CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Predicting Scientific Impact of Authors

by

Samreen Ayaz

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Computing
Department of Computer Science

2022

Predicting Scientific Impact of Authors

By

Samreen Ayaz

(DCS143006)

Dr. Fei Teng, Associate Professor

Southwest Jiaotong University, China

(Foreign Evaluator 1)

Dr. Irfan Ullah Awan, Professor

University of Bradfod, UK

(Foreign Evaluator 2)

Dr. Nayyer Masood

(Thesis Supervisor)

Dr. Nayyer Masood

(Head, Department of Computer Science)

Dr. Muhammad Abdul Qadir

(Dean, Faculty of Computing)

DEPARTMENT OF COMPUTER SCIENCE

CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

ISLAMABAD

2022

Dedicated to my parents & parents-in-law .

CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY ISLAMABAD

Expressway, Kahuta Road, Zone-V, Islamabad
Phone:+92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: https://www.cust.edu.pk

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the thesis, entitled "**Predicting Scientific Impact of Authors**" was conducted under the supervision of **Dr. Nayyer Masood**. No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the **Department of Computer Science, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Computer Science**. The open defence of the thesis was conducted on **December 16, 2021**.

Student Name : Samreen Ayaz (DCS-143006)

The Examining Committee unanimously agrees to award PhD degree in the mentioned field.

Examination Committee :

(a) External Examiner 1: Dr. Sharifullah Khan TI,
Professor
PAF-IAST, Haripur

(b) External Examiner 2: Dr. Waseem Shahzad,
Professor
FAST-NUCES, Islamabad

(c) Internal Examiner : Dr. Abdul Basit Siddiqui
Associate Professor
CUST, Islamabad

Supervisor Name : Dr. Nayyer Masood
Professor
CUST, Islamabad

Name of HoD : Dr. Nayyer Masood
Professor
CUST, Islamabad

Name of Dean : Dr. Muhammad Abdul Qadir
Professor
CUST, Islamabad

AUTHOR'S DECLARATION

I, **Samreen Ayaz (Registration No. DCS-143006)**, hereby state that my PhD thesis entitled, '**Predicting Scientific Impact of Authors**' is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my PhD Degree.

(Samreen Ayaz)

Dated: December, 2021

Registration No : DCS-143006

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled "**Predicting Scientific Impact of Authors**" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD degree and that HEC and the University have the right to publish my name on the HEC/ University Website on which names of students are placed who submitted plagiarized thesis.

(Samreen Ayaz)

Dated: December, 2021

Registration No : DCS143006

List of Publications

It is certified that following publication(s) have been made out of the research work that has been carried out for this thesis:-

1. S. Ayaz and N. Masood, "Comparison of researchers' impact indices," PloS one, vol. 15, no. 5, p. e0233765, 2020.
2. S. Ayaz, N. Masood, and M. A. Islam, "Predicting scientific impact based on h-index," Scientometrics, vol. 114, no. 3, pp. 993–1010, 2018.

Samreen Ayaz

(DCS143006)

Acknowledgement

All praise be to Allah Almighty, the most merciful, the most beneficent, who enabled me to acquire whatever knowledge that I have and to complete this degree.

I would like to express my deepest gratitude to my wonderful supervisor **Dr. Nayyer Masood** for his invaluable contribution and advice in undertaking this project. I owe a lot of thanks to him, he was always very kind and ready to guide me. Without his able guidance, this thesis would not have been possible and I shall eternally be grateful to him for his support.

I must also acknowledge the help that I got from Dr.Arshad Islam and Dr. Muhammad Tanvir Afzal, who were always prepared to guide me in achieving the knowledge that was required to complete this project. I have to offer my gratitude to Dr. Muhammad Abdul Qadir and all my teachers who have always been a guiding light for me in achieving the knowledge and completing the project. I also want to thank Mr. Adil-ur-Rehman and Ms. Raabia Mumtaz who were always helpful to me.

I am thankful to my parents, in laws, my husband, my children, siblings, colleagues and friends who extended whatever help I needed from them.

Abstract

Predicting the future impact of a researcher is a critical task, as it can be helpful in making many decisions, like, in identifying potential candidates for research grants, for job recruitment, for promotions etc. One of the key metrics used for evaluating the impact of a researcher is h-index which inherently is a field-specific metric. Prediction models of h-index for different fields have been proposed in literature. However, these models are developed for specific field, their performance is not evaluated for multiple fields. Considering the citations and other factors' variations across different fields, there may be a possibility that same model behaves differently for other field. There can be a need to apply h-index prediction model for different fields. For example, to compare two researchers from different fields and applying for the same job. As per existing approaches the future impact of researchers would be compared using different models. There is a gap to establish a model that performs well across multiple disciplines. Moreover, existing prediction models do not perform well for young researchers, i.e., researchers with low h-index or with less experience. So young researchers are excluded from experiments of prediction models in literature. These two research gaps have been addressed in this study, i.e, prediction model is proposed for the field of Computer Science, tested for the field of Physics, and evaluated for young researchers as well. We have considered several features of fundamental importance to authors that include existing feature from literature like average citations, number of publications, and we have also defined new features like citations in impact factor journals, average h-index of all the coauthors. We have used these features to predict next five years future impact of researchers. Machine learning techniques such as regression and Neural Network, are used to find the best set of parameters suitable for h-index prediction for the scientists from all career ages. R^2 and RMSE are used as performance metrics to measure the accuracy. Experimental results on a large data set of ArnetMiner achieved up to 97% R^2 and 0.27 RMSE for one year. Similarly, 90% R^2 for five years with 0.60 RMSE. Models proposed for the field of Computer Science are further evaluated for the field of Physics, on the data set acquired from Open Academic Graph (OAG). The proposed model exhibits

reasonably good results for the field of Physics as well i.e., 86% (R^2) predictive performance for one year and 66% (R^2) for five years with 0.15 (RMSE) for one year and 0.29 (RMSE) for five years. However, performance of the proposed models is not satisfactory for young researchers, R^2 for young researchers is 67% for one year and 55% for five years, which is very low as compared to full data set evaluation values. This poses a challenge for impact prediction of young researchers. Therefore, to tackle this challenge of Impact evaluation of young researcher's, a new measure 'NS-Index' is proposed in this study. According to our findings the proposed index performs well in identifying future impact of young researchers. Our experiments conclude that NS-Index for young researcher is a better reflection of their future performance up to three years. However, to predict the performance of young researchers for more than 3 years our proposed h-index prediction model performs better.

Contents

List of Figures

List of Tables

Abbreviations

IF	Impact Factor
MAPE	Mean Absolute Percentage Error
MSE	Mean Squared Error
RMSE	Root Mean Square Error

Chapter 1

Introduction

Researchers contribute to the frontiers of knowledge by establishing facts and reaching new conclusions through systematic investigations, and by subsequently publishing the outcomes of their research findings in the form of research publications. These research publications are indicative of researchers' scientific impact. Different bibliometric indices have been proposed to measure the impact or productivity of a researcher. These indices include publication count, citation count, number of coauthors, h-index, etc. The h-index, since its inception, has been ranked as the foremost impact indicator, the most commonly used and established measure to evaluate the impact of individual researchers on scientific literature [1]. İn 2005, a physicist Hirsch, J. E. proposed h-index to quantify the scientific impact of researchers [2]. Hirsch has discussed number of benefits of h-index over other bibliometric indicators including number of publications, number of citations, citations per paper, number of significant papers. According to Hirsch the above-mentioned indicators focus on one aspect, i.e. quality or quantity. Whereas h-index complements both impact/quality and quantity of publications. It combines the effect of two dimensions i.e. number of publications which represent the productive core of a scientist and the number of citations, representing the impact of that core. H-index is defined as "A scientist has index h if h of his/her Np papers have at least h citations each and the other (Np – h) papers have $\leq$ h citations each" [2]. h-index brought a revolution in the field of researcher's evaluation and bibliometrics. World has adopted it instantly and nowadays it is one of the most notable evaluation criteria for researchers. There are a number of studies which have used h-index for scientists' evaluation [3–5]. The introduction of h-index has also initiated a new research front; addressing and assessing advantages and disadvantages of h-index. h-index is criticized in literature for its shortcomings, like

- h-index is not capable of comparing scientists from different domains i.e. It is field dependent.
- It is influenced by duration of scientists career.
- It relies on electronic databases.
- h-index is influenced by self citations.
- It is not justifiable to use only one figure to reflect whole career of a scientist. [6–8]

These shortcomings resulted in a number of extensions and variants of h-index and definition of new indices[9–13] . Many researchers have compared and evaluated different variants/extensions of h-index and other bibliometric indicators [14–18] for different fields [19][20] and have found positive correlation among them [15][21]. Still h-index is the most widely used measure for researchers' evaluation.

The bibliometric indicators like h-index, number of citations etc. are quite useful for evaluation purposes, like to decide who should be given tenure, promotion or funding or who should be appointed for a certain task etc[22][23]. But they are more effective and helpful if we can use them to predict future impact of researchers [24]. For example, lets suppose a university has hired two persons/researchers, say person A and person B, having same h-index. Person B has not done any research work recently, while person A is an active researcher. After 5 years the h-index of person A increased by 4 and person B's h-index remained the same. It demonstrates that the person whose h-index increased i.e. person A, has contributed more to the research community, has benefitted his students further and has added more publications to the credit of the university as compared to person B. It means person A has contributed more in raising the ranking of the university. The scenario presented above shows that if there would be any mechanism to predict or assess any persons' h-index in advance, the university would be in a better position to hire the most suitable person, especially for a research oriented job.

In order to better assess/evaluate a researcher, it is very important to predict researcher's future scientific success i.e. there should be some procedures/ tools to support such decisions. With the growing number of applicants for tenure track or funding etc., there is a need to have combination of different parameters for better prediction [25][26]. Keeping in view the importance of predicting future performance for better current assessment, Hirsch also asserted that the h-index can be used quite effectively for its own prediction. Hirsch proved empirically that h-index has better predictive power than other bibliometric indicators including number of citations, number of papers and mean citations per paper[2]. According to Hirsch's findings, besides itself h-index was also found to be better predictor of number of publications. His claim that h-index is good predictor of itself is further supported by [27]. Different authors have explored the predictive power of h-index [28][29] and proposed different combinations of parameters for this purpose.

Some of the prominent techniques in the area have been briefly described in the coming section, followed by statement of our problem and proposed research questions. Significance of the proposed work is elaborated in objectives and significance section.

1.1 Impact Prediction

Researcher's evaluation is important in making many decisions like hiring, giving tenure etc. These decisions affect the institutions/ universities rankings, their functionalities and also have an impact on their future. For example, a faculty member on tenure track will receive a handsome amount and will occupy faculty position for a long time. Similarly other hiring's or projects affect enterprise performance and repute. So it is of an utmost importance to know the future scientific impact of applicants/researchers.

Hirsch has shown that h-index effectively represents a researchers' overall contribution. A good researcher should have high productivity and high impact, this rudimentary belief is well supported/well covered by h-index. h-index brings a balance while using two different dimensions. Hence despite the difference in number of publications or number of citations , researchers having same value of h-index are comparable. Being a good representative of scientific impact, predicting h-index can play important role in identifying potential candidates[2][30]. In this regard different studies [24][31][32] have been conducted for different fields using different data sets and techniques. Initially Acuna et al. proposed a model, predicting future h-index for the researchers from Life Sciences field[24]. This work is followed by application of proposed model on some other data set [33], the studies exposed the limited validity of model for different data set. Furthermore it was identified that performance of model declines for the very young researchers [34].

In this regard different parameters are considered depicting achievement of researchers from different perspectives. Combination of different parameters are considered to predict scientific impact of a researcher. These parameters are based primarily on citations, publications, venue and coauthors. All these parameters have significant importance in researcher's career. Publications having high citation count usually depicts their importance, quality and contribution to the field. Similarly a researcher having constantly high productivity over the years shows his commitment, devotion and consistency towards research. Coauthors having high profile impacts the researcher's reputation and profile as well. Publications in prestigious journals naturally demonstrate the quality of work along with this confidence that it would have high readership which eventually would have resulted in high citations. publications at different venues depicts the caliber, adaptability and versatility of a researcher. Different venues also have a benefit of diversity in audience. All these factors affect the scientific impact of a researcher and ultimately h-index of researcher increases with the increasing values of all these considerations.

Further to measure the impact of these parameters, different machine learning and

Table 1.1: H-index Prediction Related Studies and Considered Fields

Study	**Field**	**Data Source**	**Researchers**
Ibanez etal., 2011[36]	Spanish University Faculty members of Computer Science	Web of Science	
Acuna, Allesina and Kording, 2012[24]	Neurosurgery, Drosophila and Evolutionary	Scientists Academic Tree (http://www.academictree.org),Scopus (http://www.scopus.com)	Excluded young researchers
McCarty etal., 2013[35]	Different fields	Random sample from Web of Science	
Dong, Reid and Nitesh 2016[31]	Computer Science	Arnetminer	Excluded young researchers
Weihs and Etzioni, 2017[37]	Computer Science		Excluded young researchers
Mistele,Price and Hossenfelder, 2019[38]	Physics	Arxiv data set	Excluded Young researchers
Nikolentzos etal., 2021[39]	Computer Science	Microsoft Academic Graph	

statistical techniques are applied in literature. For this purpose, mostly used technique is regression. Acuna et al. (2012) considered 18 factors and finally left with 5 factors applying regression and using R-squared as an evaluation metric[24]. [31][32][35] and most of the studies have used regression. Moreover, R^2 is used mostly for evaluation purposes. Other evaluation metrics which are used include MAPE, RMSE and accuracy. Models for prediction of h-index is proposed for different fields. Table 1.1 shows the studies along with the field for which study was conducted and source of data. As shown in Table 1.1 , most of the studies have excluded young researchers. It is hard to evaluate/predict the performance of a researcher/person who has just joined the research community as compared to a person having achieved some milestones, having a number of publications and having feedback from research community (in the form of citations). Evaluation of a researcher is concerned with future impact of a researcher, but it relies on impact of previous work. Whereas a young researcher has still many horizons to explore, he has to achieve many milestones. In young researcher's case, the challenge is to predict the future impact with very limited information available. Current scientific impact of young researcher is also not very high usually, having low value for h-index. There should be some other mechanism to define the scientific impact for young researchers. There is dire need of further investigations in defining the scientific impact of young researchers.

Moreover h-index is believed to be field dependent, one model proposed for a field is not supposed to be applicable on some other field. Different research fields differ in the average number of references per paper and the average number of papers published by each researcher. Researchers from one area may have less number of publications but with a reasonably higher impact in their area/field. Similarly, according to [40], it is known that in absolute figures number of citations in the field of Mathematics is less than in Chemistry, but a Mathematician with a relatively low total number of citations can have higher impact in Mathematics than a Chemist with a larger number of citations in Chemistry. Collectively it is found in literature that there are significant differences in citation, number of coauthors or collaborations characteristics between different scientific fields [40–42]. Hirsch [2] also stated that there will be large differences in typical h-values in different fields. This is the reason that studies related to h-index prediction are field specific. In our study, we have explored the combinations already presented in literature and also some variations in the combinations. Current approaches work with imposing some constraints like considering specific career ages or placing limits on current h-index values. These constraints limit the type of researches whose h-index can be predicted. We intend to relax these constraints like we plan to consider whole career ages and all the authors having any h-index value. We also intend to explore the effect of different career ages and different threshold values of h-index on the prediction. For our experiments, we are considering a comprehensive data set for the field of Computer Science from Arnetminer[1].

We confess that h-index is only one of the many factors that can be helpful in determining the performance and forecasting the success in career of a computer scientist. Other factors like number of publications, citations, technical expertise etc. are also important. Still we assert and believe that ability to publish and be cited by as many researchers is the most vital factor in evaluating researchers. As h-index is combination of both these traits hence it can be considered as the most important factor in evaluation. After having studied/exploring the literature and also considering the fact that current approaches work with imposing some constraints like career ages, h-index etc. we have formulated our problem statement as follows:

1.2 Problem Statement

Many approaches for the prediction of h-index exist in literature. These approaches apply different machine learning models, use data sets from a specific field and select different features sets. The existing methods, however, predict h-index only for those researchers who have some research experience and certain h-index. Even after applying these constraints on data sets, most models exhibit poor performance in predicting h-index as the target year moves farther. Moreover, the performance of these models is not generally computed across multiple domains. There is a need to develop/propose a model for h-index prediction for a specific field, like Computer Science, and also to check its applicability on some other field. There is also a need to devise some mechanism to predict the performance of young researchers.

[1] https://cn.aminer.org/billboard/aminernetwork

1.3 Research Questions

To find the best set of parameters suitable for h-index prediction for the scientists from all career ages and without enforcing any constraint on their current h-index values for the field of Computer Science. Further evaluating these parameters for young researchers and validating its applicability for another field. This research is intended to explore the following research questions, detail discussion on these research questions is given in methodology :

Research Question 1 (RQ1): Which of the existing sets of parameters/Models for h-index prediction performs better when applied on the comprehensive data set from the field of Computer Science?

Research Question 2 (RQ2): What would be the best set of parameters suitable for the prediction of future h-index for the researchers from the field of Computer Science?

Research Question 3 (RQ3): What would be the suitable set of parameters for better h-index prediction for young researchers?

Research Question 4 (RQ4): What is the performance of proposed (in RQ2) model when applied on a data set other than Computer Science?

1.4 Objectives and Significance

To evaluate the performance of scientists/researchers the most commonly used parameter is h-index. This evaluation is helpful in many ways:

- in identifying the most influential scientist in any field
- in deciding who should get tenure
- in identifying the most suitable candidate for any funding or grant

- to decide who should be given promotion.
- for the students to get help/guidance in finding the most suitable supervisor
- for universities to hire/find the right person.

All of these activities have impact on the future of hiring organizations. Hence it is of utmost significance to evaluate the applicant/candidate not only on the basis of his/her previous accomplishments, rather there should be some mechanism to predict his/her future achievements. As hiring organization would be most affected by the performance of candidate in future and in exploring whether this candidate is able to fulfill their expectations or not. High value of h-index can be considered as an indicator that a scientist is doing well in the research. Hence the objective of this research is to predict the scientific impact of researchers from the field of Computer Science using a comprehensive data set and evaluate its performance across other fields as well. Moreover to predict the future impact of young researchers , which has not been considered in the previous approaches.

1.5 Thesis Organization

This chapter is followed by review of existing studies related to impact evaluation of researchers. In chapter 2 literature review is presented, based upon the studies related to the prediction of different impact evaluation criteria including, h-index, citation count and number of publications. Young researchers impact evaluation is discussed and research gaps are highlighted.
Chapter 3 explains the methodology adopted to answer our research questions. In this chapter we have explained the data set, techniques used in this study, experimental environment and feature sets. For each research question, methodology steps are separately described in detail.
Chapter 4 discusses the experimental results in detail. As per pattern in methodology , results are also discussed in accordance with research questions.

Chapter 5 comprises of conclusions of the thesis along with the main contribution of thesis. Some future directions of research are also eloborated.

Chapter 2

Literature Review

The chapter highlights important researches done in the context of impact prediction for researchers and the challenges in this field , indicating the need for this research. Different techniques to predict future impact focused on h-index, citations and publications are discussed. The chapter further describes the prediction of impact for young researchers, followed by summary of significant studies and observations inferred from this literature review.

Different studies related to the evolution of scientific impact are considered in [43] and it is discovered that scientific community is interested/concerned in having some mechanism to estimate future evolution using current data. Decision made today, on tenure, allocation of grants and publishing are based on these estimates. They asserted that first an unequivocal criteria to evaluate recognition needs should be finalized and that criteria should be utilized/used as a target variable. In literature to pursue this problem a number of studies have been done that compared and evaluated different variants/extensions of h-index and other bibliometric indicators [14, 18]. Recently in a similar type of study effectiveness / impact of h-index and two newly proposed indices in identifying the exceptional performers/researchers in the field of research, especially in the field of Computer Science is measured [44]. They have also proposed a variation of k-index based upon h-index. They have considered variants/modifications of h-index along with h-index and tested on comprehensive data set for the field of Computer Science.

The Award winners' data set is considered as the benchmark for the evaluation of these indices for individual researchers. It is established in scientific community that researcher's having high h-index have more impact or are scientifically highly recognized. Based on this , to measure author's scientific impact , they have proposed a variation of k-index, K_S-index. The crux of this new proposal is that papers cited by authors having high h-index value should be considered as more significant/influential papers in the domain. Idea is, to assess the researchers' performance/calibre/scientific social recognition by considering the impact of researchers who cite their papers. To measure this impact of researchers h-index is used as a measure. This newly proposed variation outperformed other measures considered in the study[44].

There are a number of studies which have used h-index for scientists' evaluation [3][4]. In our study we are also considering h-index as criteria to measure scientific impact of a researcher. With the growing number of applicants for tenure track or funding etc., there is a need to have combination of different parameters for better prediction of scientific impact [25][26]. Keeping in view the importance of predicting future performance for better current assessment, different studies focusing on impact prediction from differnet perspectives are discussed below. Impact prediction studies considering h-index as impact evaluation criteria are discussed first, followed by citation count and number of publications.

2.1 h-index Prediction

Different authors have explored the predictive power of h-index [28][45]. Acuna et al. (2012) have proposed formula to predict h-index of a small sample of researchers from life sciences field, they have considered neurosurgery, drosophila and evolutionary scientists [24]. To predict future h-index, initially they have considered 18 factors and found out that only 5 are significant .The five parameters they have identified include number of publications, current h-index, years since publishing first article, number of distinct journals published in, and number of articles in top journals. They claim that the prediction based on 5 parameters yielded better results than using only h-index for neurosurgery field. The paper is focused only on the sub fields of life sciences and within that, it yielded good results for neurosurgery. Using regression models, Acuna et al. (2012) have predicted author's h-index for five years with R^2 value of 0.66 for Neuroscientists and for Drosophila and Evolutionary scientists somehow poor prediction i.e. R^2= 0.54 and R^2= 0.61, respectively. They have considered the researchers having 5-12 years of experience and h-index greater than 4, but such constraints are hindrance in the usability of formula. Analysis of Acuna et al. (2012) shows that it should be applied on large data set of same field and on other multidisciplinary data. The formula can be recalculated for other fields and while applying on other fields it is also possible to find one or more common factors for different fields. The equations proposed by Acuna et al.(2012) were validated for Spanish psychologists including Neuroscience psychologists [33]. This study exposed the limited validity of these equations for different data sets. The equations overestimated h-index, error of prediction were high and even worse when target year moves farther [33].

In [31] Dong et al. have also proposed h-index prediction technique while considering some other parameters applying on the data set from the field of Computer Science. The parameters they have considered include current h-index, average citations per paper, number of coauthors, years since publishing first article and number of publications. According to their findings author h-index is the most important factor in predicting author future h-index followed by number of publications and coauthors. In this study they have considered only first/primary authors of a paper and also authors having h-index greater than 10. Positive correlation was found between h-index and number of papers and coauthors. They have predicted author's h-index for five years with an R^2 value of 0.92. It was found that predicting h-index for longer time frame and of those scientists who have high h-index is more difficult. They have not considered the case for young scientists or scientists having low h-index values.

Both techniques [24] and [31] have considered current scientific impact of authors to predict h-index, but considering different parameters. Penner et al. have considered small data sets (762 careers) from Physics, Biology and Mathematics domains[34] . The parameters considered were same as of Acuna et al.[24]. According to them the model exhibits better results when we consider scientists of all the career age. But its performance deteriorates when we apply some limitations on the time duration, like if we consider junior scientists only or when only certain age groups data is considered. They have emphasized to consider the career age when predicting h-index, as h-index is a cumulative measure and according to them prediction aimed models should avoid cumulative, non-decreasing measures as it would yield artificially large coefficient of determination R^2. Instead they have reformulated the problem and predicted the increase in h-index for fixed time interval and also considered different age groups, with this setting the model didn't show good results. R^2 value was found to be 0.30, 0.50 and 0.54 for Physics, Cell Biology and Mathematics respectively. They have also tested the predictability of the citation impact of a scientist based upon the number of publications, their citations, and h-index of scientist. It is emphasized that the variation in the coefficient weights across different fields and career ages should be carefully studied. Also some prediction model suitable for real world is needed.

In another research, the effect of different characteristics of coauthor network of an author on h-index is studied [35]. They have considered 594 authors' record from web of science from different fields. They have used regression models and coefficient of determination R^2 to find out the factor which explains the variability in h-index better than others. It was found that high h-index can be achieved by working with many coauthors and if some of those have high h-index it would have extra benefit. R^2 found out for this study was 0.69. The data set considered for this analysis was not very comprehensive, and it has relied on ISI web of science data only. Whereas, we are quite aware of the fact that Web of science does not index all the journals.

In another study, cost-sensitive naïve Bayes approach is considered to predict h-index [36]. They have considered university faculty members from 48 Spanish universities of three subfields of Computer Science that is Computer Architecture and Technology, Computer Science and Artificial Intelligence, and Computer Languages and Systems.They have divided professors in two categories senior and

junior teachers. Time span for the publications ranges from 1978 to 2005. Teachers having their first publication in last 3 years are put into junior category, where as those who published their first paper 8 or more years earlier are categorized as senior. They have considered total 16 parameters which included university, number of publications, total citations and designation of faculty member along with 12 variants of h-index. They have selected features on the basis of correlation value of features that is to consider those features who have high correlation with the values of the class to be predicted. University which they belong, publications, g-index and c-index are found to be most important factors/factors playing important role in h-index prediction for senior professors.
Accuracy for first year prediction for junior model was 81.31 whereas for senior model it was 69.50.for 2nd year it was 71.29 and 58.20, 3rd year 54.26 and 50.96 and for fourth year 49.65 and 50.89 with minimum 2.37 and maximum 7.9 standard deviation. The main /obvious drawback for this approach is, one has to do a lot of Calculations like to calculate 12 indices first before the prediction process. Different algorithms have been proposed to predict the impact of authors featuring h-index[32]. In this study, features from three different fields/angles i.e, attribute feature, time-series based features and heterogeneous network features have been considered. They have considered Long short-term memory method, and used the output predicted value of h-index from LSTM as time series feature. XGBoost method is found to be most successful in comparison with support vector regression, random forest, LSTM and gradient boosted regression trees. Authors have used the data set as is used by weighs et al., the data set is from the field of computer science ranges from 1975 to 2015. They have discarded the authors who have h-index less than 4 and also author's who have not published their first paper in last 5-12 years before prediction period. By using data till 2005 they have predicted h-index for next 10 years.. R^2 and MAPE (Mean Absolute Percentage Error) are used as evaluation metrics. They have done comparison with [31][37][46] and found that XGBoost outperforms all other. It was found that results of regression prediction are better than time series prediction.

To predict future h-index for next 10 years, Mistele et al. have used neural network

[38]. Publicly open access data set of Arxiv for the field of Physics is used in this study. They have considered authors who have written their first paper in the interval of 1996 to 2003 to have authors who have started their research in last 5 to 12 years similarly as [24] have done.. Another constraint which they have applied is to remove those authors from the list who have less than 5 and more than 500 publications. They have also removed papers having more than 30 authors. Finally they considered 39,412 author records. They have considered 11 inputs/features for neural network model including paper citations ,age of paper, papers pagerank, papers length, journal papers or not , Journal Impact Factor, number of coauthors, coauthors page rank, subfields of Physics, papers topic distribution, broadness of topic in arxiv. R^2 values were found to be above 0.90 for 1 and 5 years.

2.2 Citation and Publications Count Prediction

There are some prediction studies focusing on predicting citation impact of publications [47–50]. According to [51] regression models are usually considered effective in citation count prediction. In their study thay have considered content centric and author centric features to predict citation count using regression models. Content features including different variations of citation counts, scope of paper and diversity of papers are found to be most effective in the ciation count prediction.

To predict the citation count of a publication, a system based upon series of features of a particular publication is proposed by [52] . They have applied regression models and evaluation metric coefficient of determination (R^2) is used as performance evaluation metric. They have used this prediction to identify the potentially influential literature through future influence prediction (Citation Count). They succeeded in having 83.6% R^2 value using different combination of features/parameters. The distinguishing factors which make a paper more influential are found to be Author Rank (based on citation count) and Maximum Past Influence of Authors (maximum citation count for a single publication). Also citation count prediction for a longer period is found to be more accurate having 0.927 R^2.

Publication success for the young scientists is predicted by Laurance et al.[53] . The purpose of this study was to identify the long term performance indicators for young researchers, with respect to number of publications. They have considered 182 biological and environmental scientists who have completed their PhD in 2000 so that their further 10 years data would be available for exploration of features. They have considered period during their PhD studies and after their PhD. For evaluation purposes, they have considered five factors/characteristics, which are gender, language, university prestige, first publication date before or during PhD and first publication date after or in the year in which PhD was completed. According to their findings those who have research publications early in their career are found to be more productive later on as well. Other factors have nominal effect on the productivity of young academics. Number of coauthors and collaboration are found to be strong predictors of number of publication as found by lee and Bozeman[54].

Revesz presented data mining method to predict citation curve for an author for any time t in the future[55]. They proposed method to predict the citations to all the publications of individual authors. Authors have focused on nobel prize winners and considered publications data of 8 leading Physics researchers from Web of Science. Nobel prize winners or nominees are very few reseachers, and results acquired on such a small and extraordinary sample cannot be generalized. In [56] impact factor of term is proposed as new bibliometric indicator and its effectiveness to predict future impact of study or author is discussed. Number of citations is considered as future impact criteria. According to their findings, values of impact factor of terms are more stable with high number of articles with this term. Stability of term also helps in better prediction of future citation count of a paper. This issue is addressed as two class's classification problem, classes are based on that an article will be cited by any other article in next 3 years or not. Prediction of citation count of a scientific paper is considered for Computer Science domain by [57]. According to their findings citations over the year follow diverse patterns. In their study they have identified six categories of such patterns. Based on this they have adopted stratified learning approach for the prediction of citation count. First they identify the category of the target paper, that out of these 6 categories which category target paper belongs to.Then apply regression model based on the population which is included in that category, to predict the citation count for the target paper. Author centric features especially productivity of an author is found to play key role in predicting citation count.

Summary of some related studies are mentioned in the table 2.1. Studies related to prediction of h-index, citation count and number of publications are listed in this table. Regression is emerged as the mostly used technique for impact prediction. Very few studies have considered temporal dimensions and paper content for impact prediction. Accuracy values are encouraging but there is room for improvement. In next section, we have discussed the problem of young researchers impact prediction.

2.3 Young Researchers' Impact Prediction

It is evident form the literature discussed above that scientific impact prediction techniques usually work for the researchers who have spent sometime in the field. Potential of young researchers whether it is in the form of citations or h-index , cannot be predicted effectively. The young researchers are also referred as rising stars in the literature. There are number of studies addressing the problem of predicting or identifying rising stars[60–64]

Renowned international scientists for the field of biomedical judged that h-index is a very promising measure to assess the quality of work of young researchers [65]. Impact of established scholars/scientists on the career of young scientists is explored by [66]. They have considered a scientists' first three years after his first publication as young scholar period. A Scientist having highest number of total citations to his/her credit is considered as established or outstanding collaborator. It was found out that outstanding scientists have positive influence on the early stages of their young collaborating scientists' future career. Hence having supervised by or collaborating with outstanding scientist would be able to excel

Table 2.1: Summary of Most Relevant Studies

	Research Studies	Lee and Bozeman, 2005 [54]	Acuna etal., 2012 [24]	Gonçalves etal., 2014 [58]	Dong etal., 2016 [31]	Xiao etal., 2016 [48]	Weihs and Etzioni, 2017 [37]	Wu etal., 2019 [32]	Akella etal. 2021 [59]
Purpose	Predict h-index		✓		✓	✓	✓		
	predict Citation count			✓		✓	✓		✓
	predict number of publications	✓							
Parameters	Current h-index		✓		✓	✓	✓	✓	
	Publications		✓	✓	✓	✓	✓	✓	
	Citations				✓		✓	✓	✓
	Coauthor	✓		✓	✓	✓	✓	✓	
	Years since publishing first article		✓		✓		✓	✓	
	Articles published in distinct journals		✓	✓				✓	
	Collaborations	✓					✓		
	Impact of venue		✓	✓		✓	✓	✓	
	Yearly rate of publications			✓					
	Paper content			✓		✓			
	Temporal dimension			✓					
	Altmetrics								✓
	Accuracy Values	R^2=0.17	R^2=0.92(1 yr.) R^2= 0.67(5 yr.)	R^2=0.45(0-5 exp) R^2=0.78(20+ exp)	R^2=0.92(5 yrs)	MAPE=0.175(5 yrs.) Acc=0.82(5 yrs.)	R^2=0.93(1 yr) R^2=0.84(5 yr)	MAPE=0.0915 (5 yrs.) R^2=0.84 (5 yr)	Acc=0.793 (1 year) Acc=0.959 (4years)
	Techniques	Regression	Regression	Regression	Regression	Regression	Regression	Regression	Classification
	Data sets	Universeity Faculty Members(US)	Scientists Academic Tree	ArnetMiner	ArnetMiner	Microsoft Academic Graph	Computer Science(1975-2015)		Altmetric .com

the performance of young scientist. Impact of various factors on scholar popularity are studied in [58] considering the Computer Science field data, here scholar popularity is interpreted as total number of citations. The features here studied include, number of publications, yearly rate of publications, distinct publication venues, venue quality and coauthor network. Number of publications along with number of distinct venues have high correlation with scholar popularity for all career ages. They have also done regression analysis and calculated R^2 for scholar popularity prediction. According to their findings number of publications and quality of publication venues were found to explain most of the variance. Overall number of publications was found to be the most important factor for assessing scholar popularity i.e. total number of citations. Predicting the future performance of young researhcers' problem is addressed in [60]. In this study early career related factors for the field of Information science and computer science are considered to predict the performance of researchers. Number of publications and citation count are used as criteria to assess the research performance of 4102 scientists, data for this purpose is gathered from scopus. Frist publication of all the considered researchers is same, i.e. 2005. And another constraint which was applied that all the researchers must have published at least one paper between 2009 to 2012. With 13 independent factors considered from early career of scientists, separate regression model for each factor is applied. Adjusted Coefficient of determination (R^2) is used as an evaluation metric. Number of publications is found to be the most effective predictor of research performance i.e. number of publications and impact i.e. number of citations.

Rising stars prediction problem is treated as a classification task by [62]. They have proposed weighted evaluation model considering quality of citing papers and influence of coauthors. Impact score is calculated for each author and on the basis of that score an author is labelled as a rising star or not. Ultimately author, social,venue and temporal features sets are considered and different classification models are applied considering these feature sets. ArnetMiner data set is used and venue features are found to be most effective indicators for the correct classification of rising stars.

Rising stars identification problem is addressed as social influence prediction problem by [61]. They have proposed StarRank method to predict researchers rankings in the future. To evaluate the performance of the method, assumption is, Higher the number of rising stars in top ranks higher the performance of method. They have also used spearmann correlation coefficient as an evaluation metric. Citation count is used as an evaluation criteria of rising stars. PubRank algorithm is proposed by [67], which is based upon bibliography network, emphasizing social interactions of researchers. The algorithm incorporates mutual influence among researchers, venue of publications of a researcher and ability of a researcher that how quickly the researcher builds ad strong collaborative network earlier than others as factors to mine the rising stars. Regression model is built to identify the rising stars and DBLP data set is used for this purpose. Algorithm brings rising stars in top ranks , as the PubRank score increases, chances of rising star to be in top ranks increases. They have compared their top ranks with future achievements of these researchers and found promising results. [68] considered the problem of identifying rising stars with respect to citation count. They have considered number of factors and applied regression learning methods on those factors. Naïve Beysian out performed all other methods. Considering ArnetMiner data set they have divided authors under 10 different topics. based on topics and authors are ranked under their respective topics . That is they have takenthe problem to one level further , that is not whole field rather the sub topics from a field. When ranking the authors in their different categories based upon identified topics, it was identified that temporal factors play crucial role in ranking rising stars in top ranks. Author and social fators also play important role, but venue doesn't play any significant role .

Highly cited and highest relative average increase in citations are used to classify researchers as rising stars by [69].
Classification techniques are applied considering ArnetMiner data set, in total author, coauthor and venue based 11 features are considered. In total 44167 researches are considered from 1995 to 2000 . venue based features are found to be most effective in identifying rising stars.

Influence of well-known researchers on the career of young researchers is analyzed by Amjad et al. [70]. Researchers from the domain of Computer Science are considered from a comprehensive Arnetminer data set. The researchers having h-index greater than 40 are termed as authority authors in this study. Young researchers are divided in two categories: Young researchers who started their career with authority authors and young researchers who published earlier on their own and later collaborated with authority authors. Young researches are identified from 2000-2004 period and their progress is analyzed from 2005-2014. The study concludes that having collaborated with authority authors has a great positive effect on young researchers performance. Those young researchers who initially proved their worth on their own and later got an opportunity to work with authority authors are found to be more productive and successful in terms of citations and collaborations. In a detailed analysis [46] researchers are separated career age wise. According to their findings, in almost all the cases trend is same that is the R^2 values crosses 90% for 8 years career age and highest range of R^2 values is from 22 to 34 years approximately. That is h-index prediction for researchers having experience of 22 to 36 years is most accurate, for one year ahead prediction, it is above 0.99 and for five years prediction it is above 0.94. Though for researchers having experience greater than these values decline but still these are above 0.98 for one year prediction. Further data set is partitioned on the basis of current H-index of authors. It was observed that h-index for authors having low h-index value are difficult to predict. R^2 values for prediction of authors having h-index in range of 0-3 are not very encouraging , Whereas it has highest values for above 30 threshold. It was concluded that for the researches having different h-index values, researchers having low h-index were difficult to predict. In general R^2 values increase with the increase in h-index. It means it is difficult to predict the h-index for authors having low h-index value, whereas it improves for higher h-index values[46]. Considering the prediction of young researchers future , it is quite evident that in most of the cases citation count is considered as a target to achieve or in other words effectiveness of proposed techniques are evaluated against citation count. None of the technique effectively predict the future h-index of a young researcher. Moreover Existing prediction models for h-index prediction, aggregate all career data, which is not justified towards young researchers, and it would be hard/unrealistic to have a predictive model that would be fair/just to all groups of scientists [43]. In our study we have done factor analysis, we have done forward feature selection to identify the feature set which can effectively predict the future h-index for a young researcher.

2.4 Discussion

After having detailed analysis of existing approaches for h-index prediction, some important findings are mentioned below:

(a) Existing prediction methods apply some constraints on the selection of researchers like researchers having 5-12 years of experience or researchers having h-index value greater than 4.

(b) Existing methods should be applied on large/comprehensive data set of same field.

(c) Existing methods should be applied on multidisciplinary data i.e. suppose a method/solution is proposed for the field of Neurosurgery then this solution should also be tested for some other field for example for the field of Chemistry.

(d) Predicting h-index for longer time frame i.e. more than 8 years in future and of those scientists who have low h-index i.e. from 0 to 3 ,is more difficult.

(e) Case of young scientists or scientists having low h-index values is not considered .

(f) Mostly regression is used for prediction, in regression equations different variables/factors have different coefficient weights for different fields or career ages. It is emphasized in the literature that the variation in the coefficient weights across different fields and career ages should be carefully studied.

Addressing these observations or research gaps, it is concluded that existing methods should be applied on large/comprehensive data set of same field. Moreover existing methods should be applied on multidisciplinary data i.e. suppose a method-/solution is proposed for the field of Neurosurgery then this solution should also be tested for some other field for example for the field of Computer Science. Different parameters are explored in literature on different data sets, there is a need to apply these parameters on same dataset, to find out optimum set of parameters. Because of certain constraints on data sets young researchers got excluded from the data sets. Problem of prediction of future impact of young researchers should be addressed. Considering all the observations, research questions are devised and steps followed to retort those questions are explained in detail in methodology.

Chapter 3

Proposed Methodology

3.1 Introduction

In the era of Big Data, quantitative approaches should be taken for the evaluation purposes (Bertsimas et al., 2013). Keeping in view this fact, we have considered a comprehensive and large scale data set for the field of Computer Science taken from Arnet Miner[1] and explored it for h-index prediction using different combination of parameters. Initially, we have considered the parameters proposed in different models, which include [24][34] and [31]. Later we have proposed some variations in the proposed parameters and validated for the data from the field of Computer Science. From the literature review we have identified that existing models impose constraints on the selection of researchers. Moreover models derived/proposed for one field are not tested for other fields. While predicting future impact of researchers, case of young researcher's is excluded. Based on all the observations stated in section 2.4, in this study, we are exploring the following research questions. RQ1 addresses observation b, RQ2 is based on observation a and b, RQ3 addresses observation c and d and RQ4 addresses observation b. **Research Questions**: RQ1: Which of the existing sets of parameters/Models for h-index prediction performs better when applied on the comprehensive data set from the field of Computer Science?

[1]https://cn.aminer.org/billboard/aminernetwork

RQ2: What would be the best set of parameters suitable for the prediction of future h-index for the researchers from the field of Computer Science?

RQ3: What would be the suitable set of parameters for better h-index prediction for young researchers?

RQ4: What is the performance of proposed (in RQ2) model when applied on a data set other than Computer Science?

Our methodology is focused on these research questions. Initially we will discuss data sets and afterwards all the steps taken to solve the problems identified in literature review and documented as research questions. In order to find the answers to the above mentioned research questions, different activities to be performed in proposed methodology are shown in Fig. 3.1. Firstly we did through literature review related to scientific impact prediction studies, literature review is discussed in chapter 2.

When scholar started this work, the only existing work (to the best of our knowledge) for predicting h-index was for the field of Neuroscience. The parameters contributing in the prediction of scientific impact can behave differently for different fields. Keeping in view diversity of the field and its application in many other fields, we decided to predict scientific impact of researchers from our own field, which is the field of Computer Science and acquired comprehensive data collection of ArnetMiner. Parameters were identified from literature and some new parameters are also proposed. Machine learning techniques are applied and forward feature selection is done to obtain feature set that can predict future impact effectively. Feature sets are applied for young researchers and ultimately a new metric for impact prediction of young researchers is proposed. All the steps in methodology are discussed and elaborated in next sections.

3.2 Data set Description

We have considered comprehensive data set of ArnetMiner [71]. ArnetMiner

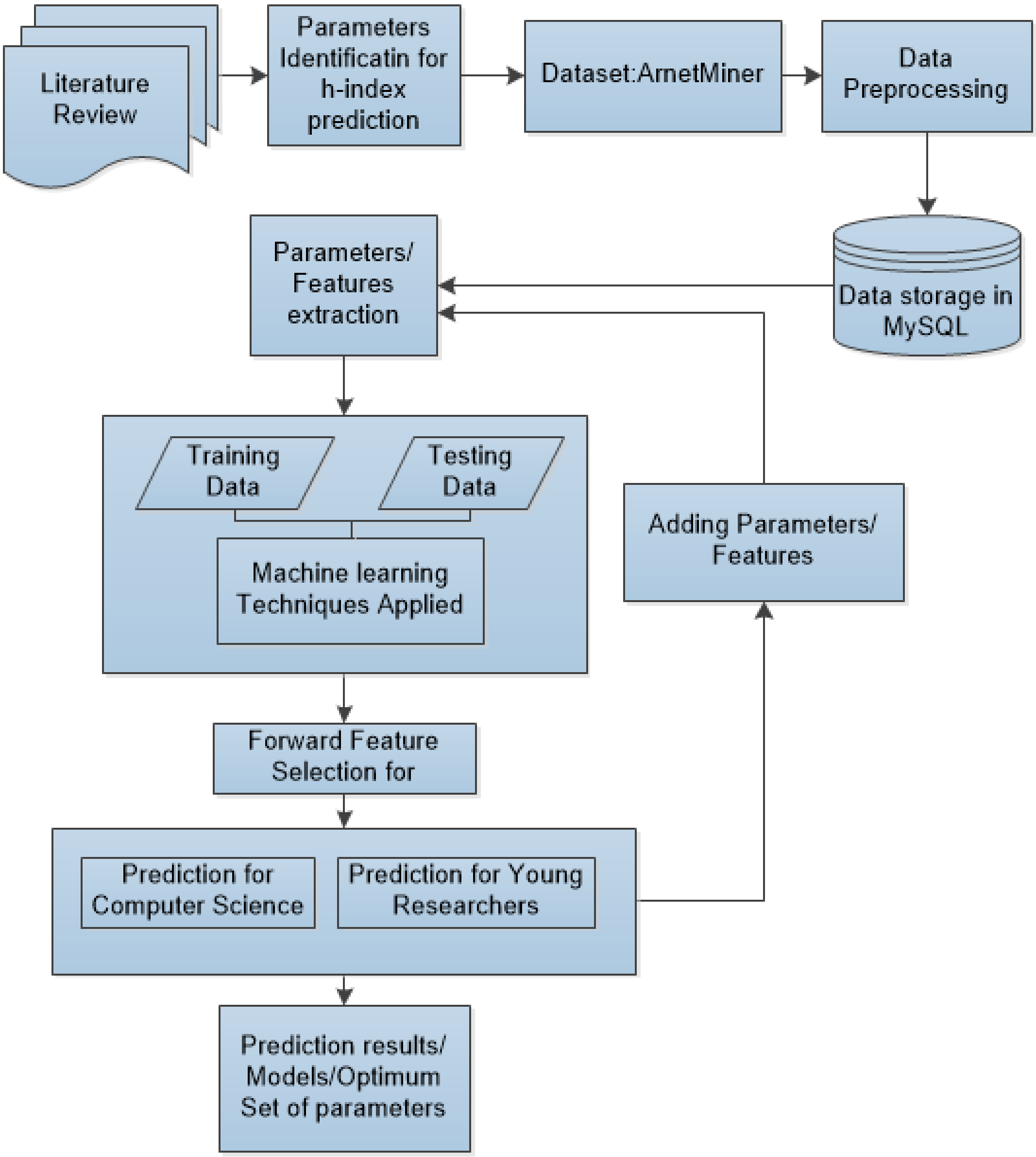

FIGURE 3.1: Proposed Methodology

data set is a collection of publications from the field of Computer science, collected from Digital Bibliography and Library Project (DBLP) bibliography[2] , Association for Computing Machinery (ACM) Digital library[3] and CiteSeer[4] . DBLP is a Computer Science bibliography website, ACM is another comprehensive bibliographic database focused exclusively on the field of computing and CiteSeer is also a repository of papers in Computer Science. [71] have developed ArnetMiner system, which extracts and mines academic social network. Researcher's profiles are

[2]http://dblp.uni-trier.de/
[3]https://dl.acm.org/
[4]citeseerx.ist.psu.edu/

automatically collected and publications data from existing libraries (mentioned above) are integrated, applying probabilistic framework for author disambiguation [71–74]. Although data set is collected from established sources for Computer Science domain and it is a large collection of data but there are some issues. This data set does not contain very recent publication records, data set has records till May 2014. Moreover, the data set is collected from multiple sources, an obvious outcome of this fact is that there maybe some duplication in the data. In spite of these factors, ArnetMiner is a widely used data set and is considered in a number of studies considering data set for the field of Computer Science[31, 62, 75–77]. Arnet Miner is originally extracted focusing on researcher profiles, whereas applications/data sets like Google Scholar and Microsoft Academic provide paper retrieval[78]. It is one of the best and well organized databases for Computer Science articles[70]. Researcher profiling technique proposed and adopted for the collection of this data set outperformed other baseline methods. A probabilistic framework is proposed and applied to disambiguate author names, performance of this framework is also quite satisfactory[71]. Moreover data tables shown in Fig. 3.2 are very useful in extracting a variety of features.

The data set contains 2,092,356 publications and 8,024,869 citations between them, also record of 1,712,433 authors and 4,258,615 collaboration relationships between authors, Data set have publications record from 1936 till May, 2014.

The tables/entities which are included in the data set are papers, authors, coauthors and author-papers. Relations and attributes are shown in Fig. 3.2; following is description of objects/schema.

Papers: [id, title, authors (separated by semicolons), affiliations (separated by semicolons, and each affiliation corresponds to an author in order), year, publication venue, the id of references of this paper (there are multiple lines, with each indicating a reference), abstract]

Authors: [id, name (separated by semicolons), affiliations (separated by semicolons), the count of published papers of this author, the total number of citations of this author, the H-index of this author, the P-index with equal A-index of this

Paper

ID	Title	authors	affiliations	Year	Publication Venue	Id of references of this paper	Abstract

Author

ID	Name	Affiliations	Publication count	Total Citations	h-index	p-index	Key terms

Coauthor

1st AuthorID	2nd AuthorID	Number of collaborations

Author-Paper

Index	AuthorID	PaperID	Author's position

Figure 3.2: Relations and Attributes

author, the P-index with unequal A-index of this author, extracted key terms of this author (separated by semicolons)]

Coauthor: [id of one author, id of another author, number of collaborations between them]

Author-Paper: [index, author id, paper id, author's position like 1st author, 2nd author etc.]

It took a lot of effort to handle such a large data set. It was imported into MySQL by using MySQL for excel add-in. Using certain queries and stored procedures we have cleaned the data i.e. removing special characters and stored the data in appropriate column in tables. A lot of time and effort is spent on first storing this data in MySQL and afterwards running different queries on it. As an example, consider the case of computing citations of papers. Total number of references in this data set is 9,268,353, so to compute citations of a single paper from this data set, it would require 9,268,353 comparisons from references table, whereas total number of papers in this data set is 2,092,356. Now to compute citations for all these papers would require thousands of millions of comparisons. Though we have used indexes and stored procedures but still a single query required sometimes 2-3 days to execute. Statistics of data set are given in Table 3.1. Data set has comprehensive coverage of publications for Computer Science. While evaluating that how many years in future we should be predicting impact of a researcher.

Table 3.1: Data Statistics

Category	Instances
Total number of authors	1,712,433
Number of authors having publication in 2007 or earlier	938,204
Total number of publications	2,092,356
Number of publications in or before 2007	1,273,731
Total Author-paper relationships	5,192,998
Papers references	9,268,353
Papers references till 2007	4,463,648

It was considered that in literature it is normally predicted for five years [31] and if we look at it objectively, a researchers future five years performance would be enough to hire him for some research oriented task. Moreover as data set has publications record till 2014, so to be on the safe side with respect to coverage of publications, we have considered publications record till 2012.

Now for prediction purposes, for our experiments we have considered the data set records till 2007. Our goal was to predict authors' h-index for next 5 years while considering authors different characteristics/parameters/features calculated in 2007. For this purpose we have considered data for all those authors whose first publication was in 2007 or before 2007 and only used data that was available till 2007. That is, on the basis of available data for an author/researcher in 2007, we have predicted his/her next five years h-index (for years 2008,2009,2010,2011 and 2012). For this purpose, along with other parameters' values (which would be stated shortly), we have calculated h-index for 2007, 2008,2009,2010,2011 and 2012 of these authors. Idea was to predict author's h-index for next five years while considering the authors' data in 2007. Hence h-index of 2008-12 is considered as target variable one by one. Number of authors having publications in 2007 or earlier were 938351. There were 146 such cases where the year of publications were not mentioned, so we discarded those records and were left with 938205 authors' records as shown in Table 3.1. Total number of authors till 2007 and onwards till 2012 are shown in Fig. 3.3. There is a smooth increase in the number of authors

FIGURE 3.3: Dataset Statistics (2007-2012) Total number of Auhtors

over years. Similarly in Fig. 3.4 total number of citations till 2007 and onwards are represented. Linear increase in the values of citations can be seen over the years. Whereas for number of publications, we have shown the publications record in these years as shown in Fig. 3.5. There are more than 100000 publications record in each of these years. This data set is used for RQ1, RQ2 and RQ3. For RQ3, we have also used random sample for young researchers from this data set. Details of acquiring the random sample and other details regarding young researchers is discussed in the relevant section.

3.3 Techniques Used for Prediction

Machine Learning algorithms are indispensable for Data Scientists with their growing number of real world applications. Most Machine learning problems fall into one of two categories: *supervised* or *unsupervised*. Fitting a model to predict the response variable and to relate the response variable to the predictor variables lies in the supervised learning domain. Variables may be quantitaive or qualitative, quantitaive can have numerical values like age , height. Where as qualitative can have values from classes or categories like gender: male or female[79].

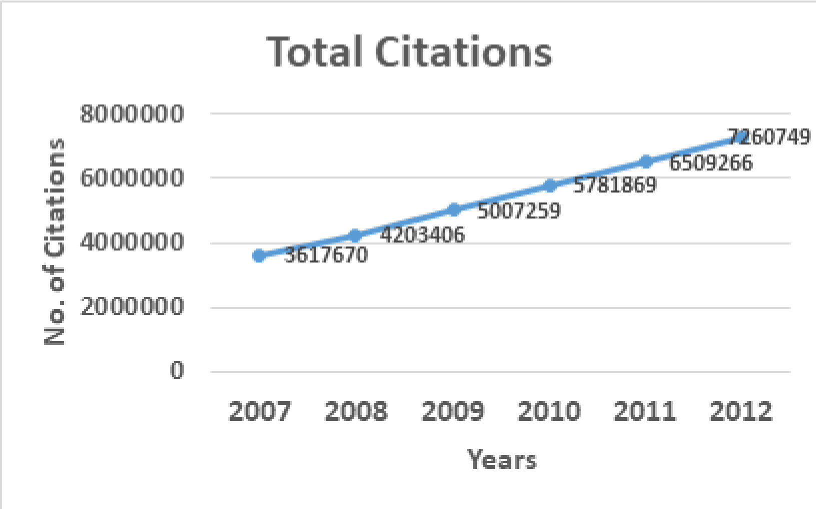

FIGURE 3.4: Dataset Statistics (2007-2012) Total number of Citations

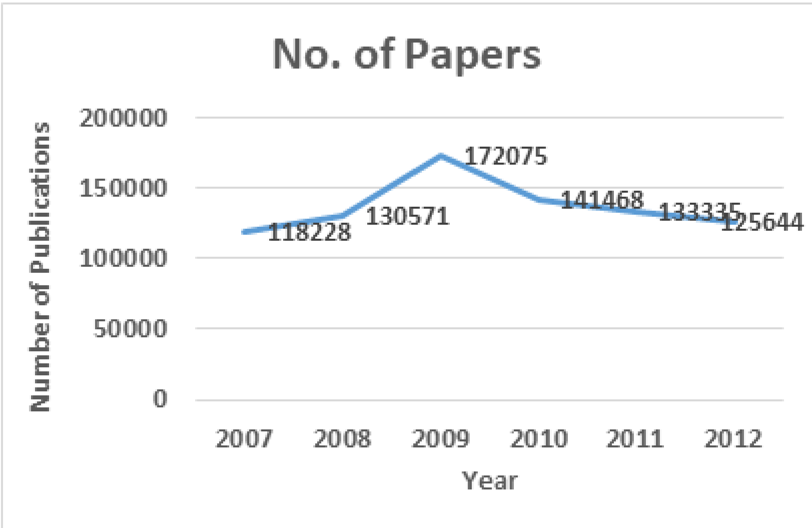

FIGURE 3.5: Dataset Statistics (2007-2012) Number of Publications in respective years

Problems with quantitaive response variables are referred as regression problems and those with quantlitaive response variables are often referred to as classification problems. The standard linear regression models provide interpretable results and works quite well on many real world problems. In contrast, unsupervised learning describes the somewhat more challenging situation in which for every observation, we observe a vector of measurements, but no associated response variable. It is not possible to fit a linear regression model, since there is no response variable to predict. In this setting, we are in some sense working blind; the situation is referred to as unsupervised because we lack a response variable that can supervise our analysis[79]. Considering the nature of our problem, we are applying regression analysis.

3.3.1 Regression Analysis

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. Regression analysis may be the most widely used statistical technique. An important objective of regression analysis is to estimate the unknown parameters in the regression model. This process is also called fitting the model to the data [80]. We have fitted model using multiple linear regression models.

In regression models/equations, coefficients show the relationship between predictor variables and the target/response. Coefficients can be with plus/positive sign or with minus/negative sign. Plus sign shows that value of target increases with the increase in predictor and minus sign indicates that value of target variable decreases with the increase in predictor. Coefficient of determination and Root Mean Square Error are used as evaluation metrics.

Coefficient of Determination, R^2

To predict the author's h-index for next five years, we have considered the values of parameters till 2007. We have fitted regression equations to predict author's h-index for next five years i.e. from 2008 to 2012. To check the validity of these regression equations Coefficient of determination, R^2 is used. Variance explained or Coefficient of determination determines that how much variation in the value of y is explained by the variation in value of x and is determined by the formula given in Eq. 3.1.

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}\ , \tag{3.1}$$

In Eq. 3.1, y represents dependent variable (actual values), $\hat{y}$ represents its predicted values. $\bar{y}$ is the mean of actual values of dependent variable. Value of R^2 ranges from 0 to 1. A value of, let us say, 0.7932 means that 79.32% of the variance in y can be explained by the changes in x. R^2 is the variation in dependent variable that is explained by model. Higher the value of R^2 smaller the differences in observed and predicted values [81].

A widely used approach to select best model would be to select the model which gives the largest value of R^2. A minor concern with the use of R^2 is that, with every additional feature R^2 value will increase. To address this issue adjusted R^2 is used. Adjusted R^2 is calculated as follows:

$$R^2 adjusted = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1} \ , \tag{3.2}$$

Where R^2 is the R-squared value, n is number of records and k is number of independent variables/predictors. Adjusted R^2 shows the strength of fitted model, it is very useful in evaluating which predictors are helping to improve the accuracy in prediction [82].

Root Mean Square Error, RMSE

Another measure of assessing the performance of fitted model is Root Mean Square Error, RMSE. R-squared is a relative measure of fit and RMSE is an absolute measure of fit. RMSE is the most important criterion for fit if the main purpose of the model is prediction[83]. It is calculated by taking square root of average of sum of squared differences in actual and predicted values. Root Mean Square Error (RMSE) measures how much error is between predicted and actual value. It will have smaller value if predicted values are very close to actual values, and will be large if for some of the observations, the predicted and actual values differ considerably.

3.3.2 Neural Networks

Neural networks are often used for statistical analysis and data modelling, in which their role is perceived as an alternative to standard nonlinear regression or cluster analysis techniques [84]. Neural networks have received considerable attention, and are considered as very promising tools for classification and prediction[85].

To predict future h-index, we have also built neural network model based on the identified features.

Experimental Environment

All the experiments are written in python. Graphlab library proposed by Turi[5] is used for fitting regression models. Graphlab is a very powerful framework for different big data applications and supports graph analytics, machine learning tasks , big data visualizations etc.

The machine learning library used for neural network is Keras[6] . Keras is an open source, powerful high level API running on top of TensorFlow. Keras is user friendly and it builds a neural network with a very few lines of code. Different split ratio is used in literature for training and testing data sets. It is observed that higher the training ratio, better the model performance. Considering this fact we have divided the data set into ratio of 80:20 for training and testing which provides best performance as compared to other split ratio[86, 87]. Relu is used as an activation function. Model is trained by using fit method on training data in batch sizes of 512 with 150 epochs. To estimate the accuracy of neural network prediction, we have used the Mean Squared Error (MSE) as loss function. Experiments are performed on a high computing CPU machine with core i7 (8th generation) processor and 16 Gigabyte of Ram and Hard Disk Drive of 500 GB.

3.3.3 Forward Feature selection

Keeping the most relevant variables from the original dataset is feature selection. When number of input variables in not large enough, forward feature selection or backward elimination techniques are used. These are also useful for linear regression models. To find the optimum features from the list of features presented in the table , we have applied forward feature selection. For forward feature selection we consider all the features one by one. We train the model with every individual

[5]https://turi.com/
[6]https://keras.io/

feature separately. We test the model for every variable. The variable having highest value for our evaluation metric is considered to be best among all of them. Our evaluation metric is coefficient of determination R^2, so the variable giving highest value of R^2 is selected as the first/starting variable. Then this process is repeated, we train the model with the chosen variable and other variables, while adding one variable at a time. Then two variables set which gives best performance is considered. We repeat this process, continue to add variables until the value of evaluation metric stops increasing.

3.4 Comparison of Models

From literature review we have identified that Acuna et al (2012) and Dong et al. (2016) have presented models for the prediction of h-index, details are given in Table 3.2. Acuna et al (2012) have considered field of Neuroscience and Dong et al have considered Computer Science. As a first step, we have decided to check the validity of equations proposed by Acuna et al. for the field of Computer Science[24]. For this purpose, we have considered the equations proposed by Acuna et al. and applied those equations on this data set to predict the h-index for next five years. But the errors in prediction were very large and R^2 values were meaningless. As also stated in [33], it has exposed limited validity of these equations for different data sets. To resolve this issue, we have fitted the regression equation for the Computer Science data set considering the parameters proposed by Acuna et al and Dong et al.[24][31]. Here we are addressing our RQ1 i.e. Whether the existing sets of parameters/Models for h-index prediction can be validated considering comprehensive data set from the field of Computer Science?

3.4.1 Experiments

The parameters which we have to consider for predicting h-index include number of publications of an author, current h-index of an author, in how many distinct journals papers are published, number of publications in impact factor journals, number of coauthors, Citations and years since starting or publishing first publication. We have calculated the value of all these parameters till 2007. For this purpose, we have separated the publications till 2007 and authors' record of those publications. We have calculated h-index of authors in 2007, also other parameters were calculated on the basis of that record. To predict the authors' h-index for next five years. We have fitted the regression equations for authors' record till 2007 considering all the above-mentioned sets separately. We have used 80% of data for training and remaining for testing of our fitted model and found out R^2 and RMSE for all the results.

Further, it was identified that both Acune et al. and Dong et al. have applied some constraints on the data set, like Dong et. al. have considered authors having h-index greater than or equal to 10[24][31]. Similarly Acuna et al (2012) have considered authors having 5 to 12 years of experience and having h-index greater than or equal to 4.

For further verification , we have planned to apply the parameters Acuna_features_set and Dong_features_set mentioned in Table 3.2 on ArnetMiner data set with following conditions:

1. Acuna's parameters with constraint of including only those scientist/ researchers having h-index greater than or equal to 4 and having experience of 5 to 12 years(Acuna et al., 2012).

2. Acuna's parameters with constraint of including only those scientist/ researchers having h-index greater than or equal to 10 (Acuna et al., 2012).

3. Dong et al.s' parameters with constraint of including only those scientists having h-index equal to or greater than 4 and having experience of 5 to 12 years (Dong et al. 2016).

4. Dong et al's parameters with constraint of including only those scientist/ researchers having h-index greater than or equal to 10(Dong et al. 2016).

TABLE 3.2: Parameters and R squared Values

Technique	**Parameters**	**Field**	R^2**(5 Yrs)**
Acuna, Stefano and Knrad 2012	current h-index square root of no of publications years since publishing first article number of distinct journals published in number of articles in top journals (Acuna_features_set)	Life Sciences, neuroscience	0.66
Dong, Johnson and Chawla 2016 (h-index>10 only)	current h-index, number of publications, number of years since first paper, average citations per paper number of coauthors (Dong_features_set)	Computer Science	0.92
Penner et al. 2013	current h-index square root of number of publications Academic age number of distinct journals published in number of articles in top journals (used small dataset)	Physics, Cell Biology, Mathematics	0.30, 0.50, 0.54

5. Applying Acuna et al. and Dong et al. parameters having no constraint imposed.

All of the features are straightforwardly adopted other than /except /apart from number of articles in top journals, we had to tune this parameter according to the field of Computer Science dataset. We have applied models using these parameters on D1, D2 and D3 with details given in Table 3.3. Acuna et al have considered top 6 journals for the field of Neuroscience. We have considered field of Computer Science and according to the field we have also considered top journals from the field and also the multidisciplinary journals considered by Acuna et al. We have considered top journals form the field of computer science as per ranked by Impact Factor (Web of Science). Multidisciplinary journals such as Nature and Science as

TABLE 3.3: Number of Author Records in data sets

No.	Dataset	Total Number of Author Records
D1	DatasetFull	938,204
D2	datasetH10	3,435
D3	datasetH4exp5-12	9,793

mentioned in Acuna et al. paper, along with Nature Communications, Proceedings of the National Academy of Sciences and PLoS ONE were also considered. In total for impact factor publications we have considered 16 journals, top 10 journals from Computer Science field and 6 multidisciplinary journals.

After applying the models proposed by Acuna et al. and Dong et al., we calculated R^2 for all the experiments' results. It gave us the comparative performance of these two previous models over the same data set, under certain assumptions/constraints and under no constraints.

3.5 Feature Selection for Computer Science

We have considered the models proposed by Acuna et al. and Dong et al. in RQ1. The common parameters among them were current h-index, number of publications and years since publishing first article, as mentioned in Table 3.2. Whereas number of distinct journals, number of publications in top journals, number of coauthors and average citations per paper are distinct parameters. Addressing our RQ2 we have considered parameters/ features/variables present in literature and also proposed some new parameters.

Initially we tried different combination of these parameters, followed by some tuned parameters. We have discussed the parameter of number of publications in top journals in RQ1, we further tuned the impact factor publications parameter by considering threshold for impact factor, i.e. 3. We considered all the publications which are published in journal from computer science field having impact factor 3 or more. For distinct publications, we have also tuned the parameter, by considering only those publications which are published in impact factor journals for the field. h-index is basically combination of two parameters, number of publications and number of citations. Keeping in view this we considered publications and citations.

Comparative analysis of different models in RQ1 have given us the comparison of two previous models over the same data set, however in this research question we want to find out the impact of some additional parameters along with some modifications in parameters of previous models. It will establish which is the optimum model/set of parameters for the prediction of future h-index for the field of computer science that is model having highest value of R^2.

3.5.1 Features Identification and Calculations/ Data set Transformations

After having comprehensive literature review, different features used for h-index prediction are identified. We have considered the features mentioned in literature , proposed some modifications in existing features and have also proposed some new features. Forward feature selection technique is applied on these features to get optimum set of parameters/features for h-index prediction. All these features are categorized as Author, Venue or Social features. Detail of these features is given below:

1. **Author Features**

   (a) **Current h-index of an author**

   h-index of an author for the year which would be considered as base year, in our case we have considered 2007 as base year. That is on the basis of data/information available for an author in 2007, we will predict future h-index of an author for next five years. We have labelled this feature as 2007_h_index.

(b) **Number of publications**

Number of publications represent productivity of an author , all the publications of an author till 2007 are considered. Total Number of publications (no_publications) and its square root (square_root_publications) are considered as features.

(c) **Career Age of an author** (years_since_start)

To find out how many years researcher has spent in research field, we consider the year of researchers first publication and gets the difference of years form current year.

(d) **Number of articles as last Author** (no_article_as_last_author)

How many articles an author has coauthored as last author. Ratio of articles as last author to total number of articles

(e) **Proportion of articles as last author** proportion_last_author

Proportion of those articles, an author has written as last authors from all the publications/articles. i.e no_articles_as_last_author divided by no_publications.

(f) **Number of articles as first Author** (no_article_as_first_author)

How many articles an author has coauthored as first author. Ratio of articles as first author to total number of articles

(g) **Proportion of articles as first author** (proportion_first_author)

Proportion of those articles, an author has written as first author from all the publications/articles. i.e no_articles_as_first_author divided by no_publications.

(h) **Average Citations** (avg_citations)

Citations represent impact of an author, hence average citations are considered. It is calculated as total number of citations divided by total number of publications of an author.

(i) **Difference in citations and h-index** (citations_diff_hindex)

h-index is directly affected by the increase in number of citations. Their might be some publications whose citations would affect h-index in near future. Keeping in view this dimension we have considered another feature which is mod of the sum of the difference of citations of a publication of an author from author's current h-index.

$$citations\_diff\_hindex = |\sum_{i=1}^{n}(citations_i - hindex)| \tag{3.3}$$

Where n is the number of publications of an individual author, citations are the citations of each paper and h-index is current h-index of the author.

(j) **Average difference in citations and h-index** (average_citations_diff_hindex)

We have also considered average of the sum of difference in citations and current h-index of an author. formula to calculate is given in eqn below:

$$average\_citations\_diff\_hindex = \frac{|\sum_{i=1}^{n}(citations_i - hindex)|}{n} \tag{3.4}$$

Where n is the number of publications of an author.

2. **Venue Features** Publishing in different and well reputed, peer reviewed venues shows the quality and diversity of an authors work. We have considered multiple venue features:

   (a) **Variety in Venues** (no_of_distinct_venues)

   To check the diversity in an author's work and his ability to satisfy diverse reviewers, another feature related to venue is considered. So this feature considers that in how many different venues an author has published his research.

   (b) **Publications in Impact Factor Journals** (no_of_IF_journals)

   Publication in impact factor journal is directly proportional to quality work, as impact journals are peer reviewed journals, so work is scrutinized by multiple people before publishing. Moreover papers published in good journals tend to attract more attention than others. Considering these, we have considered Impact factor journals from the field of Computer Science, list of impact factor journals is given in Appendix A.

(c) **Citations in Impact Factor journals** (IF_citations)

To bring the impact of the quality of publications of an author to the next level. We have also considered the quality of citations of an author's work. Quality of citations is measured by the impact of journal where it published. In this feature only those citations of all the papers of an author are considered, which are published in Impact Factor journals.

(d) **Citations in journals having Impact Factor 3 or above** journal_if_3

To measure the quality of an authors work, another venue based feature considers number of publications of an author which are published in an impact factor journal having impact factor 3 or greater.

(e) **Distinct Impact Factor Journals** (distinct_but_only_IF)

In how many distinct Impact Factor venues an author has published in.

3. **Social features**

Numerous studies have reported that scientific productivity in terms of publication and citation rate is believed to be positively associated with coauthorship[88–92]. Usually there is tendency in researchers that they cite their coauthors work[31]. Moreover Percentage of single authored paper are declining in 21st century [93], hence it would be wise to consider collaborations to examine future research performance of the researchers. Many studies have also found that collaboration between authors is also positively associated with scientific productivity[94–98], and their future success in terms of citations[49].

To further evaluate, we have considered social features including

(a) **Number of coauthors** (no_coauthors)

Number of coauthors is to sum the total number of coauthors of a specific author

(b) **Average number of coauthors** (avg_coauthors_per_article)

Average number of coauthors of an authors is achieved by dividing total Number of coauthors by number of publications of that author.

(c) **Number of collaborations** (collaborations)

If the researcher have collaborated with each other. By number of collaborations we mean how many times an author has worked with any other author or in other words how many publications of an author has more than one author?

(d) **Total h-index of all the coauthors of an author** (coauthors_total_h_index)

Impact of coauthors is also an important feature to consider. So we have considered h-index of coauthors of an author.

(e) **Average h-index of all the coauthors of an author** (average_h-index_coauthors)

To get average h-index of coauthors , first we have considered all coauthors of single paper of an author and sum their h-index divided by number of coauthors. Then we have add up all average h-index for all the publications and get them divided by number of publications.

Similarly as previous feature , we have considered average of total h-index of all the coauthors.

3.5.2 Application of Proposed Model on Sub-Data set

The regression model fitted for the above mentioned datasets is further validated for the field of Computer Science but considering different base year. Performance of proposed model is tested for authors record till 2006. That is while considering

Table 3.4: Number of Author Records till 2006

No.	Data set	Total Number of Author Records
D1_06	DatasetFull	741724

the parameters values based on the data available in 2006, we have predicted the h-index value for next five years. detail of the data set is given in Table 3.4 :

3.6 Impact Prediction For Young Researchers

Our third research question is related to the h-index prediction for young researches. Existing models have not addressed the issue of young researchers, rather they have excluded young researchers from their models. From literature we have found following definitions/meanings of young researchers:

- Having 3 years or less since his first publication [66].
- Having h-index less than 10 [31].
- Having h-index less than 4 or having 5 or less years since his first publication [24].

Our research question is
RQ3: To develop a model for better h-index prediction for young researchers.
To answer the question under focus or RQ3, we have applied existing approaches on young researcher's data from ArnetMiner, Data set statistics for young researchers is given in Table 3.5. Secondly applied the optimum solution identified as result of RQ2, on young researchers' data. We compared the performance of these models on the basis of the value of R^2. Moreover we have also identified/proposed some parameters stated below, that might be helpful in improving results obtained from the previous experiments. Description of these parameters are given below:

1. Highest h-index among coauthors (highest_hindex_coauthors) Calculate hindex for all the coauthors and consider coauthor having maximum h-index

TABLE 3.5: Number of Author Records in Young Researchers data sets

No.	Data set	No. of records	Training data	Test data
DY1	exp-lesss-than-3	257845	206195	51650
DY2	exp-lesss-than-4	306334	245339	60995
DY3	hindex-less-than-4	910242	729638	180604

2. Citations having Impact Factor 3 (IF_3_citations) How many citations of papers of an author are in journals having impact factor 3 or above.

3. Number of coauthors on second position (no_second_coauthor)

   Conisdering the authors who have worked with authors as second authors.

4. Second author Highest h-index (highest_second_author_hindex)

   For all the publications of an author, Highest h-index among all the second coauthors is considered.

5. Total h-index of coauthors on second position (sum_second_coauthor_hindex)

   Identify all the coauthors in second position , extract their all the information of publications and citations till 2007 and calculate h-index for 2007. Then add up h-index of all the coauthors of an author.

6. Average h-index of coauthors on second position (average_second_coauthor_hindex)

   To get an average of second coauthor h-index, divide sum of h-index of second coauthor (Sum_second_coauthor_hindex) by total number of second coauthors.

As stated above, we have applied equations acquired in RQ2 on DY1, Dy2 and DY3 . To compute additional parameters from such a huge population would require a lot of time and resources. Addressing this problem, we have adopted stratified sampling technique. Detail discussion on considering samples and estimating standard error with confidence interval of 95% are given in supplementary

TABLE 3.6: Random Sample for Young Researchers

	Data set	Number of Authors
Y_1	exp-lesss-than-3	12,592
Y_2	exp-lesss-than-4	14,848
Y_3	hindex-less-than-4	41,945

material. Data statistics of the sample data set for young researchers is given in Table 3.6.

3.6.1 Proposed Index

Identifying exceptional future researchers from a number of young researchers is a very crucial but difficult human resource activity [67]. Besides, it is identified in literature that with h-index, young researchers are at a disadvantage because both output productivity and impact are likely to increase with time [30]. Moreover, when a researcher has only 1 year of research experience, there are very few chances that we can predict/decide his future worth on the basis of his citations or h-index. In [99] Waltman also states that citation's information for recently published papers is not adequate, as in such a short time span these publications hardly get a chance to be cited. Hence there must be some other factors which we should consider.

Main idea: As rightfully indicated long ago by [100] that the value of a scientist's work can be derived from the fact that it is being used by other researchers to build upon or to extend. Our assumption is that a young scholar who is new in the field of research can achieve higher degree of excellence in the career if he is able to lay his foundations on strong base. That is, researcher's publications should have strong base and he should be able to continue research activity with the same standard as his mentors. He should be able to extend some good work from existing researchers and show his worth. An ontology of citation's context and reasons has been defined in [101]. They have proposed a taxonomical hierarchy of eight object properties for citations reasons. The "Extend" has been identified

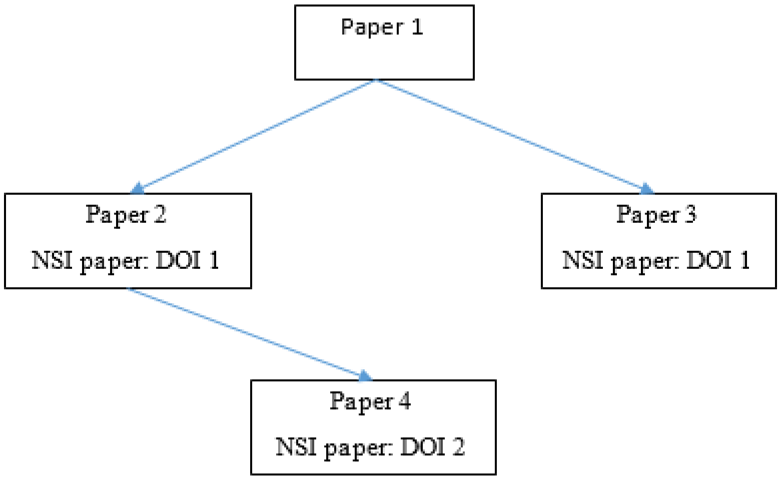

FIGURE 3.6: Extend relationship among papers (Example)

as one of the main reasons for citations [102]. By Extend it means to spread from a central research to a wider solution [101]. According to [102], to Extend someone's work is conceptual, organic, evolutionary and confirmative. That is, a reference is surely related to the concept presented in the referring paper and the reference is truly needed for the understanding, worked out the content of the paper, foundations are on the referred paper and it is correctly referring to the paper [103].

What we propose is that, if a young researcher extends some influential work, then the researcher is likely to have more potential and there are more chances of his excel in future. Based upon this idea, we have proposed a new 'NS-Index', which would be helpful in identifying future potential of young researchers. Below we have defined NS-Index for papers and authors.

NS-Index of a Paper NS-Index for a paper (NSI(P)) is defined as, "a paper has NS-Index of n if n number of papers have extended this paper." Let us consider an example, suppose paper 1 is extended by paper 2 and paper 3, then the NS-Index value for paper 1 would be '2'. In the Fig. 3.6, direction of the arrow shows that paper 1 is extended by paper 2 and paper 3. Similarly, Paper 2 is extended by paper 4, so according to our definition, NS-Index value for paper 4 is '1'.

Table 3.7: NS-Index value for papers (Example)

Paper	NS-Index Value
Paper 1	2
Paper 2	1
Paper 3	0

For future, we propose that each paper which has Extended some previous work, should include information about the extended paper. Our proposal is that as key term/keywords are part of every paper, there should be "NSI Paper" term in papers, as shown in Fig. 3.6 **(NSI Paper: DOI of paper)**. This term would refer to DOI of the paper being extended. It would make it easy to gather all the information of the papers that are extended by other papers.

Fig. 3.6 shows the levels of hierarchy, to further elaborate the proposal we can go on next levels of hierarchy. We have elaborated/explored the impact of papers on different levels. That is, NS-Index for paper 1 at level 1 is '2', as it is directly extended by 2 papers. We call it level 1 and level 1 shows the direct relationship with respect to extending the paper. On next level, paper 2 is extended by paper 4, with respect to paper 2 it is level1, and it shows indirectly extending the paper 1. Table 3.7 shows the number of direct extending relationship for the papers shown in Fig. 3.6. By going down the hierarchy levels increase so as the value of NS-Index of a paper.

NS-Index for authors NS-Index or NSI of an author is defined as, "the sum of NS-Index of all of his/her papers". Symbolically it can be represented as:

$$NS{-}Index\ of\ Author NSI(A) = \sum_{i=1}^{n} NSI(Pi(A)) \tag{3.5}$$

Where n is the total number of papers of author A, NSI (Pi(A)) shows the NS-Index value of ith paper of author A. Keeping in view limited information available for young researchers, to identify future potential of young researchers, 'NS-Index of the papers extended by young researchers' will be used.

To predict the impact of young researchers, we have to scrutinize all the papers and find out that how many papers have extended any other paper. It is a very resource consuming and time taking procedure to collect data for evaluation of this idea for large data sets due to the following reasons:

- We have to find those papers of young researchers which have extended some previous work, because it is not necessary that a young researcher writes a paper and it extends some work. For this purpose, we have to read and understand the paper to identify whether the paper extends some work and identify the paper that has been extended.
- Once we find the paper (Q) being extended in a young researcher's paper (P), it is further time consuming job to find the NSI of Q, as Q may have hundreds of citations and to find NSI(Q), we have to understand all those citations to get the count of the papers that have extended the paper Q (which will give us NSI(Q).

For these two reasons, we propose that "NSI paper" should be made a part of a research article structure like title,keywords or author affiliations. It will give us a graph of papers that have this certain relationship with each other. This will help to understand the whole chain of a concept and the things that contributed in the evolution of knowledge. Moreover, it will represent a more solid contribution of a researcher than the simple citation count which is also the basis of h-index.

In order to present the proof of our concept, we have adopted a simpler environment with two conditions/assumptions. Firstly, we have experimented on a small data set, that is, "we have randomly selected a small set of young researchers, identified their publications which have extended some previous work. Secondly, rather than trying to get the NSI of the papers being extended, we have considered following variations:

- Citation Count of those papers which have been extended in young researchers' papers
- Count of the papers which have been extended in young researhcers' papers

- Count of the papers which have extended young researchers' papers.

With these assumptions, we have performed different experiments to prove the effectiveness of the proposed index.

Experiments

Extended relationship among papers is explored by performing certain experiments. We have studied the extended relationship among papers and compared current impact of authors, future five years impact with the citation count of extended papers. For this purpose, we have randomly considered 23 researchers having h-index '1' in 2007. As mentioned earlier, researcher having h-index less than 4 is considered as a young researcher in literature [24]. So this data set of 23 researches forms our young researchers data set . We have considered all the papers written by those researchers till 2007. By carefully reviewing each paper, we have identified the papers, which have extended some previous work. We have marked the references and in the next step considered those references which were extended by these papers. All the detail of 23 authors, their papers and the papers extended by these authors are given in appendix E.

Symbolically

Let A be a young researcher, he has written some papers $\{P_1, P_2, \ldots, P_n\}$

$$\text{A writes } \{P_1, P_2, \ldots, P_n\}$$

Suppose, for some papers P_i,

$$P_i \text{ Extends } Q_i$$

P_i Extends Q_i

For all these papers Q_i which are extended by researcher A, citations data is collected from Google Scholar , all these citations are summed up , i.e.

$$\sum_{i=1}^{n} CC(Qi) \tag{3.6}$$

Let us explain this by giving an example, let us consider an author A, who has written papers P_1, P_2 and P_3 in or before 2007. Suppose in paper P_1, author A has extended a paper Q_1 and paper P_2 has extended paper Q_2, whereas paper P_3 has not extended any previous work. Using Google Scholar we found the citation count of paper Q_1 and Q_2 till 2007 and summed up the citations of both the papers.

In Table 3.8, we have given an example from data set, author having authorID 1434309 has written three papers till 2007. Publication having ID 977842 have not extended any previous work , whereas two papers have extended some previous work and the citations of those two papers till 2007 are 5 and 18. We will sum up these and finally we would have 23 citations of extended papers in total for this author.

To have data symmetrical/comparable with h-index values we have normalized total citations using min-max normalization technique. First we have compared the number of extended papers (Q_i) and their citations with future impact of researchers. For future impact of researchers, we have considered researcher's one year and 5 years h-index value, i.e. h-index value of researchers in 2008 and 2012.

Further exploring the impact to next level, we have considered those papers which have extended the papers P_i of our considered researchers A. Considering the above example author having authorID 1434309 has written three papers till 2007. His two papers have extended some previous work. Let authorID 1434309 be A, two papers which have extended some previous work be P_1 and P_2 and the work they have extended is Q_1 and Q_2,

$$P_1 \text{ extends } Q_1$$

$$\&$$

$$P_2 \text{ extends } Q_2$$

Now one of his paper say P_1 is extended by some other paper say X_1.

Table 3.8: Example of one author from data set

AuthorID &Name	**PaperID**	**Extended (ref)**	**Extended (Title)**	**No of Citations of Extended Papers (till 2007)**
1434309 (Stefan Galler)	977842	N/A	N/A	N/A
	1014810	R. Bloem, S. Galler, B. Jobstmann, N. Piterman, A. Pnueli, and M. Weiglhofer. Automatic hardware synthesis from specifications: A case study. In Proceedings of the Conference on Design, Automation and Test in Europe, 2007.	Automatic hardware synthesis from specifications: A case study.	5
	1397985	Piterman, N., Pnueli, A., Sa'ar, Y.: Synthesis of reactive(1) designs. In: Proc. Verification, Model Checking, and Abstract Interpretation, pp. 364–380 (2006)	Synthesis of reactive(1) designs.	18

$$X_1 \text{ extends } P_1$$

Now number of papers of researcher A extending some previous work is '2' and number of papers extending researcher A's work is '1'. Hence total number of Extend relationship count for author A would be '3'.

For previous experiment we considered 23 authors, but for this next level we have considered 8 authors from those 23 authors. It is necessary to mention here that to find out the papers which have extended author A's papers, we have just considered citations of 2007. Now for the 8 authors under consideration we have compared the number of papers extended by our under consideration researchers and number of papers of these researchers which were extended by some other researcher with actual future impact of researcher and predicted future impact of researcher.

3.7 Impact Prediction for Another Domain

From our previous research questions, an efficient set of parameters would be identified and the proposed model would be suitable for the field of Computer Science. Keeping in view the significance of h-index for other fields we have applied for some other field i.e. Physics. But before evaluating the features' performance for Physics, we have compared the behaviour of h-index with other indices when applied on same field. For this purpose we considered two recently proposed indices i.e. completing-h and k-index. h-index has proven to be incompatible to comparison of scientists in different domains since originally it was proposed for individual evaluation. Addrssing this shortcoming, Dienes argued that community role should be considered while evaluating an author. He points out that there is an intrinsic deficiency in basic definition of h-index. With the inclusion of community factor this deficiency can be overcome/removed and cross domain comparison is also possible[104]. Similarly Kinouchi et al. have proposed a new centrality index called K-index[11]. K-index considers the network of papers and authors. According to authors, K-index addresses many drawbacks of h-index. It is not contingent on number of publications, it not only addresses the issue of self-citations, but also has large classification range. Moreover, it is able to detect scientific counterfeits. Authors have claimed that K-index has so many advantages over h-index, but considering a small sample of researchers from field of Physics makes it debatable.

The above mentioned two recently proposed author ranking indices i.e. completing-h and K-index assure to fulfil the deficiencies of the h-index. However, there is no study that evaluates them on a common comprehensive data set. Motivated by this fact, this study compares h-index,completing-h and k-index using correlation, author rankings evaluated on award winners benchmark and a comprehensive data set that relates to the field of Computer Science. There is no standard benchmark data set available to evaluate the performance or effectiveness of different indices. There are some studies in which award winners or Nobel Prize winners of respective fields are used as benchmark [105, 106]. According to [28] high profile scientists (e.g. Nobel laureates and members of National of Academy of Sciences) generally score higher h index values. Thus according to our proposal/assumption, the index which succeeds in bringing award winners in top ranks is the most successful index.

We decided to use the awardees data set for the field of Computer Science as benchmark. In total, we have worked on 24 awards which are awarded by two well-known organizations in CS i.e. ACM and IEEE. Some of the awards that we have considered include ACM Fellow, IEEE Technical achievement Award and Turing Award. Complete list of the awards and names of award winners are given in [46].

To evaluate the performance of these three indices, we have used awardees as benchmark. The idea is to rank the authors in descending order on the basis of values of these indices and then verify which index succeeds in bringing highest number of award winners in top ranks. First, we have made separate ranked lists of authors on the basis of their completing-h, K-index and h-index values. We have marked all the award winners found in our data set and their position in these ranked lists. We have identified how many award winners are found in top 10% of these ranked lists, then in next 10–20%, followed by 20–30%, 30–40%, 40–50% and then below 50% for all the ranked lists. Spearman and Pearson correlation coefficients are determined for these three indices. The purpose of finding correlation is to check how much similar results these indices produce. Spearman correlation would find correlation in ranked lists of authors i.e. it would evaluate whether the ranked lists acquired from different indices are similar or different.

Table 3.9: Physics Data set Statistics

Category	Instances
Total Number of Physics Authors	226,373
Number of authors having publication in 2007 or earlier	113,554
Total Number of Physics publications	105,858
Number of publications in or before 2007	51,382

Table 3.10: Number Of Author Records in Physics Data sets

No.	**Data set**	**No. of Records**	**Training Data**	**Test Data**
P1	DatasetFull	113,554	90,795	22,759
P2	DatasetExp5-12	40,883	32,790	8,093

3.7.1 Features Evaluation for the Domain of Physics

To check the applicability of proposed model for other fields, that is how successful the model is for other fields, we have evaluated our proposed model on the field of Physics. It may lead us to build/find/compute a relationship/correspondence between h-index prediction of different domains. RQ4 is addressed here , i.e.

RQ4: To test the applicability of devised/proposed model for some other domain

Data set

We have used data set for the field of physics acquired from Microsoft Academic known as Open Academic Graph(OAG)[7] (Sinha et al., 2015). OAG contains data for multiple disciplines like Computer Science, Physics, Chemistry , Engineering and many other. Detail of data set acquired for physics domain is given in Table 3.9.

[7]https://www.aminer.org/open-academic-graph

3.8 Conclusion

The proposed methodology addresses the research questions which are focused on scientific impact prediction of researchers. Impact of a researcher can be predicted based on currently available information of a researcher. This information is presented in the form of different features. Seeking out the set of such features which can be helpful in effectively predicting future impact, forward feature selection is applied. Addressing the problem of young researchers impact prediction, it is proposed that such a approach should be adopted for young researchers, which can be helpful in prediction of their impact early in their career. Results of all the experiments performed to get answers for the research questions are presented and discussed in next chapter (Chapter 4).

Chapter 4

Results And Discussion

This chapter furnishes the results of different experiments performed as described in detail in the chapter of methodology. Results are organized/presented with respect to the research questions.

4.1 Results for Comparison of Models

We evaluated existing models[24, 31] found in literature on the ArnetMiner data set. Existing models were subject to some constraints on the data set. We have also applied those constraints and reevaluated them for this data set addressing our RQ1.

RQ1: Whether the existing sets of parameters/Models for h-index prediction can be validated considering comprehensive data set from the field of Computer Science?

In order to predict future h-index of authors, we have considered data set values till 2007. For all training and testing, we have only used values of features which were calculated over previous years. Like, to predict h-index for 2008, all the features values till 2007 were considered.

So our base year is 2007 and target years are 2008-2012. With existing approaches

TABLE 4.1: Data sets for RQ1 and RQ2

No.	Total Number of Records	Training Data	Test Data
D1	938,204	750,028	188,176
D2	3,435	2,757	678
D3	9,793	7,827	1,966

we have applied regression models considering different parameters/features proposed in those approaches. As discussed in methodology we have considered full data set, D1 and partitioned data sets D2 and D3.

- Data set D1 is comprised of all the authors who have published till 2007
- Data set D2 is comprised of researchers having h-index greater than 10
- Data set D3 is comprised of researchers having h-index greater than 4 and having experience of 5 to 12 years, .

Table 4.1 shows the sizes of our data set, training data and test data sets. Note that for all the training and evaluation, we only used features calculated over previous years. For example, when calculating citations for papers in 2007,all the references of the papers published till 2007 were considered. We have fitted regression equations for all Acuna and Dong features. 80% is training data and 20% testing data. R Squared values are computed for the equations fitted for all the combinations. Summary of results for these parameter combinations are given in Table 4.2. These parameters include current h-index, square root of no of publications, years since publishing first article, number of distinct journals published in,number of articles in top journals mentioned as Acuna_features_set. Where as current h-index, number of publications, number of years since first paper, average citations per paper and number of coauthors as Dong_features_set. From Table 4.2 it is quite clear that fitted models very well predicted the one year value. The model predicted future h-index for one year having R^2 value approx.0.96, but for five years the predictions are little worse than one year, for five

TABLE 4.2: RQ1 Results for DatasetFull (D1)

	Acuna_features_set			Dong_features_set		
Year	R^2	**Max_error**	**RMSE**	R^2	**Max_error**	**RMSE**
2008	0.96	3.49	0.28	0.96	3.36	0.28
2009	0.93	7.91	0.41	0.93	8.23	0.42
2010	0.91	10.5	0.51	0.9	10.9	0.52
2011	0.89	11.2	0.59	0.88	11.7	0.61
2012	0.87	11.8	0.67	0.86	13.2	0.7

TABLE 4.3: RQ1 Results for DatasetH10 (D2)

	Acuna_features_set			Dong_features_set		
Year	R^2	**Max_error**	**RMSE**	R^2	**Max_error**	**RMSE**
2008	0.97	3.14	0.78	0.97	3.06	0.78
2009	0.95	4.53	1.18	0.94	4.58	1.2
2010	0.92	7.09	1.54	0.92	7.25	1.57
2011	0.89	9.01	1.9	0.89	9.26	1.94
2012	0.86	9.89	2.22	0.86	10.2	2.26

years R^2 value is around 0.86. It is quite obvious that the fitted model is performing well in predicting short term impact but with longer periods the prediction of h-index declines. As mentioned before we have considered subset of data set having those records where h-index of an author is greater than 10. Table 4.3 comprises of the results for this data set. R^2 for one year prediction is 0.97 and for five years it is 0.86. Similarly equations are fitted for data set comprising of researchers whose h-index is 4 or less and who have 5 to 12 years of experience. Here we can see decline in performance and found that R^2 for one year is 0.93 and for five 0.78 as shown in table 4.4.

4.2 Prediction Model for Computer Science

For identification of optimum set of parameters, ideally we should calculate all possible combinations of variables to fit regression models, but it would not be

TABLE 4.4: RQ1 Results for DatasetH4exp5-12 (D3)

	Acuna_features_set			**Dong_features_set**		
Year	R^2	**Max_error**	**RMSE**	R^2	**Max_error**	**RMSE**
2008	0.93	3.33	0.59	0.93	3.34	0.6
2009	0.88	3.73	0.89	0.88	3.9	0.92
2010	0.84	6.16	1.2	0.82	6.43	1.24
2011	0.81	7.9	1.4	0.8	8.18	1.46
2012	0.78	9.52	1.65	0.77	9.98	1.71

feasible to consider all possible combination of features/parameters. So we have adopted forward feature selection/stepwise forward regression.[82, 107]

TABLE 4.5: Features and Brief Description

No.	**Features**	**Description**
1	2007_h_index	Current h-index of an author (2007)
2	no_publications	Total Number of publications till now (2007)
3	years_since_start	Number of years since first paper of an author was published
4	square_root_publications	Square root of number of publications of an author
5	no_article_as_last_author	Number of papers published as a last author
6	proportion_last_author	Proportion of papers as last author with all the papers
7	no_article_as_first_author	Number of papers as first author
8	proportion_first_author	Proportion of papers as first author with all the papers
9	no_of_distinct_venues	No of different venues papers of an author are published in

Continued on next page

Table 4.5 – continued from previous page

No.	Features	Description
10	no_of_IF_journals	How many publications of an author are in Impact factor journals
11	journal_if_3	How many publications of an author are in journals having impact factor 3 or above.
12	distinct_but_only_if	No of different but only Impact factor venues, papers of an author are published in
13	avg_citations	Average of total citations received by papers of an author
14	IF_citations	Only Impact Factor citations of papers of an author
15	citations_diff_hindex	Sum of Difference in current h-index and citations of individual papers
16	average_citations_diff_hindex	Average of sum of difference in current h-index and citations of individual papers
17	avg_coauthors_per_article	Average number of coauthors per paper of an author
18	collaborations	Number of times author has worked in at least one coauthor
19	no_coauthors	Total number of coauthors , author has published papers with
20	coauthors_total_H_index	Total h-index of all the coauthros
21	average_hindex_coauthors	Average of toal h-index of all the coauthros
22	m_index	Value of m-index

Referring to the data set comprising of all the researchers who have published a

TABLE 4.6: Regression Model Fitted on DatasetFull (D1)

features set RQ2_full	**2008 (1year)**	**2009 (2years)**	**2010 (3years)**	**2011 (4years)**	**2012 (5years)**
R^2	0.97	0.94	0.92	0.91	0.9
Max error	2.99	6.99	8.99	10.1	11.5
RMSE	0.27	0.39	0.47	0.54	0.6
intercept	0.0156	0.0286	0.0336	0.0398	0.0452
2007_h _index	0.9933	0.9914	1.0013	1.0143	1.0286
collaborations	0.004	0.01	0.02	0.02	0.03
years_since _start	-0.0051	-0.011	-0.015	-0.017	-0.02
no_coauthors	-0.0044	-0.012	-0.019	-0.026	-0.034
square_root _publications	0.0764	0.1645	0.2242	0.266	0.3001

paper till 2007,D1, we have split data set into training and testing set. Using the set of features given in table 4.5 we have applied forward selection using regression and neural networks.

4.2.1 Regression Models

For this purpose we have fitted regression equations using the parameters mentioned in table 4.5. Starting from one parameter at a time, we have moved forward till the R^2 value goes on increasing. Forward feature selection step by step for full data set D1 is given in appendix B.

From table 4.6 it is quite clear that fitted models very well predicted the one year value. The model predicted future h-index for one year having R^2 value approx.0.97, and a lot of improvement can be seen for five years, for five years R^2 value is around 0.90. It is quite obvious from the results that with longer periods the prediction of h-index declines. Values of coefficients Max error and RMSE are also given in table 4.6.

Considering the field of Computer Science in general, our model proposes that

TABLE 4.7: Regression Model Fitted on DatasetH10 (D2)

features set RQ2_H101	**2008 (1year)**	**2009 (2years)**	**2010 (3years)**	**2011 (4years)**	**2012 (5years)**
R^2	0.98	0.95	0.93	0.91	0.89
Max error	2.64	4.11	5.3	7.56	8.64
RMSE	0.75	1.12	1.43	1.75	1.98
intercept	-0.044	-0.1109	-0.23	-0.2008	-0.101
2007_h_index	1.0203	1.0357	1.0461	1.066	1.092
collaborations	0.0022	0.0056	0.0096	0.0128	0.0169
Years_since _Start	-0.002	-0.0012	0.004	-0.0002	-0.006
No_coauthors	-0.002	-0.0056	-0.01	-0.0137	-0.019
m-index	0.5448	1.2092	1.8627	2.2687	2.5554

the parameters current h-index, number of coauthors, number of collaborations, publications and experience or years since publishing first article play vital role in predicting future impact of a researcher.

Further applying certain constraints on researchers selection, we have applied forward selection approach considering data set D2, i.e. researchers having h-index greater than 10. table 4.7 comprises of the results after applying regression models for D2. Regression model fitted for D2 has same features except for publications, it is replaced by current m-index of researcher. R^2 value for one year is 0.98 but for five years performacne is slightly low i.e. 0.89.

For researchers having h-index greater than 4 and having experience of 5 to 12 years results for regression equations fittted are given in table 4.8.

Comparison of our proposed feature set with the existing approaches shows promising results. Fig. 4.1 shows the comparison based on R^2 values. Highest the value of R^2 , better the performance of model. Performance of our proposed model for five years is clearly outperforming the existing models.

Similarly in Fig. 4.2 RMSE and in Fig. 4.3 Max Error is drawn for three models. Lower the value of RMSE and lower the error, better the model is performing. It is quite evident that performance of the model proposed in this study is better for

TABLE 4.8: Regression Model Fitted on DatasetH4exp5-12 (D3)

features set RQ2_H4 exp_5-12	**2008 (1year)**	**2009 (2years)**	**2010 (3years)**	**2011 (4years)**	**2012 (5years)**
R^2	0.94	0.89	0.86	0.85	0.83
Max error	3.02	3.63	5.1	7.41	8.07
RMSE	0.57	0.86	1.12	1.27	1.46
intercept	-0.3461	-0.735	-1.058	-1.313	-1.554
2007_h_index	0.9464	0.8985	0.8547	0.8281	0.8181
Collaborations	0.0009	0.0031	0.0059	0.0087	0.0121
coauthors _total_H_index	0.0015	0.0029	0.0042	0.0055	0.0066
square_root _publications	0.1491	0.3296	0.4692	0.5816	0.6766

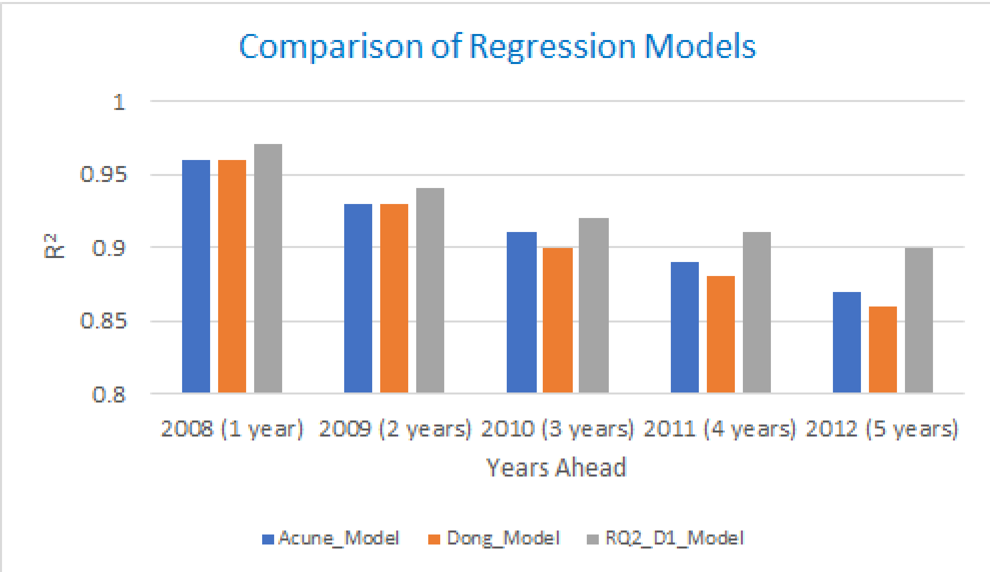

FIGURE 4.1: Comparison of Proposed Model with Existing Approaches (R^2)

long term.

4.2.2 Application of Proposed Model on Sub-Data set

We have fitted regression models considering the data of researchers available in 2007. To further evaluate the effectiveness of proposed model we have considered

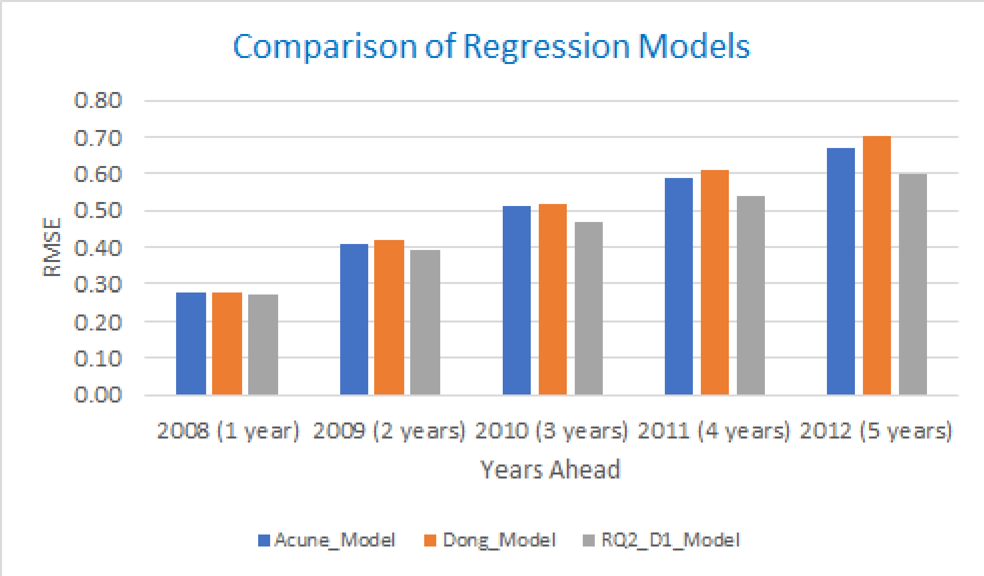

FIGURE 4.2: Comparison of Proposed Model with Existing Approaches (RMSE)

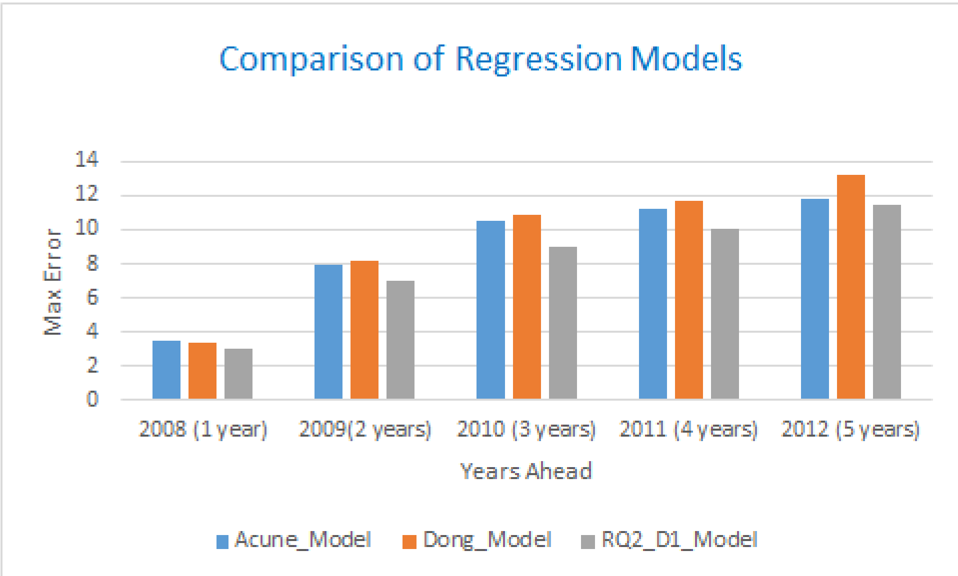

FIGURE 4.3: Comparison of Proposed Model with Existing Approaches (Max Error)

Table 4.9: Regression Model Results for D1_06

Year	R^2	**Max-error**	**RMSE**
2007	0.96	5.57	0.28
2008	0.93	7.18	0.39
2009	0.90	12.7	0.50
2010	0.87	13.3	0.56
2011	0.85	17	0.68

researchers status in 2006. Based on the credentials of researcher in 2006, we have predicted h-index for next 5 years. Results of our fitted equations are given in table 4.9. R^2 values for one year is 0.96 which is quite good and another good measure for five years i.e. 0.85.

4.2.3 Neural Networks

Considering the parameters given in table 4.5 , using forward feature selection, we have applied Neural Networks to get prediction models for h-index prediction. For this purpose we have considered data set D1 that is full data set and training and testing data are also same as described in table 4.1. MSE is used to estimate the accuracy of model. Combination of features which gave better results than others are: Current h-index, number of coauthors, square root of publications,experience of researchers,total h-index of coauthors and average citations of researchers.
With regression for same data set first four features were same, along with one additional parameter i.e. number of collaborations. Results of applying Neural Networks on D1 are given in table 4.10.
Neural Network for Acuna and Dong et al. features are given in table 4.11. Comparison of RMSE of regression model fitted and Nueral Network model fitted on DatasetFull(D1) is given in Fig. 4.4.

From Fig. 4.4 it is obvious that performance of Model fitted using regression is relatively better than Neural networks. According to Kumar, 2005 models fitted using regression performs better than neural networks for skewed data especially

TABLE 4.10: Neural Network Model Fitted for DatasetFull(D1)

Features	**2007_h_index,coauthors_total_H_index, square_root_publications,starting_year_from_2007 ,coauthors_sum,avg_citations**	
Year	R^2	**RMSE**
2008	0.94	0.36
2009	0.91	0.46
2010	0.91	0.49
2011	0.88	0.60
2012	0.90	0.66

TABLE 4.11: Existing Approaches Using Neural Networks for DatasetFull(D1)

Parameters	**Acuna_features_set**		**Dong_features_set**	
Year	R^2	**RMSE**	R^2	**RMSE**
2008	0.93	0.38	0.93	0.37
2009	0.91	0.48	0.90	0.49
2010	0.89	0.55	0.90	0.54
2011	0.87	0.63	0.88	0.61
2012	0.85	0.72	0.85	0.73

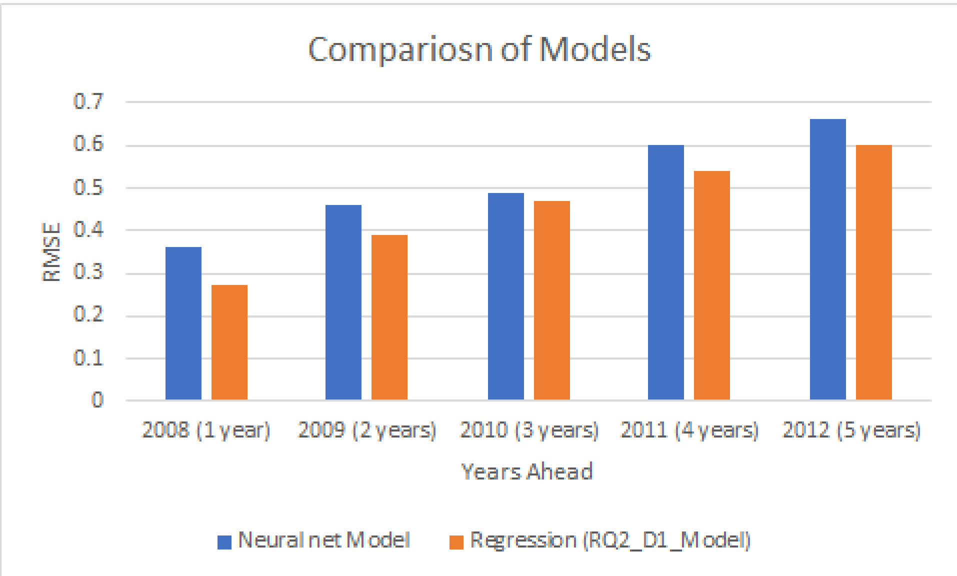

FIGURE 4.4: Comparison of Regression and Neural Networks Models

Table 4.12: Number of Author Records in Young Researchers Data sets

No.	Data set	No. of Records	Training Data	Test Data
DY1	exp-lesss-than-3	257,845	206,195	51,650
DY2	exp-lesss-than-4	306,334	245,339	60,995
DY3	hindex-less-than-4	910,242	729,638	180,604

when dependent variable is skewed. A skewed distribution is the distribution with tail on its either side. Positively skewed distribution has tail on its right side and negatively skewed distribution has tail on its left side. In this case dependent variable is h-index, histograms to represent the skewness for this data is shown in Appnedix D. There is right tail in distribution shown in Fig. D.1 to D.3 in Appendix D. It shows that the dependent variable in this case is also skewed.

4.3 Impact Prediction for Young Researchers

Keeping in view the reservations and opinions regarding the prediction of h-index for young researchers, we have considered the case of young researcher's separtely. From literature we have identified three different divisions of data sets for young researchers. One division comprises of researchers having experience less than 3 years, 2nd is having experience less than 4 years and 3rd is researchers having h-index value less than 4 . Table 4.12 shows the number of records, and number of records considered as training data and testing data.

To check the validity of proposed models for whole data set for young researchers, we have applied Acuna et al parameters, Dong et al. pararmeters and our models parameters on data set DY1, DY2 and DY3(shown in Table 4.12).

R^2 values for fitted models are shown in Table 4.13. Though the results for our fitted model is better than Acuna and Dong et al results, but still it is not satisfactory performance. Further we have decided to check th impact of coauthors of these young researchers on their future h-index prediction. keeping in view this we have considered some other parameters which are shown in Table 4.14.

As calculations for all the parameters for all the data set was quite laborious.

Table 4.13: Regression Model Fitted for young researchers

Data set	Feature Set	2008 (1year)	2009 (2years)	2010 (3years)	2011 (4years)	2012 (5years)
DY1	features_set_RQ2_full	0.67	0.56	0.54	0.54	0.55
DY1	Acuna_feature_set	0.60	0.42	0.37	0.33	0.32
DY1	Dong_feature_set	0.60	0.41	0.36	0.32	0.31
DY2	features_set_RQ2_full	0.73	0.63	0.61	0.6	0.6
DY2	Acuna_feature_set	0.67	0.51	0.46	0.44	0.41
DY2	Dong_feature_set	0.67	0.51	0.47	0.44	0.41
DY3	features_set_RQ2_full	0.89	0.81	0.78	0.76	0.74
DY3	Acuna_feature_set	0.83	0.7	0.62	0.57	0.53
DY3	Dong_feature_set	0.83	0.7	0.62	0.57	0.52

Table 4.14: Additional Features for Young Researchers

No.	Features	Description
1	highest_hindex_coauthors	highest h-index value among coauthors
2	IF_3_citations	How many citations of papers of an author are in journals having impact factor 3 or above.
3	no_second_coauthor	number of coauthors on 2nd position with an author
4	highest_second_author_hindex	Highest h-index of author on 2nd position with author
5	sum_second_coauthor_hindex	Total of h-index of author on 2nd position with author
6	average_second_coauthor_hindex	Average of h-index of author on 2nd position with author

TABLE 4.15: Number of Author Records in Sample Young Researcher's data sets

No.	Data set	No. of Records	Training Data	Test Data
Y1	exp-lesss-than-3	12592	9992	2600
Y2	exp-lesss-than-4	14848	11884	2964
Y3	hindex-less-than-4	41945	33484	8461

TABLE 4.16: Regression Model Fitted for Sample data of young researchers

Data set	Feature Set	2008 (1year)	2009 (2years)	2010 (3years)	2011 (4years)	2012 (5years)
Y1	features_set_RQ2_full	0.6048	0.4292	0.3986	0.383	0.4004
Y2	features_set_RQ2_full	0.6757	0.5214	0.4931	0.4812	0.4828
Y3	features_set_RQ2_full	0.8352	0.7238	0.6673	0.632	0.6165

So before considering additional parameters, we have considered random sample from this data set. we have selected random sample from the authors who have published till 2007. Number of authors considered for random sample were 193257. out of 193257 after applying constraints for young researchers, number of young researchers in random sample, are shown in Table 4.15. Our proposed model which we have applied on DY1,DY2 and DY3 aer also applied on Y1,Y2 and Y3 sample data sets. Results of fitting the models are shown in Table 4.16. Same pattern can be seen in these results as on full data set for young researchers. Results for young researchers are not very encouraging while considering the proposed models.

Keeping in view this shortcoming of proposed model, we have considered some other features for young researchers. Considering the features proposed in Table 4.14 along with the features identified in features_set_RQ2_full, we have applied forward feature selection.

Table 4.17 to Table 4.19 shows feature set and results acquired after having forward feature selection with some new parameters for data set Y1,Y2 and Y3 respectively.

It is quite obvious that after applying new parameters , results are not promising for young researchers. Keeping in view above results, it was realized that there is need to propose a new index which takes into account different aspects with

Table 4.17: Regression Model Fitted on Sample young researchers Data set (Y1) for features_set_RQ3_Exp_31

features set RQ3_Exp_31	**2008 (1year)**	**2009 (2years)**	**2010 (3years)**	**2011 (4years)**	**2012 (5years)**
R^2	0.61	0.43	0.4	0.39	0.41
Max error	2.75	3.31	4.97	5.66	5.36
RMSE	0.52	0.7	0.78	0.84	0.88
intercept	0.0645	0.0917	0.0808	0.0989	0.1318
2007_h_index	0.8515	0.7522	0.7142	0.699	0.6878
collaborations	0.0043	0.0116	0.0182	0.0266	0.0357
square _root _publications	0.1092	0.2574	0.3638	0.4183	0.4431
No_coauthors	-0.006	-0.0137	-0.022	-0.032	-0.043
highest _hindex _coauthors	0.0375	0.0542	0.0644	0.0656	0.0746

Table 4.18: Regression Model Fitted on Sample Young Researchers Data set (Y2) for features_set_RQ3_Exp_41

features set RQ3_Exp_41	**2008 (1year)**	**2009 (2years)**	**2010 (3years)**	**2011 (4years)**	**2012 (5years)**
R^2	0.68	0.52	0.5	0.48	0.49
Max error	2.55	3.95	3.71	5.91	8.22
RMSE	0.52	0.69	0.77	0.83	0.89
intercept	0.0148	0.0439	0.0319	0.0392	0.0525
2007_h_index	0.8616	0.7735	0.7344	0.7235	0.7127
collaborations	0.0039	0.011	0.0177	0.0253	0.033
square _root _publications	0.1459	0.2851	0.3893	0.4523	0.4897
No_coauthors	-0.005	-0.0135	-0.022	-0.031	-0.04
highest _hindex _coauthors	0.037	0.052	0.0604	0.0634	0.0736

TABLE 4.19: Regression Model Fitted on Sample Young Researchers Data set (Y3) for features_set_RQ3_h_41

features set RQ3_h_41	**2008 (1year)**	**2009 (2years)**	**2010 (3years)**	**2011 (4years)**	**2012 (5years)**
R^2	0.84	0.74	0.69	0.65	0.64
Max error	2.8	4.1	5.27	6.04	5.91
RMSE	0.39	0.54	0.63	0.7	0.74
intercept	0.0551	0.1	0.13	0.14	0.16
2007_h_index	0.9406	0.89	0.87	0.85	0.85
collaborations	0.005	0.01	0.02	0.03	0.03
square _root _publications	0.0409	0.09	0.12	0.15	0.17
No_coauthors	-0.005	0	0	0	0
average _h_index _coauthors	0.008	0.01	0.02	0.02	0.03

respect to young researchers.

4.3.1 Proposed Index (NS-Index)

As evident from the results that the prediction of h-index for young researchers does not provide promising results. Hence it would not be wise enough to use h-index as a metric/evaluation criteria for recruitment or other decisions for young researchers. Considering these findings we have proposed a new 'NS-Index'. NS-Index is based upon the 'Extend' relationship among papers. For young researchers, we are considering the NS-Index of papers which are extended by young researchers. A paper may have hundreds of citations and it would be a project in itself i.e. to scrutinize all the citations of a paper to find the 'Extend' relationship. So we have considered citations count of the papers extended by young researchers.

Experimental Results

Following are the results of experiments done to determine the usefulness of proposed idea in identifying potential young researchers. We have considered 23

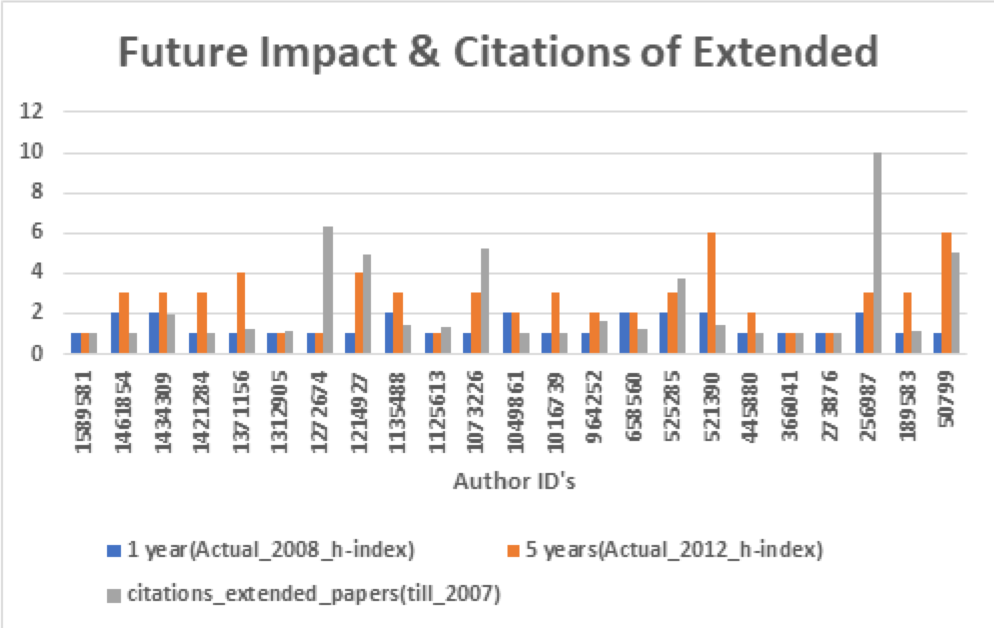

FIGURE 4.5: Future Impact and Citations of Extended Papers

young researchers having h-index value '1' in 2007 and their publications till 2007. After careful evaluation of all the references of these publications, we have identified those papers which were extended by these 23 authors. All the details of the authors, their papers and the papers extended by these 23 authors are given in appendix E. In Table 4.20 Author ID's, their future one-year h-index value (i.e. of 2008) and 5 years h-index value (i.e. of 2012) and sum of citations of the papers extended by these authors are mentioned.

We have mapped citations of those papers (till 2007) which were extended by these researchers along with their future h-index values of one year and five years i.e. h-index in 2008 and 2012 as shown in Fig. 4.5. It is evident that for most of the cases trend for citations of extended papers and h-index of researchers 5 years in future are similar.

Fig. 4.6 displays the future impact i.e. future h-index of these young researchers and simply the number of papers extended by these young researchers based on the data shown in Table 4.21. Here also trend line shows similar trend for 5 years future h-index and number of papers extended. These findings encouraged us to move in this direction and highlighted the importance of extending someone's work over just using or referring to someone's work. For further experiments we have selected 8 researchers from these 23 researchers. Researchers were selected

Table 4.20: Future Impact and Citations of Extended Papers

AuthorID	**One-year (2008 h-index)**	**Five-years (2012 h-index)**	**Total citations extended (2007)**	**Normalized Total citations extended(2007)**
1589581	1	1	0	1
1461854	2	3	1	1.022727
1434309	2	3	23	1.522727
1421284	1	3	0	1
1371156	1	4	5	1.113636
1312905	1	1	3	1.068182
1272674	1	1	131	3.977273
1214927	1	4	95	3.159091
1135488	2	3	10	1.227273
1125613	1	1	8	1.181818
1073226	1	3	104	3.363636
1049861	2	2	0	1
1016739	1	3	0	1
964252	1	2	16	1.363636
658560	2	2	5	1.113636
525285	2	3	66	2.5
521390	2	6	11	1.25
445880	1	2	0	1
366041	1	1	2	1.045455
273876	1	1	2	1.045455
256987	2	3	220	6
189583	1	3	4	1.090909
50799	1	6	98	3.227273

Table 4.21: Future Impact and No. of Extended Papers

AuthorID	**One-year (2008 h-index)**	**Five-years (2012 h-index)**	**No. of papers extended**
1589581	1	1	0
1461854	2	3	1
1434309	2	3	2
1421284	1	3	0
1371156	1	4	2
1312905	1	1	1
1272674	1	1	2
1214927	1	4	3
1135488	2	3	2
1125613	1	1	2
1073226	1	3	5
1049861	2	2	0
1016739	1	3	0
964252	1	2	4
658560	2	2	1
525285	2	3	3
521390	2	6	1
445880	1	2	0
366041	1	1	1
273876	1	1	1
256987	2	3	3
189583	1	3	1
50799	1	6	6

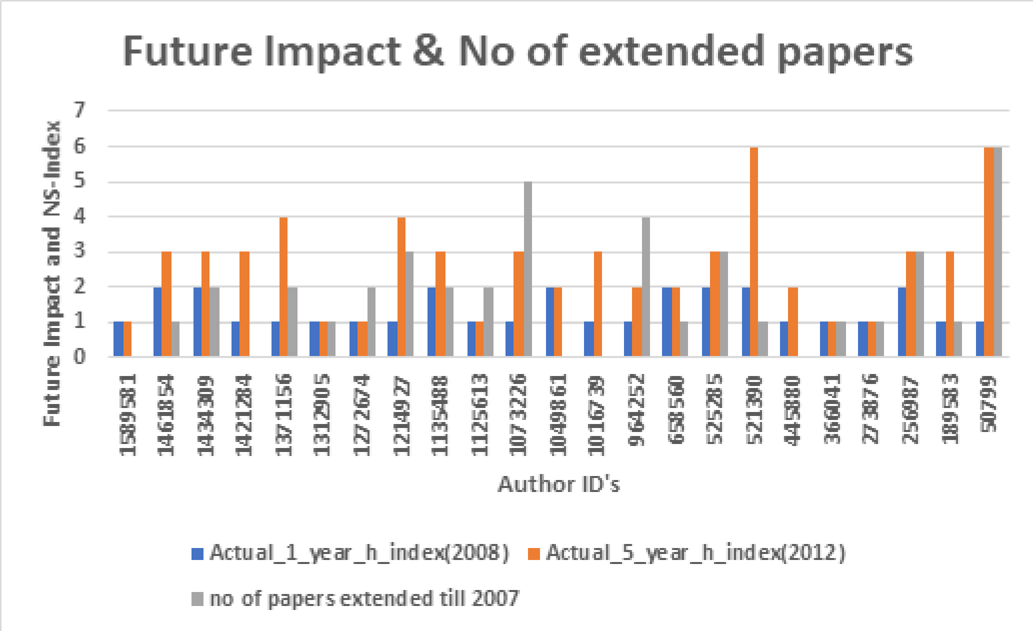

Figure 4.6: Future Impact and Extended Papers

keeping in view the diversity with respect to extending the papers and citations, e.g. author who has not extended a single paper (1589581), author whose all the papers have extended some work(50799) and so on. Further we have considered next level of extending the papers. By next level we mean by considering those papers of these researchers which are extended by someone else. Number of papers extended by under consideration 8 researchers , number of papers of these researchers which are extended by some one else and the sum of these extend relationships are shown in Table 4.22. Now in Fig. 4.7, we have mapped the total number of extended relationship papers, that is sum of the number of papers extended by young researchers and number of their papers which were extended by someone else, alongwith future impact of researchers. Fig. 4.7 shows the similar trend for 5 years future h-index and total number of extend relationship papers on both levels.

To compare the performance of our proposed Index based on Extend relationship among papers with regression models proposed earlier, we have shown trend of future actual 5 years h-index value and predicted 5 years h-index values using regression models in Fig. 4.7. It is obvious that trend of actual and predicted values for next 5 years are dissimilar.

TABLE 4.22: Future Impact and No. of Extended Relationship

AuthorID	One-year (2008 h-index)	Five-years (2012 h-index)	No papers extended by young researchers	No papers extended of young researchers	Total no. of papers extended or extended by
1589581	1	1	0	0	0
658560	2	2	1	1	2
1125613	1	1	2	0	2
1135488	2	3	2	0	2
521390	2	6	1	2	3
525285	2	3	3	2	5
1214927	1	4	3	0	3
50799	1	6	6	1	7

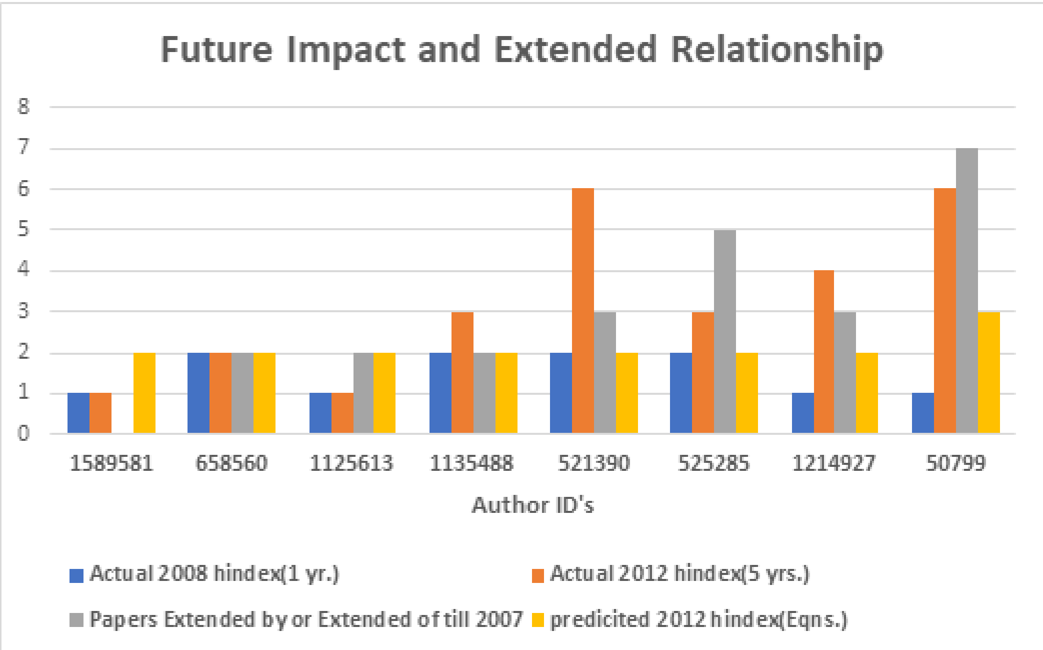

FIGURE 4.7: Future Impact and Extended Relationship

Hence, It is manifested through the results/ graphs that whether it's simply number of papers extended or the citation count of the extended papers, prediction of future potential of researchers for next five years is better represented as compared to the values calculated by using regression equations.

t-test

Table 4.23: Analysis of Regression Model and Extend Relationship

	Papers Extend	**Predicted h_index_2008**
Mean	2.285714	1.128463
Variance	3.450549	0.001186
Observations	14	14
e Pearson Correlation	0.196067	
Hypothesized Mean Difference	0	
DF	13	
t Stat	2.339144	
P($T < t$) one-tail	0.017973	
t Critical one-tail	1.770933	
P($T < t$) two-tail	0.035946	
t Critical two-tail	2.160369	

t-test is applied to check the hypothesis that the predicted values of h-index obtained after applying proposed regression model and future impact values for young researchers based on extend relationship are same or different. Outcomes of t-test applied on the predicted h-index values for one year and values obtained on the basis of extend relationship are given in Table 4.23. There was a significant difference between predicted values and extend relationship values (P value = .01 & p value=0.03).

Correlation

We have also calculated correlation of the predicted values of h-index and total extended values i.e. extended by young researchers plus their number of papers which are extended by some one till 2007 with future 5 years h-index value of young researchers as shown in Table 4.24, highest correlation of future 5 years h-index value is found with sum of number of papers extended by young researchers. Inverse and low value of correlation exists among 5 years future h-index value with its predicted value using regression equations. Correlation among predicted and

TABLE 4.24: Correlation of Actual future h-index with Extend Relationship and Regression Model predicted values

Correlation	**Extend_ Relation_Values**	**Regression_Model _predicted_values**
Actual_2008_h_index	0.297429	0.068166
Actual_2009_h_index	0.715697	0.274775
Actual_2010_h_index	0.632869	0.285762
Actual_2011_h_index	0.793169	0.272931
Actual_2012_h_index	0.749353	0.26397

TABLE 4.25: Correlation of Extend Relationship with Regression Model Fitted for Subsequent Years

Correlation	**Predicted h-index by Regression Models**			
	2008_data	**2009_data**	**2010_data**	**2011_data**
2009_h_index	0.661796			
2010_h_index	0.533388	0.92156		
2011_h_index	0.484059	0.911058	0.960373	
2012_h_index	0.553141	0.917134	0.960035	0.989167

actual values of h-index is very low, whereas Extend relationship results show significantly good performance especially for future four years h-index values. Hence, It is manifested through the correlation that with number of papers extended, prediction of future potential of researchers for next five years is better represented as compared to the values calculated by using regression equations. With young researcher having first year in field , it is quite evident that NS-Index based on Extend relationship shows better results in predicting future impact of these researchers.

Further to check after how many years regression equations performs better than extend relationship. Previously regression model was fitted using data of researchers in 2007, In these experiments, to fit regression models for subsequent years, we have considered the values of different variables using researchers data in 2008, 2009, 2010 and 2011 respectively. We have fitted regression models for the researchers data in 2008 , 2009, 2010 and 2011. Correlation of actual and predicted values are shown in Table 4.25.

From table 4.25, correlation among predicted and actual values improves after 3 years in the research field. it is inferred from this experiment that for young researchers, our proposed regression model produces better results after a researcher spends three years in a field. So in light of these experiments, it is proposed that NS-Index should be used to assess young researchers in their early career 2 to 3 years, afterwards our proposed h-index prediction regression model can be used to assess their future impact.

4.4 Impact Prediction for Physics

In this section we are presenting the findings of comparing h-index with different indices applied on same field, followed by the results of applying the proposed feature set for researchers from the field of Physics. Our findings indicate that all of these indices are highly correlated as far as Pearson Correlation is concerned. However when Spearman rank correlation was applied, the correlation among ranked lists was relatively low. It implies, that although indices are highly correlated, but the ranked lists obtained on the basis of these indices are moderately correlated. Actually, Pearson correlation represents linear relationship between two variables, whereas Spearman rank correlation measures monotonic relationship which can be nonlinear [108]. It is quite interesting to note that Spearman rank correlation of sample data set and award winner's data set between K-index and h-index is low. It implies that rankings for h-index and K-index deviate. Results of correlation among three indices are presented in Fig. 4.8

. To further evaluate the rankings by these indices we have compared against award winner's data. Authors are ranked according to their completing-h, K-index and h-index values separately. From these rankings we have evaluated the occurrence of award winners in these ranked lists. From Fig. 4.9, it is quite clear that all the indices have succeeded in identifying high percentage of award winners in top ranks, i.e. top 10 percent. For example of all the award winners found in our sample data set, completing-h succeeded in bringing 79% of award winners in top 10% researchers whereas 82% were brought by K-index and 76% by h-index. whereas for top 20% , k-index brought 92%, h-index brought 87%

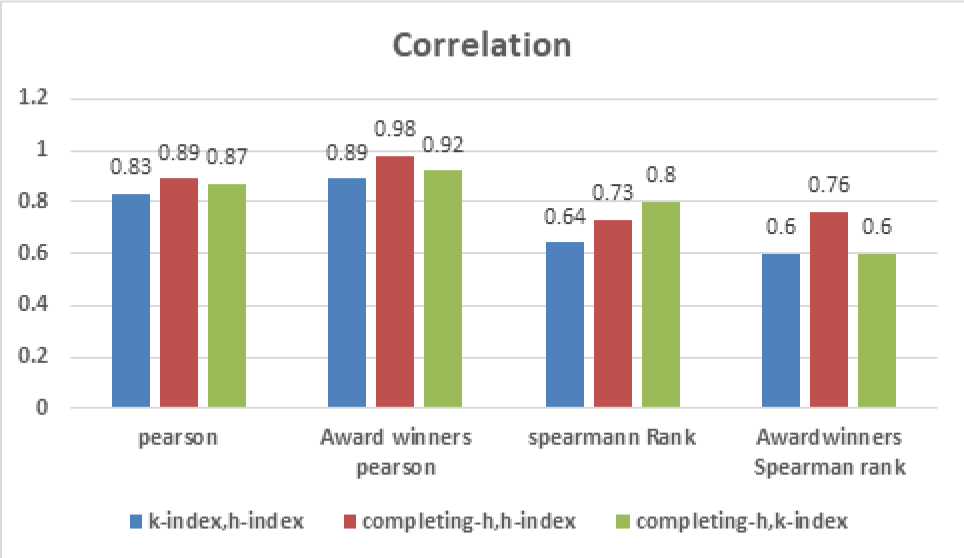

FIGURE 4.8: Correlation among indices

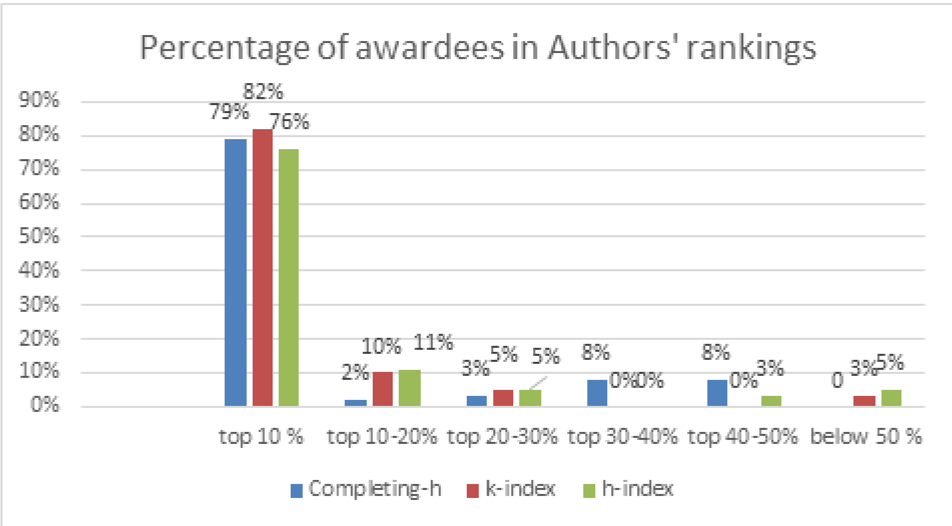

FIGURE 4.9: Occurrence of award winners in ranked lists

and completing-h brought 81% authors. Though K-index seems most successful but in broader picture performance of h-index is also good. Only completing-h has high percentage of authors in low ranks. From results it is quite evident that performance of k-index and completing-h is effectively comparable with h-index but with overhead of computation complexity. We should consider the complexity in calculations of these indices as compared to h-index. h-index is relatively simple to compute whereas these three indices require a lot of computation. Moreover when we consider completing-h, conversion factor for a community belong to the

TABLE 4.26: Data sets for Physics Data set

No.	DATA SET	Total No. of Records	Training Data	Test Data
P1	DatasetFull	113554	90795	22759
P2	DatasetExp5-12	40883	32790	8093

time/era for which it is calculated. Also we require citation data for whole community to calculate completing-h conversion factor. These factors should also be considered along with slightly better results of these indices in comparison with h-index.

Features Evaluation for the Domain of Physics

In RQ2 we have identified the parameters producing promising results to predict future impact of an author for the field of Computer science. In this research question, we have checked the validity of the proposed model for the field of Physics. We have considered Physics data set for OAG , 113554 authors record was found who have published a paper till 2007. We have considered 80% training and 20% testing data set. Number of records are shown in Table 4.26.

Equations proposed in Table 4.6 for full data set for the field of Computer Science is comprised of current h-index of an author, number of collaborations, number of coauthors, experience and square root of number of publications of an author. we have fitted regression model considering these parameters for the field of Physics. Results of applied technique are presented in Table 4.27 and Table 4.28.

Table 4.27 presents results when we have considered whole data set till 2007. Though R^2 values are low as compared to the results for the domain of Computer Science but RMSE is quite amazing, for one year RMSE is 0.15 and for five years it is 0.29. Also for researchers data set having constraints of 5 to 12 years of experience and h-index greater than 4, RMSE values are encouraging. Hence we can say that feature set identified for the field of Computer Science also shows promising results for the field of Physics.

Table 4.27: Regression model fitted for Physics Data set(P1) for features_set_RQ2_full

features_set _RQ2_full	**2008 (1year)**	**2009 (2years)**	**2010 (3years)**	**2011 (4years)**	**2012 (5years)**
R^2	0.86	0.79	0.74	0.697	0.66
Max error	1.95	2.84	3.68	4.635	4.6
RMSE	0.15	0.19	0.23	0.261	0.29
intercept	-0.131	-0.183	-0.223	-0.243	-0.261
2007_h_index	1.0638	1.1061	1.1311	1.1463	1.1575
Collaborations	-0.004	-0.007	-0.011	-0.013	-0.015
years _since _start	-0.001	-0.002	-0.003	-0.004	-0.005
No_coauthors	-1.00E-04	-2.00E-04	-3.00E-04	-3.00E-04	-4.00E-04
square _root _publications	0.1481	0.2181	0.2745	0.309	0.3382

Table 4.28: Regression Model Fitted on Physics data set (P2) for features set RQ2 H4_exp_5-12

features_set _RQ2 _H4_exp_5-12	**2008 (1year)**	**2009 (2years)**	**2010 (3years)**	**2011 (4years)**	**2012 (5years)**
R^2	0.88	0.83	0.78	0.748	0.72
Max error	1.95	2	2.51	3.352	4.25
RMSE	0.15	0.19	0.23	0.259	0.28
intercept	-0.0994	-0.173	-0.2344	-0.2705	-0.3008
2007_h_index	1.0511	1.1079	1.126	1.1362	1.1488
Collaborations	0.0034	-0.0009	-0.0061	-0.0089	-0.0115
coauthors_total _H_index	-0.0002	-0.0003	-0.0003	-0.0004	-0.0004
square_root _publications	0.0991	0.1775	0.2469	0.289	0.3246
No_coauthors	-0.0003	-0.0002	-0.0002	-0.0003	-0.0003
m-index	0.2146	0.0806	0.0456	0.072	0.0344

4.5 Conclusion

In this chapter we presented results of all the experiments based upon research questions. In research question 1 and 2, we considered three variations in data sets based on the experience and h-index of researchers. Applied existing models found in literature on these data sets. In case of all 3 variations of data set, overall Acuna's Model performs better than Dong et al's. The performance of the model proposed in this study is better than existing models. The proposed model clearly outperformed the existing models. It is identified that models are performing well for short term prediction but for longer periods the performance of the model measured in R^2 and RMSE, declines. It can be surmised that with the increase in time, variability in the h-index values also increases, so it reduces the prediction power. It is also shown/observed that performance of the model for researchers having higher h-index is more stable. Research Question 3 is focused on young researchers and it is realized that productivity and impact are likely to increase with time. H-index is based on these two measures and citation's information for recently published papers is not adequate. Hence there must be some other factors/method which we should consider. Considering this, a new NS-Index is proposed especially for young researchers. NS-Index predicts young researchers future potential better than prediction models. It is shown that NS-Index can be used for future potential prediction for initial three years of a researcher's career. After 3 years prediction models can be used to predict future impact. From predictability perspective, we propose that young researchers term should be used for researchers having 3 or less years of experience. Addressing Research Question 4, model proposed for the field of Computer Science in Research Question 2 is applied on the field of Physics. For the domain of Physics. Though R^2 values are low as compared to Computer Science, but RMSE values are really encouraging, one year RMSE is 0.15 and for five years it is 0.29. Better results for RMSE depicts less variability in the data set. Model proposed for the field of Computer Science also shows promising results for the field of Physics.

Chapter 5

Conclusion and Future work

This research aimed to identify effective scientific impact prediction model for the researchers. Based on our research questions, major conclusions are mentioned below. Future work dimensions in this field are also discussed.

5.1 Conclusion

Predicting the future impact of a scientist/researcher is a critical task. Impact of a researcher directly affects the performance of an organization/ institution. Predicting future impact is significantly important for making many decisions by an organization/institution. Knowing the future impact of researcher directly affects an organization's decision to hire a person or not, to give tenure to someone or not, to approve grant or not. To evaluate the performance of researchers, one of the most notable impact evaluation criteria is h-index.
This thesis addresses the problem of predicting future impact of researchers with focus on h-index prediction. It is identified that current h-index, experience of a researcher, number of coauthors, square root of number of publications and number of collaborations a researcher work in, contributes most for the prediction of future h-index of a researcher. From coefficients weights it is clear that highest

contribution is of current h-index ,followed by publications and number of collaborations. Our proposed prediction model shows better results than existing models , specially for next 5 years prediction. Based on the literature review it was identified that existing approaches impose constraints on the selection of researchers. Researchers having specific experience or h-index value within a specific limit are considered. Researchers having high h-index value or low h-index value are not considered. A publicly avaiable comprehensive data set of ArnetMiner is considered for this study. ArnetMiner data set comprises of the papers and authors record for the field of Computer Science.
Above mentioned feature set is applicable when whole data set is considered. Whereas for sub data set, where researchers having h-index value greater than 10 are considered, current h-index, experience of a researcher, number of coauthors, number of collaborations a researcher work in and current m-index of researcher contributes the most. For the sub data set , where researchers having h-index values greater than 4 and experience of 5 to 12 years are considered, current h-index, square root of number of publications, number of collaborations a researcher work in and sum of h-index of coauthors are considered as main contributors. R^2 and RMSE are considered as evaluation criteria.
Features which are common in all combinations are current h-index and number of collaborations a researcher work in. Addressing our RQ1 and RQ2, we can conclude that current h-index and number of times a researcher works in collaboration plays key role in predicting future h-index of a researcher.
In response to our research question 4, the proposed feature set is also applied for the field of Physics and with RMSE values of 0.15 for one year future prediction and 0.29 for 5 years , results are encouraging. Though R^2 values are comparatively low as compared to the models performance for the field of Computer Science, but still its performance is good. R^2 for Computer Science domain was 0.97 for one year and 0.90 for five years , where as for Physics these are 0.86 and 0.66 respectively.
In existing literature, young scholars future impact prediction with respect to h-index is rarely addressed. The reason behind is that in the start of a research

career very little information is available, and with limited information prediction results are not good. Anyhow corresponding to ur RQ4, we have applied our proposed regression models on young researchers' data sets, but results were not encouraging. From literature and from our own experimental results, it was concluded that we need some new measures for young researchers impact evaluation. In this thesis we have proposed a new NS-Index for researchers impact evaluation. This index is based upon the Extend relationship among the publications of authors. A publication/paper has NS-Index of n if n number of papers have extended this paper. Author's NS-Index is the sum of NS-Index of all his/her publications. This index would represent the more valuable/effective contribution of an author than simple citation count or h-index. To predict the impact of young researcher, sum of NS-Index of all the papers extended by a young researcher would be considered. To prove our idea we have considered citation count of the papers extednded by young researchers and compared the results with future h-index value of these researchers. We found that these results works well with future impact of researchers.

Considering the success of the proposed index according to our experiments, it is urged/proposed that papers should have NS Paper information in them, it should be a part of article structure. By NS paper it is meant that the paper which is extended by this paper. It would be helpful in maintaining the hierarchy of Extend relationship among papers. Main contribution of this thesis are

- Identification of main features contributing most effectively for the prediction of future h-index of researches for the field of Computer Science
- Successful application of the features identified for the field of Computer Science on the field of Physics
- A new NS-Index for young researchers is proposed.

5.2 Future Work

From our results and findings we have identified some future dimensions mentioned below:

- Conversion factor considering completing-h may be calculated for different fields. It will balance the effect of citations and publications and will be helpful in cross-domain comparison of researchers. By applying the conversion factor as proposed in completing-h [104], these feature set may have better results for all the fields.

- Proposed feature sets should be applied on some more fields after applying conversion factors, and results should be compared.

- To fully evaluate and understand the significance of the proposed NS-Index, a comprehensive data set may be prepared. The data set will be based upon the Extend relationship among papers. For example, considering Arnetminer data set, a field may be added in papers table having paperID's of those papers which would be extended by under consideration paper. Of course it would be one of the papers from references of under consideration paper. By having this information , NS-Index can be validated on large scale.

- Another direction in which NS-Index can be explored is to go on next levels of hierarchy. Impact of papers' extend relationship can be explored on different levels.

- Calculation of NS-Index for researchers will open new horizons for the research. Performance of the NS-Index can be compared with existing impact evaluation metrics like citation count, h-index. it can also be compared with different variants and extensions of h-index like g-index, k-index etc.

- One possible method to compare the performance of proposed NS-Index and other bibliometric indicators is to evaluate the rankings obtained by these. Researchers may be ranked according to these indicators/indices. These rankings may be compared with some benchmark, like prestigious award winners.

Bibliography

[1] P. N. Tyrrell, A. R. Moody, J. O. C. Moody, and N. Ghiam, "Departmental h-index: evidence for publishing less?" *Canadian Association of Radiologists Journal*, vol. 68, no. 1, pp. 10–15, 2017.

[2] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National academy of Sciences*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[3] C. Oppenheim, "Using the h-index to rank influential british researchers in information science and librarianship," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 2, pp. 297–301, 2007.

[4] L. Bornmann, R. Mutz, and H.-D. Daniel, "Are there better indices for evaluation purposes than the h index? a comparison of nine different variants of the h index using data from biomedicine," *Journal of the American Society for Information Science and technology*, vol. 59, no. 5, pp. 830–837, 2008.

[5] S. Ayaz and M. T. Afzal, "Identification of conversion factor for completing-h index for the field of mathematics," *Scientometrics*, vol. 109, no. 3, pp. 1511–1524, 2016.

[6] T. Amjad, Y. Rehmat, A. Daud, and R. A. Abbasi, "Scientific impact of an author and role of self-citations," *Scientometrics*, vol. 122, no. 2, pp. 915–932, 2020.

[7] S. G. Aoun, B. R. Bendok, R. J. Rahme, R. G. Dacey Jr, and H. H. Batjer, "Standardizing the evaluation of scientific and academic performance

in neurosurgery—critical review of the "h" index and its variants," *World neurosurgery*, vol. 80, no. 5, pp. e85–e90, 2013.

[8] L. Bornmann, "h-index research in scientometrics: A summary," *arXiv preprint arXiv:1407.2932*, 2014.

[9] Q. Wu, "The w-index: A measure to assess scientific impact by focusing on widely cited papers," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 3, pp. 609–614, 2010.

[10] F. J. Cabrerizo, S. Alonso, E. Herrera-Viedma, and F. Herrera, "q2-index: Quantitative and qualitative evaluation based on the number and impact of papers in the hirsch core," *Journal of informetrics*, vol. 4, no. 1, pp. 23–28, 2010.

[11] O. Kinouchi, L. D. Soares, and G. C. Cardoso, "A simple centrality index for scientific social recognition," *Physica A: Statistical Mechanics and its Applications*, vol. 491, pp. 632–640, 2018.

[12] M. Schreiber, "A variant of the h-index to measure recent performance," *Journal of the Association for Information Science and Technology*, vol. 66, no. 11, pp. 2373–2380, 2015.

[13] M. A. García-Pérez, "An extension of the h index that covers the tail and the top of the citation curve and allows ranking researchers with similar h," *Journal of Informetrics*, vol. 6, no. 4, pp. 689–699, 2012.

[14] L. Bornmann, R. Mutz, S. E. Hug, and H.-D. Daniel, "A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants," *Journal of Informetrics*, vol. 5, no. 3, pp. 346–359, 2011.

[15] R. Costas and M. Bordons, "The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level," *Journal of informetrics*, vol. 1, no. 3, pp. 193–203, 2007.

[16] A. P. dos Santos Rubem, A. L. de Moura *et al.*, "Comparative analysis of some individual bibliometric indices when applied to groups of researchers," *Scientometrics*, vol. 102, no. 1, pp. 1019–1035, 2015.

[17] M. Raheel, S. Ayaz, and M. T. Afzal, "Evaluation of h-index, its variants and extensions based on publication age & citation intensity in civil engineering," *Scientometrics*, vol. 114, no. 3, pp. 1107–1127, 2018.

[18] P. Vinkler, "Eminence of scientists in the light of the h-index and other scientometric indicators," *Journal of information science*, vol. 33, no. 4, pp. 481–491, 2007.

[19] G. Saad, "Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively," *Scientometrics*, vol. 69, no. 1, pp. 117–120, 2006.

[20] W. E. Schreiber and D. M. Giustini, "Measuring scientific impact with the h-index: a primer for pathologists," *American journal of clinical pathology*, vol. 151, no. 3, pp. 286–291, 2019.

[21] C. D. Kelly and M. D. Jennions, "The h index and career assessment by numbers," *Trends in Ecology & Evolution*, vol. 21, no. 4, pp. 167–170, 2006.

[22] F. Zhang, X. Bai, and I. Lee, "Author impact: Evaluations, predictions, and challenges," *IEEE Access*, vol. 7, pp. 38 657–38 669, 2019.

[23] X. Bai, H. Pan, J. Hou, T. Guo, I. Lee, and F. Xia, "Quantifying success in science: An overview," *IEEE Access*, vol. 8, pp. 123 200–123 214, 2020.

[24] D. E. Acuna, S. Allesina, and K. P. Kording, "Predicting scientific success," *Nature*, vol. 489, no. 7415, pp. 201–202, 2012.

[25] A. Jones, "The explosive growth of postdocs in computer science," *Communications of the ACM*, vol. 56, no. 2, pp. 37–39, 2013.

[26] D. E. Acuna and O. Penner, "Point/counterpoint," *Medical physics*, vol. 40, p. 110601, 2013.

[27] M. Schreiber, "Is it possible to measure scientific performance with the h-index or with another variant from the hirsch index zoo," *Journal of Unsolved Questions*, vol. 4, no. 1, pp. 5–10, 2014.

[28] A. Mazloumian, "Predicting scholars' scientific impact," *PloS one*, vol. 7, no. 11, p. e49246, 2012.

[29] M. Schreiber, "A skeptical view on the hirsch index and its predictive power," *Physica Scripta*, vol. 93, no. 10, p. 102501, 2018.

[30] A. Agarwal, D. Durairajanayagam, S. Tatagari, S. C. Esteves, A. Harlev, R. Henkel, S. Roychoudhury, S. Homa, N. G. Puchalt, R. Ramasamy *et al.*, "Bibliometrics: tracking research impact by selecting the appropriate metrics," *Asian journal of andrology*, vol. 18, no. 2, p. 296, 2016.

[31] Y. Dong, R. A. Johnson, and N. V. Chawla, "Can scientific impact be predicted?" *IEEE Transactions on Big Data*, vol. 2, no. 1, pp. 18–30, 2016.

[32] Z. Wu, W. Lin, P. Liu, J. Chen, and L. Mao, "Predicting long-term scientific impact based on multi-field feature extraction," *IEEE Access*, vol. 7, pp. 51 759–51 770, 2019.

[33] M. A. García-Pérez, "Limited validity of equations to predict the future h index," *Scientometrics*, vol. 96, no. 3, pp. 901–909, 2013.

[34] O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, and S. Fortunato, "On the predictability of future impact in science," *Scientific reports*, vol. 3, no. 1, pp. 1–8, 2013.

[35] C. McCarty, J. W. Jawitz, A. Hopkins, and A. Goldman, "Predicting author h-index using characteristics of the co-author network," *Scientometrics*, vol. 96, no. 2, pp. 467–483, 2013.

[36] A. Ibáñez, P. Larrañaga, and C. Bielza, "Predicting the h-index with cost-sensitive naive bayes," in *2011 11th International Conference on Intelligent Systems Design and Applications*. IEEE, 2011, pp. 599–604.

[37] L. Weihs and O. Etzioni, "Learning to predict citation-based impact measures," in *2017 ACM/IEEE joint conference on digital libraries (JCDL)*. IEEE, 2017, pp. 1–10.

[38] T. Mistele, T. Price, and S. Hossenfelder, "Predicting authors' citation counts and h-indices with a neural network," *Scientometrics*, vol. 120, no. 1, pp. 87–104, 2019.

[39] G. Nikolentzos, G. Panagopoulos, I. Evdaimon, and M. Vazirgiannis, "Can author collaboration reveal impact? the case of h-index," *arXiv preprint arXiv:2104.05562*, 2021.

[40] I. Podlubny, "Comparison of scientific impact expressed by the number of citations in different fields of science," *Scientometrics*, vol. 64, no. 1, pp. 95–99, 2005.

[41] E. Lillquist and S. Green, "The discipline dependence of citation statistics," *Scientometrics*, vol. 84, no. 3, pp. 749–762, 2010.

[42] P. D. Batista, M. G. Campiteli, and O. Kinouchi, "Is it possible to compare researchers with different scientific interests?" *Scientometrics*, vol. 68, no. 1, pp. 179–189, 2006.

[43] A. Gogoglou and Y. Manolopoulos, "Predicting the evolution of scientific output," in *International Conference on Computational Collective Intelligence*. Springer, 2017, pp. 244–254.

[44] S. Ayaz and N. Masood, "Comparison of researchers' impact indices," *PloS one*, vol. 15, no. 5, p. e0233765, 2020.

[45] M. Schreiber, "How relevant is the predictive power of the h-index? a case study of the time-dependent hirsch index," *Journal of Informetrics*, vol. 7, no. 2, pp. 325–329, 2013.

[46] S. Ayaz, N. Masood, and M. A. Islam, "Predicting scientific impact based on h-index," *Scientometrics*, vol. 114, no. 3, pp. 993–1010, 2018.

[47] D. Wang, C. Song, and A.-L. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, no. 6154, pp. 127–132, 2013.

[48] 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*, 2016, pp. 2676–2682.

[49] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, "Predicting scientific success based on coauthorship networks," *EPJ Data Science*, vol. 3, no. 1, p. 9, 2014.

[50] C. Stegehuis, N. Litvak, and L. Waltman, "Predicting the long-term citation impact of recent publications," *Journal of informetrics*, vol. 9, no. 3, pp. 642–657, 2015.

[51] J. Chen and C. Zhang, "Predicting citation counts of papers," in *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE, 2015, pp. 434–440.

[52] 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*, 2012, pp. 51–60.

[53] W. F. Laurance, D. C. Useche, S. G. Laurance, and C. J. Bradshaw, "Predicting publication success for biologists," *BioScience*, vol. 63, no. 10, pp. 817–823, 2013.

[54] S. Lee and B. Bozeman, "The impact of research collaboration on scientific productivity," *Social studies of science*, vol. 35, no. 5, pp. 673–702, 2005.

[55] P. Z. Revesz, "A method for predicting citations to the scientific publications of individual researchers," in *Proceedings of the 18th international database engineering & applications symposium*, 2014, pp. 9–18.

[56] M. Charnine, A. Khakimova, and A. Klokov, "Impact factor of a term: a tool for assessing article's future citations and author's influence based on pubmed and dblp collections," 2020.

[57] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee, "Towards a stratified learning approach to predict future citation counts," in *IEEE/ACM Joint Conference on Digital Libraries*. IEEE, 2014, pp. 351–360.

[58] G. D. Gonçalves, F. Figueiredo, J. M. Almeida, and M. A. Gonçalves, "Characterizing scholar popularity: a case study in the computer science research community," in *IEEE/ACM Joint Conference on Digital Libraries*. IEEE, 2014, pp. 57–66.

[59] A. P. Akella, H. Alhoori, P. R. Kondamudi, C. Freeman, and H. Zhou, "Early indicators of scientific impact: Predicting citations with altmetrics," *Journal of Informetrics*, vol. 15, no. 2, p. 101128, 2021.

[60] D. H. Lee, "Predicting the research performance of early career scientists," *Scientometrics*, vol. 121, no. 3, pp. 1481–1504, 2019.

[61] Z. Ning, Y. Liu, J. Zhang, and X. Wang, "Rising star forecasting based on social network analysis," *IEEE Access*, vol. 5, pp. 24 229–24 238, 2017.

[62] Y. Nie, Y. Zhu, Q. Lin, S. Zhang, P. Shi, and Z. Niu, "Academic rising star prediction via scholar's evaluation model and machine learning techniques," *Scientometrics*, vol. 120, no. 2, pp. 461–476, 2019.

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

[64] M. Nezhadbiglari, M. A. Gonçalves, and J. M. Almeida, "Early prediction of scholar popularity," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 2016, pp. 181–190.

[65] L. Bornmann and H.-D. Daniel, "Does the h-index for ranking of scientists really work?" *Scientometrics*, vol. 65, no. 3, pp. 391–392, 2005.

[66] M. Qi, A. Zeng, M. Li, Y. Fan, and Z. Di, "Standing on the shoulders of giants: the effect of outstanding scientists on young collaborators' careers," *Scientometrics*, vol. 111, no. 3, pp. 1839–1850, 2017.

[67] X.-L. Li, C. S. Foo, K. L. Tew, and S.-K. Ng, "Searching for rising stars in bibliography networks," in *International conference on database systems for advanced applications.* Springer, 2009, pp. 288–292.

[68] C. Zhang, C. Liu, L. Yu, Z.-K. Zhang, and T. Zhou, "Identifying the academic rising stars via pairwise citation increment ranking," in *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data.* Springer, 2017, pp. 475–483.

[69] A. Daud, M. Ahmad, M. Malik, and D. Che, "Using machine learning techniques for rising star prediction in co-author network," *Scientometrics*, vol. 102, no. 2, pp. 1687–1711, 2015.

[70] T. Amjad, Y. Ding, J. Xu, C. Zhang, A. Daud, J. Tang, and M. Song, "Standing on the shoulders of giants," *Journal of Informetrics*, vol. 11, no. 1, pp. 307–323, 2017.

[71] 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*, 2008, pp. 990–998.

[72] J. Tang, D. Zhang, and L. Yao, "Social network extraction of academic researchers," in *Seventh IEEE International Conference on Data Mining (ICDM 2007).* IEEE, 2007, pp. 292–301.

[73] J. Tang, J. Zhang, R. Jin, Z. Yang, K. Cai, L. Zhang, and Z. Su, "Topic level expertise search over heterogeneous networks," *Machine Learning*, vol. 82, no. 2, pp. 211–237, 2011.

[74] J. Tang, A. C. Fong, B. Wang, and J. Zhang, "A unified probabilistic framework for name disambiguation in digital library," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 975–987, 2011.

[75] K. Pooja, S. Mondal, and J. Chandra, "Exploiting similarities across multiple dimensions for author name disambiguation," *Scientometrics*, pp. 1–36, 2021.

[76] D. Hu and H. Ma, "Collaborator recommendation integrating author's cooperation strength and research interests on attributed graph," *Advances in Computational Intelligence*, vol. 1, no. 4, pp. 1–15, 2021.

[77] Y. Bu, S. Ni, and W.-b. Huang, "Combining multiple scholarly relationships with author cocitation analysis: A preliminary exploration on improving knowledge domain mappings," *Journal of Informetrics*, vol. 11, no. 3, pp. 810–822, 2017.

[78] X. Kong, H. Jiang, W. Wang, T. M. Bekele, Z. Xu, and M. Wang, "Exploring dynamic research interest and academic influence for scientific collaborator recommendation," *Scientometrics*, vol. 113, no. 1, pp. 369–385, 2017.

[79] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning.* Springer, 2013, vol. 112.

[80] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis.* John Wiley & Sons, 2012, vol. 821.

[81] J. Frost, "How to interpret adjusted r-squared and predicted r-squared in regression analysis," [online] Available: https://statisticsbyjim.com/regression/interpret-r-squared-regression/, [Accessed: 12-Dec-2019].

[82] D. M. Diez, C. D. Barr, and M. Cetinkaya-Rundel, *OpenIntro statistics.* OpenIntro, 2012.

[83] K. Grace-Martin, "Assessing the fit of regression models," [Online]. Available: https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/, [Accessed: 04-Aug-2018].

[84] B. Cheng and D. M. Titterington, "Neural networks: A review from a statistical perspective," *Statistical science*, pp. 2–30, 1994.

[85] K. Gurney, *An introduction to neural networks.* CRC press, 1997.

[86] R. Hassan and M. R. Islam, "Detection of fake online reviews using semi-supervised and supervised learning," in *2019 International conference on electrical, computer and communication engineering (ECCE).* IEEE, 2019, pp. 1–5.

[87] S. Verma, A. Chug, and A. P. Singh, "Application of convolutional neural networks for evaluation of disease severity in tomato plant," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no. 1, pp. 273–282, 2020.

[88] D. D. Beaver and R. Rosen, "Studies in scientific collaboration. 2. scientific co-authorship, research productivity and visibility in the french scientific elite, 1799-1830," *Scientometrics*, vol. 1, no. 2, pp. 133–149, 1979.

[89] G. Melin, "Pragmatism and self-organization: Research collaboration on the individual level," *Research policy*, vol. 29, no. 1, pp. 31–40, 2000.

[90] J. D. Adams, G. C. Black, J. R. Clemmons, and P. E. Stephan, "Scientific teams and institutional collaborations: Evidence from us universities, 1981–1999," *Research policy*, vol. 34, no. 3, pp. 259–285, 2005.

[91] K. Borner, L. DallAsta, W. Ke, and A. Vespignani, "Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams," *Complexity*, vol. 10, no. 4, pp. 57–67, 2005.

[92] J. Kaltz and R. Martin, "What is research collaboration," *Research policy*, vol. 26, pp. 1–18, 1997.

[93] P. Han, J. Shi, X. Li, D. Wang, S. Shen, and X. Su, "International collaboration in lis: global trends and networks at the country and institution level," *Scientometrics*, vol. 98, no. 1, pp. 53–72, 2014.

[94] K. Frenken, W. Holzl, and F. De Vor, "The citation impact of research collaborations: the case of european biotechnology and applied microbiology (1988–2002)," *Journal of Engineering and technology Management*, vol. 22, no. 1-2, pp. 9–30, 2005.

[95] S. Goldfinch, T. Dale, and K. DeRouen, "Science from the periphery: Collaboration, networks and'periphery effects' in the citation of new zealand crown research institutes articles, 1995-2000," *Scientometrics*, vol. 57, no. 3, pp. 321–337, 2003.

[96] Z.-L. He, X.-S. Geng, and C. Campbell-Hunt, "Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a new zealand university," *Research policy*, vol. 38, no. 2, pp. 306–317, 2009.

[97] B. F. Jones, S. Wuchty, and B. Uzzi, "Multi-university research teams: Shifting impact, geography, and stratification in science," *science*, vol. 322, no. 5905, pp. 1259–1262, 2008.

[98] P. Skilton, "Does the human capital of teams of natural science authors predict citation frequency?" *Scientometrics*, vol. 78, no. 3, pp. 525–542, 2009.

[99] L. Waltman, "A review of the literature on citation impact indicators," *Journal of informetrics*, vol. 10, no. 2, pp. 365–391, 2016.

[100] D. Lindsey, "Using citation counts as a measure of quality in science measuring what's measurable rather than what's valid," *Scientometrics*, vol. 15, no. 3-4, pp. 189–203, 1989.

[101] I. Ihsan and M. A. Qadir, "Ccro: Citation's context & reasons ontology," *IEEE Access*, vol. 7, pp. 30 423–30 436, 2019.

[102] B. Jörg, "Towards the nature of citations," in *Poster Proceedings of the 5th International Conference on Formal Ontology in Information Systems*, 2008.

[103] M. J. Moravcsik and P. Murugesan, "Some results on the function and quality of citations," *Social studies of science*, vol. 5, no. 1, pp. 86–92, 1975.

[104] K. R. Dienes, "Completing h," *Journal of Informetrics*, vol. 9, no. 2, pp. 385–397, 2015.

[105] R. K. Pan and S. Fortunato, "Author impact factor: tracking the dynamics of individual scientific impact," *Scientific reports*, vol. 4, p. 4880, 2014.

[106] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, "Quantifying the evolution of individual scientific impact," *Science*, vol. 354, no. 6312, 2016.

[107] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman, *Forecasting methods and applications*. John wiley & sons, 2008.

[108] C. Pearson's, "Comparison of values of pearson's and spearman's correlation coefficients," *Comparison Of Values Of Pearson's And Spearman's Correlation Coefficients*, 2011.

Appendix A

Impact Factor Journals

Since its introduction, h-index (Hirsch 2005) has become the most commonly used and established measure to evaluate the impact of individual researchers on scientific literature (Tyrrell, et.al. 2016). H-index combines the effect of two dimensions i.e. number of publications, representing the productive core of a scientist and number of citations, representing the impact of that core.

Following is the list of top 10 Impact Factor Journals for Computer Science according to the 2015 JCR Rankings. Additionally the journals proposed by Acune et al. starting from Sr. No. 11.

Table A.1: List of top 10 journals 2015 JCR Rankings

Sr.	**Journal Title**
1	IEEE Transactions on Fuzzy Systems
2	IEEE Communications Surveys And Tutorials
3	International Journal of Neural Systems
4	IEEE Transactions on Pattern Analysis And Machine Intelligence
5	IEEE Transactions On Evolutionary Computation
6	MIS Quarterly
7	Computer-AIDED Civil And Infrastructure Engineering
8	ACM Computing Surveys

Continued on next page

Table A.1 – continued from previous page

Sr.	Journal Title
9	Integrated Computer-Aided Engineering
10	IEEE Transactions on Cybernetics
11	Science
12	Nature Communications
13	Proceedings of the National Academy of Sciences
14	Nature
15	PLoS ONE

Table shows the journals' list satisfying the criteria of having Impact Factor equal to or greater than 3 according to 2015 JCR Rankings.

Table A.2: List of journals having at least 3 Impact Factor

Sr.	Journal Title
1	IEEE Transactions on Fuzzy Systems
2	IEEE Communications Surveys And Tutorials
3	International Journal of Neural Systems
4	IEEE Transactions on Pattern Analysis And Machine Intelligence
5	IEEE Transactions On Evolutionary Computation
6	MIS Quarterly
7	Computer-AIDED Civil And Infrastructure Engineering
8	ACM Computing Surveys
9	Integrated Computer-Aided Engineering
10	IEEE Transactions on Cybernetics
11	IEEE Transactions on Neural Networks and Learning Systems
12	Journal OF Information Technology
13	IEEE Transactions on Industrial Informatics
14	Medical Image Analysis
15	Information Fusion

Continued on next page

Table A.2 – continued from previous page

Sr.	Journal Title
16	International Journal of Computer Vision
17	ACM Transactions On Graphics
18	Archives Of Computational Methods in Engineering
19	Environmental Modelling & Software
20	IEEE Wireless Communications
21	Journal of Cheminformatics
22	Match Communications In Mathematical And In Computer Chemistry
23	IEEE Transactions on Medical Imaging
24	IEEE Transactions on Image Processing
25	Human Computer Interaction
26	Journal of Chemical Information and Modeling
27	IEEE Computational Intelligence Magazine
28	Computer Physics Communications
29	Evolutionary Computation
30	IEEE Intelligent Systems
31	Journal Of The American Medical Informatics Association
32	Pattern Recognition
33	Information Sciences
34	Artificial Intelligence
35	Knowledge Based Systems
36	Communications of The ACM
37	Neural Networks
38	Journal Of Computer-Aided Molecular Design
39	Journal of Management Information Systems
40	Internet Research

In Set 3and Set 4, for distinct Journals parameter, we have considered only those publications which were published in Impact Factor Journal. Following is the list of Impact Factor Journals from the field of Computer Science for year 2015 taken from JCR Rankings(Web of Science).

Table A.3: List of Impact Factor Journals (Computer Science)

Sr.	Journal Title
1	IEEE Transactions on Fuzzy Systems
2	IEEE Communications Surveys And Tutorials
3	International Journal of Neural Systems
4	IEEE Transactions on Pattern Analysis and Machine Intelligence
5	IEEE Transactions on Evolutionary Computation
6	MIS Quarterly
7	Computer-Aided Civil and Infrastructural Engineering
8	ACM Computing Surveys
9	Integrated Computer-Aided Engineering
10	IEEE Transactions on Cybernetics
11	IEEE Transactions on Neural Networks and Learning Systems
12	Journal of Information Technology
13	IEEE Transactions on Industrial Informatics
14	Medical Image Analysis
15	Information Fusion
16	International Journal of Computer Vision
17	ACM Transactions on Graphics
18	Archives of Computational Methods in Engineering
19	Environmental Modelling & Software
20	IEEE Wireless Communications
21	Journal of Cheminformatics
22	Match Communications in Mathematical and in Computer Chemistry
23	IEEE Transactions ON Medical Imaging
24	IEEE Transactions ON Image Processing
25	Human Computer Interaction
26	Journal of Chemical Information and Modeling

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
27	IEEE Computational Intelligence Magazine
28	Computer Physics Communications
29	Evolutionary Computation
30	IEEE Intelligent Systems
31	Journal of the American Medical Informatics Association
32	Pattern Recognition
33	Information Sciences
34	Artificial Intelligence
35	Knowledge Based Systems
36	Communications of the ACM
37	Neural Networks
38	Journal of Computer-Aided Molecular Design
39	Journal of Management Information Systems
40	Internet Research
41	Expert Systems With Applications
42	Swarm and Evolutionary Computation
43	IEEE Network
44	European Journal of Information Systems
45	Computers & Education
46	Neuroinformatics
47	Applied Soft Computing
48	Data Mining and Knowledge Discovery
49	INTERNATIONAL JOURNAL OF APPROXIMATE REASONING
50	SIAM Journal on Imaging Sciences
51	Ieee Transactions On Parallel And Distributed Systems
52	Decision Support Systems
53	Journal Of Strategic Information Systems
54	Computers & Chemical Engineering

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
55	Swarm Intelligence
56	Fuzzy Optimization And Decision Making
57	Journal Of Computational Physics
58	Ieee Transactions On Multimedia
59	Ieee Transactions On Knowledge And Data Engineering
60	Computers & Geosciences
61	Ieee Transactions On Mobile Computing
62	Journal Of The American Society For Information Science And Technology
63	Journal Of The American Society For Information Science And Technology
64	Journal Of Machine Learning Research
65	Journal Of Biomedical Informatics
66	Ieee Transactions On Information Forensics And Security
67	Future Generation Computer Systems The International Journal Of Escience
68	Computers & Structures
69	Acm Transactions On Intelligent Systems And Technology
70	Neurocomputing
71	Journal Of Statistical Software
72	Engineering Applications Of Artificial Intelligence
73	Ieee Transactions On Services Computing
74	International Journal Of Medical Informatics
75	Journal Of Network And Computer Applications
76	User Modeling And Useradapted Interaction
77	Ieee Transactions On Reliability
78	Enterprise Information Systems
79	Hemometrics And Intelligent Laboratory Systems

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
80	Structural And Multidisciplinary Optimization
81	Ieeeacm Transactions On Networking
82	Journal Of Optical Communications And Networking
83	Information & Management
84	Computeraided Design
85	Artificial Intelligence In Medicine
86	Electronic Commerce Research And Applications
87	Computer Vision And Image Understanding
88	Ieee Systems Journal
89	Journal Of Automated Reasoning
90	Computer Communications
91	Fuzzy Sets And Systems
92	Ieee Journal Of Biomedical And Health Informatics
93	Computers & Industrial Engineering
94	Scientometrics
95	Robotics And Computerintegrated Manufacturing
96	International Journal Of Geographical Information Science
97	Mathematical Programming
98	Business & Information Systems Engineering
99	International Journal Of Intelligent Systems
100	Computational Linguistics
101	Advanced Engineering Informatics
102	Journal Of Intelligent Manufacturing
103	Computational Geosciences
104	Computers & Operations Research
105	Foundations Of Computational Mathematics
106	International Journal Of Systems Science
107	Cognitive Computation

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
108	Astronomy And Computing
109	Displays
110	Sar And Qsar In Environmental Research
111	Computers And Electronics In Agriculture
112	Computers & Fluids
113	Acm Transactions On Mathematical Software
114	Acm Transactions On Mathematical Software
115	Ieee Transactions On Affective Computing
116	Journal Of Chemometrics
117	Mechatronics
118	Journal Of The Association For Information Science And Technology
119	Computer Methods And Programs In Biomedicine
120	Journal Of Computing In Civil Engineering
121	International Journal Of Electronic Commerce
122	Computer Methods In Biomechanics And Biomedical Engineering
123	Ieee Pervasive Computing
124	Information Systems
125	Information Systems
126	Journal Of The Acm
127	Journal Of The Acm
128	Ieee Transactions On Human Machine Systems
129	Medical & Biologicalengineering & Computing
130	Journal Of The Association For Information Systems
131	Statistics And Computing
132	Semantic Web
133	Computer Supported Cooperative Work The Journal Of Collaborative Computing
134	Image And Vision Computing

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
135	Wiley Interdisciplinary Reviews Data Mining And Knowledge Discovery
136	Neural Processing Letters
137	Vldb Journal
138	Ieee Transactions On Information Theory
139	Artificial Intelligence Review
140	Frontiers In Neurorobotics
141	Ieee Transactions On Computers
142	Machine Learning
143	Pervasive And Mobile Computing
144	Computers And Geotechnics
145	Journal Of Experimental & Theoretical Artificial Intelligence
146	Knowledge And Information Systems
147	Big Data
148	Computers In Industry
149	International Journal Of General Systems
150	Journal Of Molecular Graphics & Modelling
151	Advances In Engineering Software
152	Ad Hoc Networks
153	Journal Of Artificial Intelligence Research
154	Computers & Security
155	Soft Computing
156	Neural Computation
157	Robotics And Autonomous Systems
158	Journal Of Cryptology
159	Biological Cybernetics
160	Ieeeacm Transactions On Computational Biology And Bioinformatics
161	Acm Transactions On Computer Systems
162	Ieee Transactions On Systems Man Cybernetics Systems

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
163	Ieee Transactions On Dependable And Secure Computing
164	Pattern Recognition Letters
165	Journal Of Computer And System Sciences
166	Molecular Informatics
167	Information And Software Technology
168	Journal Of Realtime Image Processing
169	Journal Of Grid Computing
170	Autonomous Robots
171	Computer Graphics Forum
172	World Wide Webinternet And Web Information Systems
173	Mobile Networks & Applications
174	Journal Of Computational Biology
175	Journal Of Visual Communication And Image Representation
176	Social Science Computer Review
177	Earth Science Informatics
178	Computers In Biology And Medicine
179	Ieee Transactions On Software Engineering
180	Cluster Computing The Journal Of Networks Software Tools And Applications
181	Acm Transactions On Software Engineering And Methodology
182	Data & Knowledge Engineering
183	Personal And Ubiquitous Computing
184	Neural Computing & Applications
185	Simulation Modelling Practice And Theory
186	International Journal Of Humancomputer Studies
187	Journal Of Mathematical Imaging And Vision
188	Engineering With Computers
189	Information Systems Frontiers

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
190	Acm Transactions On Sensor Networks
191	International Journal For Numerical Methods In Fluids
192	Computer Networks
193	Journal Of Molecular Modeling
194	Multidimensional Systems And Signal Processing
195	Multidimensional Systems And Signal Processing
196	Journal Of Systems And Software
197	Autonomous Agents And Multiagent Systems
198	Multimedia Systems
199	Acm Sigcomm Computer Communication Review
200	Ieee Internet Computing
201	Ieee Transactions On Visualization And Computer Graphics
202	Information Processing & Management
203	Empirical Software Engineering
204	International Journal Of Bioinspired Computation
205	Informatica
206	Mathematical And Computer Modelling
207	Computing In Science & Engineering
208	Ieee Multimedia
209	Ieee Multimedia
210	Journal Of Complexity
211	Journal Of Functional Programming
212	International Journal Of Critical Infrastructure Protection
213	Journal Of Heuristics
214	Multimedia Tools And Applications
215	Quantum Information & Computation
216	Computer Speech And Language
217	Journal Of Parallel And Distributed Computing

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
218	International Journal Of Computer Integrated Manufacturing
219	Automated Software Engineering
220	Natural Computing
221	Acm Transactions On Computer Human Interaction
222	International Journal Of Information Security
223	International Journal Of Information Security
224	Industrial Management & Data Systems
225	Journal Of Web Semantics
226	Machine Vision And Applications
227	Computer Standards & Interfaces
228	Distributed Computing
229	International Journal Of Humancomputer Interaction
230	Ieee Access
231	Informs Journal On Computing
232	Ieee Transactions On Very Large Scale Integration (Vlsi) Systems
233	International Journal Of Web And Grid Services
234	Applied Intelligence
235	Behaviour & Information Technology
236	Digital Investigation
237	Behaviour & Information Technology
238	Digital Investigation
239	Ieee Transactions On Autonomous Mental Development
240	Cognitive Systems Research
241	Ieee Computer Graphics And Applications
242	Bell Labs Technical Journal
243	International Journal Of Modern Physics
244	International Journal Of Information Technology & Decision Making

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
245	Ieee Transactions On Computeraided Design Of Integrated Circuits And Systems
246	Journal Of Hydroinformatics
247	Computational Statistics & Data Analysis
248	Bit Numerical Mathematics
249	Journal Of Simulation
250	Current Computer Aided Drug Design
251	Online Information Review
252	Acm Transactions On Programming Languages And Systems
253	Aslib Proceedings
254	Aslib Proceedings
255	Genetic Programming And Evolvable Machines
256	Optical Switching And Networking
257	Formal Methods In System Design
258	Ieee Transactions On Learning Technologies
259	Connection Science
260	Mathematics Andcomputers In Simulation
261	Computers & Graphicsuk
262	Acm Transactions On Autonomous And Adaptive Systems
263	Computer
264	International Journal Of Machine Learning And Cybernetics
265	Requirements Engineering
266	Pattern Analysis And Applications
267	Adaptive Behavior
268	Information Technology And Libraries
269	Computer Aided Geometric Design
270	Ieee Micro
271	Journal Of Supercomputing

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
272	Computers & Electrical Engineering
273	Software Testing Verification & Reliability
274	International Journal Of High Performance Computing Applications
275	Journal Of Combinatorial Optimization
276	Journal Of Computational Science
277	Journal Of Computational Science
278	Journal Of Network And Systems Management
279	It Professional
280	Tsinghua Science And Technology
281	Acm Transactions On The Web
282	Geoinformatica
283	Visual Computer
284	R Journal
285	Artificial Life
286	Knowledge Engineering Review
287	Speech Communication
288	International Journal Of Applied Mathematics And Computer Science
289	Acm Transactions On Storage
290	Ieee Transactions On Haptics
291	Journal Of Symbolic Computation
292	Information Systems Management
293	Information Systems Management
294	Concurrent Engineeringresearch And Applications
295	Journal On Multimodal User Interfaces
296	Computational Biology And Chemistry
297	Iet Information Security
298	Random Structures & Algorithms
299	Wireless Networks

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
300	Journal Of Intelligent & Fuzzy Systems
301	Acm Transactions On Knowledge Discovery From Data
302	Computer Journal
303	Ieee Transactions On Computational Intelligence And Ai In Games
304	International Journal Of Uncertainty Fuzziness And Knowledgebased Systems
305	Journal Of Intelligent Information Systems
306	Parallel Computing
307	Programelectronic Library And Information Systems
308	Peertopeer Networking And Applications
309	Software And Systems Modeling
310	Acm Transactions On Multimedia Computing Communications And Applications
311	Acm Transactions On Information Systems
312	Language Resources And Evaluation
313	Acm Transactions On Information Systems
314	Language Resources And Evaluation
315	Theory And Practice Of Logic Programming
316	Expert Systems
317	Annals Of Mathematics And Artificial Intelligence
318	Journal Of Organizational Computing And Electronic Commerce
319	Performance Evaluation
320	Networks
321	Concurrency And Computationpractice & Experience
322	International Journal Of Fuzzy Systems
323	Computer Applications In Engineering Education
324	Journal Of Intelligent & Robotic Systems
325	Wireless Communications & Mobile Computing

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
326	Journal Of Communications And Networks
327	International Journal Of Pattern Recognition And Artificial Intelligence
328	International Journal Of Distributed Sensor Networks
329	Ieee Security & Privacy
330	Memetic Computing
331	Information Retrieval Journal
332	Interacting With Computers
333	Natural Language Engineering
334	International Journal On Document Analysis And Recognition
335	Science China Information Sciences
336	Cybernetics And Systems
337	Journal Of Information Science
339	Ai Edamartificial Intelligence For Engineering Design Analysis And Manufacturing
340	Queueing Systems
341	Information And Computation
342	Computing
343	Mobile Information Systems
344	Cincomputers Informatics Nursing
345	Acm Transactions On Computational Logic
346	Computers And Concrete
347	Optimization Methods & Software
348	Siam Journal On Computing
349	Biologically Inspired Cognitive Architectures
350	Journal Of Ambient Intelligence And Humanized Computing
351	Science Of Computer Programming
352	Graphical Models
353	Acm Transactions On Design Automation Of Electronic Systems

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
354	Ieee Software
355	Discrete & Computational Geometry
356	Security And Communication Networks
357	Distributed And Parallel Databases
358	Journal Of Computational Analysis And Applications
359	Algorithmica
360	Journal Of Computing And Information Science In Engineering
361	Presenceteleoperators And Virtual Environments
362	Software Quality Journal
363	Designs Codes And Cryptography
364	Acm Transactions On Algorithms
365	Journal Of New Music Research
366	Journal Of New Music Research
367	Minds And Machines
368	Iet Biometrics
369	Journal Of Computer Information Systems
370	Acm Transactions On Information And System Security
371	Mathematical Structures In Computer Science
372	Journal Of Statistical Computation And Simulation
373	Cryptography And Communications Discrete Structures Boolean Functions And Sequences
374	International Journal Of Unconventional Computing
375	Realtime Systems
376	Journal Of Software Evolution And Process
377	Acta Informatica
378	Computational Intelligence
379	Acta Informatica
380	Computational Intelligence

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
381	Journal Of Visualization
382	Theory Of Computing Systems
383	Acm Transactions On Embedded Computing Systems
384	Journal Of Ambient Intelligence And Smart Environments
385	Acm Journal On Emerging Technologies In Computing Systems
386	Acm Transactions On Internet Technology
387	Integration The Vlsi Journal
388	International Journal Of Quantum Information
389	Sigmod Record
390	Engineering Computations
391	Journal Of Systems Architecture
392	Ieee Design & Test
393	Journal Of Systems Architecture
394	Ieee Design & Test
395	International Journal Of Network Management
396	International Journal Of Parallel Programming
397	Aslib Journal Of Information Management
398	Frontiers Of Computer Science
399	Fundamenta Informaticae
400	Universal Access In The Information Society
401	Software Practice & Experience
402	Network Computation In Neural Systems
403	Theoretical Computer Science
404	Simulationtransactions Of The Society For Modeling And Simulation International
405	Information Visualization
406	Kybernetes
407	Ieee Annals Of The History Of Computing

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
408	Journal Of Logical And Algebraic Methods In Programming
409	Journal Of Visual Languages And Computing
410	Acm Transactions On Database Systems
411	Information Technology And Control
412	Problems Of Information Transmission
413	Advances In Mathematics Of Communications
414	Intelligent Data Analysis
415	Ai Magazine
416	Kybernetika
417	Combinatorics Probability & Computing
418	International Journal Of Computers Communications & Control
419	Ibm Journal Of Research And Development
420	Ibm Journal Of Research And Development
421	International Journal Of Data Warehousing And Mining
422	Mathematical And Computer Modelling Of Dynamical Systems
423	Computer Science And Information Systems
424	Constraints
425	Journal Of Web Engineering
426	International Journal On Semantic Web And Information Systems
427	Information Processing Letters
428	Foundations And Trends In Information Retrieval
429	Ad Hoc & Sensor Wireless Networks
430	Acm Transactions On Architecture And Code Optimization
431	Journal Of Logic And Computation
432	Acm Transactions On Architecture And Code Optimization
433	Journal Of Logic And Computation
434	Journal Of Logic And Algebraic Programming
435	Statistical Analysis And Data Mining

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
436	Iet Computer Vision
437	Logical Methods In Computer Science
438	Sustainable Computing Informatics & Systems
439	Virtual Reality
440	Neural Network World
441	Acm Transactions On Applied Perception
442	Photonic Network Communications
443	Acm Transactions On Modeling And Computer Simulation
444	Computer Languages Systems & Structures
445	Computer Animation And Virtual Worlds
446	Journal Of Universal Computer Science
447	Applied Artificial Intelligence
448	Discrete Mathematics And Theoretical Computer Science
449	Journal Of Internet Technology
450	New Generation Computing
451	New Review Of Hypermedia And Multimedia
452	Applied Ontology
453	International Journal Of Cooperative Information Systems
454	Canadian Journal Of Electrical And Computer Engineeringrevue Canadienne De Genie Electrique Et Informatique
455	Computing And Informatics
456	International Journal Of Rf And Microwave Computeraided Engineering
457	Journal Of Applied Logic
458	Formal Aspects Of Computing
459	International Arab Journal Of Information Technology
460	Turkish Journal Of Electrical Engineering And Computer Sciences
461	Iet Computers And Digital Techniques

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
462	Journal Of Signal Processing Systems For Signal Image And Video Technology
463	International Journal On Artificial Intelligence Tools
464	Journal Of Computer And Systems Sciences International
465	Acm Transactions On Reconfigurable Technology And Systems
466	International Journal Of Ad Hoc And Ubiquitous Computing
467	Acm Sigplan Notices
468	Malaysian Journal Of Computer Science
469	Journal Of Computer Science And Technology
470	Iet Software
471	Microprocessors And Microsystems
472	International Journal Of Foundations Of Computer Science
473	Ieee Computer Architecture Letters
474	Advances In Electrical And Computer Engineering
475	Scientific Programming
476	International Journal Of Sensor Networks
477	Journal Of Logic Language And Information
478	Ieee Latin America Transactions
479	Ieee Latin America Transactions
480	Compelthe International Journal For Computation And Mathematics In Electrical And Electronic Engineering
481	Journal Of Cellular Automata
482	Analog Integrated Circuits And Signal Processing
483	International Journal Of Wavelets Multiresolution And Information Processing
484	Journal Of Information Science And Engineering
485	Journal Of Zhejiang University Science Computers & Electronics
486	International Journal Of Computational Intelligence Systems

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
487	Programming And Computer Software
488	Applicable Algebra In Engineering Communication And Computing
489	Computational And Mathematical Organization Theory
490	Ksii Transactions On Internet And Information Systems
491	Ksii Transactions On Internet And Information Systems
492	Ai Communications
493	Intelligent Automation And Soft Computing
494	Computational Complexity
495	Journal Of Multiplevalued Logic And Soft Computing
496	Rairotheoretical Informatics And Applications
497	Journal Of Circuits Systems And Computers
498	Advances In Computers
499	Computer Systems Science And Engineering
500	Computer Music Journal
501	Romanian Journal Of Information Science And Technology
502	International Journal Of Web Services Research
503	Modeling Identification And Control
504	International Journal Of Web Services Research
505	Modeling Identification And Control
506	International Journal Of Software Engineering And Knowledge Engineering
507	Ieice Transactions On Fundamentals Of Electronics Communications And Computer Sciences
508	Ieice Transactions On Information And Systems
509	Cryptologia
510	Icga Journal
511	Journal Of Organizational And End User Computing

Continued on next page

Table A.3 – continued from previous page

Sr.	Journal Title
512	Journal Of Database Management
513	Design Automation For Embedded Systems
514	Infor
515	Traitement Du Signal

Appendix B

Forward Selection

Considering different parameters forward selection process was executed considering R^2 as an evaluation metric. Following are the forward selection steps for Data set D1 considering features mentioned in the table. Tables from B1 to B6 corresponds to Data set D1:

TABLE B.1: Forward feature Selection (D1)

Features	**R-Squared**		**Adj-RSquared**	
	1year (2008)	**5years (2012)**	**1year (2008)**	**5years (2012)**
2007_h_index	0.96	0.83	0.96	0.83
No_of_distinct_venues	0.63	0.61	0.626	0.61
no_of_IF_journals	0.06	0.05	0.056	0.05
Journal_IF_3	0.1	0.11	0.103	0.11
Distinct_but_only_IF	0.54	0.54	0.538	0.54
starting_year_from_2007	0.02	0	0.019	0
Collaborations	0.45	0.54	0.446	0.54
avg_citations	0.07	0.05	0.074	0.05
coauthors_total_H_index	0.6	0.62	0.602	0.62
No_coauthors	0.51	0.52	0.514	0.52

Continued on next page

Table B.1 – continued from previous page

Features	**R-Squared**		**Adj-RSquared**	
	1year (2008)	**5years (2012)**	**1year (2008)**	**5years (2012)**
average_h_index_coauthors	0.03	0.05	0.033	0.05
avg_coauthors_per_article	0	0	0	0
no_publications	0.59	0.58	0.593	0.58
no_article_as_last_author	0.39	0.37	0.385	0.37
proportion_last_author	0.17	0.17	0.166	0.17
no_article_as_first_author	0.37	0.35	0.366	0.35
proportion_first_author	0.14	0.16	0.144	0.16
square_root_publications	0.64	0.65	0.641	0.65
m_index	0.28	0.28	0.278	0.28
if_citations	0.17	0.15	0.17	0.15
Current_citations_diff_hindex	0.02	0.03	0.018	0.03
Current_citations_diff_hindex_nopub	0	0	0.005	0

Table B.2: Forward feature Selection (D1)

Features	**R-Squared**		**Adj-RSquared**	
	1year (2008)	**5years (2012)**	**1year (2008)**	**5years (2012)**
2007_h_index, No_of_distinct_venues	0.96	0.85	0.96	0.846
2007_h_index, no_of_IF_journals	0.96	0.84	0.96	0.837
2007_h_index, Journal_IF_3	0.96	0.84	0.96	0.838
2007_h_index, Distinct_but_only_IF	0.96	0.85	0.96	0.846
2007_h_index, starting_year_from_2007	0.96	0.85	0.96	0.847
2007_h_index, collaborations	0.96	0.87	0.96	0.873
2007_h_index, avg_citations	0.96	0.84	0.96	0.838

Continued on next page

Table B.2 – continued from previous page

Features	**R-Squared**		**Adj-RSquared**	
	1year (2008)	**5years (2012)**	**1year (2008)**	**5years (2012)**
2007_h_index, coauthors_total_H_index	0.96	0.85	0.96	0.854
2007_h_index, No_coauthors	0.96	0.85	0.96	0.846
2007_h_index, average_h_index_coauthors	0.96	0.84	0.96	0.84
2007_h_index, avg_coauthors_per_article	0.96	0.84	0.96	0.836
2007_h_index, no_publications	0.96	0.85	0.96	0.845
2007_h_index, no_article_as_last_author	0.96	0.84	0.96	0.839
2007_h_index, proportion_last_author	0.96	0.84	0.96	0.839
2007_h_index, no_article_as_first_author	0.96	0.84	0.96	0.839
2007_h_index, proportion_first_author	0.96	0.84	0.96	0.84
2007_h_index, square_root_publications	0.96	0.85	0.96	0.855
2007_h_index, m_index	0.96	0.84	0.96	0.839
2007_h_index, if_citations	0.96	0.84	0.96	0.837
2007_h_index, Current_citations_diff_hindex	0.96	0.84	0.96	0.837
2007_h_index, Current_citations_diff_hindex_nopub	0.96	0.84	0.96	0.838

Table B.3: Forward feature Selection (D1)

Features	R-Squared		Adj-RSquared	
	1year (2008)	5years (2012)	1year (2008)	5years (2012)
2007_h_index, collaborations, No_of_distinct_venues	0.96	0.9	0.96	0.9
2007_h_index, collaborations, no_of_IF_journals	0.96	0.87	0.96	0.87
2007_h_index, collaborations, Journal_IF_3	0.96	0.87	0.96	0.87
2007_h_index, collaborations, Distinct_but_only_IF	0.96	0.87	0.96	0.87
2007_h_index, collaborations, starting_year_from_2007	0.96	0.88	0.96	0.88
2007_h_index, collaborations, avg_citations	0.96	0.87	0.96	0.87
2007_h_index, collaborations, coauthors_total_H_index	0.96	0.87	0.96	0.87
2007_h_index, collaborations, No_coauthors	0.96	0.88	0.96	0.88
2007_h_index, collaborations, average_h_index_coauthors	0.96	0.88	0.96	0.88
2007_h_index, collaborations, avg_coauthors_per_article	0.96	0.87	0.96	0.87
2007_h_index, collaborations, no_publications	0.96	0.87	0.96	0.87
2007_h_index, collaborations, no_article_as_last_author	0.96	0.87	0.96	0.87
2007_h_index, collaborations, proportion_last_author	0.96	0.87	0.96	0.87

Continued on next page

Table B.3 – continued from previous page

Features	R-Squared		Adj-RSquared	
	1year (2008)	5years (2012)	1year (2008)	5years (2012)
2007_h_index, collaborations, no_article_as_first_author	0.96	0.87	0.96	0.87
2007_h_index, collaborations, proportion_first_author	0.96	0.87	0.96	0.87
2007_h_index, collaborations, square_root_publications	0.96	0.87	0.96	0.87
2007_h_index, collaborations, m_index	0.96	0.88	0.96	0.88
2007_h_index, collaborations, if_citations	0.96	0.87	0.96	0.87
2007_h_index, collaborations, Current_citations_diff_hindex	0.96	0.87	0.96	0.87
2007_h_index, collaborations, Current_citations_diff_hindex_nopub	0.96	0.87	0.96	0.87

Table B.4: Forward feature Selection (D1)

Features	R-Squared		Adj-RSquared	
	1year (2008)	5years (2012)	1year (2008)	5years (2012)
2007_h_index, collaborations, starting_year_from_2007, No_of_distinct_venues	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, no_of_IF_journals	0.96	0.88	0.96	0.88

Continued on next page

Table B.4 – continued from previous page

Features	**R-Squared**		**Adj-RSquared**	
	1year (2008)	**5years (2012)**	**1year (2008)**	**5years (2012)**
2007_h_index, collaborations, starting_year_from_2007, Journal_IF_3	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, Distinct_but_only_IF	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, avg_citations	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, coauthors_total_H_index	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, No_coauthors	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, average_h_index_coauthors	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, avg_coauthors_per_article	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, no_publications	0.96	0.88	0.96	0.88

Continued on next page

Table B.4 – continued from previous page

Features	**R-Squared**		**Adj-RSquared**	
	1year (2008)	**5years (2012)**	**1year (2008)**	**5years (2012)**
2007_h_index, collaborations, starting_year_from_2007, no_article_as_last_author	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, proportion_last_author	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, no_article_as_first_author	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, proportion_first_author	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, square_root_publications	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, m_index	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, if_citations	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, Current_citations_diff_hindex	0.96	0.88	0.96	0.88
2007_h_index, collaborations, starting_year_from_2007, Current_citations_diff_hindex_nopub	0.96	0.88	0.96	0.88

Table B.5: Forward feature Selection (D1)

Features	**R-Squared**		**Adj-RSquared**	
	1year (2008)	**5years (2012)**	**1year (2008)**	**5years (2012)**
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, No_of_distinct_venues	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, no_of_IF_journals	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, Journal_IF_3	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, Distinct_but_only_IF	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, avg_citations	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, coauthors_total_H_index	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, average_h_index_coauthors	0.96	0.89	0.96	0.89

Continued on next page

Table B.5 – continued from previous page

Features	R-Squared		Adj-RSquared	
	1year (2008)	5years (2012)	1year (2008)	5years (2012)
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, avg_coauthors_per_article	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, no_publications	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, no_article_as_last_author	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, proportion_last_author	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, no_article_as_first_author	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, proportion_first_author	0.96	0.89	0.96	0.89

Continued on next page

Table B.5 – continued from previous page

Features	R-Squared		Adj-RSquared	
	1year (2008)	5years (2012)	1year (2008)	5years (2012)
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, m_index	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, if_citations	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, Current_citations_diff_hindex	0.96	0.89	0.96	0.89
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, Current_citations_diff_hindex_nopub	0.96	0.89	0.96	0.89

Table B.6: Forward feature Selection (D1)

Features	R-Squared		Adj-RSquared	
	1year (2008)	**5years (2012)**	**1year (2008)**	**5years (2012)**
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, No_of_distinct_venues	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, no_of_IF_journals	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, Journal_IF_3	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, Distinct_but_only_IF	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, avg_citations	0.97	0.9	0.97	0.9

Continued on next page

Table B.6 – continued from previous page

Features	**R-Squared**		**Adj-RSquared**	
	1year (2008)	**5years (2012)**	**1year (2008)**	**5years (2012)**
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, coauthors_total_H_index	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, average_h_index_coauthors	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, avg_coauthors_per_article	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, no_publications	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, no_article_as_last_author	0.97	0.9	0.97	0.9

Continued on next page

Table B.6 – continued from previous page

Features	**R-Squared**		**Adj-RSquared**	
	1year (2008)	**5years (2012)**	**1year (2008)**	**5years (2012)**
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, proportion_last_author	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, no_article_as_first_author	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, proportion_first_author	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, m_index	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, if_citations	0.97	0.9	0.97	0.9

Continued on next page

Table B.6 – continued from previous page

Features	**R-Squared**		**Adj-RSquared**	
	1year (2008)	**5years (2012)**	**1year (2008)**	**5years (2012)**
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, Current_citations_diff_hindex	0.97	0.9	0.97	0.9
2007_h_index, collaborations, starting_year_from_2007, No_coauthors, square_root_publications, Current_citations_diff_hindex_nopub	0.97	0.9	0.97	0.9

Appendix C

Random Sample

A sample should be able to represent the whole population and it should be un biased. Random sampling is done to avoid biasness in the sample selection. Considering the nature of problem we have adopted stratified sampling technique. In stratified sampling, population is divided into groups called strata, related cases are grouped together. Within/from each group, sample is randomly selected. Initially, we have divided the entire population on the basis of h-index value. A large number of authors have the h-index value of 0 in this data set as evident from the table A1. In table A1, we have shown, in how many strata we have divided the whole population.

A good random sample should have more than 30 number of records and should be less than 10% of entire population. So we have considered sample size of approximately 5% of entire population. After dividing the whole population in 6

Table C.1: Sample for young researchers

h-index	**Number of Auhtors**	
	Population	**Randomly Chosen Sample**
0	852270	42816
1	497016	24848
2	81572	4119
3	37986	1921
4	20533	1050

TABLE C.2: Standard Error

POPULATION MEAN	0.7695
POPULATION STANDARD DEVIATION	1.588968
STANDARD ERROR	0.001285
SE*95%CONFIDENCE INTERVAL	0.011242
MARGIN OF ERROR	(0.75667,0.77915)
SAMPLE MEAN	0.76791

TABLE C.3: Young Researchers Sample

Authors Having	**Randomly Chosen Sample Size**
h-index Less than 4	42022
Experience less than 4 years	14848
Experience less than 3 years	12592

different stratum, on the basis of h-index value, we have randomly selected 5% records from each stratum.

A good random sample's point estimate (mean in this case) should satisfy/ lie in the margin of error, i.e. confidence level * Standard Error. Standard error in this case is equal to the ratio of population standard deviation and square root of sample size[1].

We have considered a number of samples and checked for standard error in this sample on 95% confidence interval. Some samples satisfied the evaluation criteria, and some didn't. We have selected one of the randomly selected sample, whose mean was within the range of margin of error. Population mean, sample mean for selected sample, standard error and the confidence interval are shown in table C2.

After having random sample from whole population, we have excluded those authors' records who had no publication before 2008, i.e. we have considered only those authors who had published any paper in or before 2007. Further, records according to the definitions of young researchers are extracted. Details are given in table C3.

[1]D.M.Deiaz, C.D. Barr and M. Cetinkaya-Rundel. OpenIntro Statistics. Create Space, 2015, p. 173.

Appendix D

Dependent Variable Distribution

Histogram showing the distribution of h-index values for 2008, 2009,2010,2011 and 2012 are shown below in figures.

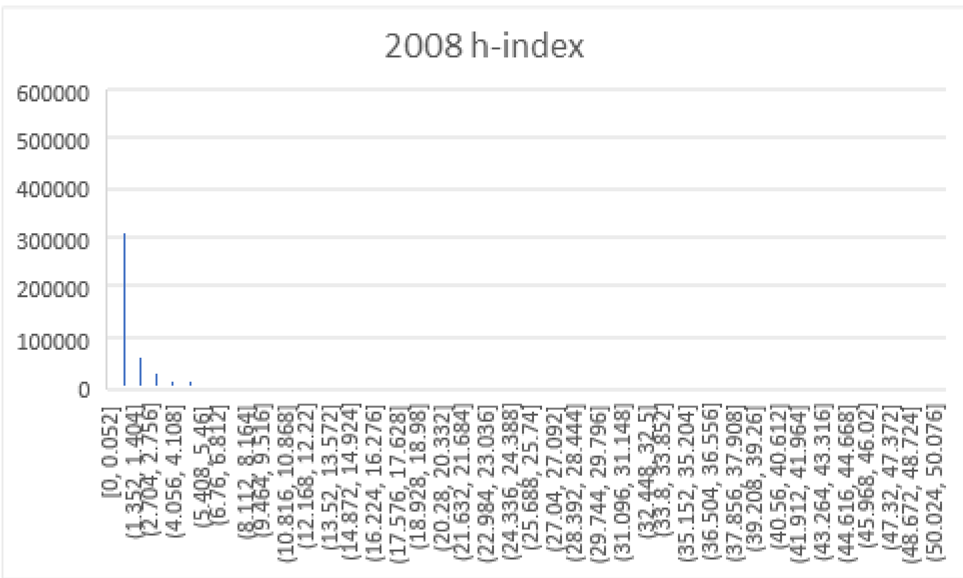

Figure D.1: distribution for h-index values in 2008

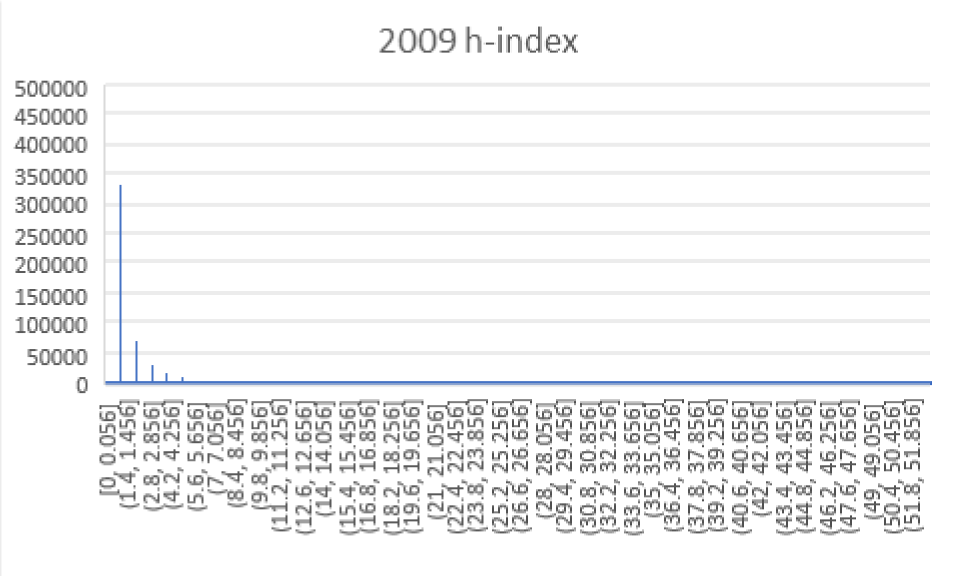

FIGURE D.2: distribution for h-index values in 2009

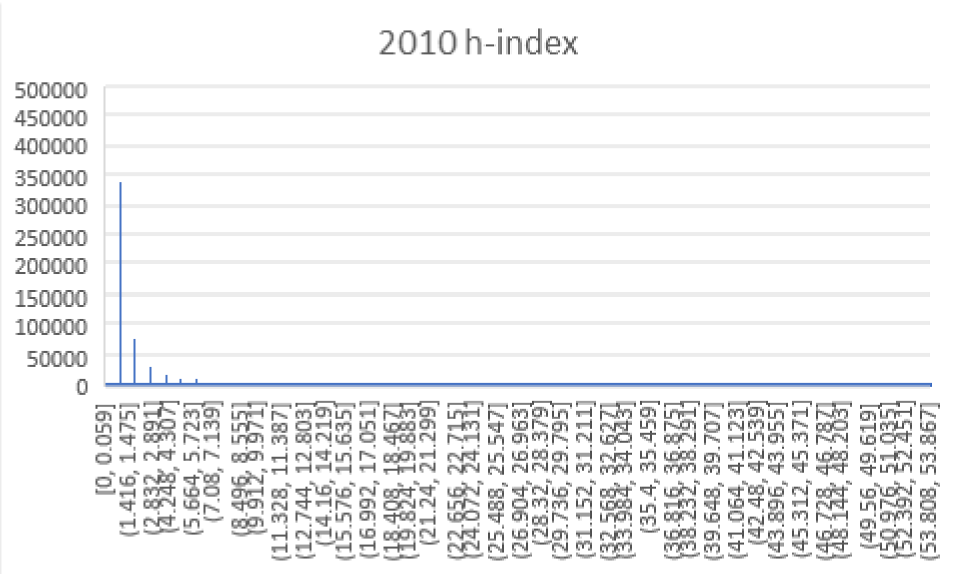

FIGURE D.3: distribution for h-index values in 2010

Appendix E

Authors and Publications Data for NS-Index Calculations

Data considered for NS-Index calculations are given in this appendix. Table E.1 shows the authors and their publications records. Detail of Papers extended by these authors' papers is given in table E.2.

Table E.1: 23 Authors record having h-index '1' in 2007

Sr.	authorID	Author Name	paperID	Paper Title
1	1589581	Alexander Adli	1031502	Piano Sound Characteristics: a Study on Some Factors Affecting Loudness in Digital and Acoustic Pianos
			1154450	A Content Dependent Visualization System for Symbolic Representation of Piano Stream
			1154451	Audio Watermarking Based on the Psychoacoustic Model and Modulated Complex Lapped Transform

Continued on next page

Table E.1 – continued from previous page

Sr.	authorID	Author Name	paperID	Paper Title
2	1461854	Ishan Vaishnavi	1033012	Media Presentation Synchronisation for Non-monolithic Rendering Architectures
			1070332	Multimedia content management support in next generation service platforms
			1293782	NeighbourCast: a synchronisation algorithm for ad hoc networks
3	1434309	Stefan Galler	977842	Interactive presentation: Automatic hardware synthesis from specifications: a case study
			1014810	Specify, Compile, Run: Hardware from PSL
			1397985	Anzu: a tool for property synthesis
4	1421284	Vinh Ninh Dao	967148	VisiCon: a robot control interface for visualizing manipulation using a handheld projector
			987624	CoGAME: manipulation using a handheld projector
			1015036	A semi-automatic realtime calibration technique for a handheld projector
5	1371156	Christian Wolter	1004760	Collaborative Workflow Management for eGovernment 1101542 A Simple, Smart and Extensible Framework for Network Security Measurement

Continued on next page

Table E.1 – continued from previous page

Sr.	authorID	Author Name	paperID	Paper Title
			1407070	Deriving XACML policies from business process models
			1415574	Modeling of task-based authorization constraints in BPMN
6	1312905	Sideny Youlou	1033218	An Efficient Parallel Algorithm for the Longest Increasing Subsequence Problem on a LARPBS
			1075292	Repetitions detection on a linear array with reconfigurable pipelined bus system
			1415633	An efficient sequence alignment algorithm on a LARPBS
7	1272674	Gaëlle Loosli	961255	Comments on the "Core Vector Machines: Fast SVM Training on Very Large Data Sets"
			1097907	Regularization Paths for SVM and SVR
8	1214927	A. Gürhan Kök	1190846	Category Management and Coordination in Retail Assortment Planning in the Presence of Basket Shopping Consumers
			1191116	Inspection and Replenishment Policies for Systems with Inventory Record Inaccuracy

Continued on next page

Table E.1 – continued from previous page

Sr.	authorID	Author Name	paperID	Paper Title
			1191121	Implementation of the Newsvendor Model with Clearance Pricing: How to (and How Not to) Estimate a Salvage Value
			1191223	Demand Estimation and Assortment Optimization Under Substitution: Methodology and Application
9	1135488	Kristene Unsworth	961073	Mobile government fieldwork: a preliminary study of technological, organizational, and social challenges
			1023947	Choices and challenges in e-government field force automation projects: insights from case studies
			1914500	E-government field force automation: promises, challenges, and stakeholders
10	1125613	Luis H. Garcia-Munoz	962425	Recovery Protocols for Replicated Databases–A Survey
			982184	Optimizing Certification-Based Database Recovery
			1409499	Reviewing amnesia support in database recovery protocols
			1409999	Improving recovery in weak-voting data replication
11	1073226	Pei-Luen Patrick Rau	1397155	Provide context-aware advertisements with interactivity

Continued on next page

Table E.1 – continued from previous page

Sr.	authorID	Author Name	paperID	Paper Title
			1397252	A survey of factors influencing people's perception of information security
			1397260	Relevance measurement on chinese search results
			1399821	Developing instrument for handset usability evaluation: a survey study
			1399822	Tips for designing mobile phone web pages for the elderly
			1399832	Design effective navigation tools for older web users
			1399840	Effects of time orientation on design of notification systems
			1399849	The impact of moving around and zooming of objects on users' performance in web pages: a cross-generation study
			1399854	Perception of movements and transformations in flash animations of older adults
			1916357	Player immersion in the computer game narrative
12	1049861	Hui Ye	1214073	Training a real-world POMDP-based dialogue system
			1265104	Agenda-based user simulation for bootstrapping a POMDP dialogue system

Continued on next page

Table E.1 – continued from previous page

Sr.	authorID	Author Name	paperID	Paper Title
			1265136	The hidden information state dialogue manager: a real-world POMDP-based system
13	1016739	Brian Allen	983245	On the beat!: timing and tension for dynamic characters
			984921	A dynamic controller toolkit
			989226	Environment-based physical motion for secondary characters
14	964252	Hildur Olafsdottir	1396452	Sparse statistical deformation model for the analysis of craniofacial malformations in the Crouzon mouse
			1396522	Robust pseudo-hierarchical support vector clustering
			1396531	A statistical model of head asymmetry in infants with deformational plagiocephaly
			1402556	A point-wise quantification of asymmetry using deformation fields: application to the study of the Crouzon mouse model
15	658560	William Cameron	945047	Towards a syllabus repository for computer science courses
			967313	Automatic syllabus classification
			1406469	Using automatic metadata extraction to build a structured syllabus repository

Continued on next page

Table E.1 – continued from previous page

Sr.	authorID	Author Name	paperID	Paper Title
16	525285	Mirco Stern	956598	DIANE: an integrated approach to automated service discovery, matchmaking and composition
			1151240	DIANE: A Matchmaking-Centered Framework for Automated Service Discovery, Composition, Binding, and Invocation on the Web
			1208265	Optimal Locations for Join Processing in Sensor Networks
17	521390	Bassem Elka-rablieh	1001171	Starc: static analysis for efficient repair of complex data
			1019208	Assertion-based repair of complex data structures
			1916712	Efficiently generating structurally complex inputs with thousands of objects
18	445880	Ravi Vaidyanathan	1016041	A Dual Mode Human-Robot Teleoperation Interface Based on Airflow in the Aural Cavity
			1064455	Semi-autonomous micro robot control and video relay through internet and iridium networks
			1785009	Tongue-Movement Communication and Control Concept for Hands-Free Human–Machine Interfaces

Continued on next page

Table E.1 – continued from previous page

Sr.	authorID	Author Name	paperID	Paper Title
19	366041	Shuangjia Chen	1403143	FBSA: a self-adjustable multi-source data scheduling algorithm for P2P media streaming
			1403144	An optimized topology maintenance framework for P2P media streaming
			1403151	QoS adaptive data organizing and delivery framework for P2P media streaming
20	273876	Shachar Fienblit	946389	Distributed desk checking: Research Articles
			978996	Architectures for controller based CDP
			1411795	The advantages of post-link code coverage
21	256987	Cristian Prisacariu	979278	Coordination by Timers for Channel-Based Anonymous Communications
			1399493	A formal language for electronic contracts
			1405216	Model checking contracts: a case study
22	189583	Tonghua Su	977989	Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text

Continued on next page

Table E.1 – continued from previous page

Sr.	authorID	Author Name	paperID	Paper Title
			1006502	HMM-Based Recognizer with Segmentation-free Strategy for Unconstrained Chinese Handwritten Text
			1006703	Skew Detection for Chinese Handwriting by Horizontal Stroke Histogram
			1398494	Gabor-based recognizer for Chinese handwriting from segmentation-free strategy
23	50799	Hannes Moser	944333	Feedback arc set in bipartite tournaments is NP-complete
			1402825	The parameterized complexity of the induced matching problem in planar graphs
			1407157	The parameterized complexity of the unique coverage problem
			1916617	Isolation concepts for enumerating dense subgraphs

Table E.2: Detail of papers extended by 23 authors

paperID	Extended (ref)	Extended (Title)	Citations of extended
1033012	I. Vaishnavi, D. Bulterman, P. Cesar, B. Gao, and J. Jansen. Neighbourcast: A synchronisation algorithm for ad hoc networks. Accepted for publication in IASTED PDCS, 2007.	A synchronisation algorithm for ad hoc networks.	1
1014810	R. Bloem, S. Galler, B. Jobstmann, N. Piterman, A. Pnueli, and M. Weiglhofer. Automatic hardware synthesis from specifications: A case study. In Proceedings of the Conference on Design, Automation and Test in Europe, 2007.	Automatic hardware synthesis from specifications: A case study.	5
1397985	Piterman, N., Pnueli, A., Sa'ar, Y.: Synthesis of reactive(1) designs. In: Proc. Verification, Model Checking, and Abstract Interpretation, pp. 364–380 (2006)	Synthesis of reactive(1) designs. In: Proc. Verification, Model Checking, and Abstract Interpretation	18

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
1004760	P. Schmitz, T. V. Cangh, and A. Boujraf, "R4eGov Deliverable WP3-D2 - Eurojust/ Europol collaboration." www.r4egov.info, 2006.		0
1101542	Cheng, F., Meinel, Ch.: Research on the Lock-Keeper Technology: Architectures, Applications and Advancements. International Journal of Computer and Information Science 5(3), 236–245 (2004)	Research on the Lock-Keeper Technology: Architectures, Applications and Advancements	5
1415633	Chen, L., Juan, C., Pan, Y.: Fast scable algorithm on LARPBS for sequence alignment. In: ISPA Workshops, pp. 176–185 (2005)	Fast scable algorithm on LARPBS for sequence alignment.	3
1097907	Hastie, T., Rosset, S., Tibshirani, R., Zhu, J.: The entire regularization path for the support vector machine. Journal of Machine Learning Research 5 (2004) 13911415	The entire regularization path for the support vector machine	125

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
	Gunter, L., Zhu, J.: Computing the solution path for the regularized support vector regression. In: NIPS. (2005)	Computing the solution path for the regularized support vector regression	6
1190846	Chen, Y., J.D. Hess, R.T. Wilcox, Z.J. Zhang. 1999. Accounting profits versus marketing profits: A relevant metric for category management. Marketing Science. 18 (3). 208-229.	Accounting profits versus marketing profits: A relevant metric for category management	43
1191116	DeCroix, G. A., V. S. Mookerjee. 1997. Purchasing demand information in a stochastic-demand inventory system. Eur. J. Oper. Res. 102 36–57.	Purchasing demand information in a stochastic-demand inventory system	23
1191121	Petruzzi, N., M. Dada. 2001. Information and inventory recourse for a two-market, price setting retailer. Manufacturing and Service Oper. Management . 3 242-263	Information and inventory recourse for a two-market, price setting retailer	29

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
961073	Smart, P.K., Brookes, N.J., Lettice, F.E., Backhouse, C.J. and Burns, N.D. A boundary-based view of product development: A feasibility study. Proceedings of the Institution of Mechanical Engineers, 216 (1). 1-12.		10
	Taylor, J.R. and Van Every, E.J. The emergent organization: communication as its site and surface. Lawrence Erlbaum Associates, Mahwah, N.J., 2000.		0
1409499	Luis H. Garcia-Munoz, J. Enrique Armendariz- Inigo, Hendrik Decker, and Francesc D. Munoz-Esco ı. Recovery protocols for replicated databases - a survey. In Workshop FINA-07, in the AINA-07 Conference. IEEE-CS Press, 2007. Accepted for Publication.	Recovery protocols for replicated databases - a survey	1

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
1409999	Armendariz, J.E., Munoz, F.D., Decker, H., Juarez, J.R., de Mendıvil, J.R.G.: A protocol for reconciling recovery and high-availability in replicated databases. In: Levi, A., Savaş, E., Yenigun, H., Balcısoy, S., Saygın, Y. (eds.) ISCIS 2006. LNCS, vol. 4263, pp. 634–644. Springer, Heidelberg (2006)	A protocol for reconciling recovery and high-availability in replicated databases.	7
1397252	Yenisey, M.M., Ozok, A.A., Salvendy, G.: Perceived security determinants in e-commerce among Turkish university students. Behaviour and Information Technology 24(4), 259–274 (2005)	Perceived security determinants in e-commerce among Turkish university students	6
1399821	Ling, C., Hwang, W., Salvendy, G.: Diversified users' satisfaction with advanced mobile phone features. Universal Access in the Information Society 5(2), 239–249 (2006)	Diversified users' satisfaction with advanced mobile phone features.	2

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
1399832	Coyne, K., Nilsen, J.: Web Usability for Senior Citizens: 46 Design Guidelines Based on Usability Studies with People Age 65 and Older. In: Nielson Norman Group Report (2002) Web Usability for Senior Citizens: 46	Design Guidelines Based on Usability Studies with People Age 65 and Older	21
1399840	McCrickard, D.S., Chewar, C.M., Somervell, J.P., Ndiwalana, A.: A model for notification systems evaluation-assessing user goals for multitasking activity. ACM Transactions on Computer-Human Interaction 10(4), 312–338 (2003)	A model for notification systems evaluation-assessing user goals for multitasking activity	74
1399849	Wang, L., Sato, H., Jin, L., Rau, P.P., Asano, Y.: Perception of Movements and Transformations in Flash Animations of Older Adults. In: 12th International Conference on HumanComputer Interaction	Perception of Movements and Transformations in Flash Animations of Older Adults	1

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
1396452	Olafsd ottir, H., Darvann, T.A., Hermann, N.V., Oubel, E., Ersboll, B.K., Frangi, A.F., Larsen, P., Perlyn, C.A., Morriss-Kay, G.M., Kreiborg, S.: Computational mouse atlases and their application to automatic assessment of craniofacial dysmorphology caused by Crouzon syndrome. Journal of Anatomy (submitted) (2007)	Computational mouse atlases and their application to automatic assessment of craniofacial dysmorphology caused by Crouzon syndrome	6
1396522	Sj ostrand, K., Larsen, R.: The entire regularization path for the support vector domain description. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, Springer, Heidelberg (2006)	The entire regularization path for the support vector domain description.	4

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
1396531	Darvann, T.A., Hermann, N.V., Tenenbaum, M.J., Govier, D., Naidoo, S., Larsen, P., Kreiborg, S., Kane, A.A.: Head shape development in positional plagiocephaly: Methods for registration of surface scans. In: proceedings: Darvann, T.A., Hermann, N.V., Larsen, P., Kreiborg, S. (eds.): "Craniofacial Image Analysis for Biology, Clinical Genetics, Diagnostics and Treatment", Workshop of the 9th MICCAI conference, Copenhagen, Denmark, pp. 59–66 (October 5) (2006)	Head shape development in positional plagiocephaly: Methods for registration of surface scans.	1

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
1402556	Olafsd ottir, H., Darvann, T.A., Ersboll, B.K., Hermann, N.V., Oubel, E., Larsen, R., Frangi, A.F., Larsen, P., Perlyn, C.A., Morriss-Kay, G.M., Kreiborg, S.: Craniofacial statistical deformation models of wild-type mice and crouzon mice. In: Pluim, J.P.W., Reinhardt, J.M. (eds.) Medical Imaging 2007: Image Processing, SPIE, vol. 6512, p. 65121C (2007)	Craniofacial statistical deformation models of wild-type mice and crouzon mice.	5

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
945047	M. Tungare, X. Yu, G. Teng, M. Perez-Qui nones, E. Fox, W. Fan, and L. Cassel. Towards a standardized representation of syllabi to facilitate sharing and personalization of digital library content. In Proceedings of the 4th International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL), 2006.	Towards a standardized representation of syllabi to facilitate sharing and personalization of digital library content	5
956598	M. Klein and B. Konig-Ries. Coupled signature and specification matching for automatic service binding. In Proceedings of the European Conference on Web Services (ECOWS 2004), Erfurt, Germany, September 2004.		25

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
	M. Klein and B. Konig-Ries. Integrating preferences into service requests to automate service usage. In First AKT Workshop on Semantic Web Services, Milton Keynes, UK, Dezember 2004.		11
	M. Klein, B. Konig-Ries, and M. Mussig. What is needed for semantic service descriptions - a proposal for suitable language constructs. International Journal on Web and Grid Services (IJWGS), 1(3/4):328–364, 2005.		30
1916712	Khurshid, S., Garciıa, I., Suen, Y.L.: Repairing structurally complex data. In: Proc. 12th SPIN Workshop on Software Model Checking (2005)	Repairing structurally complex data	11

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
1403143	Huo, L.: Study on key techniques of media streaming over the internet, Ph.D. dissertation,Graduate University of Chinese Academy of Sciences (2006)	Study on key techniques of media streaming over the internet	2
946389	Hoare CAR. Structured programming in introductory programming courses. State of the Art Report on Structured Programming. InfoTech International: Jacksonville, FL, 1976		2
979278	Hennessy, M. and J. Riely, Resource access control in systems of mobile agents, Information and Computation 173:1 (2002), pp. 82–120.	Resource access control in systems of mobile agents	206
1399493	Broersen, J., Wieringa, R., Meyer, J.J.C.: A fixed-point characterization of a deontic logic of regular action. Fundam. Inf. 48, 107-128 (2001)	A fixed-point characterization of a deontic logic of regular action.	11

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
1405216	Gordon Pae, Cristian Prisaariu, and Gerardo Shneider. Model Checking contracts – a case study. In 5th International Symposium on Automated Technology for Veriation and Analysis (ATVA'07), volume 4762 of LNCS, pages 8297, Tokyo, Japan, october 2007. Springer-Verlag.	extended and revised version	3
977989	Su, T., Zhang, T., Guan, D.: HIT–MW dataset for offline Chinese handwritten text recognition. In: The 10th International Workshop on Frontiers in Handwriting Recognition. (2006)	HIT–MW dataset for offline Chinese handwritten text recognition	4
944333	V. Conitzer. Computing Slater rankings using similarities among candidates. In Proc. 21st AAAI. AAAI Press, 2006	Computing Slater rankings using similarities among candidates	20

Continued on next page

Table E.2 – continued from previous page

paperID	Extended (ref)	Extended(Title)	Citations of extended
1402825	Alber, J., Fellows, M.R., Niedermeier, R.: Polynomial-time data reduction for dominating set. Journal of the ACM 51(3), 363–384 (2004)	Polynomial-time data reduction for dominating set	43
	5 Guo, J., Niedermeier, R.: Linear problem kernels for NP-hard problems on planar graphs. In: Arge, L., Cachin, C., Jurdzinski, T., Tarlecki, A. (eds.) ICALP2007. LNCS, vol. 4596, pp. 375–386. Springer, Heidelberg (2007)	Linear problem kernels for NP-hard problems on planar graphs	2
	Guo, J., Niedermeier, R., Wernicke, S.: Fixed-parameter tractability results for full-degree spanning tree and its dual. In: Bodlaender, H.L., Langston, M.A. (eds.) IWPEC 2006. LNCS, vol. 4169, pp. 203–214. Springer, Heidelberg (2006)	Fixed-parameter tractability results for full-degree spanning tree and its dual	4

Continued on next page

Table E.2 – continued from previous page

paperID	**Extended (ref)**	**Extended(Title)**	**Citations of extended**
1407157	E. D. Demaine, M. T. Hajiaghayi, U. Feige, and M. R. Salavatipour. Combination can be hard: approximability of the unique coverage problem. In Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2006), pages 162–171. SIAM, 2006.	Combination can be hard: approximability of the unique coverage problem	25
1916617	H. Ito, K. Iwama, and T. Osumi. Linear-time enumeration of isolated cliques. In Proc. 13th ESA, volume 3669 of LNCS, pages 119–130. Springer, 2005	Linear-time enumeration of isolated cliques	4