

StarRank: Finding Rising Stars in Academic Social Networks



THESIS SUBMITTED FOR PARTIAL REQUIREMENT OF MASTER
OF SCINECES IN COMPUTER SCIENCE

BY
Rashid Hanif Abbasi
560-FBAS/MSCS/F09

SUPERVISED BY
Dr Ali Daud

DEPARTMENT OF COMPUTER SCIENCE AND
SOFTWARE ENGINEERING
FACULTY OF BASIC AND APPLIED SCIENCES
INTERNATIONAL ISLAMIC UNIVERSITY, ISLAMABAD
PAKISTAN
2012



Accession No. TH-10065

MSZ
005-8
ABS

1- Computer networks ; Security measures

2- Internet
(Computer Network) Security measures

DATA ENTERED

Aug 8
2011/04/13

International Islamic University, Islamabad

**Faculty of Basic & Applied Sciences, Department of Computer Science
& Software Engineering**

Dated: Jun 18, 2012

FINAL APPROVAL

It is certified that we have read the thesis, entitled “**StarRank: Finding Rising Stars in Academic Social Network**”, submitted by Rashid Hanif Abbasi Reg. No. 560-FBAS/MSCS/F09. It is our judgment that this thesis is of sufficient standard to warrant its acceptance by the International Islamic University Islamabad for MS Degree in Software Engineering.

PROJECT EVALUATION COMMITTEE

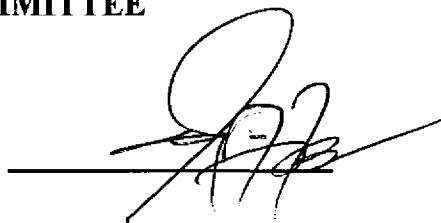
External Examiner:

Dr. Zia ul Qayyum

Professor

Department of Computing and Technology

IQRA University H-9 Islamabad



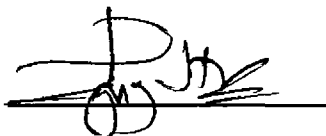
Internal Examiner:

Mr. Ayyaz Hussain

Assistant Professor

Department of CS & SE

IIU Islamabad



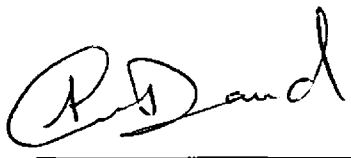
Supervisor:

Mr. Dr Ali Daud

Assistant Professor

Department of CS & SE

IIU Islamabad



Abstract

Academic social network is made up of papers, authors and publication venue nodes. Finding rising stars in these networks is interesting to know for future famous researchers. We have proposed a new technique called StarRank author contribution for author publication quality score and StarRank dynamic publication venue score Proposed method has high h-index, paper and citation result for top rising stars.

DECLARATION

I hereby declare that this work, neither as a whole nor as a part has been copied out from any source. It is further declared that I have conducted this research and have accomplished this thesis entirely on the basis of our personal efforts and under the sincere guidance of my supervisor Dr Ali Daud if any part of this project is proved to be copied out from any source or found to be reproduction of some other project, I shall stand by the consequences. No portion of the work presented in this dissertation has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.



Rashid Hanif Abbasi

560-FBAS/MSCS/F09

DISSERTATION

A Dissertation submitted to the
**Department of Computer Science
and Software Engineering**
International Islamic University Islamabad
As a partial fulfillment of requirements for the award of
The degree of
MS in Computer Science

**This thesis is dedicated to my father, who taught me that the best kind of
knowledge to have is that which is learned for its own sake. It is also
dedicated to my mother, who taught me that even the largest task can be
accomplished if it is done one step at a time**

ACKNOWLEDGEMENTS

All praise to Almighty Allah who has all the names, and who needs no name the most generous, considerate, and compassionate who has blessed mankind with this verdict to think, explore, to learn and discover the hidden secrets of this universe and helped me to broaden the veils of my thought and enabling me to get through the difficulties indulged during this project. Also admiration to our beloved Prophet Muhammad (PBUH) who is always a great source of inspiration of divine devotion and dedication to me.

I would cordially pay my special appreciations and whole heartedly considerations to my reverend supervisor Dr Ali Daud for his endless support, guidance and coordination while conducting this project. I owe him a great respect and honor and I am privileged to work under their supervision. It was their efforts, courage, moral support and endeavoring attitude that helped me to get through any problem or difficulty during each step of this project.

I would also like to pay my gratitude to all my respected teachers making me capable of what I am today due to their guidance and help.

Thanking my friends for always being there for me whenever I needed them for their help, generosity and moral support. Special thanks to Mr. Zahid Mahood Ch. Finally my beloved parents, sisters and my uncle Siddique Abbasi who deserve the credit more than I could ever express for always being completely supportive to me. They have been a constant source of advice, love and devotion to me. From moral to financial they have been blessing me with all the support that I needed up till now in my life.



Rashid Hanif Abbasi
560-FBAS/MSCS/F09

TABLE OF CONTENTS

1.	Introduction.....	1
1.1	Social Network.....	2
1.2	Academic Social Networks(ASN).....	3
1.3	Problems and Difficulties	3
1.3.1	Node Mining	3
1.3.2	Expert Finding.....	4
1.3.3	Name Disambiguation.....	4
1.3.4	Edge Finding	4
1.3.5	Social Influence.....	4
1.3.6	Advisor Recommendation.....	5
1.3.7	Community Mining.....	5
1.4	Component size analysis.....	5
2.	LITERATURE SURVEY.....	7
2.1	Rising Star in Academic Social Network.....	13
2.2	Problem Statement	13
3.	METHODOLOGY.....	15
3.1	PageRank Method	16
3.2	Existing Method PubRank	18
3.3	Proposed Methods StarRank	22
3.3.1	Author Contribution based StarRank	24
3.3.2	Dynamic Publication venue based StarRank	26
3.3.3	Entropy of venue	27
3.3.4	Composite StarRank	29
4.	EXPERIMENTS.....	30
4.1	Data set.....	31
4.2	Performance Measurements	32
4.3	BaseLine Method	33
4.4	Result and Discussion	34
4.4.1	Comparative Study	38
4.4.2	Affect of Alpha Parameter	41

4.4.3	Rising Star paper Finding	42
4.4.4	Rising Stars Venue Base StarRank	44
4.4.4	Value on different Damping factor	46
5.	CONCLUSIONS.....	49
	REFERENCES.....	51

List of Figures

2.1	Authors Graph	10
3.1.1	Authors Network Graph	16
4.3.1	Average Citations PubRank	34
4.4.1	Average Citations Author Contribution based StarRank	35
4.4.2	Average Citations Dynamic Publication venue based StarRank	36
4.4.3	Average Citations Composite StarRank	37
4.4.1.1	Overall Performance Comparison	39
4.4.2.1	Average Citation on different alpha value	42
4.4.3.1	Average Citation of Rising stars paper rank	43
4.4.4.1	Average Citation on Venue Based StarRank	44
4.4.5.1	Effect of Dumping Factor in term of average H-Index of top ten stars ranked by StarRank	46
4.4.5.2	Effect of Dumping Factor in term of average Paper of top ten stars ranked by StarRank	47
4.4.5.3	Effect of Dumping Factor in term of average Citation of top ten stars ranked by StarRank	48

List of Tables

3.1.1	Author in link and out link	17
3.1.2	PageRank Score.....	17
3.2.1	Author weight.....	18
3.2.2	Author total publication	19
3.2.3	Author Publication Quality Score	20
3.2.4	Author PubRank.....	21
3.3.1	Authors Publication.....	22
3.3.2	Author Co-authors.....	23
3.3.1.1	Author Contribution based StarRank Score	26
3.3.3.1	Entropy of venue	27
3.3.2.2	Dynamic Publication venue based StarRank Score.....	28
3.3.3.1	Composite StarRank Score.....	29
4.1.1	Author name and Publication	32
4.3.1-a	Top 10 Results “PubRank”	33
4.3.1-b	Average Citation “PubRank”	33
4.4.1-a	Top 10 Results “Author Contribution based StarRank”	35
4.4.1-b	Average Citations, “Author Contribution based StarRank”	35
4.4.2-a	Top 10 Results “Dynamic Publication venue based StarRank”	36
4.4.2-b	Average Citations, “Dynamic Publication venue based StarRank”	36
4.4.3-a	Top 10 Results “Composite StarRank”	37
4.4.3-b	Average Citations, “Composite StarRank”	37
4.4.1.1	Overall Performance Comparison	38
4.4.1.2	Top Ten Predicted Rising Stars from StarRank	40
4.4.3.1	Rising Stars paper rank and citation.....	42
4.4.4.1	Value on different Damping factor value.....	45

Chapter 1

Introduction

1-INTRODUCTION

1.1 Social Networks

The network is collection of nodes or location which are connected by means of voice, data or video communication. Social network is social structure is made up of “nodes” these individual relationships are occurring between people, group of people and organizations etc Node or people are individual actor in network and an edge is path between the actors .Social network play vital role in social community. For example; animal community, group of business community, group of different village community, different city community, political community are exist. Every node has individual community with connection to hundreds of nodes.

Some other unconnected nodes also exist in social community and make disconnected tree like structure. We search unconnected social community using different techniques as well as through clustering method detect the unconnected nodes .if the x node and y nodes are attached with a network then through the clustering algorithm we detect the nodes and their link with other nodes. In sociology every node of the network [1] is agent of the network and more than one node connection is called social interaction. Degree of node indicates the total edges connection with it and total weight of the edges is indicate the strength of that node. Social network (Class fellow, relatives, friends, other colleagues) linked with each other and share data or information, receive or send e-mail, solicit (obtained) opinions, exchange idea etc

Now a day’s many web based Social network like Face book, MySpace are growing quickly in size and millions of user (e.g. more than 150 million MySpace user in 2006 [5] several other social network are exit now a days (e.g. spount, you tube) [2, 3].Computer network[14] play main role in social network society. People link with many other people and make a computer network .Computer is best communicator for large people networks .Scientist feels that computer play vital role in social community and most important part of social community.

Web based networks increase very quickly and attracts new member. Large number of new user joins the networks which add other friend day by day .In 2006 MySpace has 150 million users and double every year. Academy network handles only business user which member is more than 100,000 in 2002.Frendster [5] social network has 32,000,000 members. Large number of other social networks like Facebook and MySpace user first enters the basic information and creates the complete profile. Scientist realizes that if some user who didn't want to fill the profile basic information in this case how we would extract the basic user profile information from the user during time by time. Social network are attract the attention of academic researchers. Academic social network are cooperating an organized body of researcher in research community.

1.2 Academic Social Networks (ASN)

Many computer science social networks present these days. In Academic social structure the author research in different topics are find most active area with own interest. A new scientists face many problem when they find a relevant expertise researcher with relevant topics (e.g. DBLP [23] and Citeseer [24], Arnetminer [25]) provide much relieve for researchers to search the citations, publications records, co-authors record and find out the most expert supervisor. This information is very useful for new researcher.

1.3 Problems and Difficulties

This section identify some problems in academic social networks from different aspect find most important node, finding the relationship between the node is called edge mining and find the cluster of the node is called community mining.

1.3.1 Node Mining

In Researchers interests is most prominent area if investigation we find the most important area of investigation. In academic social networks (ASN) [19] Researcher interests means who is writing or researching on what topics. For example, author's research different topics according to own interests and find suitable field for research.

1.3.2 Expert Finding

In expert finding “who is expert in this topic “identifying the right researcher with lot expertise with specific knowledge domain. The task of expert finding as well as “who are more expert in topic A” found the relevant person with expertise which fulfill the all recommendation tasks [20].

1.3.3 Name Disambiguation

Academic social network DBLP [23] Citeseer [24] provide much facilities for author in social community to access the researcher total publication, citation which include author name, venue, year of publication .this information is very useful when we find expert person in different field. The author’s names are inconsistent which create more problems when we get some information about author.

1.3.4 Edge Mining

The basic objective of association finding aims at discovering the relationships between different nodes. Now a day large number of online systems which explore social structure as well as networks of friends e.g. ArnetMiner [25], FaceBook researchers Association finding is formulated into further sub task which are people association finding or relationship finding, collaboration finding, and connection finding. In people association we find how different people connect with one another. the direct association between the people is email networks which provide the relationship between the sender and receiver.

1.3.5 Social Influence

Social influence of people is each other is more important.author are influenced by other for different reasons. Authors which are more actives in network more influenced to other in community. Many social network like instant message like (e.g. MSN, Skype Yahoo), video sharing web sites are Flickr, YouTube, social network social networks (e.g. MySpace, Facebook), academic collaboration networks (e.g., Citeseer [24], DBLP [23]) to refer a few, and quantifying the social influence between actor.

1.3.6 Advisor Recommendation

It is a bit hard for them as students has not more idea about having enough exposure about research things When student select advisor before they get lot of information about researcher, keep researcher profile, read researcher paper and find researcher area and other information from profile but still it is very difficult for student because not much more idea about researcher domain.

1.3.7 Community Mining

The identification of community is called community mining .in community mining now a day's face a problem with heterogeneous academic social network. Some community member is strongly attached with each other in network but some other community member which is not strongly attached with each other in network and also exits some disconnected community in network.

1.4 Component size analysis

Social network structure consist of many disconnect group are nodes which have separate structure some method is use for connected and some other is use for disconnected component of the network. Author used three procedure e.g. if a node is link with other node and all edges is connected with other node and consider the closest Node according to closeness. Centrality defines these points when a node exists in shortest path near the other network and disconnected node or graph would have zero cardinality. We mine such community and how these communities increase in size and how many stars are cooperate with each other with other communities. Social network sites provide [16] memory for user which retain our information and maintain our profile, video, photo etc in network. Many characteristics of Facebook and MySpace are same .Facbook is attach with many institute ,university while MySpace is available for general public user .In Facebook user

register using e-mail and create profile User mail account is must for using the Facebook. After register in Facebook users view to every other user in networks. It's depending upon user if some users hide all information from other users. In network student made different college community and connects with many other colleague for long time made friend of friend. Student exchange different information e.g. information about exam, information about paper, study etc

These sites will be favorable for academic because all the information about student is exist in social network which can be use for different decisions. Author has proposed [17] methods for community structure. Many other methods detect the node which attach with only one community and some node attach with more than one community. Facebook MySpace [18] handles the user differently as well as how a user makes profile how hide information from other or profile will be public or private in network. MySpace look like LinkedIn the user information is visible when they have paid account while Facebook if user belongs with same network otherwise the request is denying. User just send request to other and make more friend.

The First social website was develop in 1997 (SixDegrees.com) where user were creating our profile some other as well as (Classmates.com) user can't do this. If we detect the community form large network then before we divide the structure into part and find some node which are connected with many community in hierarchical form .SixDegree were millions of users but couldn't maintain our site in 2000 service has closed. After that many other social sites like Asian Avenue, Black Planet which provides the facility to community which makes personal and dating account. In 2007 live journal sites user manages security setting. In 2001 Ryze.com provides help for user which manages their business in social network. Frindster introduced in 2002 this is dating site which provides help to fined friend profile list .At the start restriction is too much where every user couldn't see the profile information from other user profile but when a user have four or more than four friend in network they can see all the information . In 2005 Facebook add many other feature or activity for high school level students which easily access the other friend or colleague.

Chapter 2

Literature Survey

2-LITERATURE SURVEY

There are many techniques have been used to access the node [7] Google has provided PageRank algorithm that checks efficiently the standard of the web page in large network. In redundant network large number of page are exist which is modified, updated irregularly. PageRank has presented the link structure way with give some value and uses in links from one page to other Pages hold weight when they connect with each other. A page would be more prominent which gain large number of link and rank of page or score will be more. Indefinite numbers of web site are containing more record. Author used the Breadth First search procedure through this method assign number and start from signal node to other. Billion of page are exist in network author has used sparse metric to speed up the procedure in web and efficiently measure the value or score of node. Author [4] proposes PageRank, a procedure that computes the score of all web pages through the graph .Many usages of Page Rank e.g. searching, browsing and traffic estimation. Www large number of irregular information is subsisting. Link structure plays more prominent function when we find the score of the page. Out link and back link is arduous work to indentify between the nodes so due to link structure we overcome to this problem. Some most important page (e.g. Yahoo page) has millions of back-link (citations).

Numerous search engines have counted the citation for page quality. However, numerous problems during this procedure we applied PageRank [7] then used adjacency matrix with directed graph (wecrawler program) and recover this problem with BFS method.

BFS discover all nearest neighbor and start BFS from specific node Numeric value is use to depict the page importance in pageRank. One page linked to other page and send vote to other page if a page has more votes the page would be more prominent. Google measure the value of page through PageRank algorithm. Lawrence Page and Sergey Brin [5, 7] convey the idea of PageRank algorithm. Author [10] proposed the position of Digital library is applied on social network analysis with co-Authorship. The button part of analysis on binary unidirectional model has accustomed and investigates the various established network. Author presented weight directional model about co-authorship network. A graph $G = (V, E, W)$ V is actor or author or node and E is relationship or

edges, W is weight when two author has connected with other author e.g. $W(v_i, v_j)$ we Find the magnitude of the relationship base on the two premises. (a) Occurrence of co-authorship and co-author weight (b) the whole no of joint-author in article if a paper has many author so overall co-author weight will be less. Author [6] proposed PubRank algorithm that examine the star from the community or large network. Author examine two factors when we will mine the star form the networks 1).Collaboration between scholars in social community 2).Increase the quality of publication (top most rank journal/conference paper rank will be more compared with low rank 3). A scientist which is more work together with other would be more prominent in community. Author assign weight to node e.g. author B and weight is fraction of 'C' author publication.

Author [13] proposed a procedure which mine the research paper from multi language database .Author select and examine those papers which is not consider for further discussion. In this paper we use the HITS algorithm [2], the rank of Web pages with many numbers of documents. The HITS algorithm assumes two kinds of prominent page. Authorities which contain high-quality information and hubs which are comprehensive lists of links In academic literature Author assumes two pages in algorithm the one page which is more important and which have many link with other page. Cluster technique play important role which combine all author into one community Weight graph improve the weight of co-Authorship where many author are co-authored doing collaboration very frequently and clustering algorithm gather into one group. Different institutions are gathering into large cluster and indefinite number of author co-author and important role play in research community and connected with other cluster. Co-authorship play prominent role in any community where more than one actor are connected with each other. Author use three steps in network. (a) Undirected binary method (b) directed binary graph (c) binary network Undirected binary graph is extensively use in network community. For example

Article	Author
Article 1	→ {v1, v2, v3, v4, v5}
Article 2	→ {v1, v2, v3}

In article the following author (v_1, v_2, v_3, v_4, v_5) are connected with each other and working together and second article the following author co-authored in first and second article.

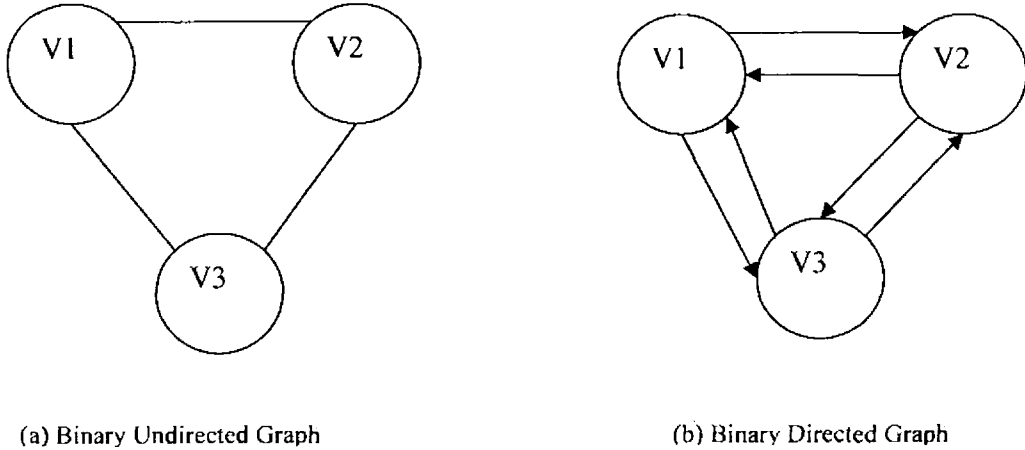


Figure 2.1 Authors Graph

Article .In first transform the undirected graph into directed graph where every adjacent edges is replaced by two directed graph edges and made weight directed graph. In this diagram three author is involve in article first and second and would have high rank or score compare with author (v_4, v_5) which has co-author in only first article .Therefore, the author (v_1, v_2, v_3) would be more prominent than other. Directed weight graph $G=(V, E, W)$ V is node of the graph E is the edges and w is the weight of an every edges In the network which network are connected with other network $w_{ij} = (v_i, v_j)$ weight of v_i and v_j Association of author v_i and v_j in article. If authors v_i and v_j are co-authors in article a_k

$$g_{i,j,k} = \frac{1}{(f(a_k)-1)} \quad (1)$$

$g_{i,j,k}$ Is the degree of author (co-authorship) between v_i and v_j

$$c_{ij} = \sum_k^m g_{i,j,k} \quad (2)$$

Frequency of co-author between two author is sum in all article count all publication if weight is more than authors will be more rank because have more co-authors paper .the weight of author relationship or co-author is sum into one consider in all document .

$$w_{ij} = \frac{c_{ij}}{\sum_{k=1}^n c_{ik}} \quad (3)$$

Dangling strongly affecting on this model If a page has no any out link path or in link with other network is called dangling problem Large number of web page is available on net and due to dangling problem we cannot downloaded So that relationship weight would be zero. any page have no outgoing link is called dangling link We have large number of web page which we can't downloaded now due to dangling factor because dangling relationship have zero weight we remove all the edges link with other node and assign a unique number or unique ID of each link through this ID we perform all operation on node and search from top level parent ID store in database and at the remove from the database. PageRank is use to calculate to rank when the entire link have been removed. Social network [10] has caused to attract more community from Distinct field Author define co-authorship relationship between different author on an ADL JCDL, digital library paper has used for social network analysis Social network directly produce the Efficient relationship between social people .through using graph .Social network graph define in two level Global graph convey the complete network while actor characteristic convey the single prosperities of node. Central node is attached with all other node.

[21] Author discusses a Technique which has used in large system based on topic sensitive PageRank. The original Page Rank algorithm is used for improving the ranking of search-query results computes a single vector by using of link structure of the Web which get the relative important on Web pages to get accurate search results. He proposed a set of Page Rank vectors by the use of inclination using a set of delegate topics, to get more precisely

sense the importance with respect to a particular topic. Page Rank algorithm counts the point's scores for pages gratify the query using the topic of the query keywords. Page Rank can be perceived as if page 'a' has a link to page 'b', then the author of 'a' is un-explicitly granting some significance to page 'b'. Author conducted a number of tests to measure the nature and qualities of topic-sensitive Page Rank. He narrates the similarity measure used to differentiate between two rankings. He inquires how the collaborated rankings differentiate based on both the topic used to bias the rank vectors, as well as the choice of the bias factor. He also present the results obtaining performance of ordinary Page Rank against topic sensitive Page Rank. He provides an initial layout at how the use of query context can be used in collaborative topic-sensitive Page Rank.

Author [22] introduces a newly open text word sense disambiguation method that define the logical evidences with Page Rank style algorithms implemented on graphs obtained from natural language documents and evaluates the correctness of un-implemented algorithm on several sense furnished or trimmed texts, and provided that it can constantly perform better against the proposed knowledge-based word sense disambiguation methods which was presenting in the past. He also find outs methods that co-work with several open text word sense disambiguation algorithms. Author discovers the implementation of Page Rank to semantic networks, and provides evidence that such graph based ranking algorithms can be deployed in language processing applications. he undergoes with a new un-superintended knowledge-based word sense disambiguation algorithm, which succeeds to identifying the meaning of all words in the open text with big margin than any other proposed knowledge based algorithms presented in past. Page Rank is a way to make a decision on the basis of importance of a vertex within a graph, by taking global information constantly repeating itself is inserted from the entire graph, rather than depending on the local vertex-specific information. It is used to allocate the nearest appropriate meaning to a double meaning word within the assumed context. It has two main methods i.e.(a) Knowledge based method(Used for word context disambiguation are usually relevant to all words in open text) (b) Corpus based method(Used for only on few selected words for which a large corpora are available)The knowledge based methods have been developed so far for word sense disambiguation as: (a) Lesk algorithms (b) Semantic

similarity (c) Selectional preferences (d) Heuristic-based methods When one vertex relates to another vertex, it is making a reason to live for the other vertex. The greater number of votes that are available for a vertex, the greater will be the status of the vertex. The importance of the vertex casting the vote determines how important the vote itself is, and this information is also stored in the ranking model. The score linked with a vertex is generated is based on the votes those are casted for it, and the score of those vertices by which these votes are being casted. Word Net is a physical or abstract knowledge base on English language that includes words, meanings, and connection between them. The basic unit of Word Net is a synset, which is a set of synonym words or word phrases, and represents a concept. Word Net defines several semantic relations between synsets, including ISA relations (hypernym/hyponym), part of relations (meronym/holonym), entailment, and others. The input for disambiguation algorithm consists of raw text. The output is text with meaning words. The algorithm consists of many step as well as (a) Preprocessing (b) Graph construction (c) Page Rank (d) Assign word meanings.

2.1 Rising Stars in Academic Social Network

In Academic social network we investigate the rising star form social network. Academic social network is organize body of people or author which co-author in different article in different time .A researcher who have build a strong collaborative network and efficiently co-author in different article would be more important in academic social community .

2.2 Problem Statement

In existing method we can't find the exact expertise of the author and rank of author. PubRank Algorithm handle the author weight not correctly according to author contribution because some author less or more contributed so not fairly give the rank to author and publication rank has assigned statically to each author publication. For example if three author have same total publication. For example, a author has 20 paper and 10 co-authored with (A_x, A_y) and 10 with (A_x, A_z) author A_x more effect to A_y and have equal weight while in first step author A_y co-author participation is more while in 2nd step A_z

participation is less. We have proposed a technique to calculate weight according to author contribution and assign weight to each author in a fair manner. Author calculates Publication Quality score with the help of static rank of publication.

Chapter 3

Methodology

3-METHODOLOGY

3.1 PageRank Method

PageRank algorithm [5, 7] was originally design by Sergey and Larry. Through PageRank algorithm we find the pages important in web. If pages have more in link that page would be more important and score would be more prominent in network. PageRank calculates the rank of every page separately.

$$PR(A) = (1 - d) + d \left(\frac{PR(T_i)}{c_i} + \dots + \frac{PR(T_n)}{c(T_n)} \right) \quad (4)$$

Where

PR (A) = PageRank of A,

PR (T_i) = PageRank to pages T_i which link to page A,

R (T_i) is Number of outbound links on page T_i

D is a damping factor value =0.85

Suppose author network web graph which contain three node W, X, Y, Z. W does not link with any node and X has linked with 'W' and 'Y', 'Y' has linked with 'W' and 'Z', 'Z' has linked with 'W', 'X', 'Z'. 'W' node have three in link which is coming from 'Y', 'X', 'Z' and out link of 'Y' is 'Z', 'X' So PageRank equation is

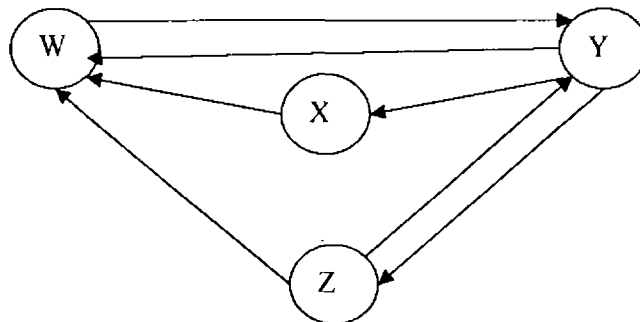


Figure 3.1.1 Authors Network Graph

$$PR(W) = (1 - d) + d \left[\frac{PR(X)}{L(X)} + \frac{PR(Y)}{L(Y)} + \frac{PR(Z)}{L(Z)} \right] \quad (5)$$

$$PR(X) = (1 - d) + d \left[\frac{PR(Z)}{L(Z)} \right] \quad (6)$$

$$PR(Y) = (1 - d) + d \left[\frac{PR(X)}{L(X)} + \frac{PR(Z)}{L(Z)} \right] \quad (7)$$

$$PR(Z) = (1 - d) + d \left[\frac{PR(Y)}{L(Y)} \right] \quad (8)$$

Table 3.1.1 Author inlink and outlink

S#	W	X	Y	Z
W	0	1	1	1
X	1	0	1	1
Y	1	0	0	1
Z	0	0	1	0

We consider following three author node calculation with many repetition steps

Table 3.1.2 PageRank Score

S#	Repetition of PR (X)	PR (Y)	PR (Z)
1	1	1	1
2	0.71667	1.30458	0.51963
3	0.66686	0.68606	0.34438
4	0.44196	0.54827	0.30534
5	0.39186	0.51758	0.29665
6	0.38069	0.51074	0.29471
7	0.37821	0.50921	0.29428
8	0.37766	0.50887	0.29418

-	-	-	-
-	-	-	-
-	-	-	-
n-2	0.37749	0.50877	0.29415
n-1	0.37749	0.50877	0.29415
n	0.37749	0.50877	0.29415

Above graph we have taken node Inlink and node outlinks and calculate PageRank through PageRank equation .After some repetition we find following result. It's very difficult to evaluate the result for large web network and easy to evaluate the result for small web network.

3.2 Existing Method PubRank

PubRank Method determines the stars from web communities. Large numbers of nodes are connecting with small or large web network Nodes describe authors and edges describe relationship or link with other node .When authors $[6] (v_k, v_l)$ are co-author in any artical we put the weight to $v_k = (v_k, v_l)$ is Fraction of v_l author which is co-author with v_k Moreover, the weight of $v_l = (v_l, v_k)$ is fraction of v_k .we have taken DBLP [23] data and calculated author co-author weight.

Table 3.2.1 Author weight

S#	Author	Co-author	weight
1	tsung-kai yang	jln-yu bai	1.0
2	tsung-kai yang	sheng-chang chen	0.5
3	tsung-kai yang	chyi-ren dow	0.1
4	tsung-kai yang	cheng-min lin	0.5
5	young-rok yang	kwang ho chun	1.0
6	young-rok yang	seung-hyun min	0.3
7	young-rok yang	myoung-jun kim	1.0
8	catherine deegan	kabita shakya	0.4

9	catherine deegan	fran hegarty	1.0
10	catherine deegan	charles markham	1.0

The following existing author weight with all author suppose the author 'michael t. orchard' weight with all other co-authored in all paper. 'michael t. orchard' is co-author with 'scott m. lepresto' in first paper further 'michael t. orchard' is co-author with 'onur g. guleryuz' in second paper . 'michael t. orchard' co –author with 'kannan ramchandran' in thrid paper .moreover , 'lawrence a. rowe' is co-author with 'radhika malpani' in first paper 'lawrence a. rowe' is co-author with 'ketan mayer-patel' in 2nd paper . 'lawrence a. rowe' is co- author with 'andrew swan' in third paper . 'lawrence a. rowe' is co-auhtor with 'joseph michiels' in fourth paper and so on...The author 'lawrence a. rowe' have total 11 publication and co-authors with many other author.

Table 3.2.2 Author Total Publication

S#	Author	Publication
1	xibei jia	2
2	joost vennekens	2
3	freacute vernier	2
4	charalambos vrasidas	1
5	robert pitts	1
6	uri zwick	12
7	naoko matsumoto	1
8	joseph m. kahn	1
9	luca dallasta	1
10	andrew cofler	1

Example 1

Author $(v_l, v_m, v_n, v_o, v_p, v_q \dots)$,

Co-authorship $(v_l, v_m), (v_l, v_n), ((v_l, v_o) \dots$

Author (v_l) publication = 15

Author v_m Publication = 10

Co-Authorship publication $(v_m, v_n)=5$

$$\lambda(v_i) = \frac{1}{|p|} * \sum_{i=1}^p \frac{1}{\alpha^{r(\text{pub})-1}} \quad (9)$$

$\lambda(v_i)$ is Publication Quality score of author v_i

We have taken data form DBLP [23] and calculated publication quality score of following author

Table 3.2.3 Author Publication Quality Score

S#	Author	Publication Quality Score
1	tsung-kai yang	0.14598
2	young-rok yang	0.16000
3	catherine deegan	0.24000
4	ammer al-khayri	0.16000
5	satoshi hada	0.12000
6	laura recalde	0.30303
7	judith s. donath	0.33333
8	james e. lewis	0.54545
9	patricia gilfeather	0.66667
10	joost vennekens	0.54545

Where

pub_i is i-th publication,

$r(pub)$ is publication rank of paper,

α Value is ($0 < \alpha < 1$)

$$pubRank(p_i) = \frac{1-d}{n} + d * \sum_{j=1}^{|v|} \frac{w(p_i, p_j) * \lambda(p_i) * pubRank(p_j)}{\sum_{k=1}^{|v|} w(p_k, p_j) * \lambda(p_k)} \quad (10)$$

Where

n is total number of scientist

$w(p_i, p_j)$ Is weight for edges (p_i, p_j)

$\lambda(p_i)$ Is publication quality score

We have taken data from DBLP [23] and above PubRank equation [6] we have calculated author PubRank score of following author.

Table 3.2.4 Author PubRank

S#	Author	PubRank
1	wei-ying ma	0.27236
2	wei wang	0.19670
3	mahmut t. kandemir	0.16833
4	philip s. yu	0.16575
5	zheng chen	0.15396
6	edward a. fox	0.14922
7	hsinchun chen	0.14386
8	aoying zhou	0.13901
9	david blaauw	0.13401
10	donald f. towsley	0.13079

The large web network increase rapidly and publish thousand of paper each year Therefore, find the pub-score for all author and find some expert or hidden star form network in community We can find and some author which have low position but in future would be more noticeable.

3.3 Proposed Methods StarRank

In this portion we have proposed new technique called StarRank .We have solved the author contribution problem with author quality score give rank to author in fair manner. less contributed author score would be less and more contributed author score would be more score. We have computed the author rank and calculated the entropy of venue .At the end we have hybridize the author contribution technique and entropy of venue in composite StarRank. We have taken data from DBLP [23]. The author name and publication data are following.

Table 3.3.1 Authors Publication

S#	Author	Publication
1	tsung-kai yang	instant messaging based multi
2	young-rok yang	soft real time guaranteed java thread mapping method
3	catherine deegan	dynamic response measurement clinical gas analysers
4	ammer al-khayri	application cepstrum algorithms speech recognition
5	satoshi hada	xml access control static analysis
6	laura recalde	reachability autonomous continuous petri net systems,continuization timed petri nets performance evaluation observation control
7	judith s. donath	telemurals linking remote spaces social catalyts

8	charles koelbel	scheduling workflow applications grads
9	james e. lewis	multiple copy distributed genetic algorithm, performance study distributed genetic algorithms,
10	skef iterum	inheritance inspired interface versioning corba

Table 3.3.2 Author Co-Authors

S#	Author	Co-Author 1	Co Author 2	Co – Author 3	Co – Author 4	Co - Author 5
1	jin-yu bai	sheng-chang chen	chyi-ren dow	cheng-min lin		
2	young-rok yang	ho chun, seung-hyun min	myoung-jun kim			
3	catherine deegan	kabita shakya	fran hegarty	charles markham		
4	ammer al-khayri	raed abu zitar	mohammed abu arqub	anwar al-shrouf		
5	satoshi hada	michiharu kudo	makoto murata	akihiko tozawa		
6	laura recalde:jorge jú	lvez, manue l silva				
7	judith s. donath	karrie karahalios				
8	charles koelbel	anirban mandal	ken kennedy	anshuman dasgupta	b. liu, gabriel marin	. johnsson
9	james e. lewis	rammohan k. ragade	anup kumar			
10	skef iterum	ralph campbell				

Many author connected with several other author and published many paper
 For example jin-yu bai is co –author with sheng-chang, chyi-ren Dow, cheng-min lin.
 We have taken following data from DBLP [23].

3.3.1 Author Contribution based StarRank

In this section we have calculated author contribution weight. Author (v_j, v_m, v_n) have published 20 total numbers of papers in 10 co-Authorship papers between (v_j, v_m). Another 10 co-Authorship publication between (v_j, v_n) in both case author v_j influence to author v_m and v_n and weight would be high because author v_j author contribution is more compare with other co-author. We have computed the author contribution [11] when a paper has more than one co-author. Therefore, author score sometime be wrong. In this case authorship is harmful for more contributed author because equal contribution score does to unfair rank. So, single author contribution is more effective for every scholar.

Authors in one paper E, F, G, H, I, J

Author rank (K) 1, 2, 3,4,5,6

$$H_n = \sum_{k=1}^n \frac{1}{k} \quad (11)$$

$$S_k = \frac{1}{(k * H_n)} \quad (12)$$

Where 'k' is author rank,

'n' is number of authors

The co-author contribution value is $1/k$ and s the sum of all value and rank of first and last author value will be different.

$$H_n = 1/1 + 1/2 + 1/3 + 1/4 + 1/5 + 1/6 = 2.45$$

$$s_E = 1 / (1 * 2.45) = 0.4081$$

$$s_F = 1 / (2 * 2.45) = 0.204082$$

$$s_G = 1 / (3 * 2.45) = 0.1357$$

$$s_H = 1 / (4 * 2.45) = 0.102041$$

$$s_I = 1 / (5 * 2.45) = 0.082$$

$$s_j = 1 / (6 * 2.45) = 0.07$$

We set the author contribution weight of the edges $AC(p_i, p_j)$ is the fraction of author p_j moreover, we set the author contribution weight of the edges $AC(p_j, p_i)$ is the fraction of author p_i . We have calculated author contribution in following example

Example

If author K have 4 totals paper and we have calculated the author individual contribution in co-author paper if author K and L co-authored in two papers.

The author K= (1= First paper) (1=rank of author in first paper), (2 =Second paper) (3= rank of author in second paper), (3= third paper) (2= rank of author in third paper), (4=fourth paper) (1=rank of author in four paper)

Value of L= 1(2), 2(2), 3(1), rank of M=1(3), 2(4), 3(4) and rank of N=1(1), 2(3), 3(2), 4(1)

$$v_l = (v_l, v_k) / v_k \quad \frac{2}{4} = .5$$

$$v_m = (v_m, v_n) / v_n \quad \frac{2}{4} = .5$$

$$v_l = \frac{(.5+.5)+(1+.33)}{(1+.33+.5+1)} = .823$$

$$v_m = \frac{(.33+.25)+(1+.33)}{(1+.33+.5+1)} = .67$$

$$StarRank(p_i) = \frac{1-d}{n} + d * \sum_{j=1}^{|v|} \frac{AC(p_i, p_j) * \lambda(p_i) * starRank(p_j)}{\sum_{k=1}^{|v|} AC(p_k, p_j) * \lambda(p_k)} \quad (13)$$

We have calculated StarRank author contribution score using above equation .we have taken data from DBLP [23].

n = number of author

Ac is author contribution

Table 3.3.1.1 Author Contribution based StarRank Score

S#	Authors	StarRank Score
1	wei wang	0.33971
2	wei-ying ma	0.21821
3	philip s. yu	0.20885
4	mahmut t. kandemir	0.16128
5	jiawei han	0.15577
6	jeffrey xu yu	0.14728
7	hongjun lu	0.14228
8	zheng chen	0.12979
9	mary jane irwin	0.12266
10	baile shi	0.12074

3.3.2 Dynamic Publication venue based StarRank

In first method we calculated the StarRank score using author contribution method and this section we have calculated dynamic publication venue based SatarRank. We have calculated entropy of venue.

Entropy is measure[12,26,27] the amount of disorder in a system if disorder is more than more entropy and less disorder mean system is better or low entropy mean it is more topic specific and high entropy means that it is less topic specific. If a process creates more disorder the entropy change of the process is positive. According to second law of thermodynamics any self generated process the entropy of universe must increase. Entropy is distinct from energy .Energy is neither created nor destroyed but any self generated process creates the entropy. If heat is flow into or out of system entropy will be change in the system. More heat involves more entropy. The total entropy will be change from the heat flow and by created a process. The high reputable venue has low entropy and less reputable venue has high entropy.

3.3.3 Entropy of Venue

we have calculated the entropy of venue .If a paper have five word ‘Searching’, ‘Rising’ , ‘Stars’ ,‘Bibliography’ , ‘Networks’ so , probability of first word is 1/5 , 2nd word have 1/5, 3rd word have 1/5 ,4th word have 1/5 ,5th word have 1/5 probability .

Entropy of venue can be calculated through following equation

$$prob(w_i) = \frac{\text{frequency of } w_i}{\text{total word of that doc}} \quad (14)$$

$$\text{Entropy of } w_i = prob(w_i) * \log(prob(w_i)) \quad (15)$$

Where w_i is the probability of $word_i$

$$\text{Entropy}(doc) = \sum_{i=1}^{doc} Entropy \text{ of } (w_i) \quad (16)$$

$$\text{Entropy of confrence} = \frac{\text{sum of entropy of doc}}{\text{number of doc in venue}} \quad (17)$$

Some High level and low level venues entropy are shown in the following table
The high reputable venue has low entropy and less reputable venue has high entropy.

Table 3.3.3.1 Entropy of venue

High Level Venues	Entropy	Normal Venues	Entropy
SAM	1.33	BIBE	1.99
SIGCSE	1.69	CBMS	1.99

SIGMOD	1.71	CODES	1.97
SIROCCO	1.69	DAGM	1.90
SODA	1.67	DSD	1.94
SPAA	1.72	CBMS	1.96
FC	1.65	IPDPS	1.93

$$\lambda(dpq) = \frac{1}{|p|} * \sum_{i=1}^p \frac{1}{\alpha \text{Entropy of venue}} \quad (18)$$

$$\text{StarRank}(p_i) = \frac{1-d}{n} + d * \sum_{j=1}^{|v|} \frac{w(p_i, p_j) * \lambda(dpq) * \text{starRank}(p_j)}{\sum_{k=1}^{|v|} w(p_k, p_j) * \lambda(dpq)} \quad (19)$$

Where

$W(p_i, p_j)$ Author weight with co-authors,

Dpq = dynamic publication quality score

In this case quality of the result produced by dynamic publication venue based StarRank score is more prominent compare with pub-Rank score. We have taken data from DBLP [23] and calculated dynamic publication venue base StarRank score using above equation. Dynamic publication venue based StarRank score of author are following.

Table 3.3.3.2 Dynamic Publication venue based StarRank score

S#	Author	StarRank Score
1	ei-ying ma	0.31667
2	jiawei han	0.26913
3	divesh srivastava	0.22418

4	philip s. yu	0.21102
5	erik d. demaine	0.16166
6	zheng chen	0.16069
7	sebastian thrun	0.15905
8	nick koudas	0.15879
9	bertram ludaumscher	0.13589
10	yufei tao	0.13481

3.3.4 Composite StarRank

In third method we have calculated the rank of author according to author contribution and dynamic publication venue based. We hybridizes the first method Author contribution and second method dynamic publication venue based and calculates the final composite StarRank score. We have used DBLP [23] data and obtained following author score.

$$StarRank(p_i) = \frac{1-d}{n} + d * \sum_{j=1}^{|V|} \frac{AC(p_i, p_j) * \lambda(dpq) * starRank(p_j)}{\sum_{k=1}^{|V|} AC(p_k, p_j) * \lambda(dpq)} \quad (20)$$

Table 3.3.4.1 Composite StarRank Score

S#	Author Name	StarRank Score
1	ying ma	0.34328
2	philip s. yu	0.22413
3	jiawei han	0.20701
4	zheng chen	0.18506
5	divesh srivastava	0.17392
6	wei wang	0.15528
7	hsinchun chen	0.15212
8	erik d. demaine	0.14415
9	sebastian thrun	0.14229
10	lee tan	0.13742

Chapter 4

Experiments

4 - EXPERIMENTS

In this chapter we have discussed the implementation scenarios and obtained the results in detail. The implementation scenario is divided into four parts. In the first part we have discussed baseline line PubRank Method and calculate the PubRank score of the authors. In second part we have discussed author contribution base rank, we have calculated the author contribution on every paper publication and calculated the StarRank score of all authors, in third part we have discussed the dynamic publication venue based rank through using entropy method and calculated the author StarRank score. In fourth part we have used composite method using dynamic publication venue based and author contribution based method to calculated StarRank score of all author.

4.1 Data set

We have taken data from Digital Bibliography and Library Project DBLP [23]. We have used the data from 1996-2000 to predict the rising stars. The data harvested form XML file which contain with in this "<ListRecords>" tag etc and used parser techniques to extract the data from xml file. We will take title of paper, author name and conference/journal where papers have published. After the complete data extraction the size of data is more and reserve more memory When data size in GBs or even MBs then consume more time and memory therefore, more memory is require for data. We read some specific tags or line and keep it as whole file from xml file. Some words which is frequently use when we applying several method these word create a problem so we remove all these word from data. For example the, these, so, have, to, are, some, when, it, that, this, of etc

```
String dirname = "F: /PROJECT/Raw Data";
```

We give directory path to main folder .the Raw Data is the folder name in which there are many other folders. We read every file one by one and extract required data from file and store the data.txt .we extract the paper name and author name using these two tags and we read the data from inside these tags.

Table 4.1.1 Author name and publication

Author Name	Paper Name
tsung-kai yang	instant messaging based multi
young-rok yang	soft real time guaranteed java thread mapping method
catherine deegan	dynamic response measurement clinical gas analysers
Ammer al-khayri	application cepstrum algorithms speech recognition
satoshi hada	xml access control static analysis
laura recalde	reachability autonomous continuous petri net systems,continuization timed petri nets performance evaluation observation control
judith donath	telemurals linking remote spaces social catalysts

4.2 Performance Measurements

No ground truth about raking of rising star available. We have taken the data from DBLP [23] and checked the author and his paper citations result from arnetminer [25]. We have performed many experiments by using Pubrank method, Author contribution based StarRank, Dynamic publication venue based StarRank and composite based StarRank. We measure the performance and evaluate the previous method and proposed method and randomly select the top ten rank of different author for all query and we check for each query and count the number of citation for top ten author. First the pervious method we have calculated and we have selected top 10 author and read the total citation, H-index, Paper form arnetminer [25] If a author have more citation,

H-index, Paper average that would be consider more better. If an author has less citation, paper, H-index that would be less important in community.

4.3 BaseLine Method

PubRank Method determines the stars from web communities. Large number of node is connected with small or large web network. Nodes describe authors and edges represent relationship. When authors (v_k, v_l) are co-author in a paper. We put in the weight to $v_k = (v_k, v_l)$ is Fraction of v_l author which is co-author with v_k . Moreover, the weight of $v_l = (v_l, v_k)$ is fraction of v_k . Author calculated the quality score and statically allocating the rank for author without any author contribution. We have calculated the previous method and calculated the rank score. We have selected top ten author and we have taken the citation, total paper publication, H-index from arnetminer [25]

Table 4.3.1-a Top 10 Results “PubRank”

S#	Author Name	PubRank	H-index	Papers	Citation
1	ying ma	0.27235	59	277	14355
2	wei wang	0.19670	49	712	9873
3	mahmut t. kandemir	0.16834	43	492	7945
4	philip s. yu	0.16575	80	658	28429
5	zheng chen	0.15396	35	162	3937
6	edward a. fox	0.14922	1	1	1
7	hsinchun chen	0.14387	53	391	8161
8	Qaoying zhou	0.13902	22	200	2157
9	david blaauw	0.13401	51	257	9259
10	donald f. towsley	0.13079	87	382	25187

Table 4.3.1-b, Average Citations “PubRank”

S#	Total	Average
H-index	480/10	48
Paper	3532/10	353.2
Citations	109307/10	10930.7

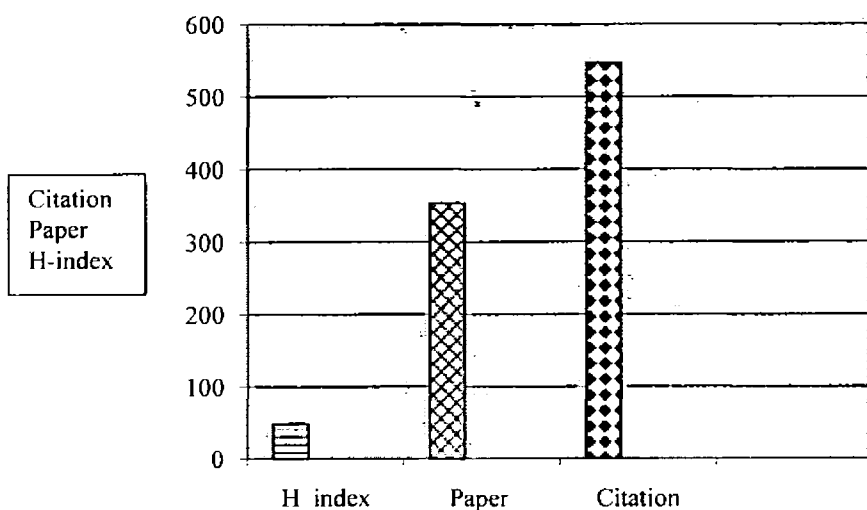


Figure 4.3.1: Average Citations, "PubRank"

4.4 Result and Discussion

We have performed several experiments to obtain the result. We have separately calculated rank of author. In first method we calculated the rank of author score through PubRank method (previous method) and obtained the citation from arnetminer[25]. In second method we observed the result through author contribution based StarRank and taken the citation, paper, H-index from arnetminer [25] for top ten authors, in third method we have calculated dynamic publication venue based StarRank and taken the citation, paper, H-index from arnetminer [25] for top ten authors. In fourth method We hybridized the author contribution based StarRank and daynamic publication venue based StarRank selected top ten authors after obtain the citation, H-index, paper form arnetminer[25] and compare the result with Base line method.

Table 4.4.1-a, Top 10 Results "Author Contribution based StarRank"

S#	Author Name	StarRank	H-index	Papers	Citation
1	wei wang	0.33971	59	277	14355
2	wei-ying ma	0.21821	49	712	9873
3	philip s. yu	0.20885	43	492	7945
4	mahmut t. kandemir	0.16128	80	658	28429
5	jiawei han	0.15577	35	162	3937
6	jeffrey xu yu	0.14728	1	1	4
7	hongjun lu	0.14228	53	391	8161
8	zheng chen	0.12979	22	200	2157
9	mary jane irwin	0.12266	51	257	9259
10	baile shi	0.12074	87	382	25187

Table 4.4.1-b, Average Citations "Author Contribution based StarRank"

S#	Total	Average
H-index	487/10	48.7
Paper	3622/10	362.2
Citations	126515/10	12651.5

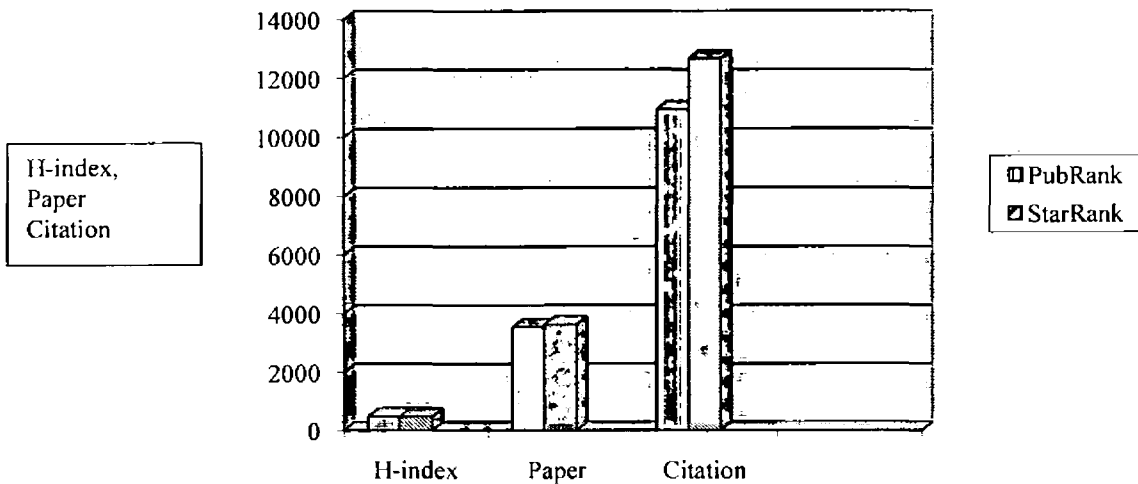


Figure 4.4.1: Average Citations, "Author Contribution based StarRank"

We have taken the data from DBLP [23] and through contribution base we have calculated StatRank. We have selected top ten authors from StarRank score and taken h-index, paper, citation from arnetminer [25] and compare the result with PubRank method in above graph.

Table 4.4.2-a, Top 10 Results Dynamic Publication venue based StarRank

S#	Author Name	StarRank	H-index	Papers	Citation
1	ying ma	0.31666	59	277	14355
2	jiawei han	0.26913	88	536	46654
3	divesh srivastava	0.22418	55	239	11520
4	philip s. yu	0.21102	80	658	28429
5	erik d. demaine	0.16166	49	379	7361
6	zheng chen	0.16069	35	162	3937
7	sebastian thrun	0.15906	84	240	29544
8	nick koudas	0.15879	46	146	7132
9	bertram ludäscher	0.13589	10	30	553
10	yufei tao	0.13481	42	116	6613

Table 4.4.2-b Average Citations “Dynamic Publication venue based StarRank”

S#	Total	Average
H-index	548/10	54.8
Paper	2783/10	278.3
Citations	156098/10	15609.8

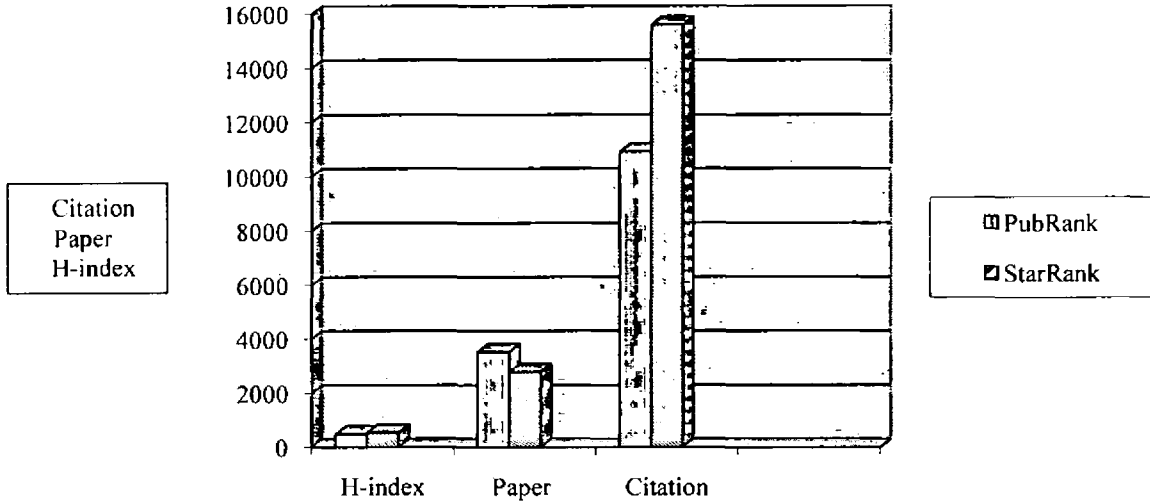


Figure 4.4.2: Average Citations, “Dynamic Publication venue based StarRank”

We have taken the data from DBLP [23] and through dynamic publication venue based we have calculated StatRank. We have selected top ten authors from StarRank score and taken h-index, paper, citation from arnetminer [25] and compare the result with PubRank method in above graph.

Table 4.4.3-a, Top 10 Results “Composite StarRank”

S#	Author Name		H-index	Papers	Citation
1	ying ma	0.34328	59	277	14355
2	philip s. yu	0.22413	80	658	28429
3	jiawei han	0.20701	88	536	46654
4	zheng chen	0.18506	35	162	3937
5	divesh srivastava	0.17392	55	239	11520
6	wei wang	0.15528	49	712	9873
7	hsinchun chen	0.15212	53	391	8161
8	erik d. demaine	0.14415	49	379	7361
9	sebastian thrun	0.14229	84	240	29544
10	lee tan	0.13742	44	285	6824

Table 4.4.3-b Average Citations “Composite StarRank”

S#	Total	Average
H-index	596	56.6
Paper	3852	385.2
Citations	166658	16665.8

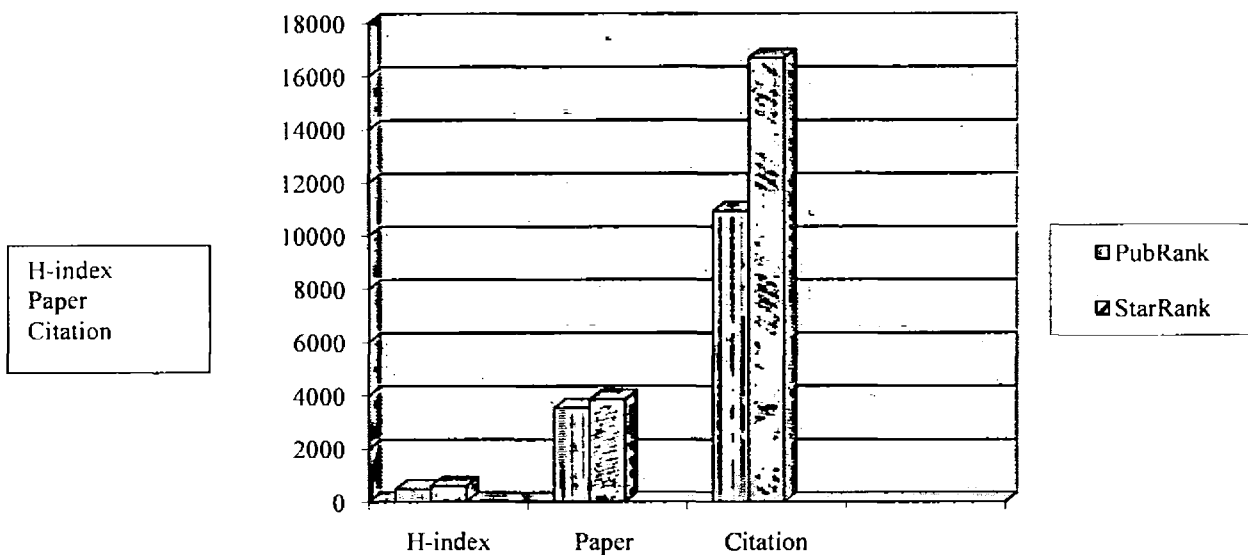


Figure 4.4.3: Average Citations “Composite StarRank”

We have taken the data from DBLP [23] and through hybridize StarRank we have calculated StatRank. We have selected top ten authors from StarRank score and taken h-index, paper, citation from arnetminer [25] and compare the result with PubRank method in above graph.

4.4.1 Comparative Study

We have performed several experiments to obtain the result and separately we have calculated PubRank method and obtained the h-index, paper and citation from arnetminer [25]. Proposed method we have calculated author contribution based StarRank, dynamic publication venue based StarRank and composite based StarRank. We have taken the data from DBLP [23] and through author contribution base we have calculated StatRank. We have selected top ten authors from StarRank score and taken h-index, paper, citation from arnetminer [25] and compare the result with PubRank method. Secondly we have calculated dynamic publication venue base method and obtained the h-index, citation, paper from arnetminer [23] and compare the result with baseline method. Composite StarRank method we have calculated StarRank score and taken h-index, paper, citation from arnetminer [25] compare the result with Baseline method. Proposed method has h-index, paper and citation compare with PubRank or previous method.

Table 4.4.1.1 Overall Performance Comparison

S#	Base line	Proposed Method		
	PubRank Average	Author contribution base StarRank Average	Dynamic publication venue base StarRank Average	Composite StarRank Average
H-index	480/10 =48	487/10=48.7	548/10=54.8	596/10=59.6
Paper	3532/10=353.2	3622/10=362.2	2783/10=278.3	3852/10=385.2
Citation	109307/10=10930.7	126515/10=12651.5	156098/10=15609.8	166658/10=16665.8

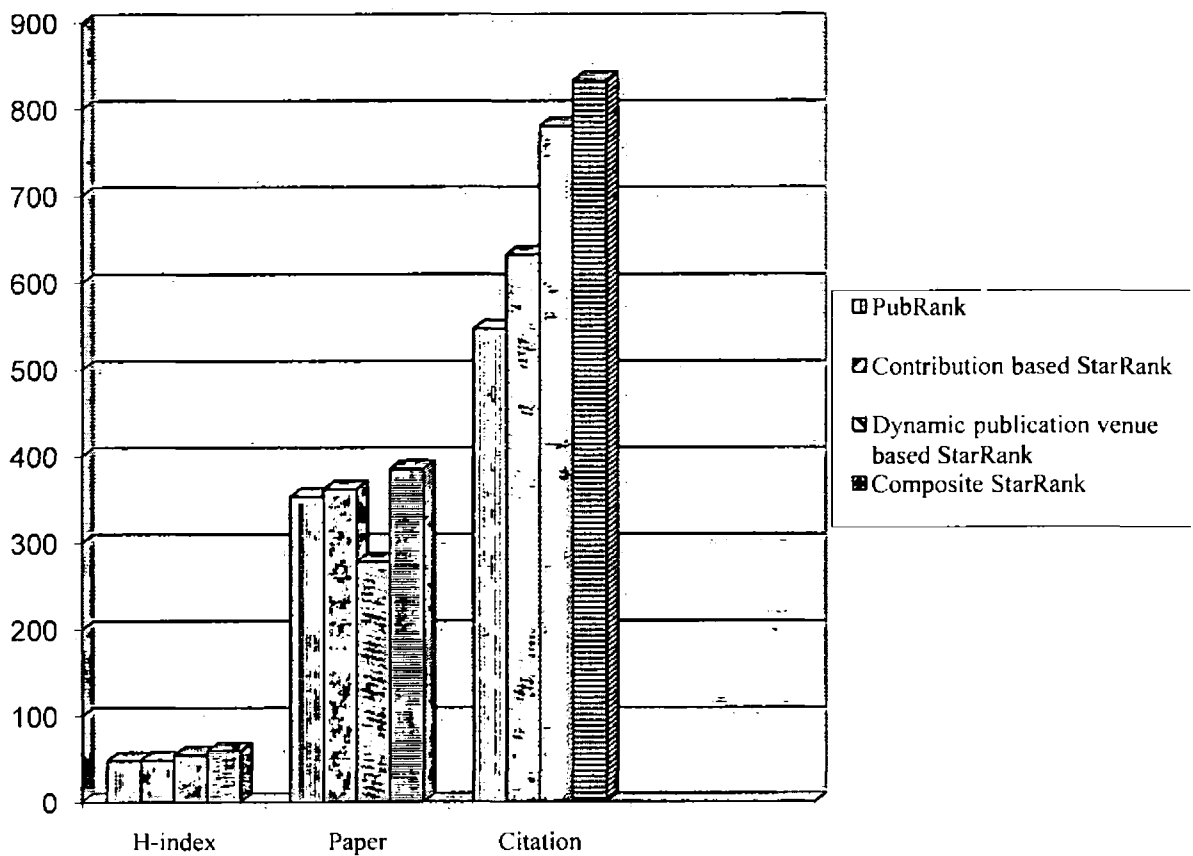


Figure 4.4.1.1: Overall performance comparison

In the above figure the following graph line shows the average of H-index, paper and citation of top ten authors in PubRank (Baseline method shown Vertical line) and else the vertical line shows the proposed method author contribution base StarRank, dynamic publication venue base StarRank and Composite StarRank. We have performed many experiments using data from Digital Bibliography and library project DBLP [23].we implemented our methods and calculate the StarRank and after we have selected the top

ten authors. We have taken citations, H-index, Paper from arnetminer [25] and compare the results with baseline method.

Table 4.4.1.2 Top Ten Predicted Rising Stars from StarRank

Author	Position	Citation
ying ma	Principal Researcher, Research Area Manager, Microsoft Research Asia	14355
philip s. yu	Professor and Wexler Chair in Information Technology, Department of Computer Science, University of Illinois Chicago	28429
jiawei han	Professor, Department of Computer Science, University of Illinois at Urbana-Champaign	46654
zheng chen	Senior Researcher , Microsoft Research Asia	3937
divesh srivastava	AT&T Labs,Inc.	11520
wei wang	Professor, University of North Carolina at Chapel Hill	9873
hsinchun chen	Professor and Director, Management Information Systems Department Eller College of Management The University of Arizona	8161
erik d. demaine	Associate Professor, Massachusetts Inst. Tech., Lab. for Computer Science	7361
bertram ludwigäscher	Professor of Computer Science, Computer Science Department Stanford University	29544
lee tan	Provost's Chair Professor, School of Computing	6824

4.4.2 Affect of Alpha Parameter

It is commonly used to measure the internal consistency or reliability of a psychometric test score. It was originally derived [7,31] by Kuder & Richardson (1937) classified into two part scored data (0 or 1) the value of alpha can take less than or equal to 1 after describe by Cronbach [32] (1951) It was first named alpha by Lee Cronbach. Globally the Value of Alpha may not exist We always observe in terms of type I errors alpha, which are always small (.1, .05, .01) The smaller alpha value gets the more tight proof that the alternative is correct, because the probability of type I error is reduced, but some case alpha high value is Caused high variance which score is wide spread value which is easily differentiate able Several investigators have set quite high values (e.g Cortina, 1993 Cronbach [32], 1951 Green, Lissitz & Mulaik, 1977 Revelle, 1979 Schmitt. 1996 Zinbarg, Yovel, Revelle & McDonald, 2006). As a result, alpha is most appropriately used when the items measure different considerable areas within a single construct. Require the reliability of alpha is 0.5 or higher (obtained on a substantial score or true) .We have calculated the StraRank author contribution base, dynamic publication venue base and Composite method we calculate the rank of author using the value of alpha is 0.1, 0.2, 0.3, and up to 0.9.first we set the alpha value is 0.1 we calculated the baseline method, StarRank first, second, third method and get author citation, paper and h-index from arnetminer [25]. when we set the alpha value 0.2 in all method little bit change in author rank , on value 0.3 author rank score is also increased .on alpha value 0.4 little bit change in all method compared with previous method on alpha value but compare with baseline proposed method value is high. When we set the 0.5 value then little bit change in rank score .author rank score value in decrease on 0.6 alpha value in baseline and proposed method .when we set 0.7 alpha value in all method little bit value is decreased in all method compared with previous alpha value and 0.8 .the value of author rank score in also decreased on 0.9 alpha value.

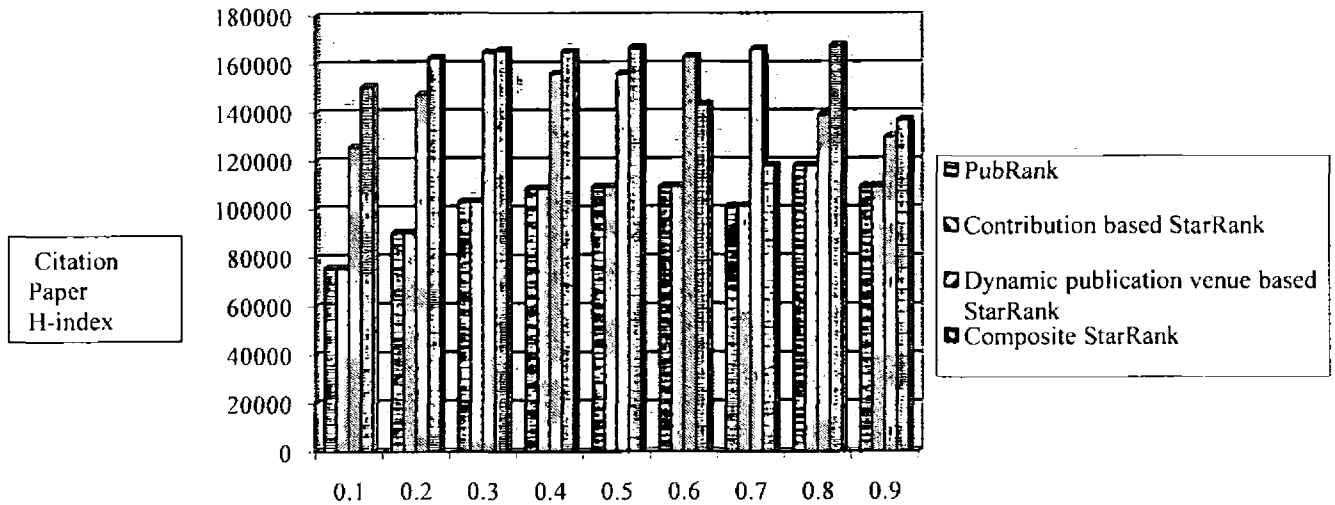


Figure 4.4.2.1: Average citation on different alpha value

4.4.3 Rising Star paper Finding

We have performed several experiments to obtain the result and calculated the rank of author with average of citation, h-index, paper of PubRank, we have further calculated rank of author on different alpha value. We have calculated the rank of paper of different author and obtained the citation, h-index and of top ten paper.

Table 4.4.3.1 Rising stars paper rank and citation

S#	Paper Rank	Citation
1	Probabilistic robotics	2579
2	An Overview from a Database Perspective	1886
3	Efficient and Effective Clustering Methods for Spatial Data Mining	1788
4	Holistic twig joins: optimal XML pattern matching	871
5	Learning to cluster web search results	442
6	Clustering by pattern similarity in large data sets	362

7	Credit rating analysis with support vector machines and neural networks	340
8	Mobile-assisted localization in wireless sensor networks	207
9	Efficient Progressive Skyline Computation	116
10	A survey of scheduling with deterministic machine availability constraints	36

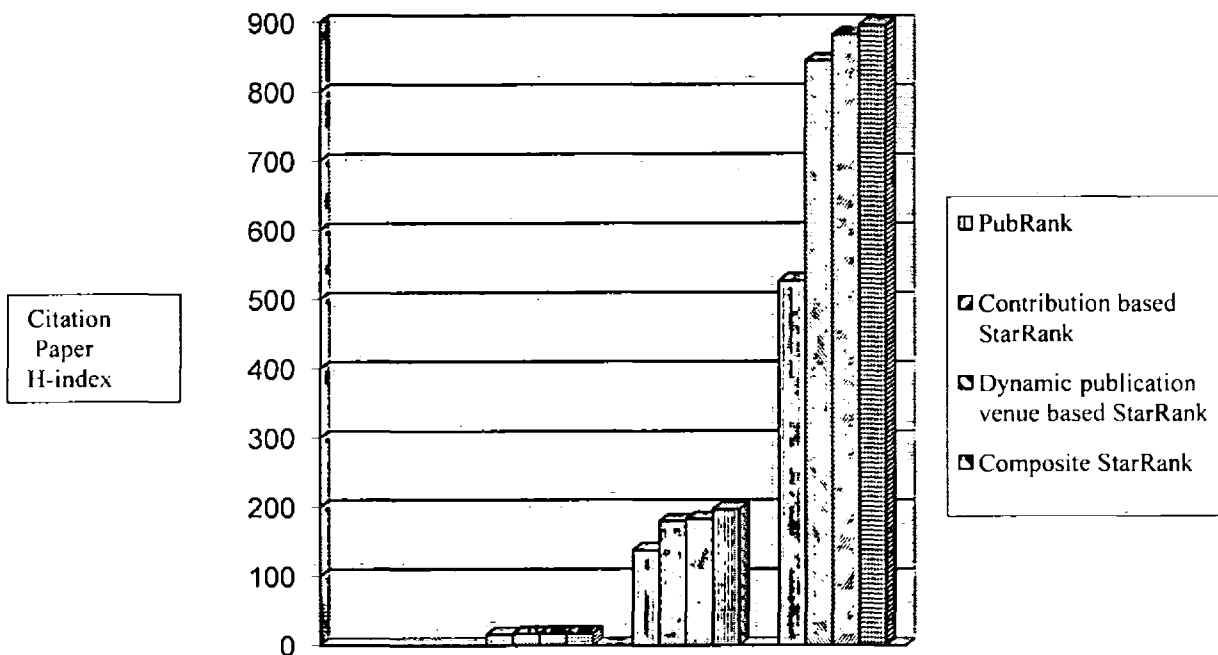


Figure 4.4.3.1: Average citation of Rising stars paper rank

We have performed several experiments to obtain the result and separately we have calculated PubRank method and obtained the h-index, paper and citation from arnetminer [25]. Proposed method we have calculated author contribution based StarRank, dynamic publication venue based StarRank and composite based StarRank. We have taken the data from DBLP [23] and through author contribution base we have calculated StatRank. We

have selected top ten authors from PubRank method, StarRank Author contributions StarRank dynamic publication venue based and composite based StarRank and we extract top ten papers which has h-index, citation and high rank of venue of that paper .we have calculated existing method PubRank and we taken high citing paper for top ten rising star. We have calculated h-index value of paper and third step we have calculated entropy of venue. If a paper published at less significant venue would be less important as compare to other who published at high significant venue. Moreover, we calculated StarRank author contribution, daynamic publication venue and composite based StarRank and finding rising star paper finding from top ten rising star.

4.4.4 Rising Stars Venue Base StarRank

We used data from digital biography and library project [23] and indentify the rising star in above experiment now we have calculated and indentify the rising star form venue based StarRank through rising star author contribution base StarRank, Dynamic Publication venue based StarRank and Composite StarRank. We have taken the top ten star and taken H-index, paper citation from Arnetminer[25].

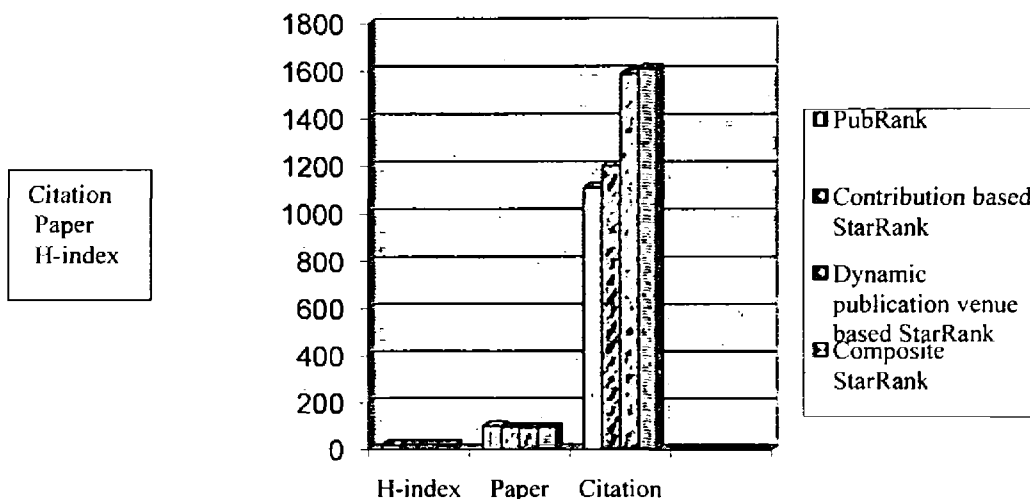


Figure 4.4.2.1: Average citation on venue based StarRank

Table 4.4.4.1 Top Ten Predicted Rising Stars from Database venue

Author	Position	Citation
david wagner	Associate Professor, Computer Science Division University of California, Berkeley	10588
steffen staab	Professor, Faculty of Computer Science of the University of Koblenz-Landau	9578
jeffrey xu yu	Professor, The Chinese University of Hong Kong	4005
Jayavel shanmuga sundaram	Associate Professor, Associate Director, School of Arts, Media and Engineering, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University	3829
huaxiong wang	Lecturer, Department of Computing Macquarie University	1484
gregory hornby	Computer Scientist, University of California University Affiliated Research Center (UARC) NASA Ames Research Center Evolvable Systems Group Intelligent Systems Division	862
tsutomu matsumoto	Professor, Division of Social Environment and Information Faculty of Environment and Information Sciences	817
wen jin	Assistant Professor at UC Irvine	684
ludovic dea cute	Associate Professor, University of Paris	253
r. brien maguire	Professor Department of Computer Science University of Regina Regina, Saskatchewan Canada S4S 0A2	108

4.4.5 Values on different Damping Factor

Authored Barin and Page [7, 31] used alpha is 0.85. Google itself use this value because it is easy to get the result and small value is not suitable because too much weight or much dampened the result and flow of the PageRank is dampened the iteration. The high damping factor means low dampened and PageRank grow higher .we have calculated the StarRank on different Damping factor. We have taken average value of h-index, paper and citation on different damping factor. We have taken high citation on 0.80, 0.85.

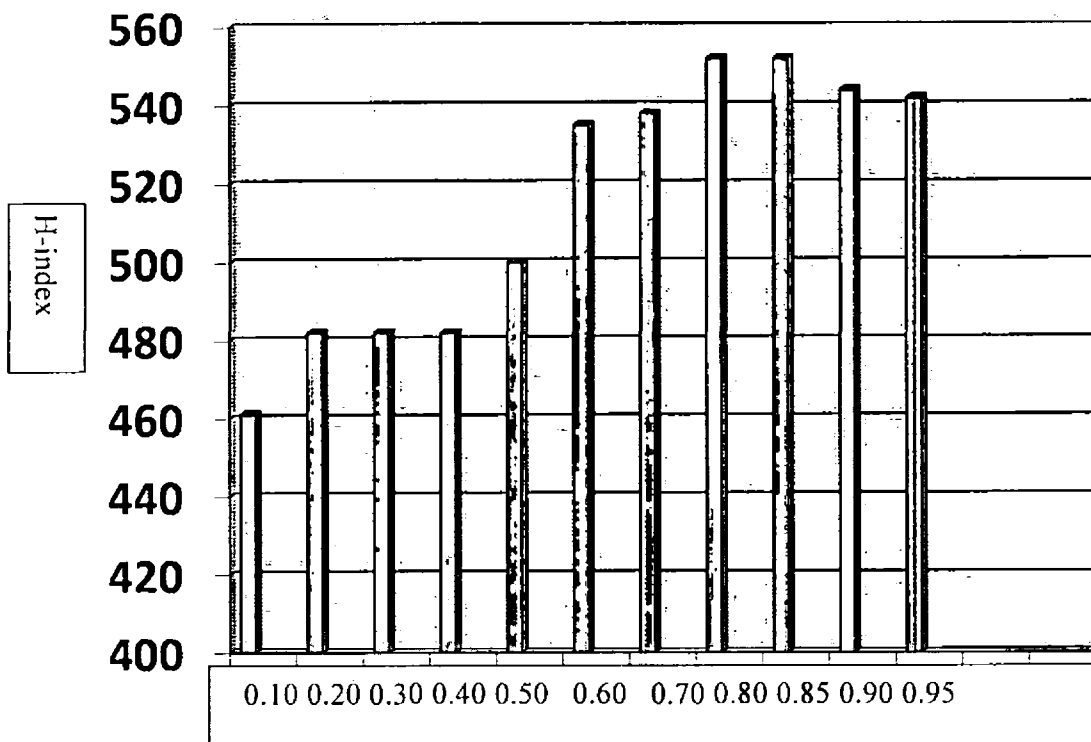


Figure 4.4.5.1: Effect of damping factor in terms of Average H-index of top ten stars ranked by StarRank

We have taken the data from DBLP [23] and we have calculated StatRank on different damping factor value. We have selected top ten authors from StarRank score and taken h-index value from arnetminer [25] in above graph. We took H-index value we set value of

$d = 0.10, 0.20, 0.40, 0.50, 0.60, 0.70, 0.80, 0.85, 0.90, 0.95$. The value of h-index is gradually increased on $0.70, 0.80, 0.85$.

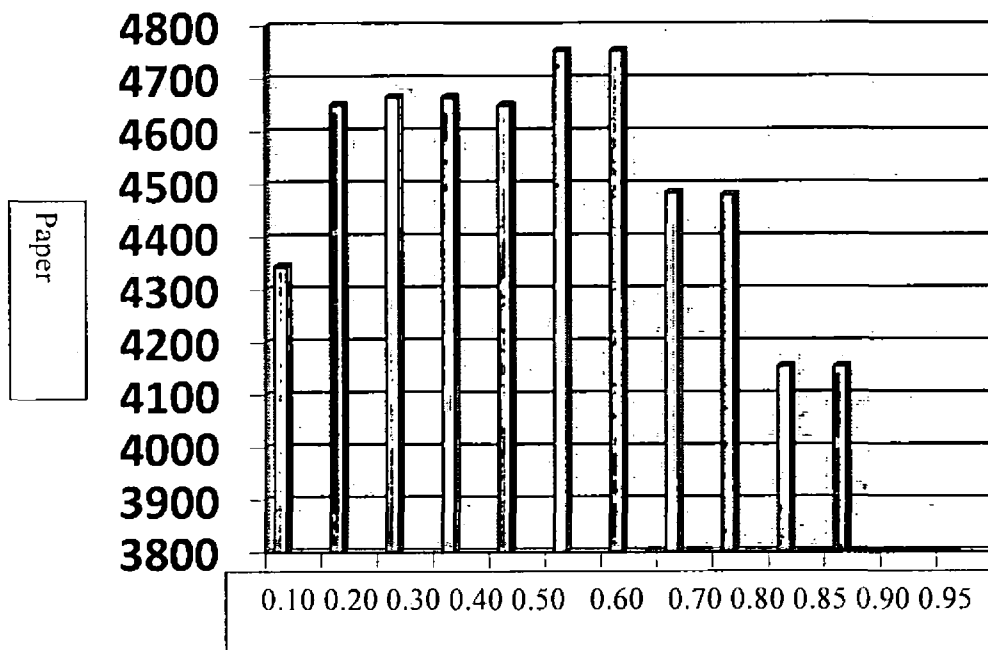


Figure 4.4.5.2: Effect of damping factor in terms of Average Paper of top ten stars ranked by StarRank

We have taken the data from DBLP [23] and we have calculated StatRank on different damping factor value. We have selected top ten authors from StarRank score and taken paper from arnetminer [25] in above graph. We have taken average paper on value of $d = 0.10, 0.20, 0.40, 0.50, 0.60, 0.70, 0.80, 0.85, 0.90, 0.95$. We have taken high average value on $0.20, 0.30$ which is gradually increased also when we set value of $d = 0.50, 0.60,$

0.70. The average paper value is decrease on 0.80, 0.85 because we have received high citation in less number of paper.

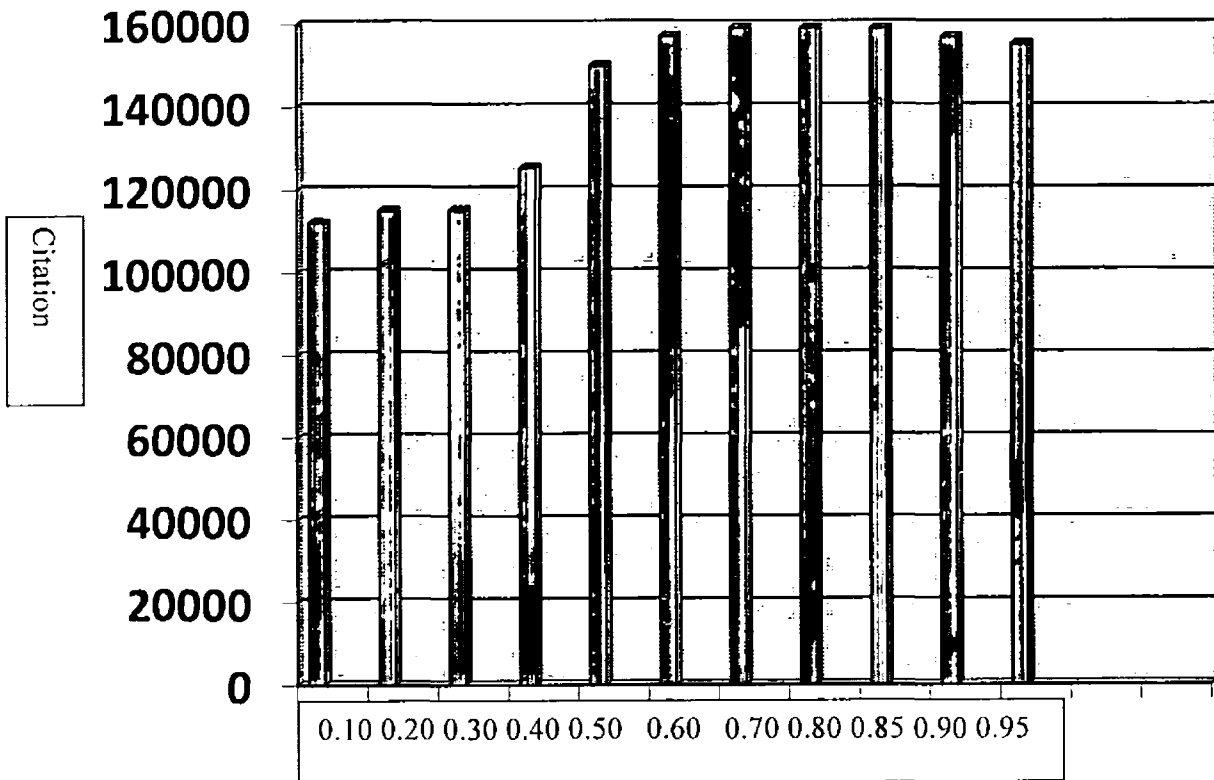


Figure 4.4.5.3: Effect of damping factor in terms of Average Citation of top ten stars ranked by StarRank

We have taken the data from DBLP [23] and we have calculated StarRank on different damping factor value. We have selected top ten authors from StarRank score and taken citation from arnetminer [25] in above graph. We have taken average citation on value of $d = 0.10, 0.20, 0.40, 0.50, 0.60, 0.70, 0.80, 0.85, 0.90, 0.95$. The citation of author is gradually increased and we have gained maximum citation on 0.80, 0.85.

Chapter 5

Conclusions

5.1-CONCLUSIONS

Social network is social structure which is made up of nodes. Node or people are individual actor in network and an edge is path between the actors. We have proposed new technique called StarRank. We have solved the author contribution the author quality score in fair manner. Entropy is most important contribution for rising stars finding. Author contribution based and dynamic publication venue based hybrid technique (Ac+EV) is more important for rising stars finding and performance is increased. We have taken the data from DBLP [23] and checked the author and his paper citations result from arnetminer [25]. We have performed many experiments and measured the performance with existing method and proposed method. we selected the top ten authors and count the number of citation for top ten authors. If an author has more citation will be more important in community and more experts in our filed. Proposed method has high h-index, paper and citation result.

REFERENCES

- [1] M.A.Porter, J.P.Onnela, P.J.Mucha, "Communities in Network," Notices of the American Mathematical Society, Vol.56, Issue 7, pp 7:13, 2006.
- [2] D.M.Boyd, N.B.Ellison, "Social network sites Definition, history, and scholarship," Computer-Mediated Communication, Vol.13, Issue 3, pp 16 - 31, 2007.
- [3] M.Brzozowski, T.Hogg, G.Szabo, "Friends and foes: Ideological social networking," In Proceedings of the SIGCHI Conference on Human Factors in Computing, ACM Press, New York, Vol.56, Issue 9, pp 817-820, 2008.
- [4] P.Lawrence, B.Sergey, M.Rajeev, M.Terry, "The PageRank Citation Ranking Bringing Order to the Web," In Proceedings of the 7th International World Wide Web Conference Brisbane, Australia, Vol.26, Issue 1999-66, pp 107-117, 1999.
- [5] S.Brin, L.Page, "The anatomy of a large-scale hyper textual Web search engine," In Proceedings of international conference on World Wide Web, Vol.30, Issue 1-7, pp 107-117, 1998.
- [6] L.L.Xiao, F.S.Chuan, T.L.Kar, N.See-Kiong, "Searching for Rising Stars in Bibliography Networks," In Proceedings of the 14th International Conference on Database Systems for Advanced Applications, Vol.5463, Issue 2, pp 288-292, 2009.
- [7] B.catherine, C.Adena, H.Emily, k.Mathew, L.kody, L.Eddery, R.john, R.Ishani, S. Michael, W.Nathaniel, "Page Rank Algorithm," In Proceedings of the 7th International World Wide Web Conference Brisbane, Amherst, Vol 2, Issue 3, pp 1-64, 2006.

-
- [8] G.Lies, D.P.Christopher, "Link Mining: A Survey," ACM SIGKDD Explorations Newsletter, Vol.7, Issue 2, 2005.
- [9] G.Jennifer, "The Dynamics of Web-based Social Networks," International Sunbelt Social Network Conference, Vol.6, Issue 1, pp 1-18, 2008.
- [10] L.Xiaoming, B.Johan, L.N.Michael, V.D.S.Michael, "Co-Authorship Networks in the Digital Library Information Processing and Management," International Journal, Vol.41, Issue 6, pp 1462-1480, 2005.
- [11] S.CH, "Quantifying coauthor contributions," Science, Vol.322, Issue 5900, pp 416- 417, 2008.
- [12] C.Amit, D.B.Khanh, S.Muthukrishnan, "Internet Mathematics Estimating Entropy and Entropy Norm on Data Streams," Taylor & Frances, Vol.3, Issue 1, pp 63-78, 2006.
- [13] N.Hidetsugu, O.Manabu, "Automatic Detection of Survey Articles," In Proceeding of ECDL 9th European conference on Research and Advanced Technology for Digital Libraries, Vol.3652, pp, 391-401, 2005.
- [14] W.Barry, "Computer Networks as Social Networks," Science, Vol.293, Issue 5537, pp 2031-2034, 2001.
- [15] T.Jie, Z.Duo, Y.Limin, "Social Network Extraction of Academic Researchers." In Proceeding of ICDM Seventh IEEE International Conference on Data Mining IEEE Comuter Society ,Washington, Vol.5 Issue 4, pp.292-301, 2007.

-
- [16] C.Jeff, "Online Social Networking Issues within Academia and Pharmacy Education," *American Journal of Pharmaceutical Education SCIENCE*, Vol.72, Issue 10, pp 233-242, 2008.
- [17] S.Huawei, C.Xueqi, C.KaiB, H.Mao, "Detect overlapping and hierarchical community structure in networks." *Physica a Statistical Mechanics and its Applications*, Vol.388, Issue 3, 1706-1712, 2008.
- [18] G.Jennifer, "The Dynamics of Web-based Social Networks: Membership, Relationships and Change," *PLOS computational biology*, Vol.12, Issue 11, pp 11-15, 2007.
- [19] T.Tang, Z.Zhang, Y.Limin, L.Juanzi, Z.Li, S.Zhong, "ArnetMiner: Extraction and Mining of Academic Social Networks," In *Proceeding of 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, NewYork, Vol.2, Issue 1, pp 990-998, 2008.
- [20] K.Balog, L.Azzopardi, M.Rijke, "Formal models for expert finding in enterprise corpora", In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, vol.369, Issue 7, pp 43-50, 2006.
- [21] H.H.Taher, "Topic-Sensitive PageRank a Context-Sensitive Ranking Algorithm for Web Search," *IEEE Transactions on Knowledge and Data Engineering*, Vol.15, Issue 4, pp 784-796, 2003.
- [22] M.Rada, T.Paul, F.Elizabeth, "PageRank on Semantic Networks with Application to Word Sense Disambiguation," In *Proceedings of the 20st International Conference on Computational Linguistics*, Vol.154, Issue 3, pp 1126-1133, 2004.

-
- [23] DBLP bibliography database, <http://www.informatik.uni-trier.de/~ley/db/>
- [24] Citeseer scientific literature digital library, <http://citeseer.ist.psu.edu/>
- [25] ArnetMinerAcademic Researcher Social Network Search, <http://www.arnetminer.org/>
- [26] G.Cheng,N. Chen, J.Chang , “Evaluation of Supply Chain Partnership Based on Entropy Theory,” In proceeding of 3rd International Conference on Information Management, Innovation Management and Industrial Engineering , Vol. 01,Issue 1, pp 494-497,2010.
- [27] M.Zhang,W.Liu, “Research on Evaluation of Equipment Procurement Organizational Structure Based on Entropy Theory, “In proceeding Of Eighth International Conference on Electronic Measurement and Instruments,” Vol.2, pp 1-55, 2007,
- [28] L.Gbor,G.Vince,” When the Web Meets the CellUsing Personalized Page Rank for Analyzing Protein Interaction Networks,”Bioinformatics,Vol.27, Issue 3, pp 405-407,2010.
- [29] J.Yushi,B.Shumeet, “Visual Rank: Applying Page Rank to Large-Scale Image Search,” IEEE Transactions on Pattern Analysis and Machine Intelligence,Vol. 30, pp. 1877-1890, 2008.
- [30] M.E Newman,M.girvan, “Mixing patterns and community structure in networks,” Statistical Mechanics of Complex Networks,Vol.625,Issue 3, pp 66-87, 2003.
- [31] B.Paolo Boldi,Massimo Santini, V.Sebastiano, “ PageRank as a Function of the Damping Factor,” In Proceedings of the 14th international conference on World Wide Web,Vol.27, pp 557 – 566 ,2005.
- [32] T.Mohsen,D. Reg, “Making sense of Cronbach’s alpha,” International Journal of Medical Education, Vol.2, pp 245-246, 2011.