rules to extract affiliations. We employ heuristics such as presence of country name, tokens like "University", "Research", "Laboratories", "Corporation", "College", "Institute", superscript character, etc.

3.2.6 Section header and body text extraction

We employ CRF model to label section headings. Differentiating features (the first token of the chunk, the second token, avg. boldness of the chunk, avg. font-size, Arabic/Roman/alpha-enumeration, etc.) are extracted from chunks to train the CRF.

3.2.7 URL extraction

We extract URLs using a single regular expression described below:

```
http[s]?://(?:[a-zA-Z]|[0-9]|[\$-_@.&+]|[!*\(\),]|(?:\%[0-9a-fA-F][0-9a-fA-F]))
```

3.2.8 Footnote extraction

Most of the footnotes have numbers or special symbols (like asterisk etc.) at the beginning in the form of a superscript. Footnotes have font-size smaller than the surrounding text and are found at the bottom of a page – the average font size of tokens in a chunk and y-coordinate were used as features for training the CRF. Moreover, footnotes are found in the lower half of the page (this heuristic helped in filtering false positives).

3.2.9 Figure and table heading extraction

Figure and table heading (caption) extraction is performed after chunking. If the chunk starts with the word "FIGURE", "Figure", "FIG.", or "Fig.", then the chunk represents a figure heading. Similarly, if the chunk starts with the word "Table" or "TABLE", then the chunk represents a table heading. However, it has been observed that table contents are also present in the chunk. Therefore, we use a feature "bold font" to extract bold tokens from such chunks.

3.2.10 Bibliography extraction

The bibliography extraction task includes extraction of citation instances and references. All the tokens succeeding the reference section are considered to be part of references and further each reference is extracted separately. Again, we employ handwritten rules to distinguish between two consecutive references. On manual analysis, we found 16 unique citation instance writing styles (Table 3.1). We code these styles into regular expressions to extract citation instances.

3.2.11 Mapping tasks

- **Connecting author name to email:** In general, each author name present in the author section associates with some email. *OCR++* tries to recover this association using simple rules, for example, the sub-string match between username and author names, abbreviated full name as username, the order of occurrence of emails, etc.

- **Citation reference mapping:** Each extracted citation instance is mapped to its respective reference. Since, there are two different styles of writing citation instances, *indexed* and *non-indexed*, we define mapping tasks for each style separately. Indexed citations are mapped directly to references with the index inside enclosed brackets. The extracted index is

Table 3.1: Generic set of regular expressions for citation instance identification. Here, AN represents author name, Y represents year, and I represent reference index within citation instance.

Citation Format	**Regular Expression**
<AN> et al. [<I>]	`([A-Z][a-zA-Z]* et al[.][\string\d\{1,3\}])`
<AN> [<I>]	`([A-Z][a-zA-Z]* [\string\d\{2\}])`
<AN> et al.<spaces> [<I>]	`([A-Z][a-zA-Z]* et al[.][ ]*[\string\d\{1\}])`
<AN> et al., <Y><I>	`([A-Z][a-zA-Z]* et al[.],\string\d\{4\}[a-z])`
<AN> et al., <Y>	`([A-Z][a-zA-Z]* et al[.][,] \string\d\{4\})`
<AN> et al., (<Y>)	`([A-Z][a-zA-Z]* et al[.][,] (\string\d\{4\}))`
<AN> et al. <Y>	`([A-Z][a-zA-Z]* et al[.] \string\d\{4\})`
<AN> et al. (<Y>)	`([A-Z][a-zA-Z]* et al[.] (\string\d\{4\}))`
<AN> and <AN> (<Y>)	`([A-Z][a-zA-Z]* and [A-Z][a-zA-Z]*(\string\d\&{4\}))`
<AN> & <AN> (<Y>)	`([A-Z][a-zA-Z]* & [A-Z][a-zA-Z]*(\string\d\&{4\}))`
<AN> and <AN>, <Y>	`([A-Z][a-zA-Z]* and [A-Z][a-zA-Z]*[,] \d\{4\})`
<AN> & <AN>, <Y>	`([A-Z][a-zA-Z]* & [A-Z][a-zA-Z]*[,] \d\{4\})`
<AN>, <Y>	`([A-Z][a-zA-Z]*[,] \string\d\{4\})`
<AN> <Y>	`([A-Z][a-zA-Z]* \string\d\{4\})`
<AN>, (<Y><I>)	`([A-Z][a-zA-Z]*(\string\d\{4\}[a-z]*))`
< multiple indices separated by commas >	`.*?[(.*?)]`

mapped to the corresponding reference. Non-indexed citations are represented using the combination of the year of publication and author's last name.

3.3 Results and discussion

Following an evaluation carried out by [110], GROBID provided the best results over seven existing systems, with several metadata recognized with over 90% precision and recall. Therefore, we compare *OCR++* against the state-of-the-art tool *GROBID*. We compare results for each of the sub-tasks for both the systems against the ground-truth dataset. The ground-truth dataset is prepared by manual annotation of title, author names, affiliations, URLs, sections, subsections, section headings, table headings, figure headings, and references for 138 articles from different publishers. The publisher names are present in Table 3.3. We divide the article set into training and test datasets in the ratio of 20:80. Note that each of the extraction modules described in the previous section also have separate training sample count, for instance, 6300 samples have been used to train the title extraction. Also, we observe that both the systems provide partial results in some cases. For example, in some cases, only half of the title is extracted or the author names are incomplete. In order to accommodate partial results from extraction tasks, we provide evaluation results at the token level, i.e., what fraction of the tokens are correctly retrieved.

Table 3.2: Micro-average F-score for GROBID and OCR++ for different extractive subtasks.

	GROBID			OCR++		
Subtask	Precision	Recall	F-Score	Precision	Recall	F-Score
Title	0.93	**0.94**	**0.93**	**0.96**	0.85	0.90
Author first name	0.81	**0.81**	**0.81**	**0.91**	0.65	0.76
Author middle name	N/A	N/A	N/A	**1.0**	**0.38**	**0.55**
Author last name	0.83	**0.82**	**0.83**	**0.91**	0.65	0.76
Email	0.80	0.20	0.33	**0.90**	**0.93**	**0.91**
Affiliation	0.74	0.60	0.66	**0.80**	**0.76**	**0.78**
Section headings	0.70	**0.87**	**0.78**	**0.80**	0.72	0.76
Figure headings	0.59	0.42	0.49	**0.96**	**0.75**	**0.84**
Table headings	0.77	0.17	0.28	**0.87**	**0.74**	**0.80**
URLs	N/A	N/A	N/A	**1.0**	**0.94**	**0.97**
Footnotes	0.80	0.42	0.55	**0.91**	**0.63**	**0.77**
Author-email	0.38	0.24	0.29	**0.93**	**0.44**	**0.60**

Table 3.2 presents comparative results for GROBID and *OCR++*. It shows that in terms of precision, *OCR++* outperforms *GROBID* in all the sub-tasks. Recall is higher for *GROBID* for some of the metadata extraction tasks. In *OCR++*, since title extraction depends on the first extracted chunk from section extraction, the errors in chunk extraction lead to a low recall in title extraction. A similar problem results in lower recall in author name extraction. Due to the presence of a variety of white space length between author first, middle and last name in various formats, we observe low recall overall in author name extraction subtasks. We also found that in many cases author emails are quite different from author names resulting in lower recall for author-email mapping subtask. *OCR++* outperforms *GROBID* in the majority of the structural information extraction subtasks in terms of both precision and recall. We observe that *GROBID* performs poorly for table heading extraction due to the intermingling of table text with heading tokens and unnumbered footnotes. A similar argument holds for the figure heading as well. URL extraction feature is not implemented in *GROBID*, while *OCR++* extracts it very accurately. Similarly, poor extraction of non-indexed footnotes resulted in a lower recall for footnote extraction subtask.

Similarly, Table 3.3 compares *GROBID* and *OCR++* for different publishing formats. Here the results seem to be quite impressive with *OCR++* outperforming *GROBID* in almost all cases. This demonstrates the effectiveness and robustness of using generic patterns and rules used in building *OCR++*. As our system is more biased towards single and double column formats, we observe lower performance on three column formats, e.g., CHI. Similarly, non-indexed

Table 3.3: Micro-average F-score for GROBID and OCR++ for different publishing styles.

	GROBID			OCR++		
Publisher	Precision	Recall	F-Score	Precision	Recall	F-Score
IEEE	0.82	0.61	0.70	**0.9**	**0.69**	**0.78**
ARXIV	0.75	0.63	0.68	**0.91**	**0.73**	**0.81**
ACM	0.69	0.49	0.58	**0.89**	**0.71**	**0.79**
ACL	0.89	0.59	0.71	**0.91**	**0.79**	**0.85**
SPRINGER	0.78	0.6	0.68	**0.85**	**0.63**	**0.72**
CHI	0.13	0.20	0.16	**0.5**	**0.36**	**0.42**
ELSEVIER	0.58	0.6	0.59	**0.82**	**0.74**	**0.78**
NIPS	0.82	0.68	0.74	**0.83**	**0.72**	**0.77**
ICML	**0.6**	**0.6**	**0.6**	0.59	0.54	0.56
ICLR	0.49	**0.55**	0.52	**0.67**	0.52	**0.59**
JMLR	0.58	0.55	0.56	**0.86**	**0.83**	**0.85**

sections format show less performance than indexed sections format.

Since citation instance annotation demands a significant amount of human labor, we randomly select eight PDF articles from eight different publishers from ground-truth dataset PDFs. Manual annotation produces 328 citation instances. We also annotate references to produce 187 references in total. Table 3.4 shows performance comparison for bibliography related tasks. As depicted from Table 3.4, *OCR++* performs better for both citation and reference extraction tasks. GROBID does not provide Citation-Reference mapping, which is an additional feature of *OCR++*.

Table 3.4: Micro-average accuracy for GROBID and OCR++ bibliography extraction tasks.

	GROBID			OCR++		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Citation	0.93	0.81	0.87	**0.94**	**0.97**	**0.95**
Reference	0.94	0.94	0.94	**0.98**	**0.99**	**0.98**
Citation-reference	N/A	N/A	N/A	**0.94**	**0.97**	**0.95**

Next, we investigate whether better formatting styles over the years lead to higher precision by the proposed tool. Also, we compare *OCR++* with *GROBID* in terms of processing time.

3.3.1 Effect of formatting style on precision

We select International Conference on Computational Linguistics (COLING) as a representative example to understand the effect of evolution in formatting styles over the years on the accuracy of the extraction task. We select ten random articles each from six different years of publications. *OCR++* is used to extract the title for each year. Figure 3.3 presents title extraction accuracy for each year, reaffirming the fact that the recent year publications produce higher extraction accuracy due to better formatting styles and advancement in converters from Word/LaTeX to PDF. We also notice that before the year 2000, ACL anthology assumes that PDFs do not have embedded text, resulting in a lower recall before 2000.

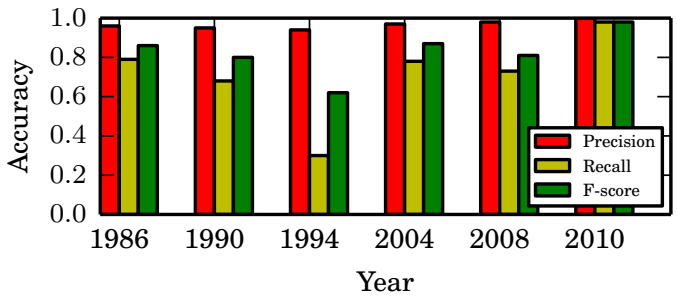


Figure 3.3: Title extraction accuracy calculated at six different years for COLING.

3.3.2 Processing time

To compare the processing times, we conducted experiments on a set of 1000 PDFs. The evaluation was performed on a single 64-bit machine, eight core, 2003.0 MHz processor and CentOS 6.5 version. Figure 3.4 demonstrates comparison between processing time of *GROBID* and *OCR++*, while processing some PDF articles in batch mode. There is significant difference in the execution time of *GROBID* and *OCR++*, with *OCR++* being much faster than *GROBID* for processing a batch of 100 articles.

3.3.3 User experience study

To conduct a user experience study, we present *OCR++* to a group of researchers (subjects). Each subject is given two URLs: 1) OCR++ server URL

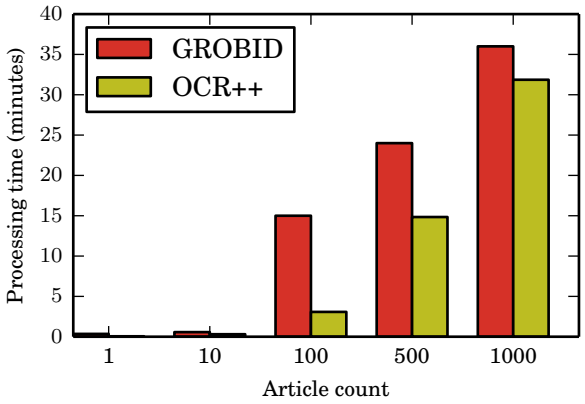


Figure 3.4: Comparison between batch processing time of *GROBID* and *OCR++*.

and 2) Google survey form[10]. A subject can upload any research article in PDF format on the server and visualize the output. In the end, the subject has to fill in a response sheet on the Google form. We ask subjects questions related to their experience such as, (a) which extraction task did you like the most?, (b) have you found the system to be really useful? (c) have you used similar kind of system before?, (d) do you find the system slow, fast or moderate?, (e) comments on the overall system experience, and (f) drawbacks of the system and suggestions for improvements.

A total of 30 subjects participated in the user experience survey. Among the most liked sub-tasks, title extraction comes first with 50% of votes. Affiliation and author name extraction tasks come second and third respectively. All the subjects found the system to be very useful. Only two of the subjects had used a similar system before. As far as the computational speed is concerned, 50% subjects found the system performance to be fast while 33% felt it be moderate. Table 3.5 presents detailed summary of survey.

Even though, OCR++ is trained on Computer Science datasets, we claim that the proposed features are sufficiently general and can potentially extract information from non-CS research articles too. However, fine-tuning experiments such as detection of several sub-sections such as background, results, conclusions, keywords, etc. within abstract section in Biomedical documents will lead to better domain-specific extraction.

[10]http://tinyurl.com/juxq2bt

Table 3.5: Detailed summary of the survey.

Total participants	30
Time-period	18-01-2016 – 21-01-2016
Most liked task	Title extraction
Second most liked task	Affiliation extraction
Third most liked task	Author name extraction
Performance speed (in %)	Fast-50, Moderate-33.33, Slow-13.33
Similar system used before	2 participants
OCR++ useful	All participants
Constructive comments	1. "Excellent work ! One of the best projects I've ever seen." 2. "The results shown are quite accurate. Might work on improving the speed of the entire process." 3. "The system is no doubt good. But the section parts should be more elaborated." 4. "I would prefer more data in a chunk, especially in the introduction and abstract so that I could get a better idea of the material." 5. "All the fields being extracted may not be necessary for a particular user. Please provide an option to keep only parts of the metadata as desired by a user."
Survey URL	`http://tinyurl.com/juxq2bt`

3.4 Use cases

3.4.1 Curation of dataset links

With the community experiencing a push toward reproducibility of results, the links to datasets in the research papers are becoming very informative sources for researchers. Nevertheless, to the best of our knowledge, we do not find any work on automatic curation of dataset links from the conference proceedings. With *OCR++*, we can automatically curate dataset related links present in the articles. In order to investigate this in further detail, we aimed to extract dataset links from the NLP venue proceedings. We ran *OCR++* on four NLP proceedings, ACL 2015, NAACL 2015, ACL 2014, and ECAL 2014, available in PDF format. We extract all the URLs present in the proceedings. We then filter those URLs which are either part of *Datasets* section's body or are present in the

footnotes of *Datasets* section, along with the URLs that consist of one of the three tokens: *datasets, data, or dumps*. Table 3.6 presents statistics over these four proceedings for the extraction task. From the dataset links thus obtained, precision was computed by the human judgement as to whether a retrieved link corresponds to a dataset. One clear trend we saw was the increase in the number of dataset links from the year 2014 to 2015. In some cases, the retrieved link corresponds to project pages, tools, researcher's homepage, etc., resulting in lowering of precision values.

Table 3.6: Proceedings dataset extraction statistics. Article count represents a total number of articles present in the proceedings. Total links and dataset links correspond to a total number of unique URLs and a total number of unique dataset links extracted by OCR++ respectively. Precision measures correct number of dataset links.

Venue	Year	Articles count	Total links	Dataset links	Precision
ACL	2015	174	345	38	0.74
NAACL	2015	186	186	18	0.50
ACL	2014	139	202	16	0.50
EACL	2014	78	141	12	0.67

3.4.2 Sectionwise citation distribution

Citation instance count plays a very important role in determining the future popularity of a research paper. An article's text is distributed among several sections. Some sections have more fraction of citations than the rest. In the second use case, we plan to study the sectionwise citation distribution. Sectionwise citation distribution refers to how citations are distributed over multiple sections in the article's text. This is an important characteristic of the citations and has recently been used for developing a faceted recommendation system [30]. We group specific sections to five generic sections, Background, Datasets, Method, Result/Evaluation, and Discussion/Conclusion. Table 3.7 shows an example mapping from specific to generic section names. Note that this mapping can be changed as per the requirement. Figure 3.5 shows citation distribution for article dataset consisting of the 138 articles mentioned earlier. A maximum number of citations are present in the method section, followed by background and discussion and conclusion. Result section comprises the least number of citations.

Table 3.7: Specific to generic section mapping.

Generic section	Specific sections
Background	Introduction, Related work, Background
Method	Methodology, Method specific names
Result/Evaluation	Results, Evaluation, Metrics
Discussion/Conclusion	Discussion, Conclusion, Acknowledgment

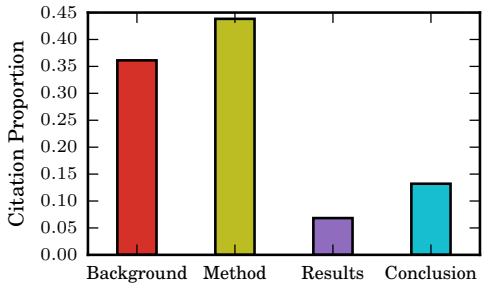


Figure 3.5: Sectionwise citation distribution in article dataset.

3.5 Deployment

The current version of *OCR++* is deployed at our research group server[11]. The present infrastructure consists of single CentOS instance[12]. We have made the entire source code available in the public domain[13].

3.6 Summary of the chapter

We develop framework for knowledge extraction from scholarly articles. Our contributions in this chapter can be summarized as below:

1. We develop an open-source information extraction framework for scientific articles using generic patterns present in various publication formats.
2. We extract metadata information, section information, and bibliography related information along with two mapping tasks, author and email mapping and citations to reference mapping.
3. Despite OCR errors and the great difference in the publishing formats, the framework outperforms the state-of-the-art systems by a high margin.

[11]CNeRG. http://www.cnergres.iitkgp.ac.in
[12]OCR++ server. http://www.cnergres.iitkgp.ac.in/OCR++/home/
[13]Github: https://github.com/mayank4490/OCR-plus-plus

CHAPTER 4

Mining performance comparisons to rank scholarly articles

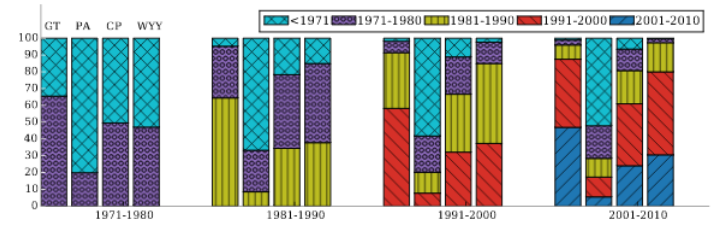
This chapter is devoted to our second objective - mining performance comparisons to rank scholarly articles. Here, we develop framework that extracts experimental performance comparisons from comparative tables.

4.1 Introduction

Comparison against best prior art is critical to publishing experimental research. With the explosion of online research paper repositories like arXiv and the frenetic level of activity in some research areas, keeping track of the best techniques and their reported performance on benchmark tasks has become increasingly challenging. *Leaderboards*, a tabular representation of the performance scores of some of the most competitive techniques to solve a scientific task, are now commonplace. However, most of these leaderboards are manually curated and therefore take a while to emerge. The resulting latency presents a barrier to entry of new researchers and ideas, trapping "wisdom" about winning techniques to small coteries, disseminated by word of mouth.

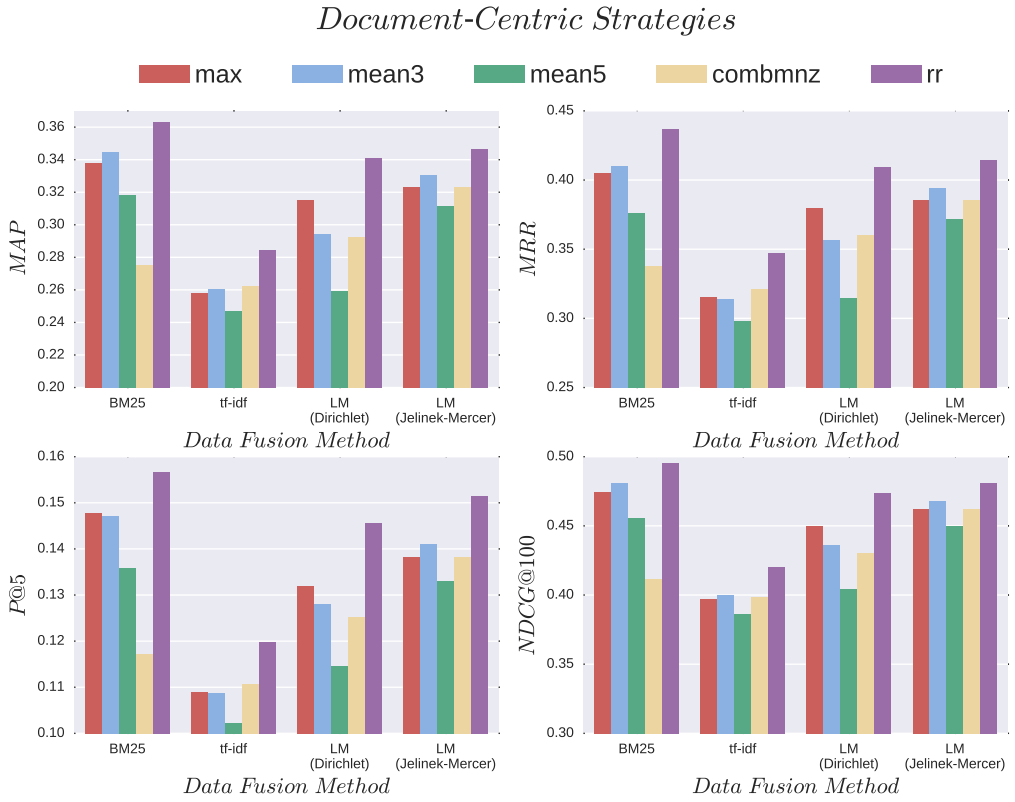
4.1.1 Limitations of conventional information extraction

The ordering of competing techniques in a leaderboard depends on a large number of factors, including the task being solved, the data set(s) used, the experimental conditions, and the choice of performance metrics. Further, there



Simulator	Distance	Turnover	Divergence
GT	–	2.70	–
PA	4.98	0.97	0.77
CP (p_{copy} = .5)	1.97	1.53	0.18
WYY (λ = .11)	1.67	2.59	0.13

(a)



(b)

Figure 4.1: Examples of challenging performance extraction cases. (a) Comparative charts and tables embedded together in a single figure [156]. (b) Multiple comparative subplots with multi-color bars representing baseline papers [39].

are often tradeoffs between various competing metrics, such as recall vs. precision, or space vs. time. In fact, an accurate extraction in conjunction to all the context details like the data sets used, the sampling protocols, the hyperparameters used, the performance metrics and their reporting units (percentage or

fraction) is almost impossible. We argue that conventional table and quantity extraction [26, 147] is neither practical, nor sufficient, for leaderboard induction. Figure 4.1 illustrates two difficult performance extraction cases [156, 39]. Figure 4.1(a) shows a combination of comparative charts and tables embedded together in a single figure. Figure 4.1(b) shows multiple comparative subplots with multi-color bars representing baseline papers.

4.1.2 Table citations

A practical way to work around the difficult extraction problem is to focus on the relatively cleaner and more structured parts of a paper, e.g., tables. Performance numbers are very commonly presented in tables. A prototypical performance table (Figure 4.2) has a first column[1] that records the name of a competing system or algorithm, along with a citation. Each subsequent column is dedicated to some performance **metric**. The rows make it simple to associate performance numbers with specific papers. In recent years, tables with citations (here, named **table citations**) and performance summaries have rapidly become popular. Experimental outcomes are also often expressed as graphs and charts. There are systems [91, 42, 6, 149] that extract tabular data from graphs and charts. Using these in our system is left for future work.

4.1.3 Performance improvement graphs

We digest a multitude of tables in different papers into a novel **performance improvement graph**. Each edge represents an instance of comparison between two papers, labeled with the ID of the paper where the comparison is reported, the metric (e.g., recall, precision, F1 score, etc.) used for the comparison, and the numeric values of the metric in the two papers. Note that every individual performance edge is extracted from a table with citations to other papers. Each such extracted edge is noisy. Apart from the challenge of extracting quantities from tables and recognizing their numeric types [26, 147], there is no control on the metric names, as they come from an open vocabulary (i.e., the column headers are arbitrary strings). Processing one table is a 'micro' reading; we must aggregate these 'micro' readings into a satisfactory 'macro' reading on an edge connecting two papers. We propose several reasonable edge aggregation strategies to simplify and featurize the performance improvement graph, in preparation for ranking papers.

[1] A transposed table style is easily identified with simple rules.

Tracker	accuracy	# failures	overlap	speed (fps)
MDNet [9]	0.5620	46	0.3575	1
EBT [41]	0.4481	49	0.3042	5
DeepSRDCF [6]	0.5350	60	0.3033	< 1 *
SiamFC-3s (ours)	0.5335	84	0.2889	**86**
SiamFC (ours)	0.5240	87	0.2743	58
SRDCF [42]	0.5260	71	0.2743	5
sPST [43]	0.5230	85	0.2668	2
LDP [12]	0.4688	78	0.2625	4 *
SC-EBT [44]	0.5171	103	0.2412	-
NSAMF [45]	0.5027	87	0.2376	5 *
StruckMK [3]	0.4442	90	0.2341	2
S3Tracker [46]	0.5031	100	0.2292	14 *
RAJSSC [12]	0.5301	105	0.2262	2 *
SumShift [46]	0.4888	97	0.2233	17 *
DAT [47]	0.4705	113	0.2195	15
SO-DLT [7]	0.5233	108	0.2190	5

Figure 4.2: Sample performance numbers in a table with citations [19]. Each row corresponds to a competing algorithm or system, which is associated with a paper cited (green highlighted link) from that row. Each column represents a performance metric.

4.1.4 Ranking papers using table citation tournaments

Ranking sports teams into total orders, on the basis of the win/loss outcomes of a limited number of matches played between them, has a long history [87, 50, 144]. We adapt two widely-used tournament solvers and find that they are better than some simple baselines. However, we can further improve on tournament solvers using simple variations of PageRank [129, 181] on a graph suitably derived from the tournament. Overall, our best ranking algorithms are able to produce high-quality leaderboards that agree very well with various manually curated leaderboards. In addition, using a popular list of papers spanning 27 different areas of Computer Science, we show that our system is able to produce reliable rankings of the state-of-the-art papers. We also demonstrate that commercial academic search systems like Google Scholar (GS)[2] and Semantic Scholar (SS)[3] cannot be used for discovering leaderboards, because of their emphasis on aggregate citations, which typically favors classic papers over latest performance leaders.

[2]https://scholar.google.co.in/
[3]https://semanticscholar.org/

The rest of the chapter is as follows. Section 4.2 presents examples of some popular and well-maintained leaderboards. Section 4.3 presents motivating discussion to show the limits of conventional table information extraction systems. Section 4.4 describes the arXiv's Computer Science dataset and details step by step extraction procedure for performance tournament graph construction. In Section 4.5, we show how to construct performance tournament graphs. In the subsequent Section 4.6, we present various schemes for ranking tournaments. We present evaluation results in Section 4.7 and conclude in Section 4.8.

4.2 Emergence of leaderboards

Leaderboards are important resources in experimental areas of science. Experts in an area are usually familiar with latest approaches and their performance. In contrast, new members of the community and practitioners need guidance to identify the best-performing techniques. This need is currently served by "organically emerging" leaderboards that organize and publish the names and the performance scores of the best algorithms in a tabular form. Leaderboards technically depend on many parameters including task, data set, other experimental settings, performance metrics, etc. Organic leaderboards are commonplace in the area of Computer Science, as in many other applied sciences. Examples of some popular and well-maintained leaderboards are noted in Table 4.4.

Table 4.1: Identification of leaderboard papers for the PASCAL VOC challenge.

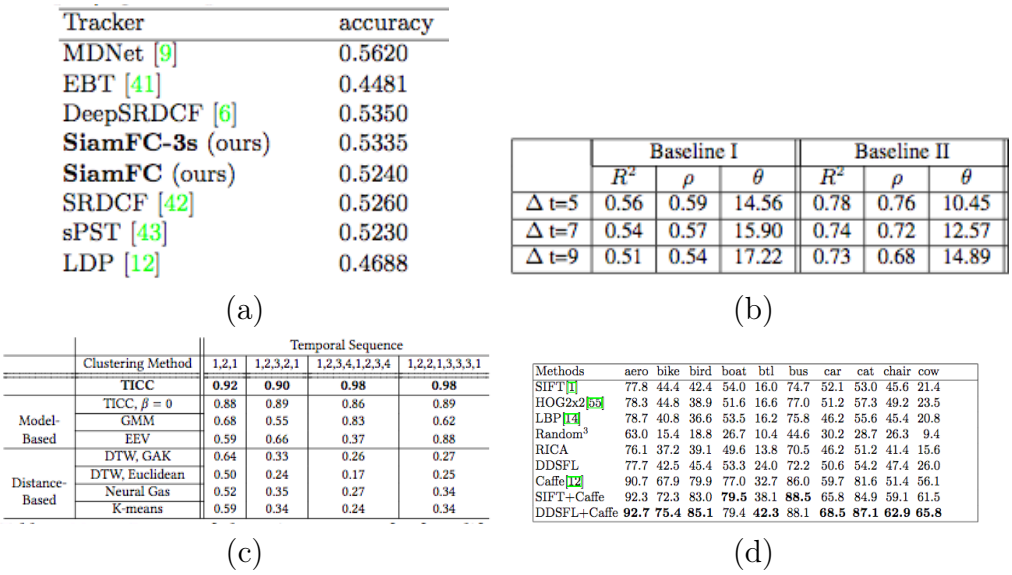
Paper	GS	SS	Our
Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation	×	×	×
Rethinking Atrous Convolution for Semantic Image Segmentation	×	×	✓
Pyramid Scene Parsing Network	×	×	×
Wider or Deeper: Revisiting the ResNet Model for Visual Recognition	×	×	×
RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation	✓	×	×
Understanding Convolution for Semantic Segmentation	×	×	✓
Not All Pixels Are Equal: Difficulty-aware Semantic Segmentation via Deep Layer Cascade	×	×	✓
Identifying Most Walkable Direction for Navigation in an Outdoor Environment	×	×	×
Fast, Exact and Multi-Scale Inference for Semantic Image Segmentation with Deep Gaussian CRFs	×	×	×
DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs	×	✓	✓
Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation	✓	×	✓
High-performance Semantic Segmentation Using Very Deep Fully Convolutional Networks	×	×	✓
Higher Order Conditional Random Fields in Deep Neural Networks	×	×	×
Efficient piecewise training of deep structured models for semantic segmentation	✓	✓	✓
Semantic Image Segmentation via Deep Parsing Network	×	✓	✓
Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform	✓	×	✓
Pushing the Boundaries of Boundary Detection using Deep Learning	×	×	✓
Attention to Scale: Scale-aware Semantic Image Segmentation	✓	✓	✓
BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation	✓	✓	✓
Learning Deconvolution Network for Semantic Segmentation	✓	✓	✓
Conditional Random Fields as Recurrent Neural Networks	×	×	×
Weakly- and Semi-Supervised Learning of a DCNN for Semantic Image Segmentation	×	×	✓
Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding	×	×	✓
Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs	×	✓	✓
Global Deconvolutional Networks for Semantic Segmentation	×	×	×
Convolutional Feature Masking for Joint Object and Stuff Segmentation	×	×	×

The prime limitation of manually curated leaderboards is the natural latency until the performance numbers in a freshly-published paper are noticed, verified, and assimilated into the leaderboards. This can induce delays in the dissemination of the best techniques to non-experts. In this work, we present development of information extraction system to extract labeled performance tournament graphs from tables. This end-to-end system helps in automating the process of leaderboard generation. The system is able to mine table citations, extract noisy performance comparisons from these table citations, aggregate the micro readings to a smooth macro reading and finally obtain rankings of papers. In Table 4.1, we show an example leaderboard generated by our system (details of the system to be discussed later in the subsequent sections) for the PASCAL VOC Challenge[4]. We observe that our system is able to find many of the papers present in this archival leaderboard. Traditional academic search systems like GS and SS do not do well in finding leaderboard entries; both GS and SS returned only seven papers (see Table 4.1) in their top 50 results retrieved for the query 'semantic segmentation'. In general, these commercial systems cannot be used for leaderboard identification since they mostly emphasize cumulative citations rather than performance scores. Citations to a paper making incremental improvements resulting in the best performance may never catch up with the seminal paper that introduced a general problem or technique.

4.3 The limits of conventional table information extraction

Performance displays are implicitly connected to a complex context developed in the paper, including the task, the data set, choice of training and test folds, hyperparameters and other experimental settings, performance metrics, etc. Millions of reviewer hours are spent each year weighing experimental evidence based on the totality of the experimental context. "Micro-reading" one table at a time is not likely to replace that. Further beyond contextual ambiguities, there are often tradeoffs between different metrics like space vs. time, recall vs. precision, etc. In summary, leaderboard induction is not merely a matter of accurately extracting performance numbers and numerically comparing them.
One way to mitigate the above challenges is to use relatively cleaner, structured parts of the papers, e.g., single tables or single charts. We focus on tables as our first-generation system. However, with advanced visual chart mining and OCR [121, 5, 91], we can conceivably extend the system to charts as well.

[4]Similar results reproducing other leaderboards can be found in https://goo.gl/YSu5CL



Tracker	accuracy
MDNet [9]	0.5620
EBT [41]	0.4481
DeepSRDCF [6]	0.5350
SiamFC-3s (ours)	0.5335
SiamFC (ours)	0.5240
SRDCF [42]	0.5260
sPST [43]	0.5230
LDP [12]	0.4688

(a)

	Baseline I			Baseline II		
	R^2	ρ	θ	R^2	ρ	θ
Δ t=5	0.56	0.59	14.56	0.78	0.76	10.45
Δ t=7	0.54	0.57	15.90	0.74	0.72	12.57
Δ t=9	0.51	0.54	17.22	0.73	0.68	14.89

(b)

		Temporal Sequence			
	Clustering Method	1,2,1	1,2,3,2,1	1,2,3,4,1,2,3,4	1,2,2,1,3,3,3,1
	TICC	**0.92**	**0.90**	**0.98**	**0.98**
Model-Based	TICC, $\beta = 0$	0.88	0.89	0.86	0.89
	GMM	0.68	0.55	0.83	0.62
	EEV	0.59	0.66	0.37	0.88
Distance-Based	DTW, GAK	0.64	0.33	0.26	0.27
	DTW, Euclidean	0.50	0.24	0.17	0.25
	Neural Gas	0.52	0.35	0.27	0.34
	K-means	0.59	0.34	0.24	0.34

(c)

Methods	aero	bike	bird	boat	btl	bus	car	cat	chair	cow
SIFT[1]	77.8	44.4	42.4	54.0	16.0	74.7	52.1	53.0	45.6	21.4
HOG2x2[55]	78.3	44.8	38.9	51.6	16.6	77.0	51.2	57.3	49.2	23.5
LBP[24]	78.7	40.8	36.6	53.5	16.2	75.8	46.2	55.6	45.4	20.8
Random[3]	63.0	15.4	18.8	26.7	10.4	44.6	30.2	28.7	26.3	9.4
RICA	76.1	37.2	39.1	49.6	13.8	70.5	46.2	51.2	41.4	15.6
DDSFL	77.7	42.5	45.4	53.3	24.0	72.2	50.6	54.2	47.4	26.0
Caffe[12]	90.7	67.9	79.9	77.0	32.7	86.0	59.7	81.6	51.4	56.1
SIFT+Caffe	92.3	72.3	83.0	**79.5**	38.1	**88.5**	65.8	84.9	59.1	61.5
DDSFL+Caffe	**92.7**	**75.4**	**85.1**	79.4	**42.3**	88.1	**68.5**	**87.1**	**62.9**	**65.8**

(d)

Figure 4.3: Feasible cases of extraction of comparative tables. (a) Explicit reference to baseline papers [19]. (b) Baseline spanning multiple columns [155]. (c) Baseline papers grouped together [74]. (d) Comparison across multiple data sets without reference to the metric [187].

Figure 4.3 shows some of the comparative tabular formats that allow for efficient extraction.
We concentrate on increasing number of tables that also cite papers, which are surrogates for techniques. Figure 4.4 shows the average number of citations in a paper p that occur in tables, against the year of publication of p. Clearly, there is a huge surge in the use of table citations in the last five years, which further motivates us to exploit them for building our system.

4.4 Data curation

In this section, we present the dataset description and describe several preprocessing and table citation extraction steps.

4.4.1 *ArXiv* dataset

We downloaded (in June 2017) the entire arXiv document source dump that includes papers from nine fields — Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science, and Economics. For the current study, we restricted

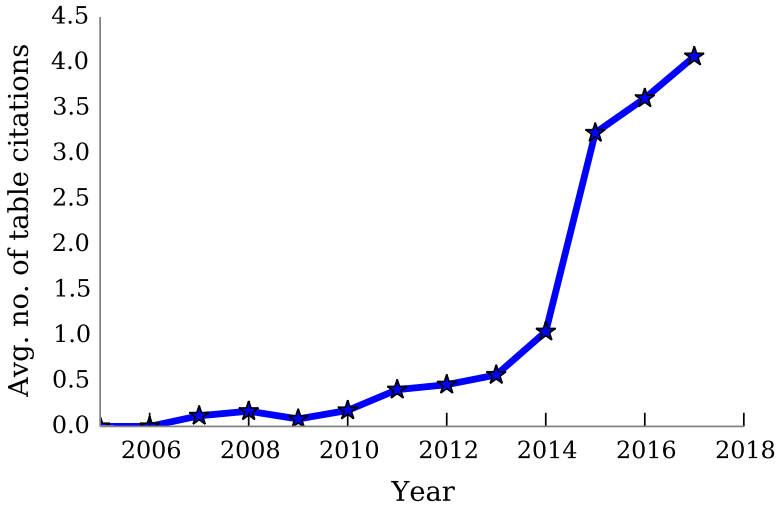


Figure 4.4: Average number of table citations made by an arXiv paper between 2005 and 2017.

ourselves to the field of Computer Science because of rapid innovations and ease of judging and interpretation of the results from our system. However, we claim that the curation process is sufficiently general and can be easily extended to non-CS research articles too. However, domain-specific fine-tuning will be required such as edge orientation (described in Section 4.4.2), identification of organic leaderboards for evaluation (described in Section 4.4.4), etc. Table 4.2 shows statistics of arXiv's Computer Science papers. ArXiv accepts four different paper formats — (i) (La)TeX, AMS(La)TeX, PDFLaTeX, (ii) PDF, (iii) PostScript, and (iv) HTML with JPEG/PNG/GIF images. However, arXiv mandates uploading the source of DVI, PS, or PDF articles generated from LaTeX code. This resulted in a large volume of papers (1,181,349 out of 1,297,992 papers) with source code.

4.4.2 Preprocessing and extracting table citations

In this section, we describe several preprocessing steps for bibliography and table extraction from arXiv's Computer Science collection. The curation process mainly involves pattern matching tasks. We also leverage popular extraction tool ParsCit [47] for title extraction from textual references. Our curation process is divided into the following sub-tasks.

Reference extraction: Since arXiv does not run $BibTeX$ in the auto-TeXing procedure, the references can only be resolved by parsing the ".bbl"

Table 4.2: General statistics about the arXiv dataset and Computer Science collection. A large fraction (91%) of papers have LaTeX code available. A significant fraction of papers contain comparative tables with citations.

Full	Year range	1991–2017
	Number of papers	1,297,992
	Number of papers with LaTeX code	1,181,349
	Total fields	9
Computer Science	Number of papers	107,795
	Year range	1993–2017
	Total references	2,841,554
	Total indexed papers	1,145,083
	Total tables	204,264
	Total table citations	98,943
	Unique extracted metrics	14,947

files. Albeit, occasionally, references are embedded in ".bbl" format within the ".tex" file itself. We, therefore, extract references from either ".bbl" or ".tex" files. Since the bibliography styles are standard and quite limited in Computer Science we were able to successfully extract all the major families like *Bibitem*, *Harvarditem*, etc. Overall, we extracted 2,841,554 references.

Reference mapping: We use a combination of regular expressions and state-of-the-art reference extraction framework ParsCit [47] to extract titles. ParsCit uses several NLP features such as number or punctuation in the reference string, token's orthographic case, location within the text, etc. Overall, we successfully extract 2,745,465 titles from 2,841,554 references (96.6%). Next, we index the arXiv paper titles and extracted reference titles which resulted in 1,145,083 unique paper titles.

Table extraction: We utilize *tabular environment* for table extraction. We successfully handle complicated writing styles like separate input table files, multirow, multicolumn, etc. We extract 204,264 candidate tables from 51,392 arXiv papers.

Collecting table citations: Thanks to normative patterns for citing tables in Computer Science papers, this step is pretty straight forward. There are three dominant ways in which a citation can take place – **explicit**, **implicit non-self**, and **implicit self**. In **explicit citation**, baseline papers are cited using *'\cite'* command within table cells. In **implicit non-self citation**, the baseline paper is implicitly referred with some keyword. We look for the first

appearance of these keywords within the main text of the paper. In **implicit self-citation**, the explicit '*\cite*' command is absent. Instead we look for a certain keyword (say 'X') in popular and standard phrases like "We propose X", "This paper proposes X", "We develop X", etc.

Performance metrics extraction: In the majority of the cases, we find that table citations and metrics are present in orthogonal locations, i.e., if table citations are present in rows then evaluation metrics are mentioned in columns and vice-versa.

Edge orientation: Metrics differ in what is defined as 'improvement': larger recall, precision, F1, and transactions per second are better, while smaller error, running time, and perplexity are preferred. This is currently hardwired into our metric meta-data manually by orienting the edges of our performance improvement graph.

In Section 4.7.1, we conduct extensive evaluation of each of the above extraction tasks.

4.4.3 State-of-the-art deep learning papers

A representative example from the rapidly growing and evolving area of deep learning is `https://github.com/sbrugman/deep-learning-papers`.
The website contains state-of-the-art (SOTA) papers on malware detection/security, code generation, NLP tasks like summarization, classification, sentiment analysis etc., as well as computer vision tasks like style transfer, image segmentation, and self-driving cars. This Github repository is very popular and has more than 2,600 stargazers and has been forked 330 times. The repository notes 27 different popular topics shown in Table 4.3. The table also shows that the SOTA papers curated by knowledgeable experts rarely find a place in the top results returned by these two popular academic search systems – GS and SS. To be fair, these systems were not tuned to find SOTA papers, but we argue that this is an important missing search feature. As fields saturate and stabilize, citations to "the last of the SOTA papers" may eclipse citations to older ones, rendering citation-biased ranking satisfactory. But we again argue that recognizing SOTA papers quickly is critical to researchers, especially new comers and practitioners.

4.4.4 Organic leaderboards

We identify manually curated leaderboards that compare competitive papers on specific tasks. The four popular leaderboards that we choose for our subsequent

Table 4.3: Recall of human-curated state-of-the-art (SOTA) deep learning papers within top-10 and top-20 responses from two popular academic search engines (Google Scholar and Semantic Scholar). Both systems show low visibility of SOTA papers.

Topics	# SOTA	GS		SS	
		Top–10	Top–20	Top-10	Top-20
Code Generation	7	0	0	0	0
Malware Detection	3	0	0	0	0
Summarization	3	0	0	0	0
Taskbots	2	0	0	0	0
Text Classification	15	0	1	0	0
Question Answering	1	0	0	0	0
Sentiment Analysis	2	0	0	0	0
Machine Translation	6	1	1	0	1
Chatbots	2	0	0	0	0
Reasoning	1	0	0	0	0
Gaming	14	0	0	0	0
Style Transfer	6	1	3	2	2
Object Tracking	1	0	0	0	0
Visual Question Answering	1	1	1	1	1
Image Segmentation	15	0	1	0	1
Text Recognition	6	0	1	0	1
Brain Computer Interfacing	3	0	0	0	0
Self Driving Cars	2	1	1	1	1
Object Recognition	30	1	1	1	1
Logo Recognition	4	0	0	0	0
Super Resolution	5	0	0	0	1
Pose Estimation	4	0	0	0	0
Image Captioning	9	1	1	1	1
Image Compression	1	0	0	0	0
Image Synthesis	9	0	0	0	0
Face Recognition	8	0	0	1	1
Audio Synthesis	6	0	1	0	0
Total	166	6 (3.6%)	12 (7.2%)	7 (4.2%)	11 (6.6%)

experiments are (i) The Stanford Question Answering Dataset ($SQuAD$)[5], (ii) Pixel-Level Semantic Labeling Task ($Cityscapes$)[6], (iii) VOC Challenge ($PASCAL$)[7], and (iv) MIT Saliency ($MIT-300$)[8]. Each leaderboard consists of several competitive papers compared against multiple metrics. For example, the $SQuAD$ leaderboard consists of 117 competitive papers com-

[5]https://rajpurkar.github.io/SQuAD-explorer/
[6]https://www.cityscapes-dataset.com/benchmarks/
[7]https://goo.gl/6xTWxB
[8]http://saliency.mit.edu/results_mit300.html

pared against two metrics 'Exact Match' and 'F1 score'. Table 4.4 describes the four leaderboards in detail. The tasks mostly include topics from natural language processing (e.g., question answering) and image processing (e.g., semantic labeling, image segmentation, and saliency prediction).

Table 4.4: Four popular leaderboards for various tasks in image processing and natural language processing. Tasks include question answering, semantic labeling, image segmentation, and saliency prediction.

Name	Papers	Metrics
The Stanford Question Answering Dataset ($SQuAD$)	117	Exact Match (EM), F1 score
Pixel-Level Semantic Labeling Task ($Cityscapes$)	76	Instance-level intersection-over-union (iIOU)
VOC Challenge ($PASCAL$)	96	Average Precision (AP)
MIT Saliency ($MIT-300$)	77	AUC, Similarity (SIM)

4.5 Performance improvement graph

In this section, we describe how to construct *Performance improvement graphs*. We also present several methods to process the raw performance improvement graph.

4.5.1 Raw performance improvement graph

The performance improvement graph $G(V, E, M)$ is a directed graph between a set of research papers V that are compared against each other. Here, M represents the set of all the evaluation metrics. Edge between two papers (A, B) (see Figure 4.5) is annotated with four-tuple ($M1$,v_1,v_2,P), where $M1 \in M$, v_1 and v_2 represent metric values for lower and higher performing papers respectively. P denotes the paper that compared A and B. The directionality of an edge e ($e \in E$) is determined by the performance comparison between two endpoints. The paper with lower performance points toward better performing paper. Figure 4.5 shows a toy example of the construction of a raw performance improvement graph from an extracted table.

One table provides just one noisy comparison signal between two papers/techniques. Although table citations allow us to make numerical comparisons, there is no guarantee of the same data set or experimental conditions

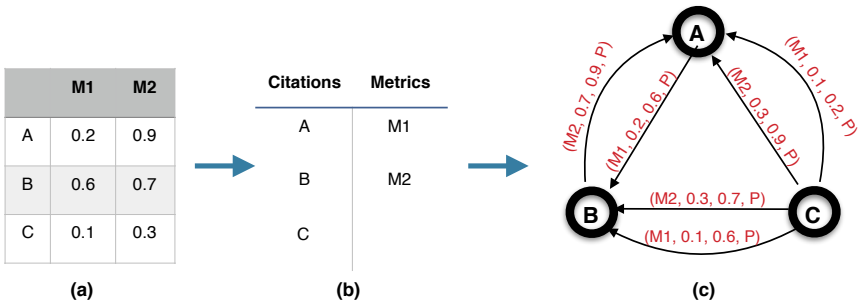


Figure 4.5: First table extraction step toward performance tournament graph construction. (a) An example table present in paper P comparing three methods, A, B and C, for two evaluation metrics, $M1$ and $M2$. (b) Unique citations to the methods as well as the evaluation metrics used are extracted. (c) an abstract performance tournament graph is constructed. Each directed edge corresponds to an improvement reported by the destination node over the source node, and is denoted by a four tuple — metric name, lower metric value, higher metric value and ID of citing paper (which might be one of the papers being compared).

across different tables, leave alone different papers. Therefore, we process the raw performance improvement graph in two steps:

Local sanitization: All directed edges connecting a pair of papers in the raw performance improvement graph are replaced with one directed edge in the sanitized performance improvement graph. This is partly a denoising step, described through the rest of this section.

Global aggregation: In Section 4.6, we present and propose various methods of analyzing the sanitized performance improvement graph to arrive at a total order for the nodes (papers) to present in a synthetic leaderboard.

4.5.2 Sanitized performance improvement graph

Relative edge improvement distribution: One unavoidable characteristic of the raw performance improvement graph is the existence of noisy edges from incomparable or botched extractions. We define the **relative edge improvement** (REI) as

$$I_m(u, v) = 100 \left(\frac{u_m - v_m}{v_m} \right) \tag{4.1}$$

where (u, v) represents a directed edge from paper u to v; u_m and v_m denote performance scores of paper u and v respectively against a metric m. As described in the previous section, v_m is lower than u_m.

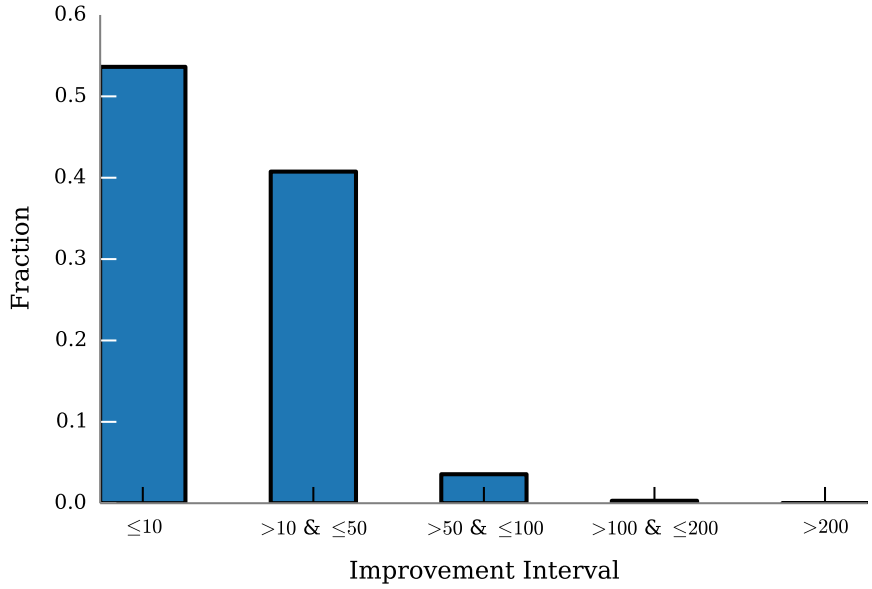


Figure 4.6: Distribution of improvement scores from four leaderboards described in Table 4.4.

Figure 4.6 shows the distribution of REIs from four leaderboards described in Table 4.4. These improvement scores are computed by considering all pairs of papers present in the respective leaderboards. We note that less than 0.5% of the edges have REI above 100%. In contrast, manual inspection of various erroneous edges revealed that their REI was much larger. Therefore, we sanitize the raw performance improvement graph by pruning edges having improvement scores larger than 100%. This simple thresholding yielded graphs as clean as by using supervised learning (details omitted) to remove noisy edges.

Sanitizing multi-edges: A pair of paper can be compared in multiple tables, resulting in (anti-) parallel edges or multi-edges. Every comparison creates a directed edge with different tuple value. Two directed edges are termed as *anti-parallel* if they are between the same pair of papers, but in opposite directions. Whereas, two directed edges are said to be *parallel* if they are between the same pair of papers and in the same direction. In Figure 4.5(c), two parallel edges exist between papers B and C and two anti-parallel edges exist between papers A and B.

Multiple strategies can be utilized to summarize and aggregate multi-edges into a clean tournament graph. We consider the following variations. Note that all of these are directed graphs. In each case, we discuss if and how a directed

edge (i, j) is assigned a summarized weight.

- **UNW — Unweighted Graph** The simplest variant preserves the directed edges without any weights. This is equivalent to giving a weight of 1 for each of these directed edge (i, j), if there is any comparison.
- **ALL — Weighted Graph (Total number of comparisons)** This variation uses the total number of comparisons between two papers p_i and p_j as the weights of the directed edge. Thus, each time an improvement is reported, it is used as an additional vote to obtain the edge weight.
- **UNQ — Weighted Graph (Unique number of metric comparisons)** This variation uses the unique number of metrics for which p_j improves upon p_i as the edge weight. The intuition for this construction is that improvements over multiple metrics should be given a higher weight.
- **SIG – Sigmoid of actual improvements on edges** This variation takes into account the sigmoid value of the actual improvement score. If paper u having a score of u_m on a specific metric m, improves upon paper v which has a score of v_m in the same table and same metric, we compute the improvement score using Equation 4.1. We then pass this score through a sigmoid function of the form:

$$sigmoid_m(u, v) = \frac{1}{1 + e^{-I_m(u,v)}} \tag{4.2}$$

  To combine the multiple improvement scores of u over v on different metrics and, thereby, obtain the edge weights, we use the following two techniques.

  Max We set the weight of the edge pointing from v to u as the maximum of all the sigmoid values of the improvement scores across the different metrics.

  Average We set the weight of the edge pointing from v to u as the average of all the sigmoid values of the improvement scores across the different metrics.

Dummy winner and loser nodes: In the tournament ranking literature that we shall discuss in the next section, the most prominent factor that guarantees convergence is that the tournament must be connected. However, performance tournament graphs are mostly disconnected due to extraction inaccuracies, incomplete article collection, etc. Therefore, we introduce a dummy node that either wins or loses over all other nodes in the graph. A dummy node has a suitably directed edge to every other node.

4.6 Mining sanitized performance improvement graphs

In this section, we explore several ranking schemes to select the most competitive papers by analyzing the sanitized performance graph. We begin with basic baselines, then explore and adapt the tournament literature, and finally present adaptations of PageRank-style algorithms. Solving an incomplete tournament over n teams means to assign each team a score or rank inducing a total order over them, and presents a natural analogy with incomplete pairwise observations. Tournament literature tries to extrapolate the anticipated outcome of a match between teams i and j (which was never played, say) in terms of the statistics of known outcomes, e.g., i defeated k and k defeated j.

4.6.1 Sink nodes

One way to be robust to numeric scores is to regard each table as comparing some papers, a pair at a time, and inserting an edge from paper p_1 to paper p_2 if the table lists a better (greater or smaller depending on metric) number against p_2 than p_1. In such a directed graph, sink nodes that have no out-links are locally maximal. Thus, the hunt for leaders may be characterized as a hunt for sink nodes. We do not expect this to work well either because our graphs contain so many biconnected components, thanks to papers being compared on multiple metrics.

4.6.2 Cocitation

An indirect indication that a paper has pushed the envelope of performance on a task is that it is later compared with many papers. We can capture this signal in a graph where nodes are papers, and an edge and its reverse edge (both unweighted) are added between papers p_1 and p_2 if they are cited by any paper. Edges in both directions are added without considering the numbers extracted from the tables.

4.6.3 Linear tournament

As described earlier, incomplete tournament presents a natural analogy to performance comparisons. Redmond *et al.* [144] started with an incomplete tournament matrix M where $m_{ij} = m_{ji}$ is the number of matches played between teams i and j. $\boldsymbol{m} = (m_i)$ where $m_i = \sum_j m_{ij}$ is the number of matches

played by team i. Abusing the division operator, let $\bar{M} = M/\boldsymbol{m}$ denote M after normalizing rows to add up to 1.
Of the m_{ij} matches between teams i and j, suppose i won r_{ij} times and j won $r_{ji} = m_{ij} - r_{ij}$ times. Then the *dominance* of i over j is $d_{ij} = r_{ij} - r_{ji}$ and the dominance of j over i is $d_{ji} = r_{ji} - r_{ij} = -d_{ij}$. Setting the dominance of a team over itself as zero in one dummy match, we can calculate the average dominance of a team i as $\bar{d}_i = \left[\sum_j d_{ij}\right] \Big/ \left[\sum_j m_{ij}\right]$, and this produces a reasonable ranking of the teams to a first approximation, i.e., up to "first generation" or direct matches. To extrapolate to "second generation" matches, we consider all (i, k) and (k, j) matches, which is given by the matrix M^2. Third generation matches are likewise counted in M^3, and so on. David *et al.* [50] showed that a meaningful scoring of teams can be obtained as the limit $\lim_{T\to\infty} \sum_{t=0}^{T} \bar{M}^t \cdot \bar{\boldsymbol{d}}$, where $\bar{\boldsymbol{d}} = (\bar{d}_i)$.

4.6.4 Exponential tournament

The exponential tournament model [87] is somewhat different, and based on a probabilistic model. Given $R = (r_{ij})$ as above, it computes row sums $\rho_i = \sum_j r_{ij}$. Let $\boldsymbol{\rho} = (\rho_i)$ be the empirically observed team scores. Again, we can sort teams by decreasing ρ_i as an initial estimate, but this is based in an incomplete and noisy tournament. Between teams i and j there are (latent/unknown) probabilities $p_{ij} + p_{ji} = 1$ such that the probability that i defeats j in a match is p_{ij}. Then the MLE estimate is $p_{ij} = r_{ij}/m_{ij}$. Jech [87] shows that there exist team 'values' $\boldsymbol{v} = (v_i)$ such that $\sum_i v_i = 0$ and

$$\rho_i = \sum_j m_{ij} p_{ij} = \sum_j \frac{m_{ij}}{1 + \exp(v_j - v_i)}. \tag{4.3}$$

Here M and $\boldsymbol{\rho}$ are observed and fixed, and $\boldsymbol{v}$ are variables. Values $\boldsymbol{v}$ can be fitted using gradient descent. Once the matrix $\boldsymbol{P} = (p_{ij})$ is thus built, it gives a consistent probability for all possible permutations of the teams. In particular, $\prod_j p_{ij}$ gives the probability that i defeats all other teams (marginalized over all orders within the other teams j). Sorting teams i by decreasing $\prod_j p_{ij}$ is thus a reasonable rating scheme.

4.6.5 PageRank

PageRank computes a ranking of the competitive papers in the (suitably aggregated) tournament graph based on the structure of the incoming links. We utilize standard PageRank implementation[9] to rank nodes in the directed weighted

[9]https://networkx.github.io

tournament graph. Consider a directed edge (u, v) with $w(u, v)$ as edge weight. Let $O(v)$ be out-neighbors of node v and $I(u)$ be in-neighbors of node u. PageRank score $PR_{t+1}(u)$ of a paper u at $(t+1)^{th}$ iteration can be computed as:

$$PR_{t+1}(u) = \alpha \sum_{v \in I(u)} PR_t(v) \frac{w(v,u)}{\sum_{i \in O(v)} w(v,i)} + \frac{1-\alpha}{n}. \tag{4.4}$$

Here α is a damping factor that is usually set to 0.85, with uniform teleport $1/n$, where n denotes total papers in the tournament graph. We run this weighted variant of PageRank on each induced tournament graph corresponding to each query. Candidate response papers are ordered using PR values. These scores can also be used for tie-breaking sink nodes.

4.7 Experimental evaluation

In this section, we evaluate experiments described in the previous section.

4.7.1 Extraction performance

Table 4.5 reports extraction accuracies of extractive sub-tasks (described in Section 4.4.2) in terms of micro-precision and micro-recall. For first three subtasks — reference extraction, reference mapping, and table extraction, we sample 20 random articles from entire ArXiv's Computer Science collection and manually evaluate them. For the next two subtasks — collecting table citations and performance metric extraction, we sample 20 random articles that consist of at least one comparative table and manually evaluate them. All subtasks performed exceptionally well, especially in terms of recall.

Table 4.5: Performance of five extractive subtasks.

	Micro-precision	Micro-recall
Reference extraction	0.95	1.0
Reference mapping	0.91	0.91
Table extraction	0.90	1.0
Collecting table citations	0.95	0.78
Performance metric extraction	0.93	0.82

4.7.2 Ranking state-of-the-art papers

Table 4.6 shows comparisons between Google Scholar (GS), Semantic Scholar (SS), and several ranking variations implemented in our testbed. Recall@10, Recall@20, NDCG@10, and NDCG@20 are used as the evaluation measures, averaged over the 27 topics shown in Table 4.3. Since our primary objective is to find competitive prior art, recall is more important in case of Web search, where precision at the top (NDCG) is paid more attention.
Given the complex nature of performance tournament ranking, our absolute recall and NDCG are modest. Among naive baselines, sink node search led to generally worst performance, which was expected. The numeric comparison is slightly better, but not much.
GS and SS are mediocre as well. Despite the obvious fit between our problem and tournament algorithms, they are surprisingly lackluster. In fact, tournament algorithms lose to simple cocitation. PageRank on unweighted improvement graphs performs beyond cocitation. However, the "sigmoid" versions of PageRank improve upon the unweighted case, almost doubling the gains beyond GS and SS, and are clearly the best choice.

4.7.3 Leaderboard generation

In this section, we demonstrate our system's capability to automatically generate task-specific leaderboards. We utilize four manually curated leaderboards for this study. Automatic leaderboard generation procedure is divided into two phases: (i) obtaining a list of candidate papers relevant to a task and (ii) ranking candidate papers by utilizing the best ranking scheme.

- **Obtaining list of candidate papers relevant to a task:** We, first, obtain a list of candidate papers relevant to a given task. We utilize textual information such as title and abstract to find relevant candidate papers. These candidate papers are further ranked by utilizing best performing PageRank schemes (described in Section 4.7.2). We consider top-50 ranked results and show comparisons between Google Scholar (GS), Semantic Scholar (SS), and top-3 high performing PageRank variations against two evaluation measures — Recall@50 and NDCG@50 — in Table 4.7. As expected, GS and SS performed poorly for all of the four leaderboards. Pagerank variations have almost double the gains beyond GS and SS and are clearly the best choice. The actual leaderboards generated can be accessed from the link provided in footnote 4.

- **Ranking candidate papers to generate leaderboard:** Next, we compute the correlation between ranks in generated leaderboards with the

Table 4.6: Comparison between several ranking schemes. Recall@10, Recall@20, NDCG@10, NDCG@20 measures are averaged over the 27 tasks (queries). Best performer is PageRank on aggregated tournament. Co-citation is surprisingly close, better than all other schemes. Tournament estimators performed worse than GS and SS. Numeric comparison and sink node search performed worse. ALL: Weighted graph (total number of comparisons); UNQ: Weighted graph (unique number of comparisons); UNW: Unweighted directed performance graph; SIG: Sigmoid of the actual performance improvement.

Ranking scheme		Top–10		Top–20	
		Recall (%)	NDCG	Recall (%)	NDCG
Systems					
Google Scholar (GS)		7.38	0.073	10.48	0.086
Semantic Scholar (SS)		7.84	0.065	10.08	0.074
Linear tournament					
Dummy Winner	ALL	5.08	0.05	13.22	0.069
Dummy Loser	ALL	3.81	0.039	13.17	0.063
Dummy Winner	UNQ	1.5	0.014	3.76	0.023
Dummy Loser	UNQ	1.5	0.014	4.0	0.024
Dummy Winner	SIG	2.21	0.024	6.71	0.036
Dummy Loser	SIG	2.21	0.027	6.71	0.039
Exponential tournament					
Dummy Winner	ALL	4.34	0.04	10.5	0.058
Dummy Loser	ALL	2.93	0.027	5.9	0.038
Dummy Winner	UNQ	4.34	0.038	9.99	0.054
Dummy Loser	UNQ	3.32	0.032	5.18	0.036
Dummy Winner	SIG	4.34	0.04	10.5	0.058
Dummy Loser	SIG	2.93	0.027	5.9	0.038
PageRank					
UNW		16.51	0.135	18.77	0.141
ALL		15.20	0.141	18.24	0.147
UNQ		15.47	0.143	18.24	0.147
SIG (Avg.)		16.78	0.151	18.77	0.155
SIG (Max.)		16.78	0.156	18.77	0.16
Sink nodes					
ALL		4.76	0.048	4.76	0.048
UNQ		4.55	0.045	4.55	0.045
Dense cocitation		10.35	0.129	16.19	0.141
Numeric comparison		7.16	0.051	11.39	0.068

ground-truth ranks obtained from the organic leaderboards. Table 4.8 presents the Spearman's rank correlation of rankings produced by PageR-

Table 4.7: Recall@50 and NDCG@50 measures for four leaderboards. Green cells indicate best scores and red cells indicate worst scores. Our PageRank variants show considerably superior performance compared to GS and SS.

Leaderboard name	GS		SS		PageRank UNW		PageRank SIG (Avg)		PageRank SIG (Max)	
	Recall (%)	NDCG	Recall (%)	NDCG	Recall (%)	NDCG	Recall (%)	NDCG	Recall (%)	NDCG
SQuAD	0	0	3.57	0.014	14.29	0.075	17.86	0.085	14.29	0.069
Cityscapes	12.5	0.067	18.75	0.159	31.25	0.198	31.25	0.237	25.0	0.205
PASCAL	13.46	0.12	13.46	0.179	30.77	0.252	26.92	0.241	30.77	0.252
$MIT-300$	21.43	0.115	7.14	0.036	21.43	0.222	21.43	0.217	21.43	0.228

ank variations, UNW, SIG (AVG), and SIG (Max), with the corresponding ground-truth rankings for the four leaderboards. $SQuAD$ shows highest correlation (0.94 for F1 and 0.88 for EM) for all of the three PageRank variations. $CityScapes$ and $PASCAL$ also exhibit impressive correlation coefficients for all the PageRank variants. For the $MIT-300$ leaderboard, while the correlation coefficient is decent for the SIM metric it is a bit low for the AUC metric. The reason for the low correlation is existence of multiple weakly connected components. A local winner in one component is affecting the global ranks across all components.

Table 4.8: Spearman's rank correlation of rankings produced by UNW, SIG (AVG), and SIG (Max) with the corresponding ground-truth rankings for the four leaderboards for various tasks in image processing and natural language processing.

Name	Nodes	Metric	UNW	SIG (AVG)	SIG (MAX)
$SQuAD$	9	F1	0.94	0.94	0.94
		EM	0.88	0.88	0.88
$CityScapes$	7	iIoU	0.69	0.7	0.7
$PASCAL$	26	AP	0.79	0.57	0.57
$MIT-300$	9	AUC	0.23	0.23	0.23
		SIM	0.54	0.45	0.45

4.7.4 Effect of graph sanitization

As described in Section 4.5.2, graph sanitization is a necessary preprocessing step. In this section, we present several real examples that resulted in greater visibility of state-of-the-art after sanitization. As representative examples, we consider two tasks "image segmentation" and "gaming" to show how graph sanitization results in noise reduction in the performance improvement graphs. We find several state-of-the-art papers that performed poorer than a compet-

itive paper with high improvement score (>700%). This anomaly resulted in the poorer visibility of the state-of-the-art papers in top ranks. However, after sanitization, the visibility gets improved. Table 4.9 shows four example high improvement edges whose removal resulted in the higher recall of the state-of-the-art papers. First two cases represent "image segmentation" papers. Next, two cases represent "gaming" papers. We observe that back-edges (anti-parallel edges) may not necessarily be present in such example scenarios.

Table 4.9: Effect of graph sanitization. The first two edges correspond to the task of "image segmentation" and the last two to the task of "gaming". Removal of these edges resulted in higher visibility of SOTA papers.

Source	Destination	Improvement %	Back-edge (Y/N)
1511.07122	1504.01013	775	Y
1511.07122	1511.00561	6597	Y
1611.02205	1207.4708	4012.3	N
1412.6564	1511.06410	928.8	N

4.7.5 Why is PageRank better than tournaments?

PageRank variations performed significantly better than tournament variations. Several assumptions of tournament literature do not hold true for scientific performance graphs; for instance, existence of disconnected components is a common characteristic of performance graphs. Unequal number of comparisons between a pair of papers in performance graphs is another characteristic that demarcates it from the tournament settings. We observe that in majority of task-specific performance graphs, tournament-based ranking scheme is biased toward papers with zero out-degrees. Therefore the tournaments mostly converge to the global sinks; in fact, we observe more than half of the tournament based top-ranked papers are sink nodes. This is the reason the recall and the NDCG reported in Table 4.6 for the above two methods are pretty close.

4.8 Summary of the chapter

We develop framework for experimental performance comparisons extraction from scholarly articles. Our contributions in this chapter can be summarized as below:

1. We introduce performance tournament graphs that encode information about performance comparisons between scientific papers.

2. Currently, these tournaments are extracted from tables with citations and performance numbers. The process of extracting tournaments is designed to be robust, flexible, and domain-independent, but this makes our labeled tournament graphs rather noisy.

3. We present a number of ways to aggregate the tournament edges and a number of ways to score and rank nodes on the basis of this incomplete and noisy information.

4. We adapt two widely-used tournament solvers and find that they are better than some simple ranking baselines. However, we can further improve on tournament solvers using simple variations of PageRank on a graph suitably derived from the tournament.

To the best of our knowledge, ours is the first framework to extract performance information from scholarly articles. It should be trivial to incorporate our system into freely accessible scholarly search systems.

CHAPTER 5

Modeling scientific growth through relay-linking phenomenon

This chapter is devoted to our third objective - modeling scientific growth through relay-linking phenomenon.

5.1 Introduction

How do actors in a evolving network pass from prominence [38, 130, 166] to obsolescence [51, 93, 106, 131, 139, 168, 171] and obscurity? Is aging intrinsic, or informed and influenced by the local network around actors? And how does the aging process affect properties of social networks, specifically, the tension between entrenchment of prominence (aka "rich gets richer" or the Matthew effect) vs. obsolescence? These are fundamental questions for any evolving social network, but particularly well-motivated in bibliometry. With rapidly growing publication repositories, understanding the networked process of obsolescence is as important to the emerging field of *academic analytics*[1] as understanding the rise to prominence.

We propose several measurements on evolving networks that constitute a *temporal bucket signature* summarizing the coexistence between entrenchment and obsolescence. **Temporal bucket signature** denotes a stacked histogram of the relative age of target papers cited in a source paper. Natural social networks (e.g., various research communities) show diverse and characteristic temporal

[1] https://en.wikipedia.org/wiki/Academic_analytics

bucket signatures. Surprisingly, many standard models of network evolution — and even obsolescence — fail to fit the temporal signatures of real bibliometric data. We present a family of **relay-linking** models that are the central contributions of this chapter, roughly speaking: to add a citation to a new paper, choose an existing paper p_0, but if it is too old, walk back along a citation link to p_1 and (optionally) repeat the process. We call this hypothesized process *triad uncompletion* and the associated generative model *relay-linking*. These proposed relay-linking models or *network influenced models* of aging mimic temporal signatures of real networks better than state-of-the-art aging models. We establish this with temporal bucket signatures and two associated novel measures: **distance** and **turnover**. Distance represents closeness (measured by L1 norm) between real and generated network's temporal bucket signatures. Whereas turnover measures decay of incoming citations as we move from one decade to the next decade. We also propose **age gap count histograms** to represent citation age distribution. Similar to temporal bucket signature, standard models fail to fit age gap count histogram of real data as well. We establish this fitness using another novel metric termed as **divergence**. Divergence represents closeness (measured by KL divergence) between real and generated network's age gap count histograms. As we shall see, simple models with $O(1)$ parameters find it very challenging to pass all these stringent tests for temporal fidelity. In sharp contrast to existing work, we avoid modeling aging as governed by network-exogenous rules or distributions (whose complexity scales with the number of nodes). Our models have only two global parameters shared over all nodes.

In Section 5.2, we describe a large-scale time-stamped bibliographic dataset. Section 5.3 presents empirical evidences of co-existence of obsolescence and entrenchment, leading to the development of the temporal bucket signatures described in Section 5.4. Section 5.5 presents description of classical evolution models and our simulation framework. In Section 5.6, we present evidences of relay and propose several relay-linking models. We compare proposed relay-linking models in Section 5.7. Section 5.8 presents an interesting application of the temporal bucket signatures. Section 5.9 summarizes the current chapter.

5.2 Dataset

Investigating the questions raised in this chapter requires rich trajectories of time-stamped network snapshots. However, such intricately detailed datasets are rare, even while there is an increasing number of new repositories being built

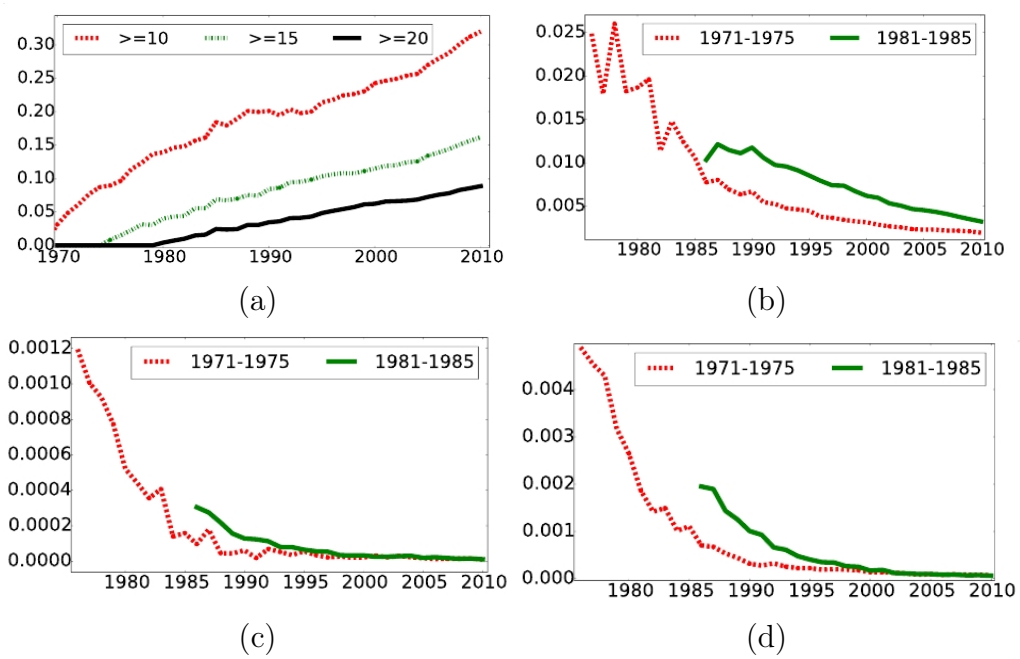


Figure 5.1: (a) For a paper written in $y \in [1970, 2010]$ (x-axis), we plot the fraction of papers it cites (y-axis) that are older than $y - t$ years, for $t = 10, 15, 20$ (red, green, black). (b) We picked a fixed set P of 100 most cited papers written in 1971–1975 (red) and 1981-1985 (green). For papers written in years $y \in [1975, 2010]$ (x-axis), we plot the fraction (y-axis) of citations made to papers in P. Unlike (a), this shows a steep decrease. (c) Replacing popular papers P with a random set R of papers written in 1971–1975 (red) and 1981–1985 (green) reduces the *absolute* y-axis but not the *relative* decay. (d) Enlarging R to 500 random papers also has no effect on the relative rate of decay.

and updated regularly[2]. Fortunately, Microsoft Academic Search[3] (MAS) provides an ideal platform for our study. MAS data includes paper titles, reconciled paper IDs, year of publication, publication venue, references, citation contexts, related field(s), abstract and keywords, author(s) and their affiliations [31]. We have filtered papers from full dataset (Table 5.1). The filtered dataset consists of papers published between 1961–2010 and have at least one outlink or one inlink (to filter isolated nodes or missing data). We call this filtered dataset as the Ground Truth dataset (GT). For each simulation initialization, we create a warmup dataset from GT having papers published between 1961–1970. De-

[2] `http://snap.stanford.edu/` is a prominent example.
[3] `http://academic.research.microsoft.com`

tailed description and the role of warmup data in the simulation framework can be found in Section 5.5.2.

Table 5.1: General statistics about the full Computer Science dataset from Microsoft Academic Search. Filtered and warmup dataset are subsets of full dataset.

	Full	Filtered	Warmup
Year range	1859–2012	1961–2010	1961–1970
Number of papers	2,281,307	1,702,471	9,568
Number of citations	27,527,432	15,791,272	7,312

To ensure that our proposed temporal signatures are generally applicable, we also experimented with papers from the biomedical domain. In this study, we use biomedical dataset that consists of 801,252 research articles[4] published between 1996-2014. All our evaluations are based on extensive experiments with the Computer Science domain dataset[5].

5.3 Entrenchment and obsolescence

Preferential attachment models without aging [88, 100] predict that older papers get more entrenched and their rate of citation acquisition can only go up. Verstak *et al.* [166] provide the support that *as a cohort* older papers are thriving: more recently written papers have a larger fraction of outbound citations targeting papers that are older by a fixed number of years. However, there is plenty of evidence [33, 168, 171] that aging counteracts entrenchment. This apparent contradiction is readily resolved by realizing that the number of papers older by a fixed number of years is growing rapidly. But the real value of the study (Sections 5.3.1 and 5.3.2) is that it leads us to the definition of new signatures of evolving networks (Section 5.4).

5.3.1 Fraction of citations to 'old' papers

Suppose that papers in our corpus, published in a year y, make C_y citations in all to older papers. Of these, say C_t citations go to papers that were published before year $y - t$, for $t = 10, 15, 20$. Figure 5.1(a) plots the quantity C_t/C_y against y, similar to the setup of Verstak et al. [166]. The plot is consistent

[4] `http://www.ncbi.nlm.nih.gov/pmc/tools/ftp`

[5] We have a comparable evaluation on biomedical papers which we omit due to space constraints.

with their claim: the fraction of citations to older papers is indeed increasing over the years y for all values of t.
However, Figure 5.1(b) paints a different picture. For each year range 1971–1975 and 1981–1985, we choose 100 most cited (through 2010) papers P. Then, for other papers written in a year $y \in [1975, 2010]$, we plotted the fraction of citations out of those papers that go to P. Clearly, this fraction decreases over time. In place of popular papers, how do *random* papers fare? Figures 5.1(c,d) show that the relative shape of decay remains stable when random paper sets of sizes 100 and 500 are picked as the targets.

5.3.2 Fraction of citations to papers in 10-year age buckets

Figures 5.1 suggests a natural and compact way to summarize citation statistics organized by age. We group papers into buckets. Each bucket includes papers published in one decade[6]. Then, for each bucket, we plot as a stacked bar-chart, the fraction of citations going to that same bucket as well as all previous buckets. Figure 5.2(a) shows the result. We note the following:

- The fraction of citations from a bucket to itself (shown as the bottom purple, yellow, red and blue bars in successive columns) decreases over time and those to all older buckets increase over time. This is consistent with Verstak *et al.*

- However, if we consider papers in a bucket as targets, the citations they receive decreases over the years. For instance, papers written in 1971–1980 (purple bars over successive columns) received 70.5% of the citations in that decade (purple) but this number reduces to $29.2, 6.4, 2.8\%$ in successive decades. Similar decay is seen for the following buckets (yellow, red) as well.

We see similar effects in Figure 5.2(b), except that papers written in 1996–2000 became obsolete much more rapidly (yellow bar) compared to papers written in 2001–2010, so there is less stationarity of the obsolescence process in the biomedical domain compared to computer science. Thus, such bar charts *simultaneously* validate Verstak *et al.* [166] and also show aging of paper cohorts, and are a succinct signature of the balance between entrenchment and obsolescence.

[6]Any suitable bucket duration can be used. We experiment with several bucket sizes, a majority of them produced similar results.

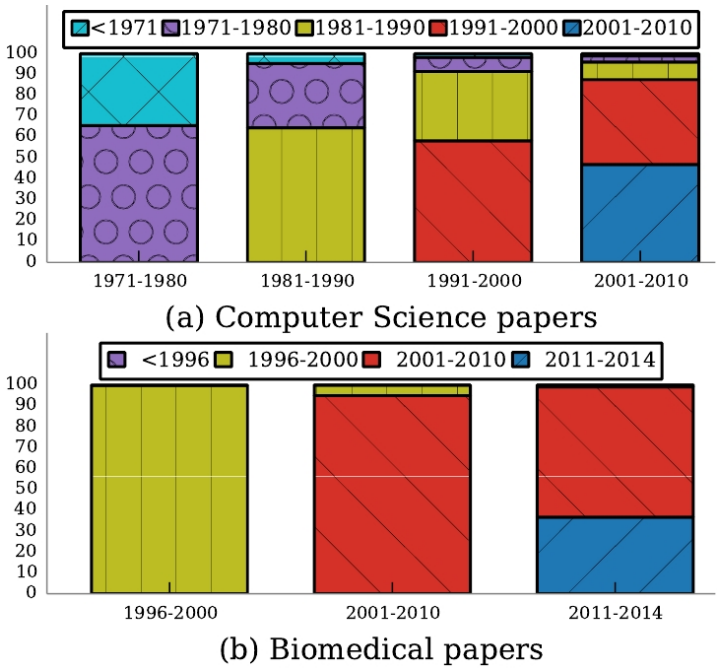
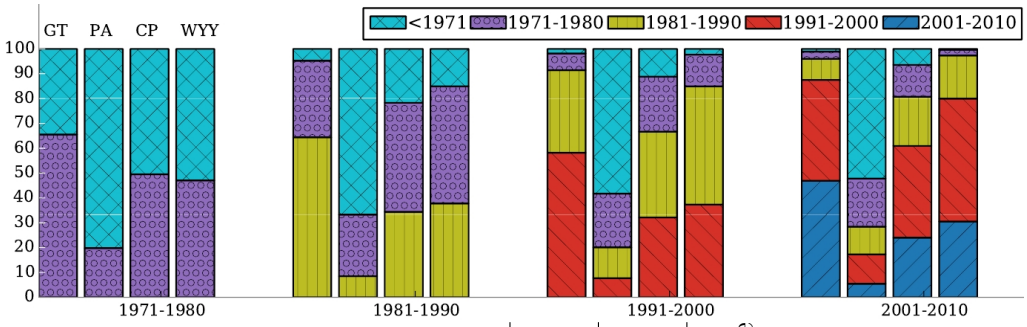


Figure 5.2: (a) Citation distribution across 10-year buckets for computer science dataset. Each vertical bar represents a decade of papers. Within each bar, colored/textured segments represent the fraction of citations going to preceding decades. The bottommost segment is to the same decade, the second from bottom to the previous decade, etc. On one hand, the volume of citations to the current decade (bottommost segment) is shrinking to accommodate " old classics" (entrenchment). On the other hand, any given color/texture shrinks dramatically over decades (most papers fade away). (b) Citation distribution of the biomedical dataset. Papers written in 1996–2000 became obsolete much more rapidly.

5.4 New signatures of evolving networks

We start with some basic notation. Time t proceeds in discrete steps (for publications, often measured in years). Sometimes we will bucket time into ranges like decades. We study an evolving graph G_t, which comprises the node set V_t and edge set E_t. Nodes are denoted by u, v, etc. Edges (i.e., citations) once added, are never removed. Also, in our bibliometric setting, edges emanating from a node v all "appear" when node v itself appears, at birth time t_v, but this assumption can be relaxed. We shall use GT as the shorthand for ground-truth data (see Section 5.2).

We introduce several natural ways to observe dynamic networks to better un-



Simulator	Distance	Turnover	Divergence
GT	–	2.70	–
PA	4.98	0.97	0.77
CP ($p_{\text{copy}} = .5$)	1.97	1.53	0.18
WYY ($\lambda = .11$)	1.67	2.59	0.13

Figure 5.3: Temporal bucket signatures comparing ground truth (GT), preferential attachment (PA) [8, 88], copying (CP) [100], and WYY [171]. Each bucket represents a decade. Ground truth turnover is 2.70. For others, distance, turnover and divergence values are shown in the accompanying table. Clearly, only WYY has even a remote similarity to ground truth.

derstand the interplay between entrenchment and obsolescence.

5.4.1 Age gap count histogram

When new paper u, born at time t_u, cites an older paper v, born at t_v, that citation link spans an *age gap* of $t_u - t_v \geq 0$. (Depending on the granularity of measuring time, $t_u = t_v$ may or may not be possible.) In case of dynamic documents where u can add citations (dropping citations is rare), we can take t_u to be the citation creation time, rather than the birth time of u. In citation data, gap g is usually expressed in whole years. For any value of g,

$$\sum_{(u,v)\in E} \begin{cases} 1, & \text{if } t_u - t_v = g, \text{ and} \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

is the number of links that span an age gap of g. As we shall see later, age gap count histograms reveal some salient dynamics of graph evolution.

Divergence

Suppose we observe age gap histograms H from real data. Each simulated model gives age gap histograms $\tilde{H}$. We assess divergence between two histograms ($\tilde{H}$ and H) by measuring Kullback-Leibler divergence. More precisely,

$$\text{divergence}(H||\tilde{H}) = \sum_{g \in H} H(g) \log \frac{H(g)}{\tilde{H}(g)} \tag{5.2}$$

A simulated model is closer to real data, if divergence $\to 0$.

5.4.2 Temporal bucket signature

Suppose we collect birth times into buckets of temporal width T (e.g., T may be 10 years). Suppose our corpus of papers P is thus partitioned into $P_1, P_2, \ldots, P_N$, based on their publication date. We pad this with sentinel bucket P_0 for all papers before P_1. Each source paper $p_s \in P_j$ may cite target papers $p_t \in P_i$, where $i \leq j$. Let the total number of citations from papers in P_j to papers in P_i be $C(i,j)$ (row=cited, column=citing). Let column sums $C(j) = \sum_i C(i,j)$ be the total number of outbound citations from papers in P_j. Let $F(i,j) = C(i,j)/C(j)$ be the fraction of outbound links from papers in P_j that target papers in P_i. The temporal bucket signature is defined as the matrix $F(i,j) : i \leq j$, i.e.,

$$F = \begin{bmatrix} F(0,1) & F(0,2) & \cdots & F(0,N) \\ F(1,1) & F(1,2) & \cdots & F(1,N) \\ 0 & F(2,2) & \cdots & F(2,N) \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & 0 & F(N,N) \end{bmatrix}, \tag{5.3}$$

where each column adds up to 1. We propose two intuitive scalar summaries of temporal bucket signatures.

Distance

Suppose we observe F from real data. We also fit a model which, upon simulation, gives bucket signature $\tilde{F}$. We propose to assess how closely $\tilde{F}$ approximates F by measuring the average row-wise L1 distance between their corresponding columns. More precisely,

$$\text{distance}(F, \tilde{F}) = \sum_{j=1}^{N} \left[\sum_{i=0}^{j} |F(i,j) - \tilde{F}(i,j)| \right]. \tag{5.4}$$

The higher the distance value, lower will be the closeness of approximation, and vice versa. Note that there is no assumption of stationarity in this definition. Communities can be in volatile and transient stages of obsolescence while replacement rates in other communities can be stable.

Turnover

Another quantity of interest summarizing F or $\tilde{F}$ is a notion of *decay* of the height of a segment of a given color from one column to the next, in the sequence $F(i,i), F(i,i+1), F(i,i+2), \ldots$ Specifically, the ratio $F(i,j)/F(i,j+1)$ (which is usually more than 1) represents how sharply citations to papers in P_i decreases from year j to year $j+1$. Because we are interested in a ratio, we aggregate these via a geometric mean:

$$\text{turnover}(F) = \left[\prod_{j=1}^{N-1}\prod_{i=0}^{j}\frac{F(i,j)}{F(i,j+1)}\right]^{\frac{2}{(N+2)(N-1)}} \tag{5.5}$$

A high value of *turnover* indicates more rapid obsolescence. Turnover can be measured on both F and $\tilde{F}$. In the later sections, we will relate the quantities we have defined with other established properties of real networks.

5.4.3 Optimization

We assume that the temporal bucket signature for GT is F and the age gap histogram is H. Similarly, for each simulated model, we denote $\tilde{F}$ and $\tilde{H}$ as temporal bucket signature and age gap histogram respectively. Note that, $\tilde{F}$ and $\tilde{H}$ are dependent on two model parameters λ and Θ (see Figure 5.4). We use $d(\cdot)$, $t(\cdot)$ and $f(\cdot)$ as shorthand for distance$(\cdot)$, turnover$(\cdot)$ and divergence$(\cdot)$ respectively. To obtain optimal set of parameters for each model, we need to solve the following optimization problem:

$$\underset{\lambda,\theta}{\text{minimize}} \quad d(F,\tilde{F}) * \Big(|t(\tilde{F}) - t(F)|\Big) * f(H||\tilde{H}) \tag{5.6}$$

Here, $|t(\tilde{F}) - t(F)|$ represents absolute difference between GT's turnover (e.g., 2.70 for one of our data sets), and relay-link model's turnover. Other combinations such as weighted sums can be considered, but product has the advantage that we do not need to manually balance typical magnitudes of the parts. To our knowledge, the above problem does not admit a tractable continuous optimization procedure. Therefore, we perform grid search and choose values for model parameters for each proposed model.

5.5 Classical evolution models and simulation results

The first generation of idealized network growth models [8, 134] generally focused on a "rich gets richer" (preferential attachment or PA) phenomenon without any notion of aging. This was followed by the vertex copying model [100]. There has been more recent work [55, 72, 168, 171, 186] on modeling age within the PA framework. We will review and evaluate some of these in Section 5.5.2.

5.5.1 Classical models

Standard preferential attachment (PA)

In Albert *et al.*'s classical PA model [8, 88], at time t, a new paper would cite an old paper p, which currently has degree $k_p(t)$, with probability $\Pi(p,t)$ that is proportional to $k_p(t)$:

$$\Pi(p,t) \propto k_p(t) \tag{5.7}$$

In their idealized model, one new paper was added at every time step, but this is easily extended to mimic and match the growing observed rate of arrival of new papers. Moreover, the number of outbound citations from each new paper can also be sampled to match real data.
If paper p arrives at time t_p, it is not hard to obtain a mean-field approximation to the degree of p at time $t \geq t_p$:

$$\tilde{k}_p(t) \propto \sqrt{t/t_p}. \tag{5.8}$$

This expression suggests that age is a monotone asset, never a liability, for any paper.

Copying model (CP)

The copying model [100] is characterized by a network that grows from a small initial graph and, at each time step, adds a new node (paper) p_n with k edges (citations) emanating from it. Let p_r be a "reference" paper chosen uniformly at random from pre-existing papers. With a fixed probability (the only parameter of the model), each citation from p_n is assigned to the destination of a citation made by p_r, i.e., p_n "copies" p_r's citations. Neither PA nor copying has a notion of aging.

Ageing model (WYY)

Wang, Yu and Yu [171] proposed modeling age within the PA framework. The probability of citing at time t a paper p that was born at time b_p, while proportional to its current degree as in PA, *decreases exponentially* with its age:

$$\Pi(p,t) \propto k_p(t)\, \exp\Big(-\lambda(t-b_p)\Big), \tag{5.9}$$

where $\lambda > 0$ is the single global parameter controlling the attention decay rate, estimated from some "warmup" data. Similar models are motivated by the measurements by Leskovec *et al.* [106]. Note, in order to avoid the huge computational overhead associated with updating probability values for each new entry, we approximate by only updating the attachment probability value once in each year. For the first 20 years, the approximate version is (a) extremely close to the original version (less than .05 L1 distance) and (b) slightly closer to the GT than the original version thus giving this baseline a small additional advantage.

5.5.2 Simulation protocol and results

We simulate the models described above for 40 years (1971–2010) and compare the results with GT ($turnover$ = **2.70**). *Warmup data* is the subset of GT generated between 1961–1970 (detailed statistics is present in Section 5.2). Warmup data consists of papers published between 1961–1970 along with the citation links formed between them. We initiate each simulation model from warmup data. The warmup data can be called as the "train data". Starting from the year 1971, for each subsequent year, we introduce as many papers in the system as the publication count of that year estimated from GT. Each incoming paper is accompanied by nine outlinks (average number references estimated from GT). This data, generated through our simulation models between 1971–2010, can be called as "test data". We simulate CP with copying probability = 0.5 (after grid search on all possible probability values) since the product of the three observables, i.e., distance, turnover, and divergence (a function similar to Equation 5.6) is the least at this value of the probability. Similarly, for WYY, we obtain through grid search $\lambda = 0.11$ that results in the lowest product of the three observables.

Results are shown in Figure 5.3. PA fits observed temporal bucket profiles very poorly. The distance score is very large (**4.98**). Neither PA nor copying has a notion of aging. Therefore, it is not surprising that CP also does not fit observed temporal bucket signatures well. The distance score is **1.97**. WYY performed best at $\lambda = 0.11$ with distance = **1.67**. As for turnover, WYY's turnover (**2.59**) is closest to that of GT (**2.70**).

5.5.3 Other related models

Forest Fire

Relay-linking has some superficial similarity to the forest fire model [107] and earlier work on random walk and recursive search based attachment processes [165]. But among many critical difference is the involvement of time and node ages. In forest fire terminology, the relative birth times of candidate source and target nodes strongly influence whether we prefer to 'burn' forward or backward edges. To our knowledge, there is no similar temporally modulated version of forest fire model that has demonstrated fidelity to bucket signatures, or age gap count histograms.

Point processes

It is attractive to think of citations as events "arriving at a node/paper" according to some temporal point process[7]. Focusing on one node, if $\mathcal{H}(t)$ is the history of the event arrivals up to time t, then the *conditional intensity function* is defined as

$$\gamma(t)dt := \Pr(\text{event in } [t, t+dt)|\mathcal{H}(t)).$$

Specifically, if $\mathcal{H}_v(t)$ comprises the points of time $t_{vi} < t$ of past arrivals at node v, then the Hawkes process [1] defines

$$\gamma_v(t) = a_v + b_v \sum_{t_{vi}<t} \exp(-|t - t_{vi}|).$$

and provides two major benefits:

1. the exponential decay term elegantly captures temporal burstiness, and
2. given $\{t_{vi}\}$, parameters a_v, b_v can be estimated efficiently [11, 60].

While Hawkes process is most suited for repeated similar events (such as messages or tweets between two people), citation happens only once between two papers. Work on coupling edge message events to network evolution itself is rare, with notable exceptions [60]. In our case, citation arrivals at different papers are not independent events but coupled to global population growth rates as well as network constraints (e.g., out-degree distribution). Given those constraints, Hawkes process provides no obvious benefits to inference or simulation. Moreover, citations are often observed in (annual) batches, but Hawkes process

[7] `https://en.wikipedia.org/wiki/Point_process`

finds simultaneous arrivals impossible. We can model arrival times as hidden and observe them in batches, but that involves a more complex EM procedure [112] to marginalize over arrivals. Even if these hurdles can be overcome, we have to estimate or sample a_v, b_v for every node, just like WSB [168], which results in too many parameters. Moreover, there is still no direct connection between declining citations and whether the network guides the diverted citations to specific targets, which is the specific goal of relay-linking models.

5.6 Proposed relay-linking models

5.6.1 Evidence of citation stealing

The central hypothesis behind the relay linking model is as follows:

At a given point in time, an old popular paper p_0 begins to lose citations in favor of a relatively young paper p_1 that cites p_0.

There are a variety of intuitive reasons why relay-linking or relay-citing can happen:

- p_1 is a journal version of a conference paper p_0,
- p_1 refutes or improves upon p_0, or
- p_1 reuses data or a procedure in p_0, and so on.

Table 5.2: Circumstantial evidence of relay-link: R_W papers acquire more citations than R_L papers. Here, r is in R_W or R_L . Higher proportion of papers belonging to R_L have zero citation count than R_W. Bold face text represents that the mean of cumulative citation count of R_W at base year T is larger than the mean of R_L. Also, R_W papers show higher increasing trend than R_L papers.

	Popularity of cited papers	#Papers	#Papers with > 0 citations	%Papers with > 0 citations	Avg #citations to r	#Recent papers with increasing trend	%Recent papers with increasing trend	Cited neighbors with decreasing trend (%)	Avg. decrease in median values
R_W	≥ 70	76082	60205	**79.13**	19.77	31749	**41.72**	48.06	5.69
R_L	≤ 10	16257	2017	**12.40**	0.31	736	**4.52**	41.39	0.41

Unlike standard preferential attachment (PA), evidence for relay-linking can only be circumstantial and in the aggregate, because the decision of p_2 to select,

Table 5.3: Circumstantial evidence of relay-link: Papers that cite fading papers gather citations at an accelerated pace. Bold face text represents that the rate at which the citations are gained by the set of R' papers is higher compared to the set of $R_W \setminus R'$ papers.

R_W			R'			$R_W \setminus R'$		
#papers in P_P	#papers in F	Avg. drop	Avg. citation count at T	Avg. citation gain in $[T, T+\delta T]$	Per-year citation gain	Avg. citation count at T	Avg. citation gain $[T, T+\delta T]$	Year-wise citation gain
21621	4962	36.41	23.48	13.92	**2.48**	11.02	11.89	**2.05**

but then *not* cite p_0, is never recorded in any form; we get to know only of the recorded citation to p_1. Here we produce such circumstantial evidence, in two parts.

Fix a base time T (2005 in our experiments). Define **popular** papers P_P as those that have at least 70 cumulative citations as of T. Define **obscure** papers P_O as those that have at most ten cumulative citations as of T. Let **recent winner** papers R_W be those that make at least ten citations[8] and at least 50% are to papers in P_P. Let **recent loser** papers R_L be those that make at least ten citations, and all are to papers in P_O.

Do R_W papers gain citations faster than R_L? We now measure the cumulative citations to each paper in R_W and R_L as of time $T + \delta T$ (say after five years) and can apply a standard test of the hypothesis that the mean of R_W is larger than the mean of R_L (see Table 5.2).

Are R_W papers stealing citations from P_P papers? Now we focus on a subset of P_P: those whose rate of acquiring citations see a sharp ($> 50\%$) drop from $[T - \delta T, T]$ to $[T, T + \delta T]$. Let this be **fading** papers $F \subset P_P$. Consider papers $R' \subset R_W$ that cite papers in F, and their rate of acquiring citations in $[T, T + \delta T]$. We investigate if this population has a significantly larger mean than a base population. Here the base population is set to the papers $R_W \setminus R'$. In Table 5.3 we observe that indeed the rate at which the citations are gained by the set of R' papers is higher compared to the set of $R_W \setminus R'$ papers.

5.6.2 Model descriptions and results

Inspired by the above experiments, we propose in Fig. 5.4, a generic template for all our relay-link models. t_u is the birth time of u. The flexible policies/

[8]To eliminate noise in extracting citations.

parameters are R, λ, Θ, D. R is either 1 (one-shot relay) or ∞ (iterated relay). D is either uniform, or as in PA, but restricted to $I(u,t)$. The λ parameter governs the time to initiate relaying while the Θ parameter governs the extent of relaying. A higher value of λ leads to relaying of citations from source paper soon after its publication and vice-versa. Similarly, Θ controls the intensity of relaying; higher values of Θ lead to higher intensity of relaying. Note that, standard PA can be achieved by keeping $\lambda = 0$. We will explore alternatives for a few design choices and that will lead us to a few variations on the basic theme.

Random relay-cite (RRC)

Our first model is obtained by setting $R = 1$ and D as the uniform distribution over $I(u,t)$. In words, we first pick a p_0 to cite, then we toss a coin with head probability = $exp(-\lambda T)$, where T is the current age of the paper p_0. If the coin turns up tail, then again, we toss a coin with head probability Θ. With coin turning up as head, we sample a paper v that links to p_0 uniformly at random and then cite v instead of p_0. Effectively, p_0 *relays* the citation to v. This version of the model thus has two parameters λ and Θ.
We simulated the model with different values of (λ, Θ). Grid search led us to the best value of $(0.19, 0.9)$ as per the optimization function defined in Equation 5.6. Figure 5.5 shows the *temporal bucket signatures* for this and the other variants described below; the best distance, turnover and divergence that RRC achieves are respectively **1.08**, **2.70**, and **0.03**.

Preferential relay-cite (PRC)

In the preferential relay-cite model, R continues to be 1, but we depart from the random relay-cite model in that D is no more a uniform distribution over the papers in $I(u,t)$. The probability of sampling v is proportional to its in-degree, as in PA. Again, we simulated this model and performed a grid search to obtain the best parameter values $(\lambda, \Theta) = (0.3, 0.9)$ as per the optimization function in Equation 5.6. We obtained the best distance score of **1.86**. The corresponding turnover and divergence scores were found to be **2.11** and **0.16**.

Iterated random relay-cite (IRRC)

In iterated random relay-cite model, we relax R to be able to follow the relay-cite hypothesis iteratively. Thus, once a paper v has sampled a paper from $I(u,t)$ based on uniform distribution, we again toss a coin with head probability = $\exp(-\lambda T')$, where T' is the current age of the paper v. In case, tail turns up,

```
1: for advancing time steps t do
2:     for each paper p_n newly added at time t do
3:         for each citation (p_n, ?) to fill do
4:             choose old paper u using PA
5:             for r = 1, 2, ... do
6:                 T = t − t_u
7:                 toss coin with head prob. exp(−λT)
8:                 if head or r > R: break
9:                 toss coin with head prob. Θ
10:                if tail: break
11:                let I(u, t) be papers that cite u
12:                    as of time t
13:                choose v ∈ I(u, t) according to
14:                    a sampling distribution D
15:                u ← v
16:            add (p_n, u) as new citation
```

Figure 5.4: Relay-linking template.

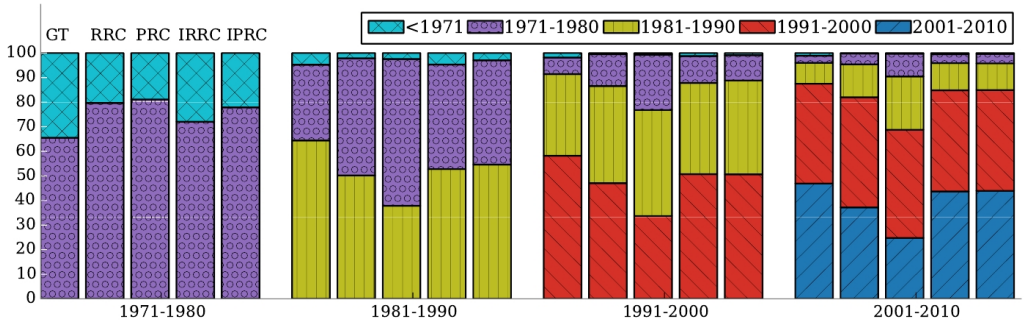
we follow this process recursively. $(\lambda, \Theta) = (0.115, 0.8)$ gives the best distance score of **0.60**, turnover of **2.67** and divergence score of **0.012**.

Iterated preferential relay-cite (IPRC)

In iterated preferential relay-cite model, once a paper v has sampled a paper from $I(u,t)$ based on PA, we again toss a coin with head probability = $\exp(-\lambda T')$, where T' is the current age of the paper v. In case, tail turns up, we follow this process recursively. We simulated the model with different parameter values, and found that $\lambda = 0.19$ and $\Theta = 0.8$ gives the best distance score of **0.72**, turnover score of **2.70** and divergence score of **0.004**.

5.6.3 Dependence on bucket size

Since divergence is computed from age gap count histograms, it does not depend on the bucket size. For distance and turnover, we observed that our observations are stable for bucket sizes 7, 8 and 9 years. For bucket sizes larger than 10 years, the number of buckets is too small to make a fair comparison.



Simulator (λ, Θ)	Distance	Turnover	Divergence
RRC (0.19, 0.9)	1.08	**2.70**	0.03
PRC (0.3, 0.9)	1.86	2.11	0.16
IRRC (0.115, 0.8)	0.60	2.67	0.012
IPRC (0.19, 0.8)	**0.72**	**2.70**	**0.004**

Figure 5.5: Temporal bucket signatures from ground truth data (GT), random relay-cite (RRC), preferential relay-cite (PRC), iterated RRC (IRRC) and iterated PRC (IPRC). λ and Θ were optimized separately for each variant using grid search. Ground truth turnover is 2.70. For others, distance, turnover and divergence values are shown in the accompanying table. Note the qualitatively better fit with ground truth compared to Figure 5.3.

5.7 Comparison between models

5.7.1 Temporal bucket signatures

Fig. 5.5 compares ground truth (GT) temporal bucket signatures against the variations of relay-linking models described above. Three out of four relay-linking models proposed above outperform the popular baseline models of network evolution in terms of all the observables, i.e., distance, turnover and divergence (see Figure 5.3 for detailed result obtained for the baseline models.) Further, note that IPRC outperforms all the other relay-linking models in at least two out of the three observables and can be considered to be the closest fit to GT. Therefore, in order to strengthen our results, we compare age gap count histograms and degree distribution of IPRC (instead of other relay-linking models) with the baseline models.

5.7.2 Age gap count histograms

Fig. 5.6 shows the age gap count histograms defined in Equation 5.1 for various simulators, compared with ground truth (over all time). Ground truth rolls down steadily after an early peak at 2–3 years age gap. As expected, the PA curve keeps going up, because aging is always an advantage. Surprisingly, but indirectly corroborating degree distribution (as well as its temporal signature in Figure 5.3), WYY does well in comparison, but its most likely gap is larger compared to real data. IPRC fits GT's decay best.

The model complexity of relay-linking is comparable to PA. Yet, we establish that relay-linking is the closest to real networks in terms of divergence, distance, and turnover.

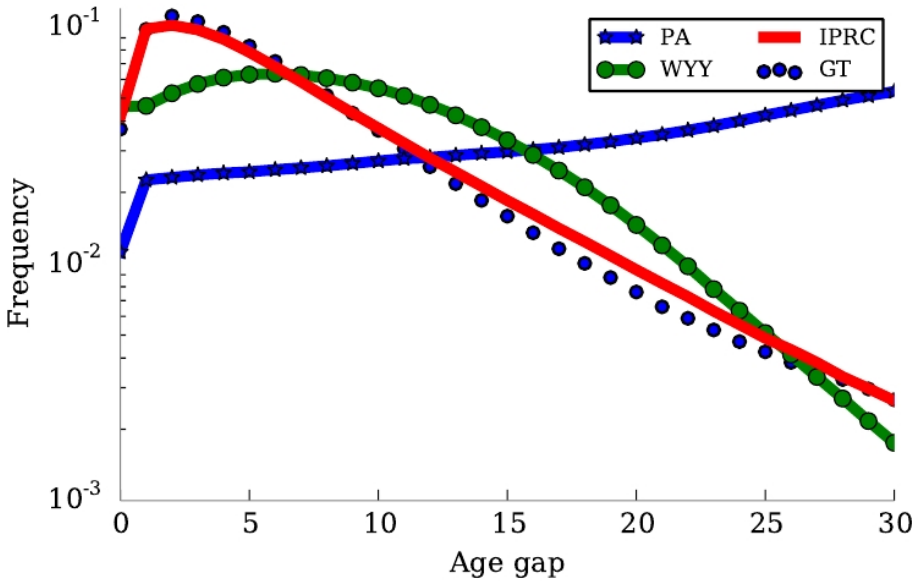


Figure 5.6: Age gap count histograms. WYY is quite close to ground-truth, but for its best choice of λ, its peak is still at too large a gap. IPRC's decay fits GT best. The divergence values are, PA: 0.77; WYY($\lambda = 0.11$): 0.13; IPRC ($\lambda = 0.19, \Theta = 0.8$): 0.004.

5.7.3 Degree distribution

In Figure 5.7 we plot the degree distribution of the network obtained by simulating IPRC. The figure shows that the distribution fits the GT quite well.

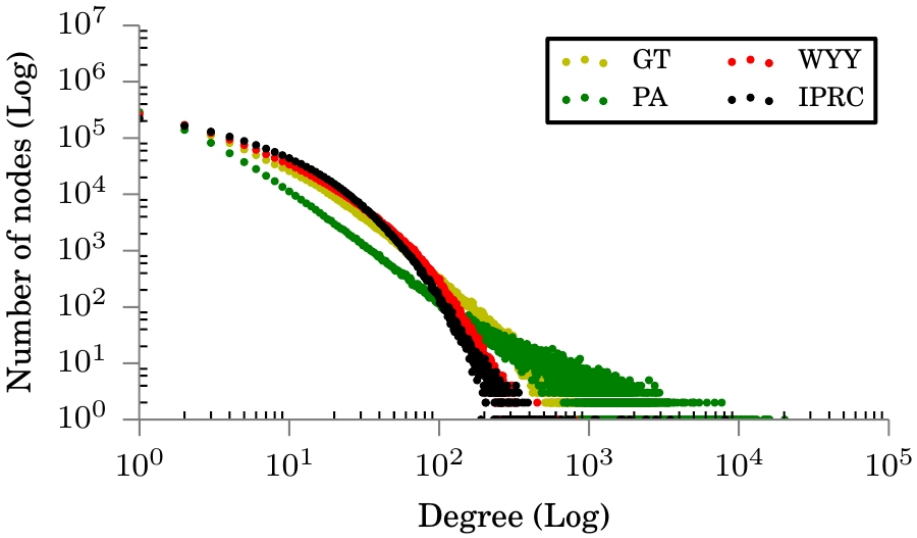


Figure 5.7: Degree distributions of ground truth (GT) and various models (PA,WYY,IPRC) at the best optimal parameters values.

5.8 Practical application

To get more insight into temporal bucket signatures, we apply these to a cross-sectional study by sub-field and conference slices. The widely quoted *impact factor* [69] (IF10) of a journal or conference is the average number of citations to recent (last 10 years) articles published there. Table 5.4 shows the turnover values we estimate against IF10 for the four conference subsets we chose. There is a clear negative correlation i.e., communities with large turnover have low IF10. Large turnover also seems associated with applied communities in a state of more intense flux.

5.9 Summary of the chapter

Our contributions in this chapter can be summarized as below:

1. We give the first plausible network-driven models for obsolescence in the context of research paper citations, based on a natural notion of relay-linking.

2. Studying large bibliographic datasets, we also propose several novel and stringent tests for temporal fidelity of evolving, aging network models.

Table 5.4: Correlation between turnover and average value of 10-year impact factor, over specific conferences as well as coherent sub-communities of computer science. Note the negative correlation between turnover and 10-year impact factor. Communities with large turnover have low IF10.

Conference Name	**Turnover**	**Avg. IF10**
SIGMOD	3.97	3.50
VLDB,ICDE	4.52	2.79
SIGIR	5.61	2.77
ICML,NIPS	6.74	1.84
Data Mining, machine learning, artificial intelligence, natural language processing and information retrieval	3.32	0.63
Distributed and parallel computing, hardware and architecture, real time and embedded systems	3.31	0.74
Algorithms and Theory, Programming Languages and Software Engineering	2.29	0.78

3. Traditional aging models do not pass these tests well, but our relay-linking models do.

4. As an interesting application, we show that estimated turnover values negatively correlate with impact factor (IF10) for the four conference subsets we chose.

To the best of our knowledge, this is the first work that attempts to leverage network-assisted link-relays to explain obsolescence in citation networks. Future extensions could possibly lead to a formal analysis of properties of relay-linking or tractable variations.

CHAPTER 6

Estimating long-term scientific impact

This chapter is devoted to our fourth objective - applications of curated scholarly information. We present long-term scientific impact prediction as an interesting application through two studies–

1. The role of citation context in predicting **long-term scientific impact**.
2. Impact of early citers on **long-term scientific impact**.

6.1 Introduction

Citation count of a publication is among the most commonly accepted metric by the research community for evaluating the impact and quality of a research article. Citation count refers to the number of citations received by an article within a specified time-period [22]. Highly-cited works remain as one of the most important criteria for the various organization (e.g. companies, universities and governments) to identify the best talents, especially at their initial stages. An early estimate would help in identification of promising articles that could accelerate research and dissemination of new knowledge. This has motivated the interest in the field of future citation prediction [136, 182].
Prediction of future citation counts, however, is difficult because of the nature and dynamics of citations [61, 70]. The citation ranges for the papers published by the same authors or the same venues show a lot of variation. The same can be said about the field of the papers as well. A recent study [34] has shown that

all the scientific papers do not follow the same trajectory and found 6 different citation patterns.
The existing works have used various venue and author-centric features, along with the citation information from the initial years for the task of citation prediction. Toward this objective, we conduct two important studies, as described below –

6.1.1 The role of citation context in predicting long-term scientific impact

In this work, we argue that the features extracted from the citation contexts can be extremely helpful for the future prediction. Citation context refers to textual descriptions of a given scientific paper found in other papers in the document collection which cites it [9]. A citation context is, in principle, a set of sentences where a paper is referred to. The intuition behind using the citation context features comes from the hypothesis that citation contexts reflect the opinion of the scientific community about the particular work. We show that even using some very simplistic features extracted from the citation context can boost the performance of a citation prediction system significantly.
We use a massive dataset consisting of more than 26 million citation contexts for nearly 1.5 million research papers in the Computer Science domain, crawled from Microsoft Academic Search (MAS)[1]. We extract two features from the citation contexts – average countX (number of times a paper is cited within the same article, averaged over all the citing papers) and average citeWords (number of words within the citation context, averaged over all the citing papers). A high value of countX implies that cited paper is referred multiple times by citing paper and thus, cited paper might be quite relevant for citing paper. Similarly, a high value of citeWords implies that cited paper has been discussed in more details by citing paper and therefore, cited paper might be relevant for citing paper. We show that these features are quite discriminative and exhibit different trends not only for different citation ranges but also for the citation categories identified in [34]. We then append these features along with various other features in an earlier framework based on *stratified learning* [31]. Experimental results show that addition of these two features gives a R^2-correlation of 0.84, 0.81, and 0.78 toward predicting the citation count at 5, 7, and 9 years after publication, improving the prediction accuracy by 8-10% on average over the nearest baseline. Specifically, these features help in predicting the long-term citation behavior of the research papers. We would like to stress here that this study brings forth the tremendous potential of the content of a scientific

[1] http://academic.research.microsoft.com/

article in predicting future citation counts; the huge success of only two very simple content related features proposed here indicates that deeper analysis of the content can lead to further significant improvements in the related areas of research.
Section 6.2.1 describes the citation context dataset used for this experimental study. The two citation context features utilized for our study have been described in Section 6.2.2. The citation prediction model has been described in Section 6.2.3. The experiments to evaluate our system under different settings have been reported in Section 6.2.4 along with a detailed comparison and feature analysis. Finally, conclusions have been presented in Section 6.5.

6.1.2 Impact of early citers on long-term scientific impact

Next, we aim to better understand the complex nature of the *early citers* (EC) and study their influence on long-term scientific impact. EC represents the set of authors who cite an article early after its publication (within 1–2 years). We investigate three characteristic properties of EC and present an extensive analysis to answer three interesting research questions:

- Do early citers influence the future citation count of the paper?
- How do early citations from influential authors impact the future citation count compared to the non-influential ones?
- How do citations from co-authors impact the future citation count compared to the others (influential as well as non-influential)?

We present a large-scale empirical study to answer these questions. We empirically show that early citations might not be always beneficial; in particular, early citations from influential EC negatively correlates with the long-term scientific impact of a paper. Motivated by the empirical observations, we incorporate the EC features in a popular citation prediction framework proposed by Yan et al. [182]. We discuss the prediction outcomes and show that our extended framework outperforms the original framework by a high margin.
Section 6.3.1 presents detailed definitions of early citers. Section 6.3.2 describes two datasets used for this experimental study. In Section 6.3.3, we present a large-scale empirical study to answer these questions. Motivated by the empirical observations, in Section 6.3.4, we incorporate the EC features in a popular citation prediction framework proposed by Yan et al. [182]. In Section 6.3.5, we

discuss the prediction outcomes and show that our extended framework outperforms the original framework by a high margin. Finally, conclusions have been presented in Section 6.5.

6.2 The role of citation context in predicting long-term scientific impact

6.2.1 Datasets

In this study, we use two Computer Science datasets, both crawled from Microsoft Academic Search (MAS)[2]. First dataset (bibliographic dataset) consists of bibliographic information of papers, the title of the paper, a unique index for the paper, its author(s), the affiliation of the author(s), the year of publication, the publication venue, references, citation contexts, the related field(s)[3] of the paper, the abstract and the keywords of the papers [31]. Second dataset (citation context dataset) consists of more than 26 million citation contexts pre-processed and annotated with cited and citing paper information. Even though, we leverage CS datasets for this study, the results are generally applicable. However, availability of such intricately detailed time-stamped and rich bibliographic datasets in non-CS domains is rare. Table 6.1 details various statistics for both the datasets.

Table 6.1: General information about the datasets.

Dataset I	Year range	1960-2010
	Number of Computer Science fields	21
	Number of publications	1,359,338
	Number of authors	138,923
	Avg. number of papers per author	5.43
	Avg. number of authors per paper	2.40
Dataset II	Number of citation contexts	26,197,440
	Avg. number of citation contexts per paper	19.27
	Avg. number of words per citation context	26
	Number of papers having at least one citation context	1,279,104

The data crawled by Chakraborty *et al.* [31] had several inconsistencies that were removed through a series of steps. First, few forward citations were removed which point to the papers published after the publication of the source

[2]http://academic.research.microsoft.com

[3]Note that the different sub-branches like Algorithms, AI, Operating Systems etc. constitute different "fields" of Computer Science domain.

paper. These forward citations appear because there are certain papers that are initially uploaded in public repositories (such as http://arxiv.org/) but accepted later in a publication venue. Further, they considered only those papers published in between 1970 and 2010 because this time period seemed to be most consistent since majority of the articles published at that time period are available in the dataset. Only those papers are considered that cite or are cited by at least one paper (i.e., isolated nodes with zero in-degree and zero out-degree have been removed). An advantage of using this dataset is that the problem arising due to the ambiguity of named-entities (authors and publication venues) has been completely resolved by MAS itself, and a unique identity has been associated with each author, paper and publication venue. Some of the authors were found missing in the information of the corresponding papers which were resolved by the DOI (Digital Object Identifier) of the publications. We double checked the filtered papers having the author and metadata information from DOI and kept only the consistent ones. Some of the references that pointed to such papers absent in the dataset (i.e., dangling references) were also removed.

Definition I: We will call a paper P *cites* paper C, if paper P refers to paper C in the text. P is termed as **citing paper** while C is termed as **cited paper**. P can refer to C at many places in the text. In our present work, we only consider the sentence as citation context where the reference to the paper is explicitly present.
For an example, Chakraborty *et al.* [31] cites Yan *et al.* [182] as
*Recently, Yan et al. conduct two similar experiments [**25**, 26], to study features covering venue prestige, content novelty and diversity and authors' influence and activity. They also account for the temporal dynamics by taking a recent version of each feature calculated on a limited time window. To the best of our knowledge, this is the latest and the most accurate future citation count prediction model and therefore serves as the baseline system in this paper. We conduct an extended examination of all these factors related to citation counts, with many new features added.*

Although the above context consists of four sentences, we only consider the first sentence as the citation context since it explicitly refers to Yan *et al.* [182].

Definition II: countX for a cited paper C with respect to a citing paper P is defined as the number of citation contexts, when a paper P cites paper C. Citation context count for a paper C denotes the sum of countX from all the citations for C.
Each paper has a specific citation context count. Figure 6.1 shows the distribution of papers having specific citation context count in our dataset. Long

tail depicts that many papers have less number of citation context count while a small number of papers have high citation context count.

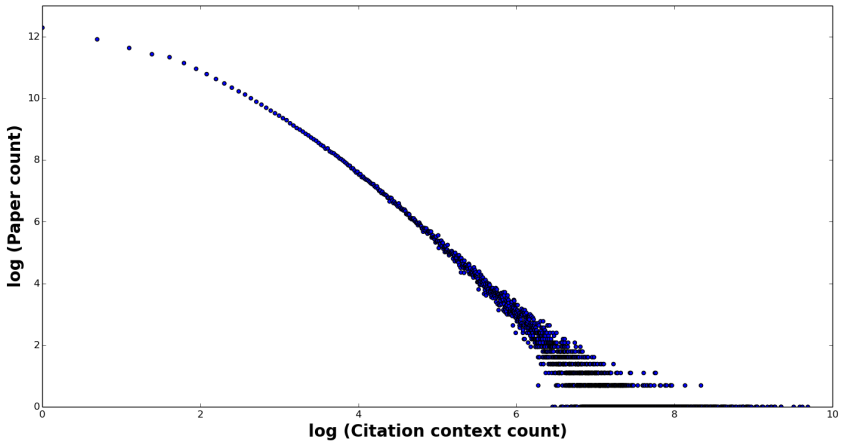


Figure 6.1: Distribution of citation context count in our dataset.

Definition III: citeWords for a cited paper C with respect to a citing paper P is defined as the number of words in the citation context, when a paper P cites paper C. If multiple papers are cited within the same citation context, the number of words is equally divided among all the cited papers. If a paper is cited multiple times within the same paper, citeWords is computed by summing over the words in all the citation contexts.
In the next section, we discuss in detail how the average values of countX and citeWords behave for papers with various citation ranges.

6.2.2 Average countX and citeWords

After identifying countX and citeWords as two features from the citation contexts, we study in detail as to whether these features are discriminative with respect to the number of citations. To normalize with respect to various ranges of citations, we only used the average values of countX and citeWords for a publication in each year starting from its publication year.
As a working example of how these features are computed, Table 6.2 presents citation contexts for paper *P* titled as "*On Relaxed Dynamic Programming in Switching Systems*", published in 2005. The first column gives the citer IDs, which refer to MAS identifier for the papers citing paper *P*. Publication year of the citing paper is shown in column 2. Column 3 notes the exact citation context(s) in the citing paper for paper *P*. Below, we describe as to how the

average countX and average citeWords features are computed for P over the years.

Table 6.2: Example citation contexts for paper (P) titled as *On Relaxed Dynamic Programming in Switching Systems*, published in 2005. Citer ID represents MAS identifier of the paper citing paper P. Publication year represents the year of publication of citing paper. Finally, context column contains the citing sentence. There are several instances where a paper is cited more than once in a citing paper. Also, a citing sentence might cite more than one paper. Bold face text represents a cited paper reference.

Citer ID	Publication year	Context
5330841	2010	Our approach relies on the following result from relaxed dynamic programming [**12**, 15], which is a straightforward generalization of proposition [5, Proposition 2.4], cf. [7] for a proof
6899965	2009	In [18, **19**] a relaxed dynamic programming procedure is proposed
		The existence of a solution is assumed in [18,**19**], which is different from the objective in this paper; to obtain a sub-optimal solution only when the minimum does not exist
5179891	2009	Recently, this has been studied by Lincoln and Rantzer in [**11**,17]
6006644	2009	Our approach relies on results on relaxed dynamic programming [9], [**13**] already used in an MPC context in [7] which we adapt to our variable control horizon setting
6413388	2008	Inequalities of such type have been used frequently in the optimal control literature, however, a systematic study seems to have performed only recently in [**14**, 18]
		The approach we take in this paper relies on recently developed results on relaxed dynamic programming [**14**, 18]
5052733	2008	These general algorithms are also used to study switched systems in [**11**], [12]
5433268	2007	Some are based on a newly elaborated condition of optimality see e.g., [1], [**2**],[3], others are more related to semi-classical approaches see e.g., [4], [5], [6], [7]
4971068	2007	A novel approach to overcome some of the difficulties mentioned above was recently proposed in [**4**], [5], [3], see also [14] for examples from switching systems
12659162	2006	For further details on the theoretical foundations, the reader is referred to [**13**]]
		Further discussion of the implicit algorithm is given in [**13**]
50488928	2006	In a recent work, it is shown that the optimal control problem can be reformulated as an approximate linear-quadratic problem, whose complexity grows only polynomially [**10**]

Average countX

For a citation edge from paper Q to paper P (i.e., Q citing P), countX denotes the number of times paper P is cited in paper Q. A high value of countX implies that paper P is cited multiple times by paper Q and thus, P might be quite relevant for paper Q. Possibly, Q has cited P for its different aspects. Our hypothesis is that if we consider all the papers citing a paper P and find the average value of countX for P, it may serve as a very strong feature to measure the importance of P.

Let us assume that the papers $Q_1, Q_2, \ldots, Q_n$ are citing paper P for $N_1, N_2, \ldots, N_n$ times respectively in the t^{th} year after publication of P. We define the average countX metric for paper P for the t^{th} year as

$$average\ countX(P,t) = \frac{\sum_{j=1}^{n} N_j}{n} \tag{6.1}$$

Using the example in Table 6.2, average countX value for paper *P* for the first year after publication (year 2006) can be calculated as:
$average\ countX(P,1) = \frac{2+1}{2} = 1.5$. This is because there are two citing papers in year 2006, one of which cites P twice within the same paper while the other cites it only once.

Average citeWords

For a citation edge from paper Q to paper P, citeWords denotes the number of words in the citation context(s), where P has been referred to. Since more than one paper might be cited within the same citation context, the number of words is divided among all the cited papers. Similar to countX, a high value of citeWords implies that paper P has been discussed in more details by paper Q and therefore, paper P might be relevant for paper Q. Dividing by the number of papers cited within the citation context takes care of the fact that the words in citation contexts have been used to describe multiple papers. Similar to countX, our hypothesis is that finding the average number of words that other papers use to describe P could be indicative of the importance of paper P.
Let us assume that paper P is cited by another paper Q_i in m different citation contexts, $S_1, \ldots, S_m$. For this citation edge, citeWords is computed as

$$citeWords(P,Q_i) = \sum_{i=1}^{m} AW(S_i,P,Q_i) \tag{6.2}$$

where $AW(S_i,P,Q_i)$ denotes the average number of words used in sentence S_i to describe P. In general, if $k \geq 1$ papers are cited within the sentence S_i,

the average words for each of these k papers (including P) is given by:

$$AW(S, P, Q_i) = \frac{Len(S)}{k} \tag{6.3}$$

where $Len(S)$ denotes the length of a sentence S and is simply computed by counting the number of words appearing in it. Now, assume that the papers $Q_1, Q_2, \ldots, Q_n$ are citing paper P in the t^{th} year after publication of P. We define the average citeWords metric for paper P for the t^{th} year as:

$$Average\ citeWords(P, t) = \frac{\sum_{j=1}^{n} citeWords(P, Q_i)}{n} \tag{6.4}$$

Using Table 6.2, average citeWords value for paper *P* for the third year after publication (year 2008) can be calculated as:

$$(citeWords(P, 6413388) + citeWords(P, 5052733))/2$$

To compute $citeWords(P, 5052733)$, we see that paper 5052733 cites P in one citation context where a total of two papers are cited. Thus

$$citeWords(P, 5052733) = \tfrac{11}{2} = 5.5,$$

where 11 is the length of the citation context.
Similarly, paper 6413388 cites paper P twice but in both the citation contexts, two papers are cited. Therefore,

$$citeWords(P, 6413388) = \frac{25}{2} + \frac{16}{2} = 20.5$$

Thus, $Average\ citeWords(P, 3) = \frac{20.5+5.5}{2} = 13$.

Correlation between citation counts and citation content features over the years

We investigate whether the average countX and average citeWords values over the years are correlated with the number of citations a paper receives. We reiterate that both average countX and average citeWords are normalized with respect to the number of citations received by the paper. We divide the set of papers in our dataset into 6 buckets based on the following criterion on the number of citations.

Bucket 1: Top 0.1% papers – citations 389-7859

Bucket 2: Top 0.1 - 1% papers – citations 95-389

Bucket 3: Top 1 - 5% papers – citations 29-95

Bucket 4: Top 5 - 10% papers – citations 16-29

Bucket 5: Top 10 - 25% papers – citations 6-16

Bucket 6: Rest of the papers – citations 0-6

For each of the citation buckets, we plot the temporal profile for the average countX values, averaged for all the papers within that bucket, in Figure 6.2. The $X-$axis denotes the year after publication of the paper, ranging from 0 (same year as publication) to 10 (10^{th} year after publication). While averaging for a citation bucket for a particular year, we consider only those papers which have non-zero citations in that year. The minimum value of countX can be 1 for any citation edge. Interestingly, as per our hypothesis, various citation ranges show differences in terms of the average countX values. Some important observations from Figure 6.2 are:

1. There is an increase in the value of countX in initial years irrespective of the citation bucket, and it further decreases continuously over the years. A slight increase is observed for the 10^{th} year after publication.

2. Highly cited papers are cited more number of times in a single paper.

We clearly see a correlation between the number of citations and the average countX profiles of the papers. Further, we investigate whether the countX values can discriminate between the 6 citation categories identified in [34]. Accordingly, we divided the set of papers into 6 categories mentioned in [34]. For readability, the six categories are described below:

(i) PeakInit: Papers whose citation count peaks within 5 years of publication followed by an exponential decay.

(ii) PeakMul: Papers having multiple peaks in different time periods of the citation history.

(iii) PeakLate: Papers having very few citations at the beginning and then a single peak after at least 5 years of the publication followed by an exponential decay in citation count.

(iv) MonDec: Papers whose citation count peaks in the immediate next year of the publication followed by a monotonic decrease in the number of citations.

(v) MonIncr: Papers having a monotonic increase in the number of citations from the very beginning of the year of publication till the date of observation.

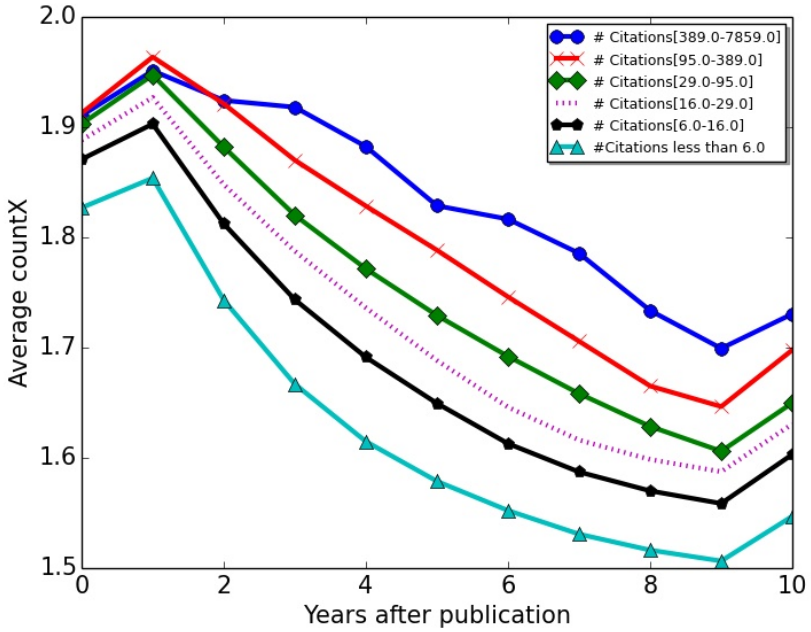


Figure 6.2: Average countX: temporal profiles for six citation buckets over the publication age.

(vi) Oth: Papers not belonging to any of the above-mentioned categories belong to this category.

Figure 6.3 presents the temporal profile of average countX values for each of these 6 categories. Again, we can see that the average countX values are the highest for the $MonIncr$ and $PeakLate$ categories, which have been identified as having the categories corresponding to a high number of citations in [34]. Similarly, average countX values are the lowest for the $MonDec$ and $Others$, which have been identified as the categories corresponding to the low number of citations (see [34] for details).

We now plot the temporal profile for the average citeWords values for the six citation buckets in Figure 6.4. Similar to average countX, while averaging for a citation bucket for a particular year, we consider only those papers which have a non-zero citation in that year. Average citeWords also shows a very similar trend as that seen with the average countX values, an initial increase and then a decreasing trend over the years. Interestingly, differences are observed between various citation ranges with the papers having the highest citations also earning a high number of average citeWords over the years.

We further use six citation categories to plot the temporal profiles in Figure 6.5. The trends are again very similar to those observed for the case of average

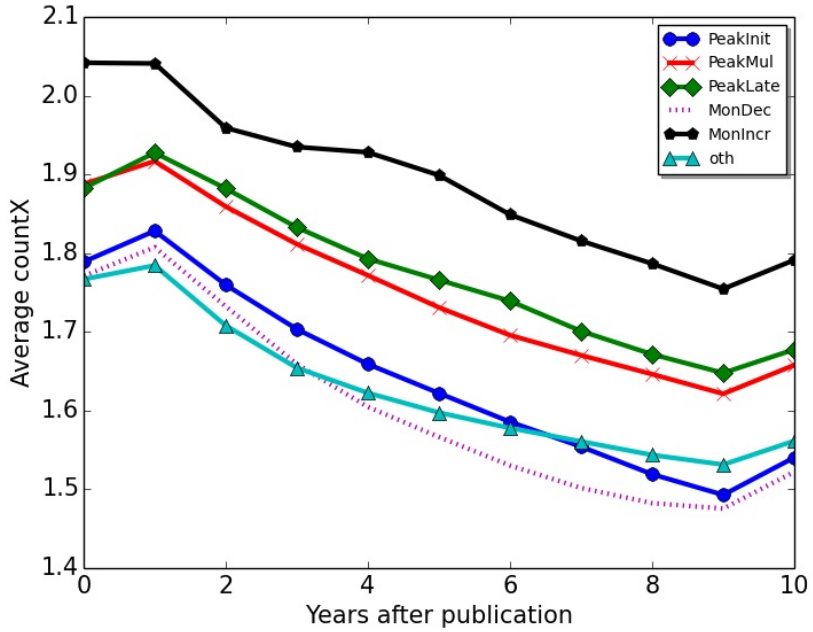


Figure 6.3: Average countX: temporal profiles for the six citation categories [34] over the publication age.

countX values, with the MonIncr and PeakLate categories having a higher value of average citeWords than the other categories and MonDec category having the lowest values.

Correlation between citation counts and citation context features for the initial years

To motivate the importance of average countX and citeWords as features for future citation prediction, Table 6.3 shows some specific examples of papers having the same citation count in the first two years after publication but different average countX and citeWords values. What we observe is that in both the cases, among the papers having the same citation count, the paper having a high countX (and citeWords) value in the initial two years receives a much higher citation count in the future. Thus, the average countX feature from the initial years of publication can serve as an important feature towards predicting future citations.

We further study whether the average countX and average citeWords values from the initial two years after publication can serve as discriminating features to predict citations at a later point in time. We, therefore, divide all the papers

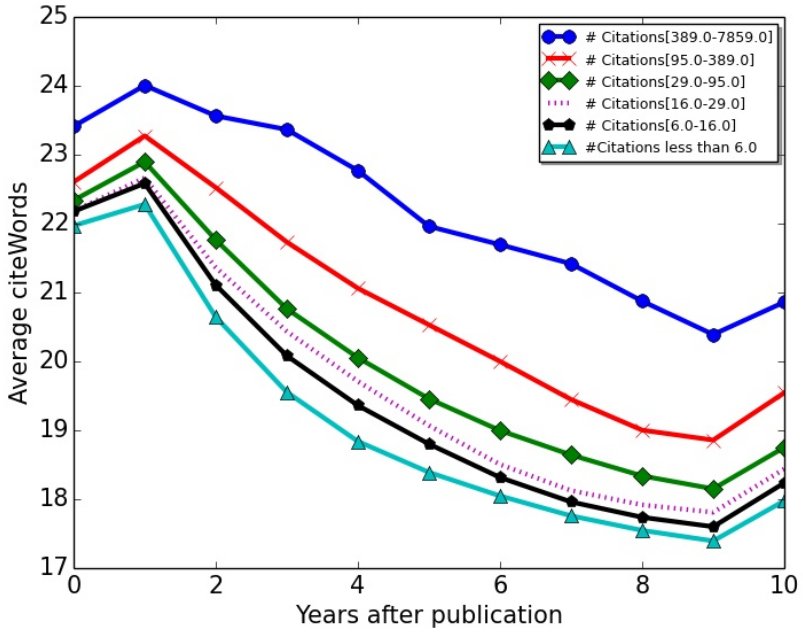


Figure 6.4: Average citeWords: temporal profiles for the six citation buckets over the first 10 years of publication age.

Table 6.3: Example paper-pairs having a similar citation count in the initial 2 years of publication but different countX values.

Paper ID	Initial citation count	Initial Avg. countX	Initial Avg. citeWords	Final Citation count
349111	4	1.75	41.75	140
25	4	1	10.58	47
1911	7	3.29	47.9	155
349954	7	1.42	16.35	38

in 3 ranges as per the average countX values ($\{1\}$, $(1, 1.5]$ and $(1.5, -)$) and as per the average citeWords ($(0, 10.5]$, $(10.5, 16.5]$ and $(16.5, -)$ in the initial two years of publication. We call these ranges as low, medium and high respectively. We now take citation counts of the papers for the time points, corresponding to 5 and 9 years after publication. For each such time point, we create 6 different citation buckets (top 0.1% etc.) and plot the distribution of the papers falling into these 6 citation buckets on various countX and citeWords ranges. For example, 5 years after publication, 75% of the papers in the lowest citation category have an average countX value=1 (see Figure 6.6(a)). On the other

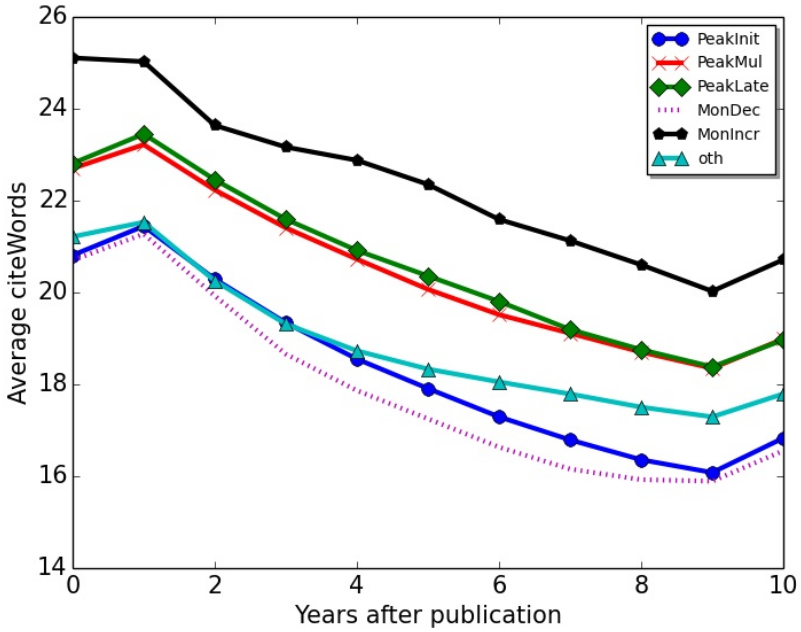


Figure 6.5: Average citeWords: temporal profiles for the six citation categories [34] over the first 10 years of publication age.

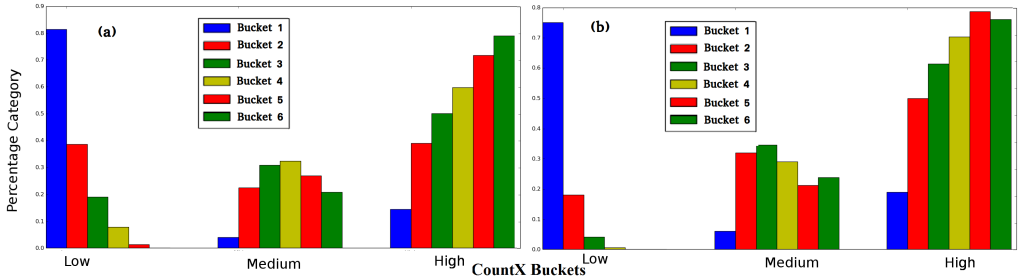


Figure 6.6: Correlating citation count and countX buckets. (a) Correlation at 5 years after publication. (b) Correlation at 9 years after publication. The six citation buckets are defined in Section 6.2.2.

hand, more than 75% of the papers in the top two categories (top 0.1% and top 0.1-1%) have a countX value ≥ 1.5. The trend becomes much more prominent for 9 years after publication (Figure 6.6(b)), with the probability of a paper having countX ≥ 1.5 increasing with increasing citation counts.
Very similar trends are observed for the average citeWords as well (see Figure 6.7). From these figures as well as examples in Table 6.3, it is clear that information from average countX and average citeWords in the initial years of

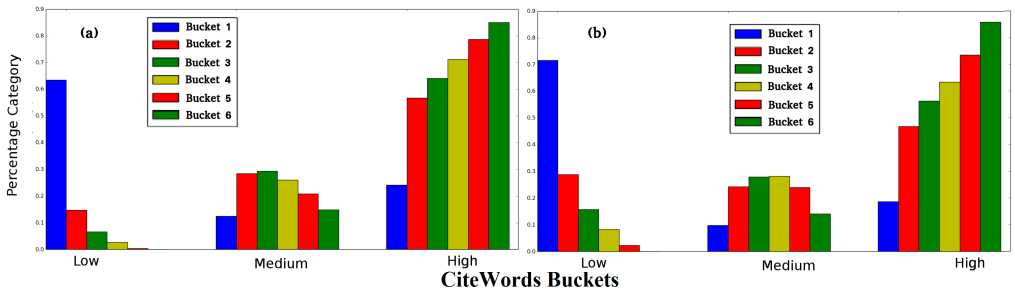


Figure 6.7: Correlating citation count and citeWords buckets. (a) Correlation at 5 years after publication. (b) Correlation at 9 years after publication. The six citation buckets are defined in Section 6.2.2.

publication act as a discriminating factor for the future citation counts, such that most of the highly cited papers have high values of average countX and citeWords in the initial years, which is not true for the low-cited papers.

Motivated by these examples, we now use these citation context features for the task of future citation prediction. The model is described in the next section.

6.2.3 Citation prediction model

We extend the two-stage stratified learning framework proposed in [31] with the addition of three features. In the first stage, a query paper is classified into one of the citation profile category using Support Vector Machine (SVM) learning model. Further, for each category, a Support Vector Regression (SVR) model is learned for predicting citation counts. Thus, given a query paper, we first classify it into one of the six citation profile categories. Post classification, category based SVR is used to predict citation count. Our citation prediction model uses features at the time of publication, along with the citation information from the first 2 years after publication. Features from the time of publication are the same as reported in [31]. These features can be divided into three categories: features based on the paper content, features based on author information and features based on venue information. We use several NLP features within paper content features. These include n-gram diversity feature (keyword diversity) and topic (inferred using Latent Dirichlet Allocation) diversity feature. For the sake of completeness, we describe these features in brief below. For more details, the reader is requested to refer to [31].

Features based on paper content

We used five paper-centric features as proposed in [31]. Last three among these are entropy-based features.
(a) Team-size (Team): The number of authors in a paper.
(b) Reference count (RefCount): The number of references mentioned in the reference section of a paper.
(c) Reference diversity (RDI): RDI measures the diversity in the fields of the referred papers. A paper citing papers of various fields has a high value of RDI.
(d) Keyword diversity (KDI): Keyword diversity refers to diversity in the keywords mentioned in the paper.
(e) Topic diversity (Topic): Each paper is assigned a set of probable topics inferred from LDA. Topic diversity gives a diversity of these probable topics.

Features based on author information

The author of a publication plays an important role in its popularity. The following four author-centric features were used for citation prediction.
(a) Author h-index (HIndex): H-index is a standard measure of author productivity and impact. This feature measures average h-index of the authors at the time of publication.
(b) Author productivity (ProAuth): Author productivity refers to the count of her publications. A more productive author will produce more. The feature is an average of the productivity of all the co-authors of a paper.
(c) Author diversity (AuthDiv): Author diversity refers to the diversity in the research fields of author publications. A highly diverse author of the paper will publish in different domains. The feature is an average of all the authors taken together.
(d) Sociality of the author (NOCA): This feature counts the number of co-authors in all the publications of each author present in the paper.

Features based on venue information

We also use certain features based on the prestige as well as the diversity of the venue, where the paper has been published. These features are described in detail below.
(a) Short-term venue prestige (VenPresS): Short term venue prestige measures the average number of citations for the papers published in a venue during the two preceding years.

(b) Long-term venue prestige (VenPresL): Long-term venue prestige measures the average number of citations for the papers published in a venue so far.
(c) Venue diversity (VenDiv): This feature measures the diversity in the research fields of the papers published in a venue.

Features after the publication year

In addition to these features, we also utilize the two features derived from the citation context, the average countX, and average citeWords for the first two years after publication, as well citation count received after the first two years of publication.
In the next section, we report the experiments using our citation prediction model.

6.2.4 Experiments

We perform experiments using the stratified learning framework for citation prediction. We selected papers having at least 10 years of history and published in between 1970 - 2005. We divided this dataset into training and testing sets. For training, we consider papers published in between 1970 - 2000. For testing purpose, we took the range as 2001 - 2005. First, we learn a stage-I classification model using our training dataset. We also learn separate regression models for each citation category, for each time point, for which the citation count is to be predicted. Given a query paper, first, the classification model is used to assign a citation category (stratum) to it (stage I). In stage II, a regression model trained on the assigned category is used for citation count prediction for the specified time periods. We use all the features described in Section 6.2.3. We have used three different time points $\Delta t = 5$, 7, and 9 for prediction.
We evaluate our model on two baselines. The first baseline [182] (baseline I) is similar to our model except that it does not include the classification stage. Thus, all the features are directly used in a regression model for citation prediction. We use Chakraborty *et al.* [31] as the second baseline (baseline II). While the authors conducted experiments both with and without the initial year of publication information, we use the citation count of first two years for their method for a fair comparison. Thus, this baseline is very similar to our model with the only difference being that we use two citation context features identified in this thesis, average countX and average citeWords, for the t^{th} year after publication, with $t = 0, 1, 2$.

Evaluation metrics

We use the following three metrics for evaluating our results. We do not use Kullback Leibler (KL) divergence because it measures distance between two probability distributions. In our setup, the two distributions (actual and predicted citation counts) are not probability distributions. Moreover, the chosen metrics are the standard metrics in citation prediction works [31, 182].

1. **Coefficient of determination (R^2):** Coefficient of determination (R^2) [28] is a number that indicates how well data fits a statistical model of future outcome prediction. It measures the variability introduced by the statistical model. It is defined as the proportionate reduction in uncertainty, due to the inclusion of regressors. Let d be the document in test document set D_T, we calculate R^2 as:

$$R^2 = \frac{\sum_{d \epsilon D_T} \left(C_{T_{ccp}}(d) - C_T(D_T)\right)^2}{\sum_{d \epsilon D_T} \left(C_T(d) - C_T(D_T)\right)^2} \tag{6.5}$$

   Here, $C_{T_{ccp}}(d)$ denotes the predicted citation count for document d. $C_T(D_T)$ denotes the mean of observed citation counts for documents in D_T. $C_T(d)$ denotes actual citation count for document d. R^2 values range between from 0 to 1. A larger value indicates better performance.

2. **Pearson correlation coefficient (ρ):** Pearson correlation co-efficient (ρ) [105] measures the degree of linear dependence between two variables. It is defined as the covariance of the two variables divided by the product of their standard deviations.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{6.6}$$

$$cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{6.7}$$

   Here, $cov(X,Y)$ denotes covariance between X and Y, σ_X and σ_Y denote standard deviation values for X and Y respectively. Similarly, μ_X and μ_Y denote mean values for variables X and Y respectively. E represents the expected value. ρ ranges from -1 to 1, where $\rho = 1$ corresponds to a total positive correlation, 0 corresponds to no correlation, and -1 corresponds to total negative correlation. A larger value indicates better performance.

3. **Mean squared error (θ):** Mean squared error (θ) measures the expected value of the squared error loss in estimation. It is a risk function

corresponding to the expected value of the squared error loss. For n number of observations, we define mean squared error as:

$$\theta = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - Y_i\right)^2}{n} \tag{6.8}$$

Here, $\hat{Y}$ and Y denote the vectors of predicted and actual values respectively. A smaller value indicates better performance.

Comparisons with the baseline models

Next, we compare the performance of the two baselines with our model. We also present performance statistics for stage I (classification) and stage II (prediction). Along with performance analysis, we compare categories and analyze results.

Table 6.4 compares the performance of these baselines with our model. Columns 2-4 in Table 6.4 show the predictive performance for baseline I using three metrics, while columns 5-7 show the predictive performance of baseline II. Columns 8-10 show the performance of our model.

We observe that for all the three systems, performance deteriorates with the increase in the time period for prediction, with the best performance achieved for $\Delta t = 5$. While baseline I performs the worst among the three models, the R^2 value of 0.56 obtained for $\Delta t = 5$ is in itself significantly better than some previous works. For example, Kulkarni *et al.* [99] achieved an R^2 value of 0.2 using 328 medical articles. Baseline II performs better than baseline I for all of the three time-periods. This performance improvement can be credited to the stratified learning approach used in baseline II, as was established in [31]. Our model performs better than both the baselines for all the three time-periods. While the improvements over the first baseline are almost over 50% in terms of R^2, improvement of the order of 8-10% is achieved over baseline II as well. Improvement in terms of θ are of the order of 20-25% over the baseline II. Since the only difference between baseline II and our model are the average countX and average citeWords features, this improvement can be credited to the use of initial year information from the citation context of the paper.

Category-wise performance analysis

Since we use the six categories as strata, we further analyze the prediction results for each of these categories. Table 6.5 presents category-wise performance metrics (except the category *Oth*) values for the three time-periods. Figure 6.8 gives the scatter plots for each category for the prediction task for the three

Table 6.4: Performance comparison between baseline I, baseline II, and our model. Three evaluation metrics θ, R^2 and ρ are used. A low value of θ and high values of R^2 and ρ represent an efficient model. Prediction is made over three time periods – $\Delta t = 5$, $\Delta t = 7$ and $\Delta t = 9$.

	Baseline I			Baseline II			Our Model		
	R^2	ρ	θ	R^2	ρ	θ	R^2	ρ	θ
Δ t=5	0.56	0.59	14.56	0.78	0.76	10.45	**0.84**	**0.79**	**7.86**
Δ t=7	0.54	0.57	15.90	0.74	0.72	12.57	**0.81**	**0.75**	**9.70**
Δ t=9	0.51	0.54	17.22	0.73	0.68	14.89	**0.78**	**0.74**	**12.43**

time periods. $X-$axis denotes the actual citation count, while the $Y-$axis denotes the predicted citation count.
From Table 6.5, we observe that for $\Delta t = 5$, the performance is the best for the *PeakLate* category on all the three metrics. Figure 6.8 also confirms this observation with most of the points densely accumulated around $x = y$ line. For $\Delta t = 7$, *PeakLate* performs the best on ρ, while *MonDec* and *MonIncr* perform well on R^2 and θ respectively. For $\Delta t = 9$, *MonIncr* performs the best among all the categories for all the three evaluation metrics. Overall, *PeakLate* and *MonIncr* categories perform the best. This is very crucial for the citation prediction model, as these categories correspond to the highly cited papers [31]. From Figure 6.8, we observe that for $\Delta t = 5$, all categories show roughly the same pattern. Majority of the papers lie below the line, which denotes that in the initial years after publication our model slightly under-estimates the citation counts. The only cases of over-estimation are for the *PeakMul* category, $\Delta t = 9$ (majority papers above the line) and for the *MonIncr* category for $\Delta t = 7$.

Table 6.5: Category-wise prediction accuracies using three metrics.

	Δ t=5			Δ t=7			Δ t=9		
	R^2	ρ	θ	R^2	ρ	θ	R^2	ρ	θ
PeakInit	0.76	0.81	7.09	0.77	0.72	9.91	0.74	0.75	14.44
PeakMul	0.79	0.73	8.25	0.78	0.76	9.78	0.78	0.73	13.40
PeakLate	**0.89**	**0.83**	**1.96**	0.81	**0.78**	9.88	0.79	0.75	13.32
MonDec	0.88	0.78	12.20	**0.89**	0.77	9.86	0.79	0.75	13.32
MonIncr	0.79	0.79	11.51	0.80	0.76	**9.22**	**0.79**	**0.79**	**12.61**

SVM classification analysis

The first stage SVM model classifies each paper into one of the six categories. Table 6.6 presents the confusion matrix of SVM classification. Each entry in

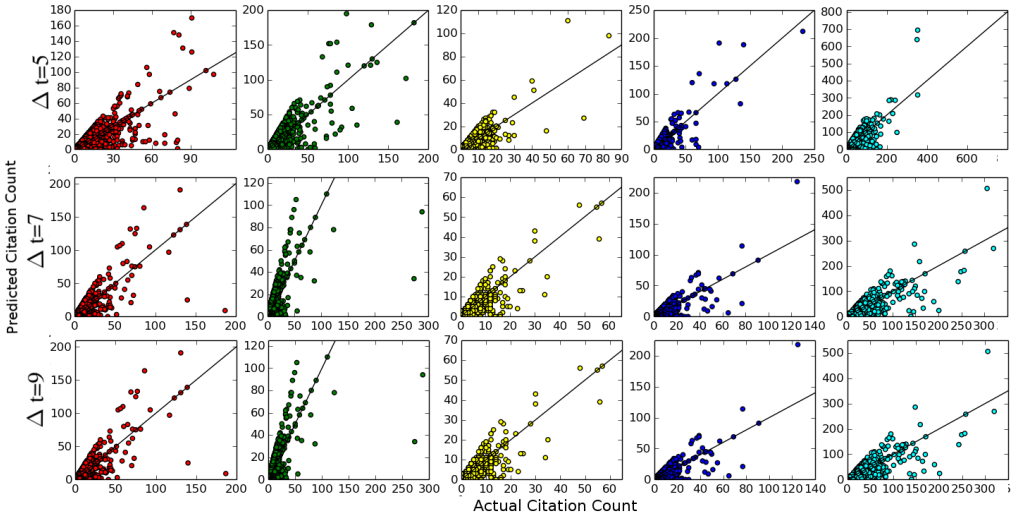


Figure 6.8: Change in prediction over the time-periods for each category. Each scatter plot shows relation between actual citation count with predicted citation count. Here, from left to right, red color represents *PeakInit*, green color represents *PeakMul*, yellow color represents *PeakLate*, blue color represents *MonDec* and cyan color represents *MonIncr*. Black color line represents $x = y$ line passing through origin.

the first column represents a ground truth category of the paper. Similarly, each entry in the first row represents predicted category. We observe that around 50% of *Oth* category paper are wrongly classified into *PeakMul*. While *MonDec* has the highest accuracy (0.989), more than 29% *PeakInit* are classified into *MonDec*, which in turn decreases the accuracy for *PeakInit* category. As our dataset is highly biased towards *Oth* category (highest % of papers), SVM overestimates *Oth* category in the classification. Classification inaccuracy in the first stage decreases prediction accuracy in the second stage, with the *Oth* category playing a significant role in lowering the precision.

Paperwise analysis

Table 6.7 presents one best representative paper from each of the five categories. For each paper, we calculate the absolute difference between actual citation count and predicted citation count for our model and baseline II for three time periods. As observed from Table 6.7, our model is closest to the actual values in terms of citations at any time instance. Baseline I shows worse results than baseline II. We, therefore, omit baseline I results due to space constraints.

Table 6.6: SVM classification confusion matrix. Column 1 represents the ground truth categories, column 2 represents total number of papers in each of these categories, columns 3-8 represent the predicted categories and column 9 presents the accuracy values for each category. Correct classification results are highlighted in bold font from column 3-8. In column 9, highlighted bold font represents both the highest and lowest accuracy values.

	No. of papers	PeakInit	PeakMul	PeakLate	MonDec	MonIncr	Oth	Accuracy
PeakInit	15178	**10987**	12	134	3245	43	757	0.724
PeakMul	30969	6	**27554**	1	1	0	3407	0.889
PeakLate	8946	49	0	**7298**	23	0	1665	0.815
MonDec	5263	1	22	0	**5207**	0	55	**0.989**
MonIncr	4010	1	64	1	0	**3005**	1003	0.749
Oth	142792	13	70618	23	1	0	**72138**	**0.494**

Table 6.7: A best representative paper for each category. each paper is mapped to its MAS paper ID. Column 3 gives the actual citation count for the paper for 3 time points. Columns 4-6 and 7-9 give the absolute difference between the actual citation count and the predicted citation count for the two systems for three different time-periods. Bold font represents the best predictions for each time period in each category. Values in parenthesis indicate predicted citation count.

Category	MAS paper ID	Actual citation count (Δt=5,7,9)	Baseline II			Our model		
			Δt=5	Δt=7	Δt=9	Δ t=5	Δt=7	Δt=9
PeakInit	73939	35,20,5	10(25)	**1(21)**	10(15)	**5(30)**	2(22)	**5(10)**
PeakMul	1447048	37,43,41	9(48)	9(52)	14(55)	**5(42)**	**5(48)**	**3(38)**
PeakLate	837621	30,33,36	8(38)	9(42)	19(55)	**5(35)**	**5(38)**	**9(45)**
MonDec	23419	8,6,3	1(7)	1(7)	3(6)	**0(8)**	**1(7)**	**1(4)**
MonIncr	9871	18,20,26	3(21)	3(23)	4(30)	**2(20)**	**1(21)**	**2(24)**

Feature Analysis

We now study as to how various features correlate with the actual citation counts. Accordingly, we divide our features into 6 different sets and compute Spearman's correlation for the three time-periods in Table 6.8. We can see from the table that the last three features, namely average countX, average citeWords and 2-year citations, show a much higher correlation than the other three feature sets. While the correlation for 2-year citation feature is slightly higher than average countX for $\Delta t = 5$, correlation is the highest for average countX for $\Delta t = 9$. Thus, average countX serves as the most important feature for predicting the long-term citation behavior of the papers.

Table 6.8: Average Spearman's rank correlation of each feature category (column 1) with the actual citation count without categorization for Δ t=5,7 and 9 years after publication.

Feature category	Δ t=5	Δ t=7	Δ t=9
Author centric	0.387	0.342	0.317
Venue centric	0.343	0.309	0.285
Paper centric	0.429	0.417	0.392
Average countX	0.569	0.543	0.521
Average citeWords	0.512	0.499	0.481
2 year citation	0.571	0.543	0.502

Table 6.9: Comparing related works in citation prediction: column 1 presents the title of the paper, column 2 presents the size of the dataset used in the paper, column 3 lists year range of test papers, column 4 presents the time periods used for prediction, column 5 lists the method/model used for prediction and column 6 presents the R^2 values reported in the paper for a time period, comparable across different methods. Papers are arranged in the increasing order of R^2 values.

Title of paper	Dataset size	Year range	Time period(s)	Method	R^2 (time period)
Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study [114]	1274	2005	3.5	Decision Trees	0.14(3.5)
Characteristics Associated with Citation Rate of the Medical Literature [99]	328	1999 - 2000	5	Linear Regression	0.2(5)
Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals [117]	204	1991	2	Multiple Regression	0.60(2)
Towards a Stratified Learning Approach to Predict Future Citation Counts [31]	1,549,317	2001-2005	1,2,3,4,5	SVR	0.71(5)
		1996 - 2000	1,2,3,4,5	SVR	0.74(5)
Citation Count Prediction: Learning to Estimate Future Citations for Literature [183]	1,558,499	1960 - 2011	5	CART	0.752(5)
The role of citation context in predicting long-term citation profiles: an experimental study based on a massive bibliographic text dataset [our model]	1,359,338	2001-2005	5,7,9	SVR	0.84(5)

Comparison with past works

The experimental results clearly confirm that the proposed method for citation prediction outperforms the other baselines for various time-periods. Further, we wanted to put this work in perspective of the previous related works for this problem. Table 6.9 lists five other works and compares them to the size of the dataset used for the study, year-ranges of the test papers, the method used by the papers, as well a time period for which the R^2 values have been reported. Our dataset size is comparable to the other datasets reported in the literature. Also, we achieve a better R^2 value on this massive dataset than the ones reported earlier in the literature. Our prediction time period ($\Delta t = 9$) is the maximum among all these works.

6.3 Impact of early citers on long-term scientific impact

6.3.1 Early (non-)influential citers

The term *early citations* refers to citations accumulated immediately after the publication. In the literature, although, there seems to be no general definition of 'early', the majority of the works kept it within ~ 2 years after publication [2]. Multiple previous works assert that early citation count helps in better prediction of the long-term scientific impact ($LTSI$) [2, 23, 31]. Although these approaches are interesting, they fail to capture the existence of different types of *early citations* leading to more complex influence patterns on $LTSI$.
Given a candidate paper P published in the year T, we are interested in the citation information generated within δ year(s) after publication, i.e., within the time interval $[T, T + \delta]$. For example, for $\delta = 2$, if an article is published in the year 2000, we look into the citation information generated till 2002. *Early citation count* $ECC_\delta(P)$ refers to the total number of citations received by the paper P from other articles within δ years after publication. Note, $ECC_\delta(P)$ quantitatively measures the early popularity of the paper P. However, $ECC_\delta(P)$ fails to capture the inherent nature of the individual early citations; for example, there exists no distinction between:

- originators (authors, journals etc.) of early citations.
- good (substantiating) and bad (criticizing) citations.
- self and non-self citations.

To incorporate some of the above distinctive characteristics in $ECC_\delta(P)$ and to better understand the inherent nature of the individual citations, we present the following three definitions:

Early citers ($EC_\delta(P)$): $EC_\delta(P)$ represents the set of authors that cite paper P within δ years after its publication. Figure 6.9 shows schematic representation of $EC_\delta(P)$ on a temporal scale. Here, $EC_\delta(P)$ consists of all authors that cite paper P within δ year after its publication. Further, we divide this set into two subsets – i) *influential* and ii) *non-influential* early citers.

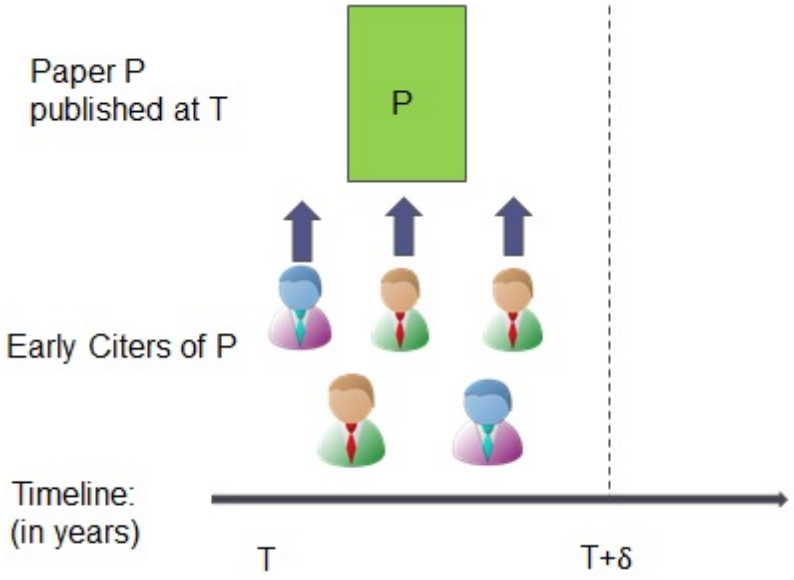


Figure 6.9: Schematic representation of early citers on a temporal scale. Early citers consist of all authors that cite paper P within δ year(s) after its publication. The set of early citers is divided into two subsets, namely, a) influential and b) non-influential. Influential early citers are represented in purple color (online) whereas non-influential early citers are represented in green color.

Influential early citers ($IEC_\delta(P)$): This is a subset of $EC_\delta(P)$ in which each author either has a high publication count or a high citation count or both at the time of the citation. Note that, in the current work, we consider top $\sim 5\%$ authors as influential early citers, both in terms of publication and citation counts. Empirically (from dataset described in Section 6.3.2), we find that top $\sim 5\%$ consists of authors who have authored at least 21 publications or acquired at least 250 citations or both. In Figure 6.9, for paper P, $IEC_\delta(P)$ are represented in the purple color.

Non-influential early citers ($NEC_\delta(P)$): Early citers that are not influential constitutes the set of non-influential citers, i.e.

$$NEC_\delta(P) = EC_\delta(P) \setminus IEC_\delta(P) \tag{6.9}$$

As described before, NEC_δ consists of the remaining $\sim 95\%$ of the authors in $EC_\delta(P)$. In Figure 6.9, $NEC_\delta(P)$ authors are represented in green color. To study the impact of influential and non-influential EC on citations gained at a later point in time, we define long-term scientific impact as:
Long-term scientific impact ($LTSI_\Delta(P)$): Given a paper P, it represents cumulative citation count of P after Δ years of its publication. Section 6.3.3 demonstrates the effect of influential and non-influential EC on $LTSI$. Next, we describe the dataset we employ for the large-scale experimental study and for the extended prediction framework.

6.3.2 Dataset description

In this work also, we utilize the same two open source computer science datasets, used in the previous study (see Section 6.2) both crawled from the Microsoft Academic Search. We filter the datasets by removing papers with incomplete information about the title, the abstract, the venue, the author(s), etc. Since the current study entirely focuses on early citers, we only include papers that consist of at least one citation within $\delta(= 2)$ years after publication. We term this dataset as *filtered dataset*. Table 6.10 outlines various statistics for both the datasets. For the rest of this chapter, we conduct all our experiments on the filtered dataset unless otherwise stated.

Table 6.10: General information about the datasets. We combine the two separately crawled datasets – (a) the bibliographic dataset and (b) the citation context dataset into a single compiled dataset. We create the filtered dataset after removing incomplete information from the compiled dataset. Note, the filtered dataset consists of articles that have at least one citation within $\delta(= 2)$ years after publication.

	Compiled dataset	Filtered dataset
No. of publications	2,473,147	949,336
No. of authors	1,186,412	535,543
Year range	1859–2012	1970–2010
No. of citation contexts	26,037,804	11,532,780

6.3.3 Empirical study

In this section, we plan to empirically investigate how the early citers impact the $LTSI$ of a paper. The section begins by introducing three properties of early citers, namely, the publication count, the citation count and the co-authorship distance. We describe each property in detail and present correlation (using Pearson Correlation) statistics along with representative examples.

General Setting: Given a candidate paper P, we construct a set of early citing papers C_P that cite P within δ year(s) after publication. For the current study, we keep $\delta = 2$. From the definition presented in Section 6.3.1, $EC_\delta(P)$ consists of all authors that have written papers present in C_P. Next, for each paper $c \in C_P$, we select one representative author among all co-authors based on different selection criterion. More specifically, each selection criterion refers to one distinguishing property of EC. Further, we construct a representative author subset $REC_\delta(P)$ from the selected authors and present correlation statistics of this newly constructed subset with $LTSI$. Note that $REC_\delta(P) \subseteq EC_\delta(P)$. Next, we define the three key properties of EC that assist in distinguishing early citations.

Publication count

Publication count of an early citer refers to the number of articles written by her before citing the paper P. High publication count denotes high productivity of an early citer. For each paper $c \in C_P$, we select the author with the maximum publication count. The authors so selected constitute the set $REC_\delta(P)$. Note that in our experiments, authors with minimum, average and median publication counts have not shown significant correlations. Further, we aggregate early citers' publication counts (PC_P) by averaging over the set of selected authors $REC_\delta(P)$. For each paper P present in our dataset, we compute PC_P and P's cumulative citation count at five later time periods after publication, $\Delta t = 5, 8, 10, 12, 15$. We utilize the definitions of influential and non-influential early citers described in Section 6.3.1, i.e., a paper P is cited by a set of influential early citers, if $PC_P >= 21$. Therefore, we split the entire paper set into two subsets: i) papers cited by non-influential EC ($PC_P < 21$), and ii) papers cited by influential EC ($PC_P >= 21$). Figure 6.10 compares these two subsets correlating PC values with cumulative citation counts at five later time periods.

Observations: Figure 6.10 presents few interesting observations. Papers with lower value of $PC(< 21)$ exhibit positive correlation. However, as Δt progresses, this positive correlation starts diminishing. Surprisingly, papers with higher values of $PC(>= 21)$, show negative correlation and this effect be-

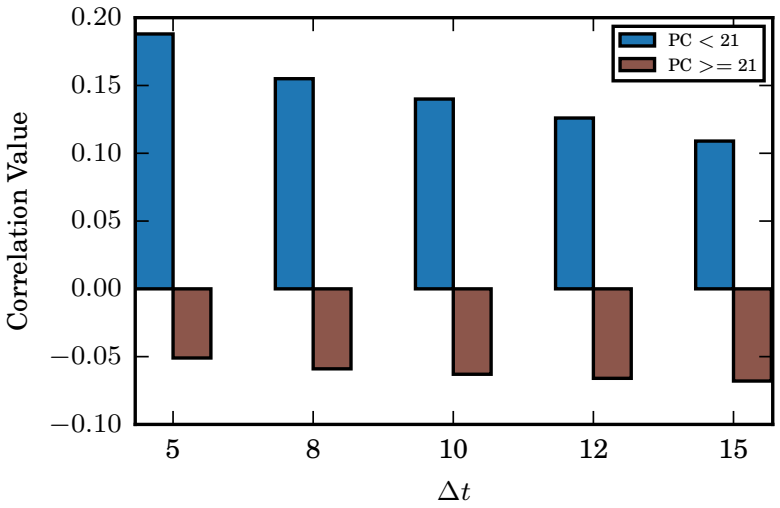


Figure 6.10: Correlation between EC publication count and cumulative citation count at five later time periods after publication, $\Delta t = 5, 8, 10, 12, 15$. Papers with lower value of $PC(< 21)$ exhibit positive correlation diminishing over the time. Papers with high value of $PC(>= 21)$ show an opposite trend. The overall separation decreases over time.

comes more profound as Δt progresses. Thus, the overall separation between the two subsets decreases over time.

This study illustrates the fact that influential EC negatively affect the long-term citations. A plausible explanation could be that in general, researchers tend to cite works written by influential authors. Therefore, once an influential author cites an article, researchers tend to cite the influential author's paper, instead of the original paper. The attention from the original paper moves to the paper written by the influential citer in the very beginning of the life-span of the original paper. Therefore, instead of flourishing, the long-term citation count of the original paper gets negatively affected. This phenomenon of attention relaying from the less popular article to the more popular article is described as *attention stealing* [176]. In case of non-influential EC, the citation count of the candidate paper exhibits a positive correlation with PC. However, with the passage of time, this positive correlation diminishes due to ageing effect associated with paper's life span [168]. In case of influential EC, same ageing effect leads to increase in the negative correlation over the passage of time.

Table 6.11 shows some specific examples of papers having the same early citation count in the first two years after publication but different PC values. In both cases, the paper having a low PC value receives a much higher citation

count in the future.

Table 6.11: Example paper-pairs having a similar early citation count in the initial two years of publication but different PC values.

Paper ID	Early citation count	Early citer PC	Later citation count
726084	13	18.9	79
140790	13	36.5	34
1663998	8	19.17	109
150167	8	65	38

Citation count

Citation count of an early citer refers to the number of citations received by her before citing paper P. High citation count denotes higher popularity of the early citer. Again, for each paper $c \in C_P$, we select the author with maximum citation count. Here again, the authors so selected constitute the set $REC_\delta(P)$. Further, we aggregate early citers' citation counts (CC_P) by averaging over the set of selected authors $REC_\delta(P)$. For each paper P present in our dataset, we compute CC_P and P's cumulative citation count at five later time periods after publication, $\Delta t = 5, 8, 10, 12, 15$. Similar to previous section, we again split the entire paper set into two subsets: i) papers cited by non-influential EC ($CC_P < 250$), and ii) papers cited by influential EC ($CC_P >= 250$). Figure 6.11 compares these two subsets by correlating CC values with the cumulative citation counts at five later time periods.

Observations: Figure 6.11 presents similar observations as reported in Figure 6.10. Papers with a lower value of $CC(< 250)$ exhibit positive correlation diminishing over the time. Papers with a high value of $CC(>= 250)$ show an exactly opposite trend. Here also, the overall separation decreases with time. The results again confirm the existence of *attention stealing*, i.e., a popular citer steals the attention from a newly arrived paper by citing it. The temporal increase and decrease in correlation values of influential and non-influential early citers respectively relate to the ageing effect as discussed in the previous section.

Table 6.12 shows some specific examples of papers having the same early citation count in the first two years after publication but different CC values. Similar to publication count, here also, we observe that in both the cases, the paper having a low CC value receives a much higher citation count in the future.

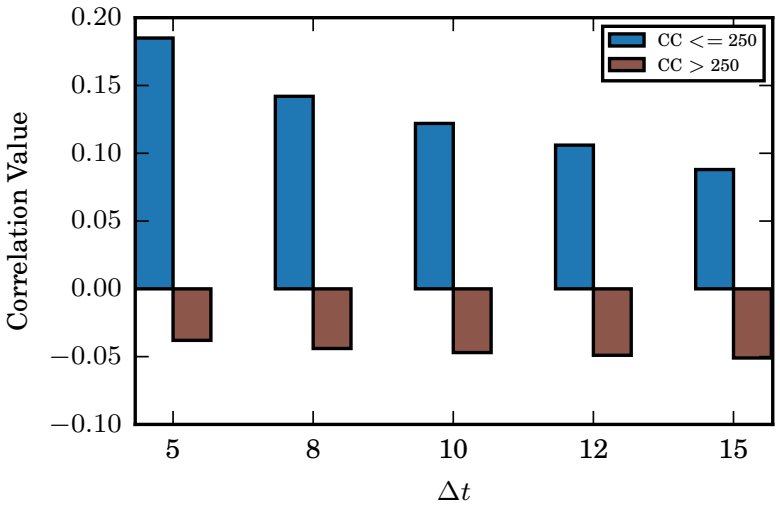


Figure 6.11: Correlation between EC citation count and cumulative citation count at five later time periods after publication, $\Delta t = 5, 8, 10, 12, 15$. Papers with lower value of $CC(< 250)$ exhibit positive correlation diminishing over the time. Papers with high value of $CC(>= 250)$ show an opposite trend. The overall separation decreases over time.

Table 6.12: Example paper-pairs having a similar early citation count in the initial two years of publication but different CC values.

Paper ID	Early citation count	Early citer CC	Later citation count
2025205	4	124.75	51
287142	4	456	13
269672	18	74.45	61
1695635	18	623.17	29

Co-authorship distance

We construct a collaboration graph $G(V, E)$ to understand the effect of the co-authorship distance between EC and the authors of candidate paper P on $LTSI$. Here, V is the set of vertices representing authors and an edge $e \in E$ between two authors denotes that they have co-authored at least one article. We define the co-authorship distance (CA) between two authors as *the shortest distance between the two in the co-authorship network*. Again, for each paper $c \in C_P$, we select the author with the lowest CA from the authors of candidate paper P. The authors so selected constitutes the set $REC_\delta(P)$ here. Note that in our experiments, authors with highest, average and median co-authorship distance have not shown better correlations. We aggregate the co-authorship

distance (CA_P) by averaging over the set of selected authors $REC_\delta(P)$. To understand the effect of co-authorship distance on $LTSI$, we divide CA into three buckets:

- **Bucket 1**: $0 \leq CA < 1$
- **Bucket 2**: $1 \leq CA < 2$
- **Bucket 3**: $CA \geq 2$

Note, $CA = 0$ represents self-citations, i.e., one of the early citer is the author of the candidate paper P. The authors at $CA = 1$ are the co-authors of the authors in the candidate paper. Hence, **Bucket 1** mainly consists of authors of the candidate paper itself. **Bucket 2** mainly consists of the immediate co-authors of the author set of the candidate paper while **Bucket 3** mainly consists of co-authors of co-authors (distant neighbours) of the author set of the candidate paper.

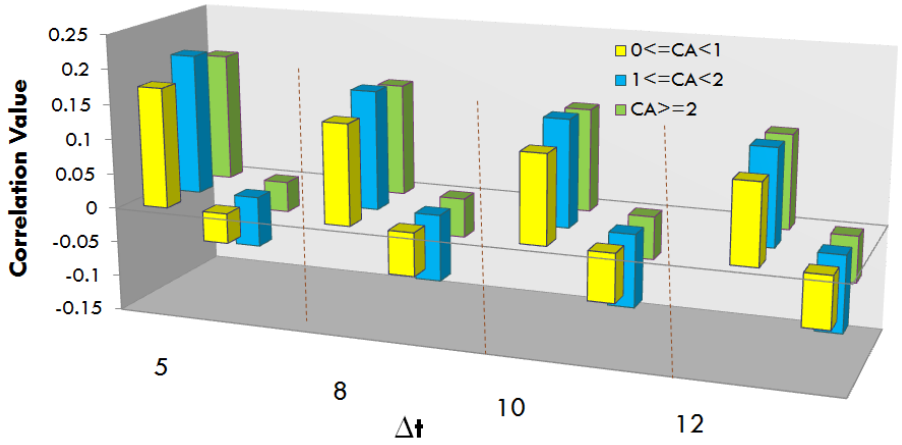


Figure 6.12: Correlation between EC's publication count and cumulative citation count for three co-authorship buckets at four later time periods after publication, $\Delta t = 5, 8, 10, 12$. For each time period, first three bars represent correlation for non-influential EC ($PC_P < 21$) whereas the next three bars represent correlation for influential EC ($PC_P >= 21$). Influential immediate co-authors (Bucket 2) seem to badly affect the citation of the candidate paper P in the long term.

For each bucket, we present correlation statistics of EC's publication count and citation count with $LTSI$. Figure 6.12 illustrates, for each bucket, correlation between EC's publication count and cumulative citation count at four later time periods after publication, $\Delta t = 5, 8, 10, 12$. For each time period, the first three bars represent correlation for non-influential EC ($PC_P < 21$) whereas the next three bars represent correlation for influential EC ($PC_P >= 21$).

Observations: For each CA bucket, we observe similar trends as before, influential EC negatively affect the $LTSI$ while non-influential EC affect positively. The most striking observation from this experiment is the effect of immediate co-authors (**Bucket 2**) on $LTSI$. Even though both influential or non-influential immediate co-authors maximally correlate with $LTSI$, influential immediate co-authors negatively affect the citation of the candidate paper P in the long term due to intensified *attention stealing* effect.

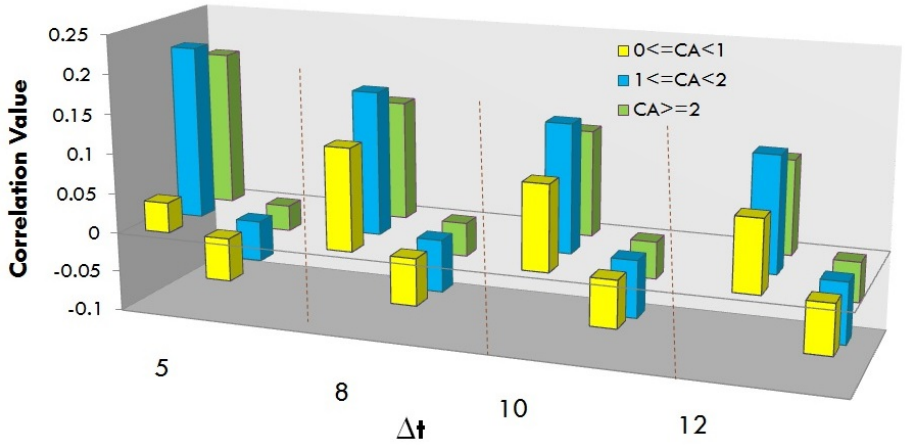


Figure 6.13: Correlation between EC's citation count and cumulative citation count for three co-authorship buckets at four later time periods after publication, $\Delta t = 5, 8, 10, 12$. For each time period, first three bars represent correlation for non-influential EC ($CC_P < 250$) whereas next three bars represent correlation for influential EC ($CC_P >= 250$). Influential immediate co-authors (bucket 2) badly affect the attention of candidate paper P in long term.

Similarly, Figure 6.13 illustrates correlation between EC's citation count and cumulative citation count at four later time periods after publication. For each time period, the first three bars represent correlation for non-influential EC ($CC_P < 250$) whereas the next three bars represent correlation for influential EC ($CC_P >= 250$).

Observations: In this case, the observations are very similar to the previous case. Motivated by these empirical observations, we incorporate the EC properties in a well-recognized citation prediction framework as described in the next section.

6.3.4 Citation prediction framework

As an intuitive use case, we extend the long-term citation prediction framework proposed by [182] by including the three EC properties discussed in the previous sections. In addition, we also include two citation context-based features proposed in the previous section. Given a candidate paper, we predict its cumulative citation count at five different time-points ($\Delta t = 3, 5, 7, 9, 11$) after publication. Our citation prediction framework employs a set of features that can be computed at the time of publication plus a set of features that can be extracted from the citation information generated within two years after publication (Section 6.3.4). We train four predictive models for comparative study, namely, linear regression, Gaussian process regression, classification and regression trees and support vector regression. We discuss each model briefly in Section 6.3.4. We compare our proposed prediction framework with three baselines in Section 6.3.4 using evaluation metrics outlined in Section 6.3.4.

Feature definition

As described before, we utilize features available at the time of publication along with the features available within two years after publication. The feature set consists of 20 different features, out of which 14 features are available at the publication time, while the other six features utilize citation information generated within two years after publication. Features[4] available at the time of publication are the same as reported in [182]. Similarly, early citation count and citation context features available after publication are same as reported in previous section. The entire feature set can be divided into seven categories: (i) features based on early citer properties, (ii) early citation count, (iii) features based on paper information, (iv) features based on author information, (v) features based on venue information, (vi) paper recency, and (vii) features based on citation context. Similar to Section 6.2.3, here also, paper information includes two NLP features (i) n-gram diversity (keyword diversity), and (ii) topic (inferred using Latent Dirichlet Allocation) diversity. Given a candidate paper P published in the year T, we compute the following features:

- **Early citer centric features:** Early citer centric features are computed within two years after the publication. Given a set of early citing papers C_P, we compute three features:

    1. **Publication count (ECPC)**: For each early citing article, we select the author with the maximum publication count. ECPC is computed

[4]Some of these features might appear correlated; however, we use all of these in order to have a faithful reproduction of the model proposed in [182].

by averaging this maximum publication count over all the early citing articles.

2. **Citation count (ECCC)**: Here, for each early citing article, we select the author with the maximum citation count. ECCC is then computed by averaging this maximum citation count over all the early citing articles.
3. **Co-authorship distance (ECCA)**: Here, we select the author with the minimum co-authorship distance from the authors of the candidate paper P. ECCA is computed by averaging this minimum co-authorship distance over all the early citing articles.

- **Early citation count (ECC):** This feature simply includes the citation counts of paper P generated within the first two years after publication.
- **Paper centric features:**
  1. **Novelty (PCN)**: Novelty measures the similarity between paper P and the other publications in the dataset. It is computed by measuring Kullback-Leibler Divergence of an article against all its references. We assume that low similarity means high novelty and more novel article should attract more citations.
  2. **Topic Rank (PCTR)**: Topics are inferred from the paper title and abstract using unsupervised LDA. Each paper is assigned a topic and further each topic is ranked based on the average citations it has received.
  3. **Diversity (PCD)**: Diversity measures the breadth of an article inferred from its topic distribution. We measure the diversity of an article by computing the entropy of the paper's topic distribution (see [182] for more details).
- **Author centric features:**
  1. **H-Index (ACHI)**: H-index attempts to measure both the productivity and the impact of the published work of a researcher [81]. Yan *et al.* [182] observed a high positive correlation between h-index and average citation counts of publications.
  2. **Author rank (ACAR)**: Author rank determines the "fame" of an author. Each author is assigned an author rank based on her current citation count. High-rank authors have high citation counts.
  3. **Past influence of authors (ACPI)**: We measure the past influence of authors in two ways: previous (1) maximum citation counts,

and (2) total citation counts. Previous maximum citation count of an author represents the citation count of author's most popular publication. Previous total citation count represents the sum of the citation counts of all the author's publications.

4. **Productivity (ACP)**: The more papers an author has published, the higher average citation counts she could expect. Productivity refers to the total number of articles published by an author.
5. **Sociality (ACS)**: A widely connected author is more likely to be cited by her wide variety of co-authors. Sociality, thus, can be computed from the co-authorship network graph employing a formulation in a recursive form as in the PageRank algorithm.
6. **Authority (ACA)**: A widely cited paper indicates peer acknowledgments, and hence indicates the 'authority' of its authors. We compute authority of paper in citation network graph using the similar recursive algorithm as proposed for the sociality feature. The paper authority then is transmitted to all its authors.
7. **Versatility (ACV)**: Versatility represents the topical breadth of an author. We measure the versatility of an author by computing the entropy of the author's topic distribution. Higher versatility implies large volumes of the audience from various research fields.

- **Venue centric features:**

1. **Venue rank (VCVR)**: The reputation of a venue relates to the volume of citations it receives. Similar to author rank, we rank venues based on its current citation count. High-rank venues have high citation counts.
2. **Venue centrality (VCVC)**: We create a venue connectivity graph $G(V, E)$ where V denotes the set of venues and the edges $e \in E$ denote the citing-cited relationships between venues. The in-degrees measure how many times a venue is cited by papers from other venues. Finally, venue centrality can be measured using a PageRank algorithm.
3. **Past influence of venues (VCPI)**: Past influence of a venue is computed similar to the past influence of authors. As in the case of authors, we measure the past influence of venues in two ways: previous (1) maximum influence of venues, and (2) total influence of venues.

- **Recency (PR):** Recency describes the temporal proximity of an article. It measures the age of a published article. The longer an article is published, the more citations it may receive.

- **Citation context centric features:**

  1. **Average countX (CCAC)**: A high value of countX implies that the cited paper is referred to multiple times by the citer paper in different sections of its text. Thus, cited paper might be quite relevant for citing paper. In the previous section, we argued that highly cited papers are cited more number of times in a single text.
  2. **Average citeWords (CCAW)**: Similar to countX, a high value of citeWords implies that the cited paper has been discussed in more details by the citer paper and therefore, cited paper might be quite relevant for the citing paper.

Predictive models

In this section, we describe four regression models. Each model is trained on features described in the previous section. All models are trained using available implementations from the Weka toolkit [73].

- **Linear regression (LR):** Linear regression is an approach to model the relationship between the dependent variable Y and one or more independent (explanatory) variables X. It attempts to model this relationship by fitting a linear equation to observed data. A linear regression line has an equation of the form:

$$Y = wX^T + b, \tag{6.10}$$

  where Y is the dependent variable, X^T is a vector of explanatory variables, w is a vector of weights (parameters) of the linear regression and b represents the error. In the current work, we consider publication's predicted citation count to be the dependent variable and features (described in Section 6.3.4) are considered to be the explanatory variables.

- **Gaussian process regression (GPR):** Due to the complex nature of the long-term citation impact estimation, it might well be the case that the dependent variable is a non-linear function of all the features used to represent the data. Gaussian processes [143] provide formulations by which the prior information about the regression parameters can be easily encoded. This property makes them convenient for our problem

formulation. Given a vector of input features X, the predicted citation counts $C(d)$ of the document d is:

$$C(d) = K(X, X^T)[K(X^T, X^T) + \sigma^2 I]^{-1} C(d^T), \tag{6.11}$$

where X^T is a matrix of feature vectors of the training set, K is a kernel function, I is the identity matrix, σ is the noise parameter and $C(d^T)$ is the vector of citation counts of the training set. Note, in our experiments, we keep $\sigma = 0.5$.

- **Classification and regression trees (CART):** Classification and regression trees [24] are obtained by recursively partitioning the training data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. Regression trees are built for dependent variables (citation count in the present context) that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values.

- **Support vector regression (SVR):** Support vector regression [157] is derived from statistical learning theory and they work by solving a constrained quadratic problem where the convex objective function for minimization is given by the combination of a loss function with a regularization term. Support vector regression is the most common application form of SVMs. In the current study, we employ LIBSVM[5] with default parameter settings. The best results were obtained for the linear kernel.

Baselines

- **Baseline I:** The first baseline [182] is similar to our model except that it does not include any information generated after the publication. It includes paper, author and venue centric features along with recency.

- **Baseline II:** The second baseline is similar to Baseline I plus one more feature – early citation counts. Chakraborty *et al.* [31] showed that inclusion of early citation counts enhances prediction accuracies mostly for the higher values of Δt.

- **Baseline III:** In the third baseline, we include citation context centric features introduced in the previous section to Baseline II. Thus, baseline III consists of paper, author, venue and citation context centric features along with recency and early citation count.

[5] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

Evaluation metrics

We employ two evaluation metrics to compare our model with baselines (described in Section 6.3.4). The metrics are Coefficient of determination (R^2) and Pearson correlation coefficient (ρ) (refer to Section 6.2.4 for detailed description).

6.3.5 Prediction analysis

Experimental setup

Our experimental setup bears a close resemblance to [182]. We randomly select 10,000 training sample papers published in and before the year 1995. We opted for a small sample size because of associated computational complexities. Since our prediction framework utilizes information generated within first two years after publication, we perform prediction task from 1998 – 2010. The reason behind choosing 1998 as the start year is to counter information leakage due to the training papers published at 1995 since prediction framework utilizes early citation data till 1997 for papers published in the year 1995. To evaluate, we select three random sets of 10,000 sample papers (published between 1998 – 2010). Note that for $\Delta t = 11$, we can only consider papers published between 1998 – 1999, for $\Delta t = 9$, we can consider papers published between 1998 – 2001 and so on. Given a candidate paper, we predict its cumulative citation count at five different time-points after publication, $\Delta t = 3, 5, 7, 9, 11$. For example, given a candidate paper P published in 1998, $\Delta t = 3$ represents prediction at 2001, $\Delta t = 5$ represents prediction at 2003 and so on. In the next section, we present a comprehensive analysis of our proposed framework.

Prediction results

1. **Comparison between predictive models:** To begin with, we incorporate all features described in Section 6.3.4 for the prediction task (includes early citer centric, paper-centric, author-centric, venue centric, citation context centric features plus early citation count and recency features). However, we observe marginal performance gain in all models after removing the citation context-based features. Therefore, it was decided that the best framework (hereafter '*our model*') for this prediction task would consist of all features except the citation context-based features. Table 6.13 compares the four predictive models (LR, GPR, CART and SVR) at five different time-points after publication, $\Delta t = 3, 5, 7, 9, 11$. Overall, SVR achieves the best performance, while GPR seems to have

the worst performance. As expected, in all the models, the performance diminishes as Δt increases.

2. **Comparison with the baseline models:** Next, we compare the performance of the three baselines (described in Section 6.3.4) with our model. Due to high-performance gain discussed in the previous section, we use SVR for modeling the three baselines as well as our model. Table 6.14 compares Baseline I, Baseline II and Baseline III with our model. Prediction is made over five time periods, $\Delta t = 3, 5, 7, 9, 11$. Each cell represents mean and standard deviation (in parenthesis) of the metric values for the three random samples. Even though, as highlighted, our model by far outperforms all three baselines at each time period for both metrics, it slightly under estimates LTSI (see Figure 6.14).

3. **Effect of different early time periods:** So far, we have performed experiments for a fixed early time period ($\delta = 2$). In this section, we experiment with $\delta = 1, 2, 3$ for estimating the early citer features[6]. Table 6.15 compares the prediction results for the SVR model using three different values of δ. The table presents an interesting finding that increasing the value of δ does not always improve prediction accuracy. R^2 values at $\delta = 2$ always outperform $\delta = 1, 3$ in the later time points.

Table 6.13: Performance comparison among the four predictive models – LR, GPR, CART and SVR. Two evaluation metrics $R^2$and ρ are used. A high value of $R^2$and ρ represent an efficient prediction. Prediction is performed over five time periods, $\Delta t = 3, 5, 7, 9, 11$.

Model	$\Delta T = 3$		$\Delta T = 5$		$\Delta T = 7$		$\Delta T = 9$		$\Delta T = 11$	
	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2
LR	0.95	0.82	0.91	0.79	0.84	0.74	0.81	0.68	0.75	0.61
GPR	0.83	0.57	0.80	0.55	0.71	0.48	0.66	0.47	0.64	0.30
CART	0.95	0.73	0.87	0.68	0.79	0.62	0.75	0.55	0.63	0.43
SVR	**0.97**	**0.84**	**0.91**	**0.82**	**0.88**	**0.76**	**0.82**	**0.69**	**0.76**	**0.65**

Feature analysis

We now study how the various features correlate with the actual citation counts. As described in Section 6.3.4, our model is trained on 18 features out of 20

[6]Note that the early citation count however is obtained using $\delta = 2$ as suggested in the literature.

Table 6.14: Performance comparison among baseline I, baseline II, baseline III and our model. Two evaluation metrics ρ and R^2 are used. A high value of both metrics represent an efficient model. Prediction is made over five time periods, $\Delta t = 3, 5, 7, 9, 11$. Each cell represents mean and standard deviation (in parenthesis) of the metric values for three random samples. Bold numbers in the table indicate the best performing model for a given time period. Our model by far outperforms all three baselines at each time period for both metrics.

Δt	Baseline I		Baseline II	
	ρ	R^2	ρ	R^2
3	0.793 (0.003)	0.654 (0.019)	0.856 (0.021)	0.724 (0.001)
5	0.745 (0.021)	0.644 (0.006)	0.792 (0.007)	0.699 (0.012)
7	0.691 (0.016)	0.593 (0.003)	0.752 (0.004)	0.688 (0.019)
9	0.543 (0.008)	0.588 (0.015)	0.646 (0.009)	0.639 (0.002)
11	0.591 (0.015)	0.544 (0.002)	0.633 (0.010)	0.542 (0.006)
Δt	Baseline III		Our model	
	ρ	R^2	ρ	R^2
3	0.895 (0.012)	0.769 (0.017)	**0.971** (0.002)	**0.841** (0.001)
5	0.814 (0.019)	0.788 (0.001)	**0.915** (0.015)	**0.819** (0.019)
7	0.754 (0.023)	0.690 (0.026)	**0.877** (0.007)	**0.765** (0.013)
9	0.684 (0.002)	0.643 (0.001)	**0.819** (0.003)	**0.687** (0.021)
11	0.675 (0.008)	0.582 (0.021)	**0.758** (0.005)	**0.651** (0.016)

Table 6.15: Performance of the model assuming different values of δ. Prediction is made over three early time periods, $\delta = 1, 2, 3$, and at three later time points, $\Delta t = 5, 7, 9$. Best results are obtained at $\delta = 2$. The added information does not always improve prediction accuracy.

ΔT	$\delta = 1$		$\delta = 2$		$\delta = 3$	
	ρ	R^2	ρ	R^2	ρ	R^2
5	0.882	0.68	**0.915**	**0.82**	0.911	0.76
7	0.841	0.61	0.877	**0.77**	**0.884**	0.72
9	0.765	0.58	0.819	**0.69**	**0.822**	0.64

features (described in Section 6.3.4); therefore, we perform feature analysis for 18 features. We train SVR with individual features and rank them based on Pearson's correlation values of each feature with the actual citation count for $\Delta t = 3$ years after publication in descending order. Table 6.16 reports ranked list of features at $\Delta t = 3$. We can observe from the table that the first six

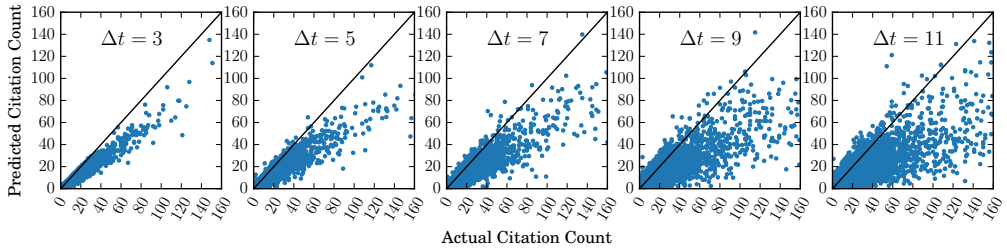


Figure 6.14: Change in prediction results over five time-periods. Scatter plots showing correlation between SVR predictions with real citation count values at $\Delta t = 3, 5, 7, 9, 11$. The black color line represents $y = x$ line passing through origin. Our model performs best for $\Delta T = 3$ with majority of the points on $y = x$ line. It performs worst for $\Delta T = 11$ with high divergence from the line. Our model under estimates $LTSI$ as majority of the points lie below the line. However, this prediction is considerably better than all the other baselines.

in the rank list consists of all the three EC features, indicating the importance of the EC features. As expected, early citation count is the most distinctive feature.

Table 6.16: Ranked list of features based on Pearson's correlation values between the predicted citation count and the actual citation count for $\Delta t = 3$ years after publication. Each SVR model is trained with individual feature.

1	ECC	6	ECCA	11	ACAR	16	PCN
2	ECCC	7	ACHI	12	ACP	17	ACV
3	ECPC	8	VCVR	13	PCTR	18	VCVC
4	VCPI	9	ACS	14	PR		
5	ACPI	10	PCD	15	ACA		

Figure 6.15 presents cross-correlation between features. Diagonal entries have maximum positive correlation (self) values = 1. Overall, features seem to be not much correlated with each other except a few cases. Interestingly, we observe that the EC features negatively correlate with the early citation count feature, the two being very distinct sources of information. Thus, including the EC features enhances the prediction performance significantly over and above the early citation count feature.

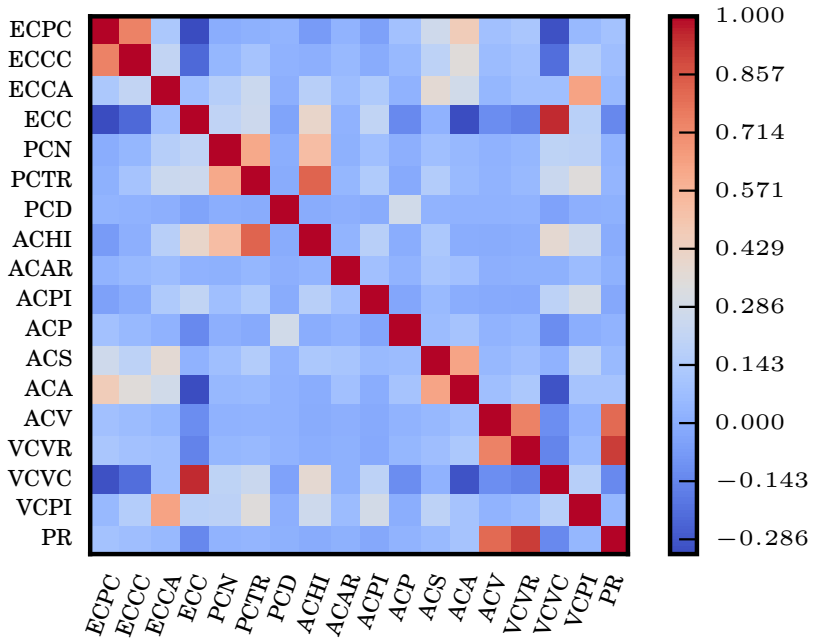


Figure 6.15: Cross correlation between features: Red color represents highly correlated features (=1). Blue represents uncorrelated to weakly negatively correlated features. Diagonal entries have maximum correlation (self) values = 1.

6.4 Online portal

We have also built an online portal to showcase the different results from our current work. Given a query paper present in our dataset, the portal displays different statistics related to the paper; in particular, each query result is accompanied by the statistics of the EC properties and other paper details. In addition, the portal also presents with a visualization comparing the actual and the predicted citation count of the paper. The current system is hosted on our research group server and can be accessed at `http://www.cnergres.iitkgp.ac.in/earlyciters/`.

6.5 Summary of the chapter

We present an interesting applications of curated scholarly knowledge. Next, we list specific contributions of this chapter:

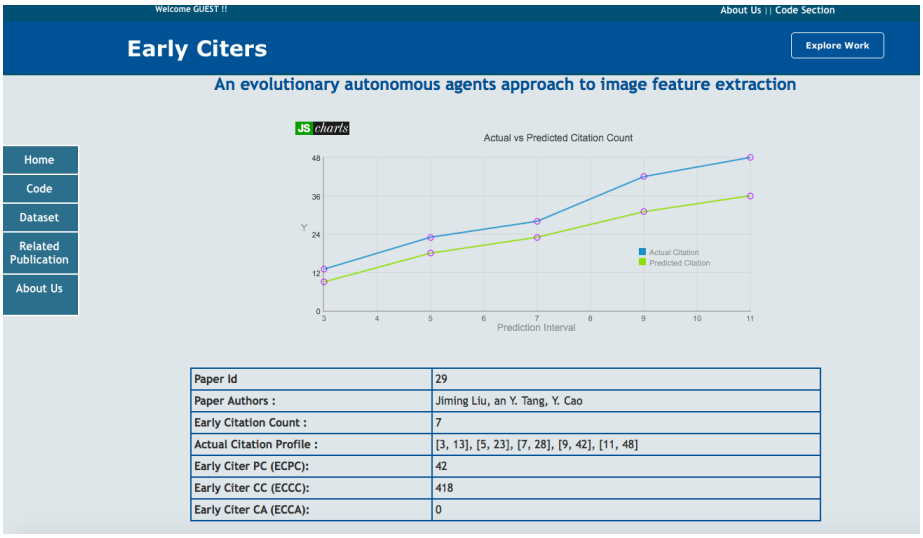


Figure 6.16: Snapshot of online portal. For input candidate paper, the portal presents visualization of prediction results along with EC statistics. It compares SVR predictions with real values at $\Delta t = 3, 5, 7, 9, 11$ years after publication.

1. We have used a massive dataset of citation contexts to show that the features extracted from the citation contexts of the papers, in the immediate years after publication, play a vital role for the task of future citation prediction.

2. We introduced two new features, average countX, and average citeWords, and feature analysis showed that these citation context features are highly correlated with the actual citation counts, specifically for the long-range citation prediction.

3. We have also been successfully able to provide empirical evidence that early citers (EC) play a significant role in determining the long-term scientific impact.

4. We empirically show that early citations might not be always beneficial; in particular early citations from influential EC negatively correlates with the long-term scientific impact of a paper. We have provided further evidence that the negative impact is more intense when EC is closer to the authors of the candidate article in the collaboration network.

To the best of our knowledge, this is the first work that attempts to use citation context-based features in the citation prediction problem. We use a massive

dataset of more than 26 million citation contexts from computer science research articles toward this goal.

CHAPTER 7

Conclusion and Future Work

In this chapter, we summarize important contributions of this thesis and finally wrap it up pointing to certain future research directions which this thesis has opened up.

7.1 Summary of Contribution

7.1.1 Knowledge extraction from scholarly articles

We develop *OCR++*, an open-source knowledge extraction frameworks for scientific articles. It extracts metadata, structural and bibliographic information from PDF research articles. We next summarize the contributions below –

- OCR++ performs a variety of information extraction tasks from scholarly articles including extraction of metadata (title, author names, affiliation, and e-mail) structure (section headings and body text, table and figure headings, URLs and footnotes) and bibliography (citation instances and references).

- We show that hand-written rules and heuristics produce better results than previously proposed machine learning models.

- Despite OCR errors and the great difference in the publishing formats, OCR++ outperforms the state-of-the-art systems with high margin, both in terms of accuracy (around 50% improvement) and processing time (around 52% improvement).

- A user experience study conducted with the help of 30 researchers reveals that the researchers found this system to be very helpful.
- OCR++ is online with the entire source code publicly available.

7.1.2 Mining performance comparisons to rank scholarly articles

We also develop another scholarly framework that robustly mines experimental performance from papers embedded within comparative tables. The key contributions are –

- We show that freely accessible academic search systems fail miserably at automatic leaderboard discovery and at finding state-of-the-art papers.
- To remedy this, we introduce performance tournament graphs that encode information about performance comparisons between scientific papers.
- We develop a robust, flexible, and domain-independent framework to extract these tournaments from tables with citations and performance numbers.
- We present a number of ways to aggregate the tournament edges and a number of ways to score and rank nodes on the basis of this incomplete and noisy information.
- The resulting scholarly search system is very simple to implement, but beats freely accessible search services by a large margin.

7.1.3 Modeling scientific growth through relay-linking phenomenon

Idealized network evolution models that explain entrenchment of prominence are abundant, but the only ones that model aging depend on post-hoc distribution-fitting (data collapse) and externality (fitness) parameters. We present the first plausible network-driven models for obsolescence in the context of research paper citations, based on a natural notion of *relay-linking*. The key contributions are –

1. We propose several measurements on evolving networks that constitute a temporal bucket signature summarizing the coexistence of entrenchment and obsolescence.

2. We show how graph structure can be utilized to predict where incoming citations to aging papers are likely to be redistributed.

3. We propose several novel and stringent tests for temporal fidelity of evolving, aging network models. Traditional aging models do not pass these tests well, but our relay-linking models do.

4. As an interesting application, we show that estimated turnover values negatively correlate with impact factor (IF10) for the four conference subsets we chose.

7.1.4 Estimating long-term scientific impact

As a final objective of this thesis, we leverage curated scientific knowledge to improve scientific impact prediction. In this regard, we contribute by utilizing early crowd-sourced textual and citers information to predict long term scientific impact of a paper. We utilize early information generated soon after publication to improve supervised machine learning frameworks for long-term scientific impact. The contributions can be summarized as:

1. We create a massive dataset consisting of more than 26 million citation contexts from computer science articles.

2. We present empirical evidence of a high correlation between two textual features from the citation contexts and three different characteristic properties of early citers with long-term citation counts of the paper.

3. We show that influential early citers have a negative impact while non-influential EC have a positive impact on a paper's long-term scientific impact. The negative impact is more intense when EC is closer to the authors of the candidate article in the collaboration network.

4. We then append these features along with various other features available at the time of publication in an earlier framework based on stratified learning [31] improving the prediction accuracy of state-of-the-art baselines with high margin.

7.2 Future direction

Here, we discuss some new research issues that have been opened up by this thesis.

7.2.1 Information extraction from scholarly articles

Some important future directions from this study are summarized as follows –

- One can aim to extend current frameworks by extracting information present in figures and tables. Figures and tables present concise statistics about dataset and results.
- OCR++ can be extended to extract and parse reference strings. Each reference string consists of set of fields such as author names, title, year of publication, venue name, volume, pages, organization, etc.
- Another possible extension would be to support the functionality for non-English articles.
- Another possibility is to develop a similar framework for patent articles.

7.2.2 Performance based scholarly ranking

Some important future directions from this study are summarized as follows –

- Performance information present in PDF articles can be extracted by utilizing several textual and image processing techniques.
- One can also aim to connect the relevant parts of the paper with comparative tables and charts.
- Another possible challenging extension can be performance extraction from textual paragraphs.

7.2.3 Modeling scientific growth through relay-linking phenomenon

Some important studies that can be taken up in future as a consequence of this work are summarized below –

1. The general applicability of relay-linking models may be investigated by extending it in the context of other citation networks, for example, legal precedence citation networks which consist of citation links between court cases, patent citation networks which consist of citation links between granted patents, etc.

2. Our proposed relay models do not consider area/author information which might be relevant in deciding the relay citation. An immediate future goal could be to introduce area/author-specific bias during link formation stages.

3. Future extensions could possibly lead to a formal analysis of properties of relay-linking or tractable variations. This would include deriving a relay-linking based mathematical formulation that governs the degree distribution.

4. Another possibility is to model inter-field citation evolution process.

7.2.4 Estimating long-term scientific impact

Some important future directions from this study are summarized as follows –

1. More features such as distribution of POS tags, sentiment score, etc., based on the textual analysis of citation contexts can be investigated.

2. Another possibility is to explore hedge count as a potential discriminating feature in the current study. Hedge identification is well researched problem in Biomedical domain [109, 65, 3].

3. The citation context may be further analyzed to obtain additional insights and properties such as common phrases, popular sentiment words, method and task identification within citation text, etc.

4. Another future direction would be to extend the current work to other scientific research fields.

5. Similar study can be performed in patent datasets to understand the effect of early citers on the overall innovation potential.

6. Another extension can lead to mathematical modeling of early citers' influence.

Many of the interesting problems in text can be solved by just looking into the data and inferring distinctive patterns, style, format, etc. We do not need complex ML models for every problem. We find extremely less applicability of AI in scientometry, thus, there is good scope in future. We should focus on better ranking schemes by looking into the textual data of research paper and not just by mere citation counts or publication age.

Bibliography

[1] O. Aalen, O. Borgan, and H. Gjessing. *Survival and event history analysis: a process point of view*. Springer, 2008.

[2] J. Adams. Early citation counts correlate with accumulated impact. *Scientometrics*, 63(3):567–581, 2005.

[3] S. Agarwal and H. Yu. Detecting hedge cues and their scope in biomedical text with conditional random fields. *Journal of biomedical informatics*, 43(6):953–961, 2010.

[4] C. C. Aggarwal and C. Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[5] R. A. Al-Zaidy and C. L. Giles. Automatic extraction of data from bar charts. In *Proceedings of the 8th International Conference on Knowledge Capture*, page 30. ACM, 2015.

[6] R. A. Al-Zaidy and C. L. Giles. Automatic extraction of data from bar charts. In *Proceedings of the 8th International Conference on Knowledge Capture*, K-CAP 2015, pages 30:1–30:4, 2015.

[7] R. A. Al-Zaidy and C. L. Giles. A machine learning approach for semantic structuring of scientific charts in scholarly documents. In *AAAI*, pages 4644–4649, 2017.

[8] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[9] B. Aljaber, N. Stokes, J. Bailey, and J. Pei. Document clustering of scientific texts using citation contexts. *Information Retrieval*, 13(2):101–131, 2010.

[10] F. F. Babatunde, B. A. Ojokoh, and S. A. Oluwadare. Automatic table recognition and extraction from heterogeneous documents. *Journal of Computer and Communications*, 3(12):100, 2015.

[11] E. Bacry, S. Gaïffas, I. Mastromatteo, and J.-F. Muzy. Mean-field inference of hawkes point processes. *arXiv/1511.01512*, 2015.

[12] M. Baker. Reproducibility crisis? *Nature*, 533:26, 2016.

[13] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[14] C. Basu, H. Hirsh, W. W. Cohen, and C. G. Nevill-Manning. Technical paper recommendation: A study in combining multiple information sources. *J. Artif. Intell. Res.(JAIR)*, 14:231–252, 2001.

[15] J. Beel, B. Gipp, S. Langer, M. Genzmehr, E. Wilde, A. Nürnberger, and J. Pitman. Introducing Mr. Dlib, a machine-readable digital library. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 463–464. ACM, 2011.

[16] J. Beel, B. Gipp, A. Shaker, and N. Friedrich. Sciplore xtract: Extracting titles from scientific pdf documents by analyzing style information (font size). In *Research and Advanced Technology for Digital Libraries*, pages 413–416. Springer, 2010.

[17] C. G. Begley and J. P. Ioannidis. Reproducibility in science: improving the standard for basic and preclinical research. *Circulation research*, 116(1):116–126, 2015.

[18] C. T. Bergstrom, J. D. West, and M. A. Wiseman. The eigenfactor? metrics. *The Journal of Neuroscience*, 28(45):11433–11434, 2008.

[19] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

[20] S. M. Bhattacharjee and F. Seno. A measure of data collapse for scaling. *J. Physics A: Mathematical and General*, 34(33):6375, 2001.

[21] T. Bogers and A. Van den Bosch. Recommending scientific articles using citeulike. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 287–290. ACM, 2008.

[22] L. Bornmann and H.-D. Daniel. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80, 2008.

[23] L. Bornmann, L. Leydesdorff, and J. Wang. Which percentile-based approach should be preferred for calculating normalized citation impact values? an empirical comparison of five approaches including a newly developed citation-rank approach (p100). *Journal of Informetrics*, 7(4):933–944, 2013.

[24] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

[25] T. Brody, S. Harnad, and L. Carr. Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060–1072, 2006.

[26] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. WebTables: exploring the power of tables on the Web. *PVLDB*, 1(1):538–549, 2008.

[27] M. Callaham, R. L. Wears, and E. Weber. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *Jama*, 287(21):2847–2850, 2002.

[28] A. C. Cameron and F. A. Windmeijer. An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2):329–342, 1997.

[29] C. Castillo, D. Donato, and A. Gionis. Estimating number of citations using author reputation. In *String processing and information retrieval*, pages 107–117. Springer, 2007.

[30] T. Chakraborty, A. Krishna, M. Singh, N. Ganguly, P. Goyal, and A. Mukherjee. FeRoSA: A faceted recommendation system for scientific articles. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 528–541. Springer, 2016.

[31] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee. Towards a stratified learning approach to predict future citation counts. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 351–360. IEEE Press, 2014.

[32] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee. Towards a stratified learning approach to predict future citation counts. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 351–360. IEEE Press, 2014.

[33] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee. On the categorization of scientific citation profiles in computer science. *Commun. ACM*, 58(9):82–90, Aug. 2015.

[34] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee. On the categorization of scientific citation profiles in computer sciences. *CoRR*, abs/1503.06268, 2015.

[35] S. Chatterjee et al. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[36] S. Z. Chen, M. J. Cafarella, and E. Adar. Searching for statistical diagrams. *Frontiers of Engineering, National Academy of Engineering*, pages 69–78, 2011.

[37] D. Chester and S. Elzer. Getting computers to see information graphics so users do not have to. In M.-S. Hacid, N. V. Murray, Z. W. Raś, and S. Tsumoto, editors, *Foundations of Intelligent Systems*, pages 660–668, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[38] J. Cho, S. Roy, and R. E. Adams. Page quality: In search of an unbiased Web ranking. In *SIGMOD conference*, pages 551–562. ACM, 2005.

[39] P. Cifariello, P. Ferragina, and M. Ponza. Wiser: A semantic approach for expert finding in academia based on entity linking. *arXiv preprint arXiv:1805.00150*, 2018.

[40] D. T. Citron and P. Ginsparg. Patterns of text reuse in a scientific corpus. *Proceedings of the National Academy of Sciences*, 112(1):25–30, 2015.

[41] M. Cliche, D. Rosenberg, D. Madeka, and C. Yee. Scatteract: Automated extraction of data from scatter plots. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 135–150. Springer, 2017.

[42] M. Cliche, D. Rosenberg, D. Madeka, and C. Yee. Scatteract: Automated extraction of data from scatter plots. In M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 135–150, 2017.

[43] A. Constantin, S. Pettifer, and A. Voronkov. Pdfx: fully-automated pdf-to-xml conversion of scientific literature. In *Proceedings of the 2013 ACM symposium on Document engineering*, pages 177–180. ACM, 2013.

[44] T. Cook. Archival science and postmodernism: new formulations for old concepts. *Archival science*, 1(1):3–24, 2001.

[45] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[46] E. Cortez, A. S. da Silva, M. A. Gonçalves, F. Mesquita, and E. S. de Moura. Flux-cim: Flexible unsupervised extraction of citation metadata. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07, pages 215–224, New York, NY, USA, 2007. ACM.

[47] I. G. Councill, C. L. Giles, and M. Kan. Parscit: an open-source CRF reference string parsing package. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, 2008.

[48] I. G. Councill, C. L. Giles, and M.-Y. Kan. ParsCit: an open-source CRF reference string parsing package. In *LREC*, 2008.

[49] F. Dalfovo, S. Giorgini, L. P. Pitaevskii, and S. Stringari. Theory of bose-einstein condensation in trapped gases. *Reviews of Modern Physics*, 71(3):463, 1999.

[50] H. A. David. Ranking from unbalanced paired-comparison data. *Biometrika*, 74(2):432–436, 1987.

[51] D. J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[52] O. DeMasi, K. Kording, and B. Recht. Meaningless comparisons lead to false optimism in medical machine learning. *PloS one*, 12(9):e0184604, 2017.

[53] A. Dengel and F. Dubiel. Clustering and classification of document structure-a machine learning approach. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 2, pages 587–591. IEEE, 1995.

[54] F. Didegah and M. Thelwall. Which factors help authors produce the highest impact research? collaboration, journal and document properties. *Journal of Informetrics*, 7(4):861–873, 2013.

[55] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks with aging of sites. *Physical Review E*, 62(2):1842, 2000.

[56] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[57] F. Esposito, D. Malerba, and G. Semeraro. A knowledge-based approach to the layout analysis. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 466–471. IEEE, 1995.

[58] J. A. Evans and J. Reimer. Open access and global participation in science. *Science*, 323(5917):1025–1025, 2009.

[59] J. Fan, A. Kalyanpur, D. C. Gondek, and D. A. Ferrucci. Automatic knowledge extraction from documents. *IBM Journal of Research and Development*, 56(3.4):5–1, 2012.

[60] M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song. COEVOLVE: A joint point process model for information diffusion and network co-evolution. *CoRR*, abs/1507.02293, 2015.

[61] D. G. Feitelson and U. Yovel. Predictive ranking of computer scientists using citeseer data. *Journal of Documentation*, 60(1):44–61, 2004.

[62] M. Franceschet. The difference between popularity and prestige in the sciences and in the social sciences: A bibliometric analysis. *Journal of Informetrics*, 4(1):55–63, 2010.

[63] L. D. Fu and C. Aliferis. Models for predicting and explaining citation count of biomedical articles. *PMC*, 2008:222–226, 2008.

[64] R. Gaizauskas, J. Foster, Y. Wilks, J. Arundel, P. Clough, and S. Piao. The meter corpus: a corpus for analysing journalistic text reuse. In *Proceedings of the corpus linguistics 2001 conference*, pages 214–223, 2001.

[65] V. Ganter and M. Strube. Finding hedges by chasing weasels: Hedge detection using wikipedia tags and shallow linguistic features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 173–176. Association for Computational Linguistics, 2009.

[66] W. Gardner. The electronic archive: Scientific publishing for the 1990s. *Psychological Science*, 1(6):333–341, 1990.

[67] E. Garfield. Journal impact factor: a brief review. *Canadian Medical Association Journal*, 161(8):979–980, 1999.

[68] E. Garfield. Citation indexes for science. a new dimension in documentation through association of ideas. *International journal of epidemiology*, 35(5):1123–1127, 2006.

[69] E. Garfield. The history and meaning of the journal impact factor. *JAMA*, 295(1):90–93, 2006.

[70] L. Getoor. Link mining: a new data mining challenge. *ACM SIGKDD Explorations Newsletter*, 5(1):84–89, 2003.

[71] M. Granitzer, M. Hristakeva, K. Jack, and R. Knight. A comparison of metadata extraction techniques for crowdsourced bibliographic metadata management. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 962–964. ACM, 2012.

[72] K. B. Hajra and P. Sen. Aging in citation networks. *Physica A: Statistical Mechanics and its Applications*, 346(1):44–48, 2005.

[73] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[74] D. Hallac, S. Vare, S. Boyd, and J. Leskovec. Toeplitz inverse covariance-based clustering of multivariate time series data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 215–223. ACM, 2017.

[75] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, E. Fox, et al. Automatic document metadata extraction using support vector machines. In *Digital Libraries, 2003. Proceedings. 2003 Joint Conference on*, pages 37–48. IEEE, 2003.

[76] S. Harnad and T. Brody. Comparing the impact of open access (oa) vs. non-oa articles in the same journals. *D-lib Magazine*, 10(6), 2004.

[77] S. Harnad, T. Brody, F. §. o. ValliÃ¨ res, L. Carr, S. Hitchcock, Y. Gingras, C. Oppenheim, H. Stamerjohanns, and E. R. Hilf. The access/impact problem and the green and gold roads to open access. *Serials review*, 30(4):310–314, 2004.

[78] H. Hashimoto, K. Shinoda, H. Yokono, and A. Aizawa. Automatic generation of review matrices as multi-document summarization of scientific papers. In *Workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)*, volume 7, pages 850–865, 2017.

[79] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles. Detecting topic evolution in scientific literature: How can citations help? In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 957–966. ACM, 2009.

[80] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430. ACM, 2010.

[81] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, pages 16569–16572, 2005.

[82] J. E. Hirsch and G. Buela-Casal. The meaning of the h-index. *International Journal of Clinical and Health Psychology*, 14(2):161–164, 2014.

[83] W. Huang, C. L. Tan, and W. K. Leow. Model-based chart image recognition. In *International Workshop on Graphics Recognition*, pages 87–99. Springer, 2003.

[84] M. Hurst and S. Douglas. Layout & language: Preliminary experiments in assigning logical structure to table cells. In *Proceedings of the fifth conference on Applied natural language processing*, pages 217–220. Association for Computational Linguistics, 1997.

[85] K. Hyland. *Hedging in scientific research articles*, volume 54. John Benjamins Publishing, 1998.

[86] A. Z. Jacobs and A. Clauset. A unified view of generative models for networks: models, methods, opportunities, and challenges. *arXiv preprint arXiv:1411.4070*, 2014.

[87] T. Jech. The ranking of incomplete tournaments: a mathematician's guide to popular sports. *The American Mathematical Monthly*, 90(4):246–266, 1983.

[88] H. Jeong, Z. Néda, and A.-L. Barabási. Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61(4):567, 2003.

[89] A. E. Jinha. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010.

[90] Y. Jo, J. E. Hopcroft, and C. Lagoze. The web of topics: Discovering the topology of topic evolution in a corpus. In *Proceedings of the 20th*

International Conference on World Wide Web, WWW '11, pages 257–266, New York, NY, USA, 2011. ACM.

[91] D. Jung, W. Kim, H. Song, J.-i. Hwang, B. Lee, B. Kim, and J. Seo. Chartsense: Interactive data extraction from chart images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 6706–6717, 2017.

[92] P. Katerattanakul, B. Han, and S. Hong. Objective quality ranking of computing journals. *Communications of the ACM*, 46(10):111–114, 2003.

[93] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, page 201424329, 2015.

[94] M. Khabsa and C. L. Giles. The number of scholarly documents on the public web. *PloS one*, 9(5):e93949, 2014.

[95] S. Kharazmi, F. Scholer, D. Vallet, and M. Sanderson. Examining additivity and weak baselines. *ACM Transactions on Information Systems (TOIS)*, 34(4):23, 2016.

[96] T. Kieninger and A. Dengel. Applying the t-recs table recognition system to the business letter domain. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pages 518–522. IEEE, 2001.

[97] T. G. Kieninger. Table structure recognition based on robust block segmentation. In *Document Recognition V*, volume 3305, pages 22–33. International Society for Optics and Photonics, 1998.

[98] A. V. Kulkarni, J. W. Busse, and I. Shams. Characteristics associated with citation rate of the medical literature. *PloS one*, 2(5):e403, 2007.

[99] A. V. Kulkarni, J. W. Busse, and I. Shams. Characteristics associated with citation rate of the medical literature. *PLoS ONE*, 2(5):e403, 05 2007.

[100] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Random graph models for the web graph. In *FOCS*, pages 57–65, 2000.

[101] C. Labbé. Ike antkare one of the great stars in the scientific firmament. *ISSI newsletter*, 6(2):48–52, 2010.

[102] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[103] S. Lawrence. Online or invisible. *Nature*, 411(6837):521, 2001.

[104] S. Lawrence and C. L. Giles. Searching the web: General and scientific information access. In *Internet Technologies and Services, 1999. Proceedings. First IEEE/Popov Workshop on*, pages 18–31. IEEE, 1999.

[105] J. Lee Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[106] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *SIGKDD Conference*, pages 462–470, 2008.

[107] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *SIGKDD Conference*, pages 177–187, 2005.

[108] L. Li and H. Tong. The child is father of the man: Foresee the success at the early stage. *arXiv preprint arXiv:1504.00948*, 2015.

[109] M. Light, X. Y. Qiu, and P. Srinivasan. The language of bioscience: Facts, speculations, and statements in between. In *HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*, 2004.

[110] M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp. Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 385–386. ACM, 2013.

[111] Y. Liu, K. Bai, P. Mitra, and C. L. Giles. Tableseer: Automatic table metadata extraction and searching in digital libraries. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07, pages 91–100, New York, NY, USA, 2007. ACM.

[112] Y.-Y. Liu, S. Li, F. Li, L. Song, and J. M. Rehg. Efficient learning of continuous-time hidden markov models for disease progression. In *NIPS*, pages 3599–3607, 2015.

[113] A. Livne, E. Adar, J. Teevan, and S. Dumais. Predicting citation counts using text and graph mining. In *Proc. the iConference 2013 Workshop on Computational Scientometrics: Theory and Applications*, 2013.

[114] C. Lokker, K. A. McKibbon, R. J. McKinlay, N. L. Wilczynski, and R. B. Haynes. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. *BMJ*, 336(7645):655–657, 2008.

[115] P. Lopez. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries*, pages 473–474. Springer, 2009.

[116] M.-T. Luong, T. D. Nguyen, and M.-Y. Kan. Logical structure recovery in scholarly articles with rich document features. *Multimedia Storage and Retrieval Innovations for Digital Library Systems*, page 270, 2012.

[117] C. M, W. RL, and W. E. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA*, 287(21):2847–2850, 2002.

[118] D. M. Markowitz and J. T. Hancock. Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology*, 35(4):435–445, 2016.

[119] A. McCallum, K. Nigam, and L. H. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–178. ACM, 2000.

[120] J. Mingers. Exploring the dynamics of journal citations: modelling with s-curves. *Journal of the Operational Research Society*, 59(8):1013–1025, 2008.

[121] P. Mitra, C. L. Giles, J. Z. Wang, and X. Lu. Automatic categorization of figures in scientific documents. In *Digital Libraries, 2006. JCDL'06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on*, pages 129–138. IEEE, 2006.

[122] S. Mohammad, B. Dorr, M. Egan, A. Hassan, P. Muthukrishan, V. Qazvinian, D. Radev, and D. Zajic. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 584–592. Association for Computational Linguistics, 2009.

[123] R. Mohemad, A. R. Hamdan, Z. A. Othman, and N. M. Noor. Automatic document structure analysis of structured pdf files. *International Journal of New Computer Architectures and their Applications (IJNCAA)*, 1(2):404–411, 2011.

[124] G. Nagy, S. Seth, and M. Viswanathan. A prototype document image analysis system for technical journals. *Computer*, 25(7):10–22, July 1992.

[125] H. T. Ng, C. Y. Lim, and J. L. T. Koo. Learning to recognize tables in free text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 443–450. Association for Computational Linguistics, 1999.

[126] K. Nishida, K. Sadamitsu, R. Higashinaka, and Y. Matsuo. Understanding the semantic structures of tables with a hybrid deep neural network architecture. In *AAAI*, pages 168–174, 2017.

[127] P. Orbanz and D. M. Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):437–461, 2015.

[128] E. Oro and M. Ruffolo. Trex: An approach for recognizing and extracting tables from pdf documents. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 906–910. IEEE, 2009.

[129] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Manuscript, Stanford University, 1998.

[130] S. Pandey, S. Roy, C. Olston, J. Cho, and S. Chakrabarti. Shuffling a stacked deck: the case for partially randomized ranking of search engine results. In *VLDB conference*, pages 781–792, 2005.

[131] P. D. B. Parolo, R. K. Pan, R. Ghosh, B. A. Huberman, K. Kaski, and S. Fortunato. Attention decay in science. *Journal of Informetrics*, 9(4):734 – 745, 2015.

[132] D. Parra-Santander and P. Brusilovsky. Improving collaborative filtering in social tagging systems for the recommendation of scientific articles. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 136–142. IEEE, 2010.

[133] R. D. Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

[134] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles. Winners don't take all: Characterizing the competition for links on the web. *PNAS*, 99(8):5207–5211, 2002.

[135] D. Pinto, A. McCallum, X. Wei, and W. B. Croft. Table extraction using conditional random fields. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 235–242. ACM, 2003.

[136] N. Pobiedina and R. Ichise. Predicting citation counts for academic literature using graph pattern mining. In *Modern Advances in Applied Intelligence*, pages 109–119. Springer, 2014.

[137] J. Polchinski, S. Chaudhuri, and C. V. Johnson. Notes on d-branes. *arXiv preprint hep-th/9602052*, 1996.

[138] V. Prabhakaran, W. L. Hamilton, D. McFarland, and D. Jurafsky. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1170–1180, 2016.

[139] D. D. S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

[140] P. Pyreddy and W. B. Croft. Tintin: A system for retrieval in text tables. In *Proceedings of the second ACM international conference on Digital libraries*, pages 193–200. ACM, 1997.

[141] V. Qazvinian and D. R. Radev. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 689–696. Association for Computational Linguistics, 2008.

[142] D. R. Radev, E. Hovy, and K. McKeown. Introduction to the special issue on summarization. *Computational linguistics*, 28(4):399–408, 2002.

[143] C. E. Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[144] C. Redmond. A natural generalization of the win-loss rating system. *Mathematics Magazine*, 76(2):119–126, 2003.

[145] F. Ronzano and H. Saggion. Knowledge extraction and modeling from scientific publications. In *International Workshop on Semantic, Analytics, Visualization*, pages 11–25. Springer, 2016.

[146] S. Ross. Digital preservation, archival science and methodological foundations for digital libraries. *New Review of Information Networking*, 17(1):43–68, 2012.

[147] S. Sarawagi and S. Chakrabarti. Open-domain quantity queries on Web tables: Annotation, response, and consensus models. In *SIGKDD Conference*, 2014.

[148] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer. Revision: Automated classification, analysis and redesign of chart images. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 393–402. ACM, 2011.

[149] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer. Revision: Automated classification, analysis and redesign of chart images. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 393–402, 2011.

[150] H. Sayyadi and L. Getoor. Futurerank: Ranking scientific articles by predicting their future pagerank. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 533–544. SIAM, 2009.

[151] M. N. Schmidt and M. Morup. Nonparametric bayesian modeling of complex networks: An introduction. *IEEE Signal Processing Magazine*, 30(3):110–128, 2013.

[152] J. M. Schwartz and T. Cook. Archives, records, and power: The making of modern memory. *Archival science*, 2(1-2):1–19, 2002.

[153] P. K. Shah, C. Perez-Iratxeta, P. Bork, and M. A. Andrade. Information extraction from full text scientific articles: where are the keywords? *BMC bioinformatics*, 4(1):20, 2003.

[154] J. H. Shamilian, H. S. Baird, and T. L. Wood. A retargetable table reader. In *Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on*, volume 1, pages 158–163. IEEE, 1997.

[155] M. Singh, V. Patidar, S. Kumar, T. Chakraborty, A. Mukherjee, and P. Goyal. The role of citation context in predicting long-term citation

profiles: An experimental study based on a massive bibliographic text dataset. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1271–1280. ACM, 2015.

[156] M. Singh, R. Sarkar, P. Goyal, A. Mukherjee, and S. Chakrabarti. Relay-linking models for prominence and obsolescence in evolving networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1077–1086, New York, NY, USA, 2017. ACM.

[157] A. Smola and V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

[158] C. Stegehuis, N. Litvak, and L. Waltman. Predicting the long-term citation impact of recent publications. *Journal of informetrics*, 9(3):642–657, 2015.

[159] D. I. Stern. High-ranked social science journal articles can be identified from early citation information. *PLOS ONE*, 9:1–11, 11 2014.

[160] K. Sugiyama and M.-Y. Kan. Scholarly paper recommendation via user's recent research interests. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 29–38. ACM, 2010.

[161] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 990–998, New York, NY, USA, 2008. ACM.

[162] X. Tang, X. Wan, and X. Zhang. Cross-language context-aware citation recommendation in scientific articles. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 817–826. ACM, 2014.

[163] S. Teufel and M. Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

[164] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, and Ł. Bolikowski. Cermine: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(4):317–335, 2015.

[165] A. Vazquez. Disordered networks generated by recursive searches. *Europhysics Letters*, 54(4):430–435, 2001.

[166] A. Verstak, A. Acharya, H. Suzuki, S. Henderson, M. Iakhiaev, C. C. Lin, and N. Shetty. On the shoulders of giants: The growing impact of older articles. *CoRR*, abs/1411.0275, 2014.

[167] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456. ACM, 2011.

[168] D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[169] D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[170] J. Wang. Citation time window choice for research impact evaluation. *Scientometrics*, 94(3):851–872, 2013.

[171] M. Wang, G. Yu, and D. Yu. Effect of the age of papers on the preferential attachment in citation networks. *Physica A: Statistical Mechanics and its Applications*, 388(19):4273 – 4276, 2009.

[172] Y. Wang and J. Hu. A machine learning based approach for table detection on the web. In *Proceedings of the 11th international conference on World Wide Web*, pages 242–250. ACM, 2002.

[173] Y. Wang, I. T. Phillips, and R. M. Haralick. Table structure understanding and its performance evaluation. *Pattern recognition*, 37(7):1479–1497, 2004.

[174] Y. Wang, Y. Tong, and M. Zeng. Ranking scientific articles by exploiting citations, authors, journals, and time information. In *AAAI*, 2013.

[175] M. C. Waumans and H. Bersini. Genealogical trees of scientific papers. *PloS one*, 11(3):e0150588, 2016.

[176] M. C. Waumans and H. Bersini. Genealogical trees of scientific papers. *PLOS ONE*, 11(3):1–15, 03 2016.

[177] J. Willinsky. *The access principle: The case for open access to research and scholarship*. Cambridge, Mass.: MIT Press, 2006.

[178] J. Wu, J. Killian, H. Yang, K. Williams, S. R. Choudhury, S. Tuarob, C. Caragea, and C. L. Giles. Pdfmef: A multi-entity knowledge extraction framework for scholarly documents and semantic search. In *Proceedings of the 8th International Conference on Knowledge Capture*, K-CAP 2015, pages 13:1–13:8, New York, NY, USA, 2015. ACM.

[179] S. Xiao, J. Yan, C. Li, B. Jin, X. Wang, X. Yang, S. M. Chu, and H. Zha. On modeling and predicting individual paper citation count over time. In *IJCAI*, pages 2676–2682, 2016.

[180] Z. Xie, Z. Ouyang, P. Zhang, D. Yi, and D. Kong. Modeling the citation network by network cosmology. *PloS one*, 10(3):e0120687, 2015.

[181] W. Xing and A. Ghorbani. Weighted pagerank algorithm. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 305–314. IEEE, 2004.

[182] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li. To better stand on the shoulder of giants. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 51–60. ACM, 2012.

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

[184] B. Yildiz, K. Kaiser, and S. Miksch. pdf2table: A method to extract table information from pdf files. In *IICAI*, pages 1773–1785, 2005.

[185] X. Yu, Q. Gu, M. Zhou, and J. Han. Citation prediction in heterogeneous bibliographic networks. In *SDM*, volume 12, pages 1119–1130. SIAM, 2012.

[186] H. Zhu, X. Wang, and J.-Y. Zhu. Effect of aging on network structure. *Physical Review E*, 68(5):056121, 2003.

[187] Z. Zuo, G. Wang, B. Shuai, L. Zhao, and Q. Yang. Exemplar based deep discriminative and shareable feature learning for scene image classification. *Pattern Recognition*, 48(10):3004–3015, 2015.

Dissemination of the work

Following is a list of all the publications by the candidate including those on and related to the work presented in the thesis. The publications are arranged in chronological order and in three sections – (i) conferences, (ii) journals, and (iii) workshops/research colloquiums/Magazines.

Conferences

1. **Mayank Singh**, Rajdeep Sarkar, Atharva Vyas, Pawan Goyal, Animesh Mukherjee, and Soumen Chakrabarti. "Automated Early Leaderboard Generation From Comparative Tables". In Proceedings of the 41st European Conference on Information Retrieval (ECIR), 2019.

2. **Mayank Singh**, Pradeep Dogga, Sohan Patro, Dhiraj Barnwal, Ritam Dutt, Rajarshi Haldar, Pawan Goyal, and Animesh Mukherjee. "CL Scholar: The ACL Anthology Knowledge Graph Miner". In Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2018.

3. **Mayank Singh**, Rajdeep Sarkar, Pawan Goyal, Animesh Mukherjee, and Soumen Chakrabarti. "Relay-Linking Models for Prominence and Obsolescence in Evolving Networks". In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 1077–1086. ACM, 2017.

4. **Mayank Singh**, Ajay Jaiswal, Priya Shree, Arindam Pal, Animesh Mukherjee, and Pawan Goyal. "Understanding the Impact of Early Citers on Long-Term Scientific Impact". In Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 1-10. ACM, 2017.

5. **Mayank Singh**, Abhishek Niranjan, Divyansh Gupta, Nikhil Angad Bakshi, Animesh Mukherjee, and Pawan Goyal. "Citation Sentence Reuse Behavior of Scientists: A Case Study on Massive Bibliographic Text Dataset

of Computer Science". In Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 277-280. IEEE Press, 2017.

6. **Mayank Singh**, Barnopriyo Barua, Priyank Palod, Manvi Garg, Sidhartha Satapathy, Samuel Bushi, Kumar Ayush, Krishna Sai Rohith, Tulasi Gamidi, Pawan Goyal, and Animesh Mukherjee "OCR++: A Robust Framework for Information Extraction from Scholarly Articles". In the 26th International Conference on Computational Linguistics (COLING), pp. 3390–3400. 2016.

7. Tanmoy Chakraborty, Amrith Krishna, **Mayank Singh**, Niloy Ganguly, Pawan Goyal, and Animesh Mukherjee. "Ferosa: A faceted recommendation system for scientific articles". In Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp. 528-541. Springer, Cham, 2016.

8. **Mayank Singh**, Vikas Patidar, Suhansanu Kumar, Tanmoy Chakraborty, Animesh Mukherjee, and Pawan Goyal. "The Role of Citation Context in Predicting Long-Term Citation Profiles: An Experimental Study Based on a Massive Bibliographic Text Dataset". In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM), pp. 1271-1280. ACM, 2015.

9. **Mayank Singh**, Tanmoy Chakraborty, Animesh Mukherjee, and Pawan Goyal. "ConfAssist: A Conflict Resolution Framework for Assisting the Categorization of Computer Science Conferences". In Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), pp. 257-258. ACM, 2015.

Journals

1. **Mayank Singh**, Tanmoy Chakraborty, Animesh Mukherjee, and Pawan Goyal. "Is this conference a top-tier? ConfAssist: An assistive conflict resolution framework for conference categorization." Journal of Informetrics (JOI) 10(4), pp. 1005-1022, 2016.

Workshops/Research Colloquiums/Magazines

1. **Mayank Singh**. "Understanding Popularity of Academic Entities: From Papers to Authors to Venues". IEEE-TCDL 2018

2. **Mayank Singh**, Soham Dan, Sanyam Agarwal, Pawan Goyal, and Animesh Mukherjee. "AppTechMiner: Mining Applications and Techniques from Scientific Articles". In Proceedings of the 6th International Workshop on Mining Scientific Publications, pp. 1–8. ACM, 2017.

3. **Mayank Singh**, Soumajit Pramanik, and Tanmoy Chakraborty. "PubIndia: A Framework for Analyzing Indian Research Publications in Computer Science". D-Lib Magazine 21, no. 11/12, 2015.