CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Evaluation of Textual and Topological Similarity Measures for Citation Recommendation

by

Abdul Samad

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2019

This study is wholeheartedly dedicated to my beloved parents, who have been my source of inspiration and gave me strength when I thought of giving up, who continually provide their moral, spiritual, emotional, and financial support.

To my teachers, who shared their words of advice and encouragement to this study.

And lastly, I dedicated this thesis to the Allah Almighty, thank you for the guidance, strength, power of mind, protection and skills and for giving me a healthy life. All of these, I offer to you.

CERTIFICATE OF APPROVAL

Evaluation of Textual and Topological Similarity Measures for Citation Recommendation

by

Abdul Samad

MCS151006

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Zahid Halim	GIKI, Topi, KPK
(b)	Internal Examiner	Dr. Muhammad Tanvir Afzal	CUST, Islamabad
(c)	Supervisor	Dr. Muhammad Azhar Iqbal	CUST, Islamabad

Supervisor Name
Dr. Muhammad Azhar Iqbal
March, 2019

Dr. Nayyer Masood
Head
Dept. of Computer Science
March, 2019

Dr. Muhammad Abdul Qadir
Dean
Faculty of Computing
March, 2019

Author's Declaration

I, **Abdul Samad** hereby state that my MS thesis titled "**Evaluation of Textual and Topological Similarity Measures for Citation Recommendation**" 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/abroad.

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

(Abdul Samad)

Registration No: MCS151006

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "***Evaluation of Textual and Topological Similarity Measures for Citation Recommendation***" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been dully 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 MS Degree, the University reserves the right to withdraw/revoke my MS 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 work.

(Abdul Samad)

Registration No: MCS151006

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. **Samad, A.**, Islam, M. A., Iqbal, M. A., Aleem, M., & Arshed, J. U. (2017, December). Evaluation of features for social contact prediction. In 2017 13th International Conference on Emerging Technologies (ICET) (pp. 1-6). IEEE.

Abdul Samad
(MCS151006)

Acknowledgements

I am thankful to my Creator **Allah** Subhana-Watala to have guided me throughout this work. Then I heartedly admire the true concern and best guidance of my respected supervisor **Dr. Muhammad Azhar Iqbal**. Words are not enough to express the gratitude towards him. I want to pay all my thanks to my best mentor **Dr. Muhammad Arshad Islam**, who encouraged me to give my best in every circumstance. I am highly indebted to my parents and my family, for their expectations, assistance, support and encouragement throughout the completion of this Master of Science degree. They form the most important part of my life. After ALLAH (S.W.T) they are the sole source of my being in this world.

I also want to express my thankfulness for **Shafiq Ur Rahman**, **Muhammad Asif Malik**, **Muhammad Yasir Noman** and **Jawad Usman** for their ethical backing. I am highly thankful to **Ishrat Nawaz(Stpd)** for her heart touching words for me. I would also like to thank **PCN** research group for being on my thesis guidance.

Finally, I have learn that, "*Only I can change my life. No one can do it for me.*"

Abdul Samad (Alvi)

Abstract

Researchers and scientists cite papers in order to connect the new research ideas with previous research. For the purpose of finding suitable papers to cite, researcher spend considerable amount of time and make effort. Due to huge collection of research publications, sometimes researcher are unable to find the related articles for citations. The purpose of citation recommendation system is to reduce the time they spend and present them the related citation papers they are not aware of.

Past studies on citation recommendation systems generally compare articles on the basis of their content, likes of the researcher, collaboration of the researchers and recommend similar articles for citations. The limitation of these studies is that they do not consider the importance of recommended papers from citation perspective. In this study, we argue that citation network can be used to identify papers that are not only relevant but also important to be cited.

The fundamental objective of this thesis is to evaluation of textual and topological similarity measures for citation recommendation system, which recommends similar as well as important papers for citation. To achieve this objective, this work analyzed textual and topological similarity measures (i.e., *Jaccard* and *Cosine*) to check which is better to find similar papers? on one hand, this work analyzed two textual parameters (i.e., *Title* and *Abstract*) and one topological features (neighbors of the paper in citation graph). On the other hand, to find the importance of papers, we compute centrality measures (i.e., *Betweeness*, *Closeness*, *Degree* and *Pagerank*).

After evaluation, it is found that, topological-based similarity via *Cosine* achieved 85.2% citation links and using *Jaccard* obtained 61.9%. On the other hand, textual-based similarity via *Cosine* on *abstract* obtained 68.9% citation links and using *Cosine* on *title* achieved 37.4%. Likewise, textual-based similarity via *Jaccard* on *abstract* obtained 35.4% and using *Jaccard* on *title* achieved 28.3%.

Contents

List of Figures

List of Tables

Abbreviations

WCC	Weakly Connected Components
SCC	Strongly Connected Components
IR	information retrieval
CBF	Content-based filtering
CF	Collaborative filtering

Symbols

a	Channel epi-layer thickness
a_{eff}	Effective available channel
b	Buffer-layer thickness
c	Speed of light
C_{gs}, C_{gsm}	Conventional and modified gate-to-source capacitance
C_{gd}, C_{gdm}	Conventional and modified gate-to-drain capacitance
C_{ds}, C_{dsm}	Conventional and modified drain-to-source capacitance
c_1, c_2	Cognitive and social parameter

Chapter 1

Introduction

1.1 Overview

The purpose of research is to introduce new ideas through scientific discourse. Large volume of research articles are publishing every year [1]. It becomes difficult for researcher to identify relevant research articles of their interest. Furthermore, it is also non-trivial to keep up-to-date with new research publications and to associate them to previously published papers. With the digitization of research publications, there has been a move to use computers to augment the search for related articles which are relevant to a researcher′s field of interest. Such systems are known as recommender systems. A recommender system can be most considered as a system that takes as input some characteristics (i.e, type of items this user like) from a user which are processed in order to identify items which are most relevant to the users interests. Item is the general term used to denote what the system recommends to users. The type of matching used commonly categorizes the approach into either a content-based approach, a collaborative filtering approach, a co-occurrences approach, or a stereotyping approach shown in Figure1.1.

Content-based filtering (CBF) is one of the most widely used recommendation approaches[2]. One main thing of CBF is the user modeling, in which the interests

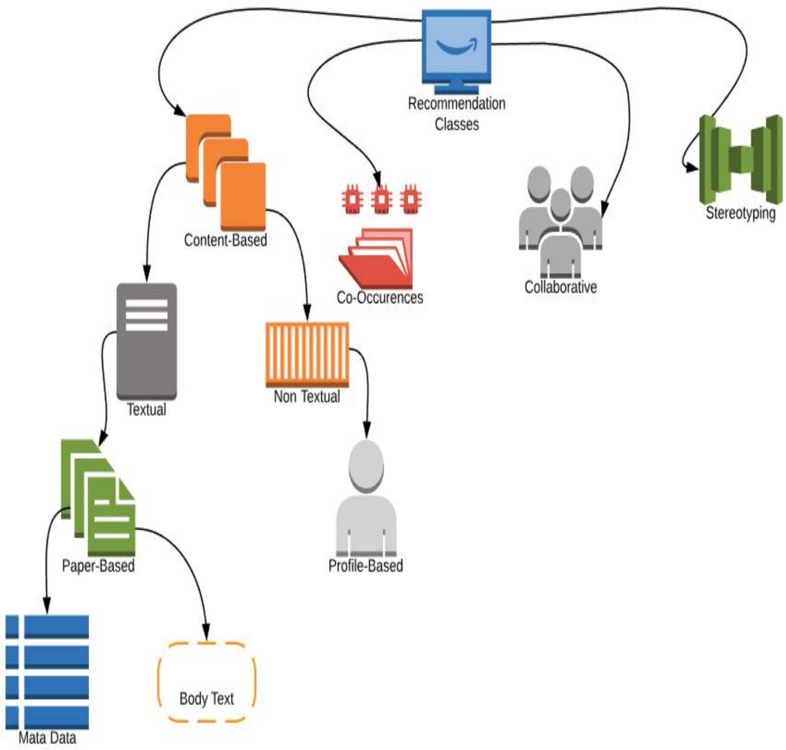


FIGURE 1.1: Recommendation Classes

of users are inferred from the items that user interacted with. Items are usually textual, such as emails or web pages. Interaction is established through actions, such as downloading, buying, authoring or tagging an items.

Unlike content-based approaches, which use the content of items previously rated by a user u, **Collaborative filtering (CF)** approaches [3] [4] rely on the ratings of u as well as other users in the system. The key idea is that the rating of u for a new item i is likely to be similar to that of another user v, if u and$v$ have rated other items in a similar way. Likewise, u is likely to rate two items i and j in a similar fashion, if other users have given similar ratings to these two items.

Co-Occurrence recommendations, those items are recommended that frequently co-occur with some source items.One of the first applications of co-occurrence was co-citation analysis introduced by [5].

Stereotyping is one of the earliest user modeling and recommendation class. It was introduced by *Rich* in the recommender system *Grundy*, which recommended novels to its users [6].*Rich* was inspired by stereotypes from psychology that allowed psychologists to quickly judge people based on a few characteristics. Majority of the researchers used content-based approaches in their research work. So that, the main focus of this thesis is on **Content-Based Filtering** approach.

Citation recommendation addresses the task of providing recommendations based on an abstraction of the users profile or contents of paper. In 1998, Giles et al. introduced the first research-paper recommender system as part of the CiteSeer project [7]. More than 200 research articles regarding research-paper recommendation systems have been published in 16 years until 2015, and there have been more new systems introduced since then which have been described in chapter 2. Since the yearly number of articles steadily increases: 66 of the 217 articles (30 percent) were published just in 2012 and 2013 alone (Figure 1.2) [8]. The amount of literature and approaches represents a problem for new researchers because they do not know, which of the articles are most relevant? Which recommendation approaches are most promising? Which paper have worth in their field of interest?

Even researchers familiar with research-paper recommender systems would find it difficult to keep track of the current developments. A move towards the recommendation of paper is becoming state-of-the-art now a day. This can be used to suggest the relevant papers for citation as well as for the topic of interest. It helps new researchers to explore the work which is already been done in their respected fields. Although, majority of research paper recommendation recommender systems are in working, but no one satisfying the need of researcher. These systems only consider the similarity of documents and recommend the similar papers to author.

The problem arises a question that how we can find the worthy papers? The main focus of the thesis is to recommend the similar but important papers for citation. For recommendation of worthy papers, centrality metrics are used in this thesis which are degree[9], closeness[10], betweeness[9] and papgerank[10].

Citations signify intellectual linkages between academic works and this link structure can be followed, backwards as well as forwards, to search for relevant papers;

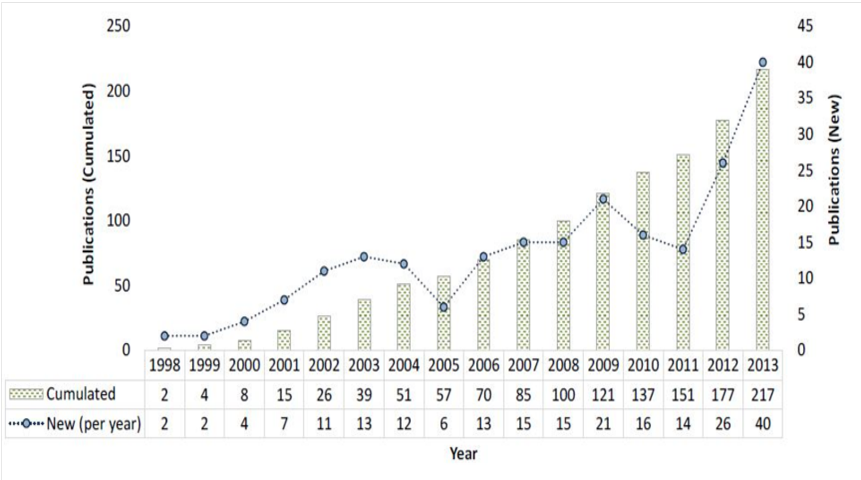


FIGURE 1.2: Articles Regarding Research Paper Recommendation

this is the basic premise of citation indexing. Two core citation analyses are bibliographic coupling [11], where documents are said to be coupled if they share one or more references, and co-citation analysis [5], where the similarity between documents A and B is measured by the number of documents that cite both A and B. The theory behind bibliographic coupling is that documents that are similar in subject have similar references; the theory behind co-citation analysis is that documents that are similar are more likely to be cited by the same other documents. These principles each provide a means of quantifying document similarity or relatedness using citations. Consequently, both bibliographic coupling and co-citation analysis have commonly been put to use in Information Retrieval(IR) over the years. There is, in fact, a tradition in IR of using methods based on statistical citation information, which continues today. For instance, [12] use co-citation data as one feature in a system that, given a document as a 'query', retrieves documents to be recommended for citation by that document [13].

In this study, we have consider bibliography and citations as topological features for finding and recommending similar papers. Moreover, metadata (i.e., titel and abstract) of the paper also considered as a textual feature to find similar papers. For the computing similarity between papers, this thesis used *jaccard* similarity and *cosine* similarity measures on textual and topological features. To recommend relevant and similar papers, topological features (i.e., citations and bibliography) are used in citation network.

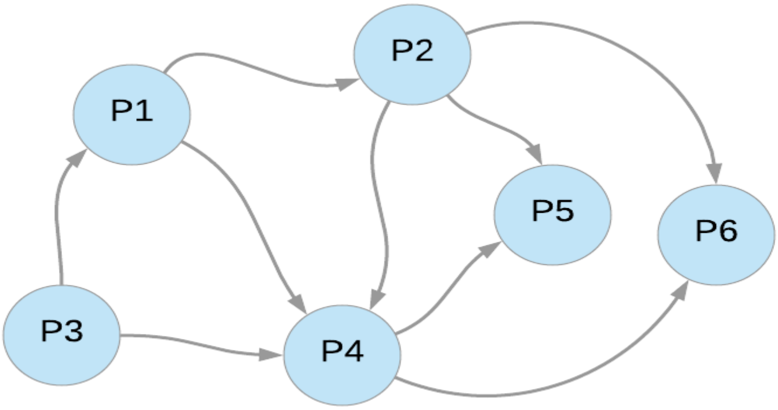


FIGURE 1.3: Citation Network

Citation Network is basically a social network. Egghe et al in [14], explain when a document d_i cites a document d_j, we can show this by an arrow going from the node representing d_i to the document representing d_j. In this way the documents from a collection D form a directed graph, which is called a citation graph or citation network. Citation network is helpful for the evaluation of publication and authors [15]. Citation network is known as directed network in which one publication cites another publication. Lets take an example as shown in Figure1.3, where P_1 cites P_2 and P_4, P_2 cites P_4, P_5, and P_6 and so on. These collectively make a citation network.

1.2 Research Objective

There can be large number of related research papers, therefore it is difficult to decide which paper should be cited. Moreover, search relevant papers for researcher takes too much time, because most of the researchers unable to find the required papers that they need. This work would helpful to search most relevant research papers. As discussed above in introduction, the only thing is textual similarity, which is considered in previous research paper recommender systems. Therefore, the purpose of this thesis is to explore the topological features to find and recommend similar as well as worthy paper. To find the importance papers through citation network, four centrality measures are used.

1.3 Scope

The aim of this thesis is to recommend similar as well as important papers to the readers. The results of this study will be useful for the researchers in finding out the required and relevant research articles in a timely fashion.

1.4 Problem Statement

A number of techniques are discussed in the literature to recommend citation paper. Most of the recommendation approaches uses textual similarities between documents and recommend the papers. These recommender systems considers similarity of the paper, but do not consider the importance of the recommended papers in that field. This is the problem, researcher are facing now a days. Therefore we are going to recommend similar as well as important papers. Furthermore, we are aiming to explore other ways to find similar papers instead of textual similarity. This has led us to explore the answers for the following questions:

RQ1: How accurate are textual similarity measures (Jaccard and Cosine) for correct identification of citation link ?

RQ2: How accurate are Topological similarity measures (Jaccard and Cosine) for correct identification of citation links ?

RQ3: Are topological similarity measures better than textual similarity measures to predict a citation link ?

1.5 Applications

- **Citation recommender system**: This study will be helpful for researchers to find papers for citation.Citing a paper requires a deep knowledge about researcher topic and it is important to cite the most relevant and important articles from literature.

- **Document retrieval system**: Most of the time, readers augment knowledge by reading the ideas related to their interest, but unable to find relevant material. This thesis will help the readers to update himself/herfself with the relevant articles.

1.6 Organization of the Thesis

The work presented in this thesis draws together ideas from information retrieval (i.e. finding relevant paper) and graph theory(i.e. finding important paper using centrality measures). Chapter 2, provides an overview of related research and our work therein. In Chapter 3, methodology is presented. Chapter 4 covers the relevant experiments along with related discussions. Chapter 5 concludes the paper with a brief discussion of future possibilities and experiments.

Chapter 2

Literature Review

In this chapter, relevant recent research work has been discussed that provide recommendations for scientific papers. In scientific research, refer others work is considered as important thing so that the previous work can be further improved[16]. Therefore, it is a very big problem to get content similar to the given paper, because lots of material related to research is publishing every day [17]. Experienced as well as new researchers are also facing this problem. Most researchers use the citation recommendation system in view of this matter. Recommendation system recommends research papers to authors, related to their research, on the base of a query paper. Recommendation system use textual as well as topological similarity to recommend research articles. Generally, the recommendation system works on prediction. On the base of this prediction, it suggests which paper should be cited?

2.1 Recommender Systems

A recommender system can be taken as a black box which takes input in the form of user profile and matches it against a candidate set of items in order to suggest previously unseen items for a user [8]. These items are considered to be the most relevant recommendations for that user. Recommender system is defined as a decision making strategy for users under complex information environments[18].

Approaches (i.e content-based filtering, collaborative filtering, co-occurrence and stereotyping) used in recommender systems can be categorized into following by [8].

2.1.1 Content-Based Filtering(CBF)

CBF, which is defined by [2], is used to match the items similar to the items that user liked in the past. A content model having the features represents items [2]. Features can be *textual* or *non-textual* e.g. layout information, writing style and XML tags. In the research community, almost 55% researcher publication on recommender system using CBF [19] [15]. Interaction between users and items was established through authorship [8] e.g. adding social tags [20] and browsing papers [19] etc.

2.1.2 Collaborative Filtering(CF)

In CF, recommendations are given on the base of interaction of other users in the systems[4][3]. Recommendations in CF is based on user similarity [21] instead of item similarity. From the existing literature, less than 20% used CF [22]. According to [23], users were too lazy to provide ratings when they were accepted to do so. To address this problem, authors in created synthetic ratings in their work. The main problem of CF is that CF requires user participation, but the motivation to participate is too low. This problem is referred to as the cold-start problem, which may occur in three situations (i.e., new users, new items and new communities or disciplines) [24].

2.1.3 Co-Occurrence

Co-occurrence recommendation approach recommends those items which co-occurs frequently. Authors in [5], proposed that the papers that frequently co-cited supposed to be related to each others. Many recommender systems implemented the same concept. For example, Amazon, customers of Amazon who bought item i also bought item j when i and j co-occurs [25]. The advantage of co-occurrence recommendation is that relatedness is focused instead of similarity. Relatedness expresses how closely two items are. For instance, two papers sharing the same characteristics are similar. Likewise, paper and pen are not similar but related, because both are required for writing letters. Hence, co-occurrence recommendations provide more unforeseen recommendations and comparable to CF as well.

2.1.4 Stereotyping

Stereotyping which is introduced in 1979 by *Rich* [6], recommends items by determining the characteristics of user. Stereotypes are collections of characteristics as defined by *Rich* [6]. In the domain of research-paper recommender systems, only [26] applied stereotyping. The authors assume that all users of their software are students or researchers. Therefore, recommendation (i.e., papers/books) are made according to the interest of researchers and students [27].

2.2 Citation Recommendation Systems

Because many papers are published in the last decade [16], so it is a difficult task to process them manually and find the most relevant and similar papers for citation. Authors in [16] have proposed a recommender system called RfSeer, which recommend papers on the topic based as well as citation context. This system is very helpful for reviewers to validate references. According to [16], for topic-based model, authors used contents of papers that are parsed. They also extracted sentence in which citation is made, furthermore authors extracted sentence before and after the citation sentence and made a citation context using these three sentences. After getting the query, their system picks top 5 topics using topic-based model, and recommend a list of citations. According to [16], topic-based citation recommendation is effective because the list of recommended citation is made through topics, and in this way, these recommended citations are clustered. In the citation context method, the context of the citation is the source and all the references are target. In the citation context, according to [16], after getting the query and using words of the query this system will assigned a score to all references. Then authors calculate term-frequency-inverse-context-frequency (TF-ICF) to check the need of citation. In the experiments, they found that citation context recommendation gets 50% recall, whereas precision for both topics-based and citation context-based indicate that 1 recommendation is correct out of 10 recommendations. The global recommendation which is topic-based and local recommendation which is context-based, can only tell us the relevant paper but it does not tell how much its importance.

Most recommender systems work on bibliography and reviewers assignment [15]. For reviewers assignment authors require reviewers profile and abstract of the papers, whereas for citation recommendation authors require partial citations and authors profiles [15]. According to [15], research in the last decade worked on partial citations to predict more citations list. Where d is the query document and l is the partial citation list to predict the complete citation list which is l'. As [15] worked on citation context, in the same way [7] also worked on citation context. They build a prototype of CiteSeerX[7]. Their system requires a title, abstract and citation context as an input. Here citation context is a place, surrounding by citation sentence, where user wants to make a citation. In their experiments, they found that global recommendation has recall of 0.45 on top 25 recommendations. As the recommendations increased, recall also increased. At 250 documents, the recall was 0.65. Local recommendations results were also like this. The maximum recall was 0.6.

Recommendation of research papers is being considered as the main issue of the current era because a huge amount of research papers are being published. And find new articles related to your work has also become a challenge. According to [28], from 1998 to 2014, almost 120 recommender systems have been published. But it still does not know which recommender system gives good results [28]. Authors in [28], also tried to make the recommender system using similarity measures.

They used three similarity measures, which are bibliographic coupling, co-citation coupling and two variants of cosine similarity. According to them, content-based similarity measures do not produce good results. Because the content of some papers does not available freely. Therefore they limit their selection to network-based similarity measures. They compare these network-based similarities [28] on mathematical as well as empirical level. In mathematical comparison, they found that co-citation similarities produced the results that are less or equal to cosine similarity using columns of the adjacency matrix. Similarly, bibliographic similarities produced the results less or equal to cosine similarity using rows of the adjacency matrix. Further, authors concluded that there is a linear relationship in the computed similarity values.

In 2015, Hanyurwimfura [29] proposed a citation recommendation systems for non-profile users. His methodology was helpful to new user for whom data is not available to build their profile. He used content-based filtering approach, and take long queries as well as short queries as input. Long queries are taken from title and abstract, whereas short queries taken from the body of paper as well as from the title of paper. The similarity is calculated using cosine and made recommendations. For the evaluation of their recommendation systems, one paper per researcher is used for recommendation and each recommendation rated for its relatedness to their field of work. In their work, they found that query generation methods are main thing for the best performance of their recommendation system.

The authors, Xue et. al., aim to solve recommendation as a supervised ranking problem [30]. They split the corpus into two parts based on a time-frame. The older papers form the training set and the new ones are the validation/test set.The authors choose to construct features such as the page rank for paper, author and venue, the age of the paper, content similarity between titles, abstracts etc. Using these features, they train a Ranking SVM model. Evaluation was done against a few baseline approaches such as a CF and CBF. In the offline evalution, which was done on a Social Scholar dataset of 730,605 papers for 10,000 authors, it was reported that PaperTaste system outperformed the others in terms of the NDCGk value.

Philip and others in a 2014 paper [31] use a keyword-based vector space model to make article recommendations for digital libraries. They build a system with user interactions in order to build a user profile. They model papers by their keywords using a *tfidf* approach and used the cosine similarity measure to find relevant articles to recommend articles based on an input query. No evaluation of their framework was provided in this paper.

Tin Huynh and others in 2015, presented a recommender system that recommends scientific research articles using co-citation and co-reference factors in citation network [32]. They used the seed papers of citation network in order to recommends research articles. Moreover, they used *CCIDF*(Common Citation Inverse Document Frequency) algorithm and proposed its modified version named *CCIDF+*. *CCIDF* algorithm is used to compute relatedness of give document A to all other documents in the database.

Naoki et.al in [33] uses citation network to predict the existence of citation links. They have used the supervised machine learning model on 11 different features. Among these 11 features, cosine similarity, jaccard coefficient and Betweeness centrality highly affect the citation predictions results. In the end, they found that F values were between 0.74 to 0.82. Morover, they concluded that different research areas require different type of models and researchers must consider typology of targeted areas while predicting citation links in citation network.

Laura et.al in [28] Analysed network based similarity measures for research papers recommendation. They have used bibliographic coupling, cosine similarity and co-citation coupling as a similarity measures in citation network. The comparisons are conducted on empirical and mathematical level. In case of empirical comparison, they concluded that bibliographic coupling and one variant of cosine produced the same ranking. On the other hand, in case of mathematical evaluation, co-citation coupling and second variant of cosine produced the same ranking. Hence, if ranking consider than both measures are interchangeable.

2.3 Summary of Literature

To the best of our knowledge, the major problem with stereotypes is that they may pigeonhole users, and making stereotypes is manual work. As the items typically need to be manually classified for each stereotype, this limits the number of item recommendations. CBF has many number of advantages as compared to Stereotyping. CBF allows user modeling so the recommender system can judge the best recommendations items for each individual user. In case of research paper, features of paper(such title, abstract etc) are publicly available. So it recommends items as similar as possible to ones a user already knows. As per my knowledge, Collaborative requires rating of users, because users are too lazy to rating the item, this situation create a cold start problem. The cold start can occur in two situations, first, when new user comes and second is the arrival of new items. If new user rate very few items or no items, then recommender system cannot find like-minded user and cannot recommend items. If item is new and cannot rated yet by atleast anyone user, it cannot be recommends. In citation recommendation systems, the main disadvantage of Co-occurrence is that, it focus on the relatedness of papers instead of similarity of papers. In co-occurrence, papers can only be recommended if they co-occur at least once with another paper. For finding the citation papers, co-occurrence approach is not suitable.

As per our knowledge, citation recommendation in the literature ignores the quality and popularity of research articles[34]. For instances, two papers may be considered equally relevant if they share the same terms. This relevancy might not be justified, for example if one paper is written by expert (with ordinal results) in the field and have some worth, while another paper is written by a student(paraphrases the results of other research papers) have no worth in that field.

Another major problem, to the best of my knowledge, in the literature is that, existing citation recommendation techniques uses user profile and paper collection which is not available sometime (not all users have registered with their profile). Specially, this thing is not good for new the users.

Table 2.1: Critical Analysis of Existing Citation Recommendation Techniques in Literature

Ref	Focus point	Technique	Strength	Weakness
[29]	1)Non-profile user 2)Short queries 3)Long queries	1) Content-based Filtering 2) Textual Similarity	best for new users, because sometime new user not registered with profile	1) Long queries in document retrieval can degrade the results 2) Did not consider the worth of recommended papers
[31]	1) User interaction to make user profile 2) cosine similarity to compute similarity for input paper	1) Content-based Filtering 2) Textual Similarity	for an expert research, who interacted with the system most of the time, is helpful	1) Require well build user profile 2) Not-Consider the importance of recommended papers
[16]	1) Relevancy of paper 2) citation-context	1) Content-based Filtering 2) Textual Similarity	it can find relevant topics of the paper	1) Results were not good 2) Do not consider the importance of recommended papers
[15]	1) Find citation using partial citation 2) citation-context	1) Content-based Filtering 2) Textual Similarity	It recommend similar documents for citation context	1) Work was base on user profile, which is not available most the time 2) Do not consider the importance of recommended papers

Ref	Focus point	Technique	Strength	Weakness
[30]	Construct features such as the page rank for paper, author and venue, the age of the paper, content similarity between titles, abstracts etc	1) Content-based Filtering 2) Textual Similarity	Recommendation using supervised learning.	1)Getting similar paper by applying classification using features is not suitable 2) Not consider the worth of recommended papers
[35]	Research paper recommendation	1) Content-based Filtering 2) Textual-based similarity	Find Relevant and irrelevant Papers	Did not consider the importance of papers
[36]	Profile-based	1) Content-based Filtering 2) Textual-based similarity	Users preferences (likes)	1) Based On User Profile Which Is Not Available Most Of The Time. 2) Not consider the importance of recommended papers
[37]	1) Citation context 2) Vector representation by combining author and venue 3) Personalized citation	1) Content based filtering 2) Graph-based 3) LSTM network	1) Recommends relevant papers for citation context using neural network	1) Setting weights for parameter regularization my influence the recommendation performance 2) Ignores the importance of recommended papers

Ref	Focus point	Technique	Strength	Weakness
[38]	1) Content-based graph representation 2) Author-based graph representation 3) Personalized citation	1) Content based 2) Graph-based 3) GAN network	1) Recommends research papers for citation using neural network	2) Finding similar papers by combining network structure information can degrades the results 3) Setting weights for parameter regularization my influence the recommendation performance
[39]	1) Three layer graph using paper, author and venue 2) Mutual reinforcement 3) Personalized citation	1) Content-based 2) Graph-based 3) Clustering approach	1) Citation recommended via mutual reinforcement on layered graph	1) High computational complexity due to large size graph 2) Does not consider the importance of recommender papers
[40]	1) Combine citation analysis and network analysis 2) Multi-level citation network 3) Personalized citation	1) Content-based 2) Graph-based	1) Recommend research article by inspecting structural information	1) Although citation information is important, it may be insufficient for appropriate papers 2) Find relation between papers on multi-level can degrades results

Ref	**Focus point**	**Technique**	**Strength**	**Weakness**
[41]	1) Bibliographic network 2) Combining authors, papers, venues 3) Personalized citation	1) Content-based 2) Graph-based	1) Recommends research papers for citation	1) Most of the time researchers aims to find similar documents for which this is not suitable 2) Does not consider the important of recommended papers
[42]	1) Heterogeneous bibliographic network 2) Personalized recommendation 3) Edge prediction model	1) Content-based 2) Graph-based	1) Prediction and recommend citation using network representation based model	1) Finding similar papers by using exploring multi type of links in heterogeneous environment is not suitable 2) Complicated network representation by combining multiple type of links
[43]	1) Heterogeneous bibliographic network 2) Personalized citation	1) Content-based 2) Graph-based 3) GAN network	1) Citation recommendation using neural network	1) Makes the network complicated by combining sparse structural information 2) Ignores papers similarity 3) Manually parameter regularization may influence recommendation performance

Ref	Focus point	Technique	Strength	Weakness
[44]	1) Knowledge graph 2) Expand semantic features of given abstract	1) Content-based 2) Graph-based 3) Machine learning	1) Citation recommendation using semantic features of abstract	1) Results were not good 2) More feature are require for well build modal
[45]	1) Graph embedding 2) Neighborhood construction strategy 3) Distributed representation of papers	1) Graph-based	1) Rank candidate papers for citation recommendation	1) Graph embedding my influence structural information 2) Does not consider the worth of papers while ranking
[46]	1) Citation network 2) Semantic network 3) Co-relations with two networks	1) Content-based 2) Graph-based	1) Recommendation is based on computing similarity using top features	1) Mapping two network may influence network structure information 2) Required more close feature to compute similarity
[47]	1) Topic modeling 2) Feature extraction	1) Content-based 2) Graph-based	3) Citation is recommended by topic modeling	1) More effective feature require to find topic distribution 2) ignores the importance of recommended papers

Ref	Focus point	Technique	Strength	Weakness
[48]	1) Bibliographic Network 2) Personalized citation 3) Mutual reinforcement 4) Multi-layered graph	1) Content-based 2) Graph-based	1) Exploiting diversified link information in bibliographic	1) More relations exploration can enhance the results 2) Does not consider the importance of recommender papers
[49]	1) Clustering on citation network 2) Classic and expert recommendation	1) Content-based 2) Graph-based	1) Research paper recommendation for citation using hierarchal clustering	1) Results were not good 2) It may be difficult to recommend when seed paper have not enough citation information

Chapter 3

Research Methodology

The study in the previous chapter shows that researchers have proposed recommender systems which are based on textual similarity. Textual similarity based recommender systems find similar research papers through the text of research papers, but do not consider the importance of recommended papers. This thesis focuses network centrality-based methodologies to retrieve important papers that can be recommended to the readers.

In this chapter, we have discussed the detailed methodology of proposed recommender system. In our proposed approach, textual as well as topological similarities have been utilized to find most similar papers. Furthermore, the results have been verified using a citation dataset [50]. To check the accuracy of model, accuracy measure has been used. To compute the importance of paper, four centrality measures, i.e., Degree [9], Closeness [10], Betweeness [9] and Page Rank [10] have been computed on the citation graph extracted from the dataset. Figure 3.1 shows a graphical representation of the recommender system. Detail about each part of recommender system is given below.

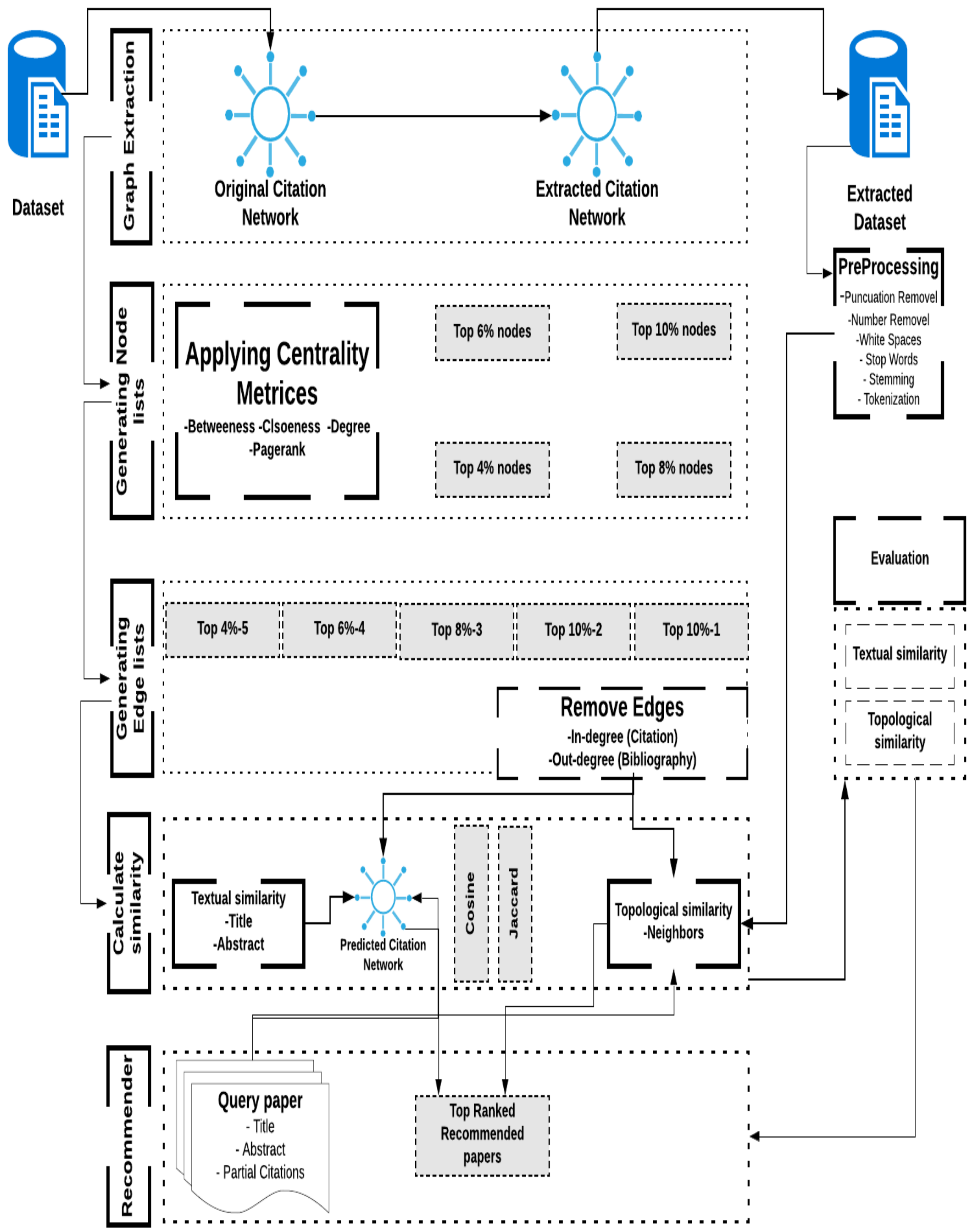


Figure 3.1: Framework Proposed Recommender System

3.1 Dataset

We have used Arxiv HEP-TH (high energy physics theory) [50] dataset in thesis experiments. The data was originally released as a part of *2003 KDD Cup* [51].*KDD Cup 2003*, a knowledge discovery and data mining competition held in conjunction with the 9th Annual *ACM SIGKDD* Conference. Dataset covers all the citations of 27,770 papers with 352,807 edges. If a paper i cites paper j, the graph contains a directed edge from i to j. If a paper cites, or is cited by, a paper outside the dataset, the graph does not contain any information about this.

3.1.1 Parameter Extraction

After getting profile list of all papers, next step is to extract information from these profiles for more experiments. Profile of every paper contains different parameters (shown in Figure 3.2). The format of profile is divided into two sections. First section contains metadata about the paper (i.e., *paper id*, *primary author*, *published date*, *paper title*, *co-authors*, *comment* about paper and *journal reference*). Second section contains abstract of the paper.

```
\\
Paper: hep-th/9201001
From: zuber@poseidon.saclay.cea.fr (C. Itzykson)
Date: Tue Dec 31 23:54:17 MET 1991 +0100   (37kb)

Title: Combinatorics of the Modular Group II: the Kontsevich integrals
Authors: C. Itzykson and J.-B. Zuber
Comments: 46 pages
Subj-class: High Energy Physics - Theory; Quantum Algebra
Journal-ref: Int.J.Mod.Phys. A7 (1992) 5661-5705
\\
  We study algebraic aspects of Kontsevich integrals
as generating functions for intersection theory over moduli space
and review the derivation of Virasoro and KdV constraints.
 1. Intersection numbers
 2. The Kontsevich integral
  2.1. The main theorem
  2.2 Expansion of Z on characters and Schur functions
  2.3 Proof of the first part of the Theorem
 3. From Grassmannians to KdV
 4. Matrix Airy equation and Virasoro highest weight conditions
 5. Genus expansion
 6. Singular behaviour and Painlev'e equation.
 7. Generalization to higher degree potentials
\\
```

FIGURE 3.2: Format of Paper Profile

Furthermore, to continue experiments, *titles* and *abstracts* are extracted and saved in separate files. To extract *title* and *abstract*, *TM* and *STRINGR* libraries of *R* tool have been used. In Figure 3.3 title and in Figure 3.4 abstract is shown.

```
" Domain Walls and Massive Gauged Supergravity Potentials"
```

FIGURE 3.3: Extracted Title

```
"An assessment of the present status of the theory, some immediate tasks which are suggested
thereby and some questions whose answers may require a longer breath since they relate to
significant changes in the conceptual and mathematical structure of the theory. "
```

FIGURE 3.4: Extracted Abstract

Next task is the **Pre-Processing**. Set of pre-processing steps will be performed to clean the extracted data (i.e., *title* and *abstract*). Lets describe all steps in detail.

- **Punctuation removal**: The characters, such as brackets, full stop and comma are called punctuation. These are used to separate sentences, words and clarify the meanings of sentences. In this step, these punctuation will be removed to clean the data.

- **Numbers removal**: Number removal is the process to remove digits from text. For calculating the text similarity of documents, numbers from titles and abstracts are removed. These numbers can be date, scores or something else, which is not helpful in our experiments.

- **White space removal**: Normally, document contains lots of white spaces, which are meaningless and not helpful in text mining process. As we know, similarity measures works only with non empty terms, therefore it is required to be removed white spaces from titles and abstracts. *Tm* package will be used to remove these white spaces.

- **Stop words removal**: In English language, there are multiple words that are called stop words (i.e., the, is, a, which, at, in etc). These words occur frequently, but are meaningless. These words are used to combine others words and do not contribute in content of text document. Here in the title and abstract, these words are often found. Therefore, it is important to remove these words from title and abstract to get the unique words. In this step, stop words will be removed from text data (i.e., *title* and *abstract*) by matching through available stop word list in *tm* package.

- **Stemming**: Stemming is the process to convert the words to their root term. For example, the words Presentation, Presented and Presenting would be converted into Present. Stemming is mostly used in text mining for information retrieval based on assumption that generating a query with Presenting will implies in all documents containing words Presentation and Presented. Advantage of stemming is that, it may reduce indexing size up to 50%.

- **Tokenization**: Text is collection of sequence of symbols. Mostly, before any text processing, text needs to be separated into chunks (i.e. numbers, words, alphanumeric etc). This process is named as tokenization. Finally, these tokens will be used to make a term document matrix through *tm* package.

3.1.2 Graph Extraction

Arxiv HEP-TH dataset contains a citation graph, which contains nodes and edges. The nodes represents the research papers and the edges between them represents the citations (shown in Figure 3.5).

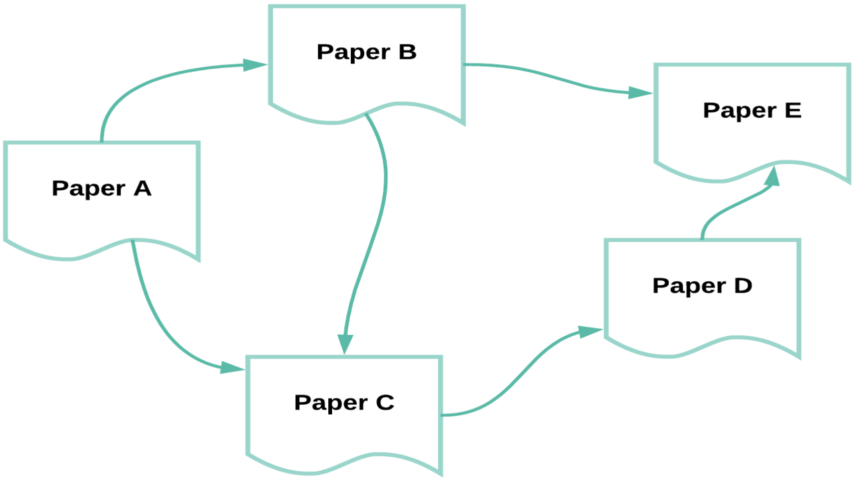


FIGURE 3.5: Citation Graph

This citation graph of *Arxiv HEP-TH* contains 27,770 nodes and 3, 52,807 edges. Profile of every research paper has been given, which includes the paper title, paper id, author name and abstract etc. This dataset contains research papers from January 1992 to 2003. The complete detail of dataset is given in Table 3.1.

Table 3.1: Summary of Dataset

DATASET STATISTICS	**Values**
$Nodes$	27770
$Edges$	352807
$NodesinlargestWCC$	27400
$EdgesinlargestWCC$	352542
$NodesinlargestSCC$	7464
$EdgesinlargestSCC$	116268
$Diameter(longestshortestpath)$	13
$Numberoftriangles$	1478735
$Numberoftriangles$	0.3120

This summary of dataset shows set of attributes that represent some characteristics of citation graph. Dataset contained a citation graph in the form of edge list. The excel file (which contained edge list), contains two columns. First column name is *From Node* and second column name is *To Node*, where *From Node* column contains list of citing papers and *To Node* column contains list of cited papers. We named this citation graph as G. Edge list of G is shown in Figure 3.6. This edge list shows that which paper is citing to which paper.

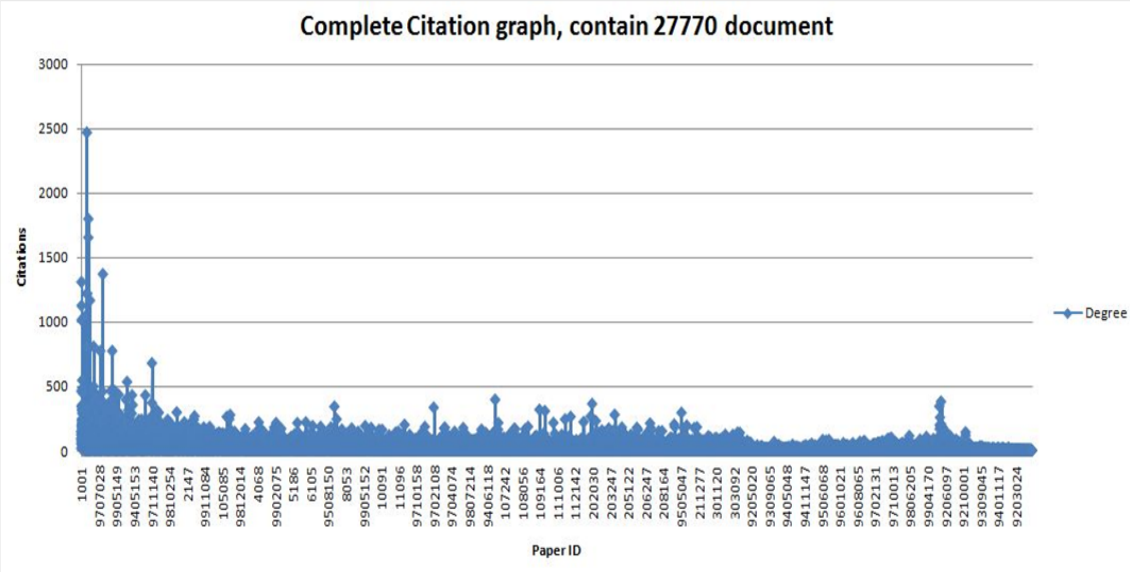


FIGURE 3.6: Edge List of Citation Graph G

The citation graph G (contained 3, 52,807 edges) was very sparse and was taking too much time in experiments. Therefore, we have decided to extract new citation graph from G. The new citation graph $\acute{G}$ is then extracted from G. This citation graph $\acute{G}$ contains 8,179 nodes and 1, 43,906 edges. In this new citation graph $\acute{G}$(shown in Figure 3.7), we have included only those papers which have 10 or more than 10 citations. The edge list of $\acute{G}$ contains two columns named *From* and *To*, where *From* represents the citing paper and *To* represents the cited paper.

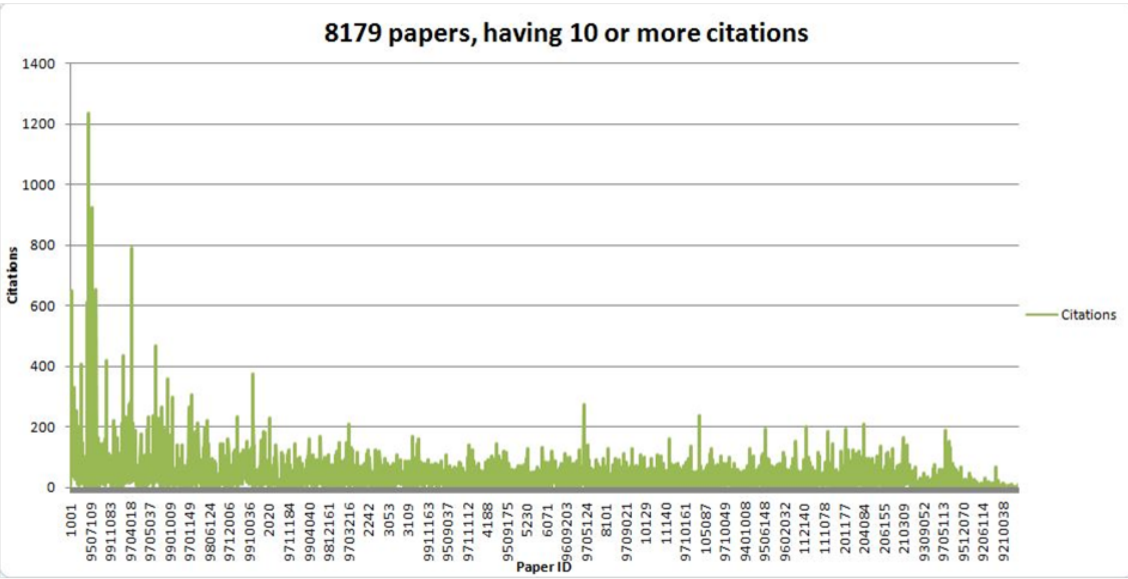


FIGURE 3.7: Edge List of Citation Graph $\acute{G}$

For extraction of $\acute{G}$, *igraph* library is used in R tool.

3.2 R Tool

R is software that gives us programming platform to execute statistical analysis on the data. Programmer or data analyst use *R* for data mining and get the required output after experimentation. R supports *Igraph* library which offers the researchers convenient tools for network sciences. R facilitates the programmer with an open source library which enables to create a graph of millions of nodes and edges. It also facilitates different file format (i.e. .xls, .csv, .txt, .sas and .xml)[52].

3.2.1 *Igraph* Library

Igraph is a library that is used for network analysis. It contains routines for simple graphs and network analysis. It can handle large graphs very well and provides functions for generating random and regular graphs, graph visualization, centrality methods and much more. The main goals of the *igraph* library is to provide a set of data types and functions for

- pain-free implementation of graph algorithms,
- fast handling of large graphs, with millions of vertices and edges,
- allowing rapid prototyping via high level languages like R.

3.3 Centrality Metrics

This thesis worked with four commonly used centralities such as Closeness, Degree, PageRank, and Betweeness [53]. To use these centrality matrices, we used R tool which supports *igraph* library. By using these centralities, citation network is then placed into *igraph* and centrality metrics have been computed. After applying centralities, now we obtained four lists of nodes in descending order.

- **Degree Centrality**: Degree centrality is defined as the number of edges that a node shares with others and it signifies the importance of the node in a network. Degree centrality [9] of a node i determines its connectivity in the network and is represented as:

$$CD(n_i) = deg(n_i) \tag{3.1}$$

  In this formula, ni shows the current paper whose degree is to be computed. For directed networks, two measures of degree centrality are represented i.e. In-degree and Out-degree .

  - **In-degree**:In a network, In-degree represents the count of the number of edges directed towards the node [9].
  - **Out-degree**: In a network, Out-degree represents the number of edges that node directs to others [9].

- **Closeness Centrality**: The closeness of the node is measured by the average length of the shortest paths between the node and all other nodes. In a citation network, the value of closeness indicates the average number of papers to be followed via references of other papers to traverse from single paper to any other paper in the network. The formula to calculate closeness is as follows [10]:

$$C_c = \sum_{i=1}^{N} \frac{1}{d(ni, nj)} \tag{3.2}$$

  In this formula, the total sum is computed for all the average length of shortest paths between nodes with all other nodes and then its reciprocal shows the value of Closeness. ni shows current paper whose closeness is computed and *d(ni,nj)* represents the shortest path between each pair of papers.

- **Betweeness Centrality**: Betweeness centrality defines the range in which a specific node lies between other nodes in a network. It is described by Xue et al. in [33] first time. A node is said to be more influential if it is on the shortest paths joining many node pairs or maybe it is in that position where node acts as a bridge between these pairs. Betweeness of node i represents the ratio of all shortest paths passing through it [9].

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3.3}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v.

- **PageRank Centrality**: PageRank is an algorithm which is generally used ranking for Web pages. Normally PageRank is calculated by the number of pages associated with the main website. PageRank of a node determines the nodes comparative importance within the whole set of nodes in the network. The formula to calculate PageRank is as follows [10]:

$$PR(Pi) = \frac{1-d}{N} + d \sum p \in M(pi) \frac{PR(Pj)}{L(Pj)} \tag{3.4}$$

In Equation 3.4:

- N represents a number of edges/pages,
- d represents dumping factor and an arbitrary weighting factor,
- $PR(Pi)$ is the PageRank of node/page,
- $L(Pj)$ is the number of outgoing edges from the node,
- $M(pi)$ is the set of links.

3.3.1 Generating Nodes Lists

Further, degree centrality is applied and then sorted the nodes in descending order. Then we have picked 4 set of nodes (top10%,top8%,top6%, and top4%) from the top of list and made another 4 lists. These extracted lists of papers further explored for similarity computation. After applying betweeness, closeness and pagerank, we obtained other 12 lists. The extracted lists are explain in Table 3.2.

Table 3.2: Getting Lists of Nodes After Applying Centrality Measures (i.e., Degree, Betweeness, Closeness and Pagerank)

List	**Nodes**
$TotalNodesinDataset$	8179
$Top10\%$	818
$Top8\%$	654
$Top6\%$	490
$Top4\%$	327

3.4 Similarity Computation

" Similarity: Comparison of commonality between different objects "

Similarity has been a subject of great interest in human history since a long time ago. Even before computers were made, humans have been interested in finding similarity in everything. Similarity computation is the process of compute similarity of items and then to select the most similar items. The basic idea in similarity computation between two items i and j is to first make a list of parameters which belongs to these items and then to apply a similarity computation technique to determine the similarity of i and j. Here, in this thesis, similarity between papers is computed on textual as well as topological parameters.

3.4.1 Textual Similarity

Textual Similarity approaches play an important role in text related research activities and applications. Textual similarity is widely used in information retrieval, text classification, document clustering, topic detection, topic tracking and others [54]. Finding similarity between words is a fundamental part of text similarity which is then used as a primary stage for sentence, paragraph and document similarities. Text Similarity is calculated between documents and web pages on the base of text which is given in that. In this thesis, we compute text similarity between set of papers using *Title* and *Abstract*. Cosine similarity and Jccard [55] similarity are used to compute similarity of papers, because these measures are usually used to measure similarity between two vectors[56].

Title Similarity

Title similarity is calculated between title of the citing and cited papers. Title similarity is calculated using Cosine and Jaccard index. Equation 3.5 is the Cosine, while Equation 3.6 represents Jaccard index. Jaccard index which is also known Jaccard similarity coefficient, is used to compare sample sets. For example, consider a set $A = link, prediction, social, network$ and $B = social, network, ties$. Both sets A and B have 3 common terms and 5 unique terms. The similarity of set A and B using Jaccard index in Equation 3.6 is $J(A, B) = 3/5 = 0.6$

$$Cos(d1, d2) = \frac{\vec{d1} \cdot \vec{d2}}{|d1||d2|} \tag{3.5}$$

$$Jac(A, B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{3.6}$$

In Equation 3.5, $d1$ and $d2$ are representing the set of terms. While A and B in Equation 3.6 represent the set of terms.

Abstract Similarity

The abstract of research article describes the purpose, hints that idea is adopted from someones work and briefly demonstrates overall outcome of the article. If high similarity exists between abstract of research articles, this increases the chances that current work extends the previous work. Based on this assumption, the abstract similarity between paper-citation pairs is calculated. The similarity is computed by using Cosine similarity of *tf-idf* scores. The Cosine similarity between two terms or documents on the vector space is a measure that calculates the cosine of the angle between them.

In this thesis, the similarity is computed by using cosine similarity and jaccard similarity using abstract of citing and cited papers. The formula to calculate cosine similarity is given in Equation 3.5 and jaccard similarity in Equation 3.6.

3.4.2 Topological Similarity

Topological similarity is calculated between two pair of nodes(i.e. Documents) in graph(i.e. Citation Graph). It is based on the simple idea: the more similar the pair is, the more likelihood a link between them, and vice versa. It can be measured by the similarity, in which each non-connected pair of nodes (d1; d2) is assigned a score signifying similarity between d1 and d2. A high score indicates high probability that d1 will cite to d2, while a low score also indicates high probability that d1 will not cite d2. Therefore, using the rank of similarity scores, we can predict and recommend citation for a document. In a citation network, paper can have many cited papers or citing papers. Here cited papers represent the bibliography(i.e. *out-degree* of paper) and citing papers represent the citations(i.e. *In-degree* of paper). Citation represents the situation where one papers is cited by other papers, while bibliographic occurs when paper cites other papers. Both bibliography and citations are the two topological features of the citation network and this thesis used these two topological features (shown in Figure 3.8) to calculate the similarity of papers.

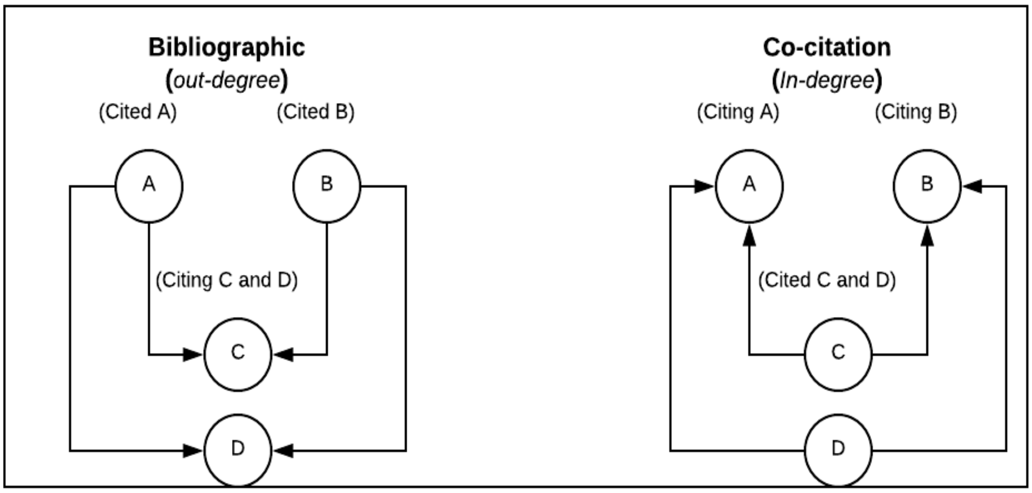


FIGURE 3.8: Bibliographic vs Co-citation

3.4.2.1 Citation-Based Similarity

Co-citation is a similarity measure for documents that makes use of citation relationships. Co-citation is defined as the frequency with which two documents are cited together by other documents. If at least one other document cites two documents in common these documents are said to be co-cited. The more co-citations two documents receive, the higher their co-citation strength, and the more likely they are semantically related. The concept of co-citation is illustrated in Figure 3.8, where documents C and D cite documents A and B. Here in Figure 3.8, documents A and B are co-cited.

3.4.2.2 Bibliography-Based Similarity

Bibliographic-based similarity is used to establish similarity relationship between documents. Two documents are bibliographically similar, if they both cite one or more documents in common. The Figure 3.8 illustrates the concept of bibliography. In the Figure 3.8 documents A and B both cite documents C and D. Thus, documents A and B have similarity.

3.5 Evaluation

In order to evaluate the proposed technique, accuracy measure is used. Model Accuracy is the ratio of number of correct predictions to the total number of input samples. Here in this thesis, the input is edges of the citation graph.

3.5.1 Accuracy

For the evaluation we have devised a model to compute the accuracy score between real graph and predicted graph. The accuracy score for the predicted graph Gp and real graph Gr is calculated using the following measure 3.7.

$$Accuracy = 1 - \frac{E(G_1) + E(G_2) - 2E(G_1 \bigcap G_2)}{Max(E(G_1), E(G_2))} \tag{3.7}$$

In Equation 3.7:

- E represents the Edges of the citation graph,
- G_1 is the original citation graph,
- G_2 is the predicted citation graph,
- Max function will return the maximum number of edges from original and predicted citation graph.

Chapter 4

Experiments and Results

This chapter provides details related to the experimental setup and analysis of proposed technique. Moreover, comparison of textual similarity with topological similarity for citation recommendation is presented in the last section of this chapter.

4.1 Experimental Setup

The experiments according to methodology are performed step-by-step. Dataset Arxiv HEP-TH (High Energy Physics Theory) is used for the experiments. Initially, this dataset contains a citation graph and profiles of papers in the period from 1993 to 2003. The citation graph contained 27770 papers and 352807 edges. First, the initial step was extraction of the dataset. This experiment performed with the extracted portion of dataset, which was contained 8179 papers and 143906 edges, because it was taking too much time in experiments using the original dataset. This extracted dataset contained only those papers which have 10 or more than 10 citations. Second, *title* and *abstract* are extracted. Third, degree, closeness, betweeness and page rank centrality metrics are applied on the citation graph. After applying the centrality metrics, lists of nodes are made (See section 3.3.1). Fourth, in order to compute similarity using co-citation and bibliography, in-degree and out-degree edges are picked for making edge lists. After picking these edges, we have removed these edges from citation graph and make another citation graph. Finally, textual similarity and topological similarity is computed between papers and evaluated the results.

4.2 Generating Edges Lists

After applying centrality measures, we obtained total 16 set of nodes where 4 sets belong to each centrality measure (as shown in Table 3.2). The next step is to get lists of edges in order to compute similarity. For making lists of edges, following steps are performed.

- First we Picked up four lists (i.e., top10%, top8%, top6% and top4%) of degree centrality measure (as shown in Table 3.2).
- Using top10% list, we randomly pick one indegree edge from each node and make edge list called top10%-1. Considering Table 3.2, top10% list contains 818 nodes, so the extracted edge list contains 818 edges.
- For making second edge list, using top10% list, randomly two indegree edges picked from each node and made another edge list top10%-2. This list contains 1634 edges.
- For the third edge list, we used top8% list, then we pick randomly 3 indegree edges from each node and make top8%-3 edge list. Here, in this list, number of edges are 1962.
- To make the fourth edge list, we used top6% list. Here, randomly 4 indegree edges from every node are picked and made top6%-4 edge list. This list contained 1960 edges.
- For the fifth edge list , top4% list used. Here, we pick randomly 5 indegree edges from each node. Then make another list called top4%-5. This list contain 1635 edges.
- Finaly, the 10 iterations are performed on the above 5 steps. In this way, 50 edge lists are computed just for the degree centrality.

After applying above 6 steps for the degree centrality, we have 50 edge lists of 5 different kinds. The same steps are performed for betweeness,closeness and Pagerank. Uptill now, indegree (citation) edges are picked and 200 edge lists (50 for each centrality measure) are made. The same procedure (which is applied on indegree edges) is then applied in order to pick outdegree (bibliography) edges. In the end, we have 400 edge lists (200 for each indegree and outdegree). Furthermore, statistics of edges lists are shown in Table 4.1.

Table 4.1: Edge Lists for Each Centrality Measure (i.e., Degree,Betweeness,Closeness and Pagerank)

Edge List	**Edges**	**Nodes**	**Titles**	**Abstracts**
$Top10\% - 1$	818	1634	1634	1634
$Top10\% - 2$	1634	3268	3268	3268
$Top8\% - 3$	1962	3924	3924	3924
$Top6\% - 4$	1960	3920	3920	3920
$Top4\% - 5$	1635	3270	3270	3270

Table 4.2: Edge Lists of 10 Different Iterations for Each Centrality Measure (i.e., Degree,Betweeness,Closeness and Pagerank)

Edge List	**Edges**	**Titles**	**Abstract**
$Top10\% - 1$	8180	16340	16340
$Top10\% - 2$	16340	32680	32680
$Top8\% - 3$	19620	39240	39240
$Top6\% - 4$	19600	39200	39200
$Top4\% - 5$	16350	3270	3270
Sum	80090	160180	160180

After performing 10 iterations for every list, statistics of edge lists are shown in Table 4.2.

4.3 Bibliographic-Based Similarity Computation

As discussed above, 200 edge lists are computed using *outdegree* edges. These *outdegree* edges are the bibliography of the papers. In this section, bibliographic-based similarity is computed and results are presented. In the bibliography, two types of similarities have been computed. First is textual similarity, which is calculated using *title* and *abstract* of the paper. Second is topological similarity, which is calculated using neighbor nodes of the paper in citation graph. In the end, both (textual and topological) similarities are evaluated in order to identify the correct citation links.

Table 4.3: Textual Similarity and Topological Similarity of Documents

Term	**Defination**
Tjac	Textual Jaccard similarity using Titles of documents
Tcos	Textual Cosine similarity using Titles of documents
Ajac	Textual Jaccard similarity using Abstract of documents
Acos	Textual Cosine similarity using Abstract of documents
Topjac	Topological Jaccard similarity using neighbors of nodes(documents) in citation network
Topcos	Topological Cosine similarity using neighbors of nodes(documents) in citation network

4.3.1 Textual Similarity

For the textual similarity, experimentation is done on two parameters, which are *title* and *abstract* of the paper. As mentioned above, 200 edge lists are used in order to compute bibliographic-based textual similarity. These edge lists are of

5 different kinds (i.e., *Top10%-1, Top10%-2, Top8%-3, Top6%-4 and Top4%-5*) and made up from 4 different set of nodes (i.e., *Top10%, Top8%, Top6% and Top4%*).

Title Similarity

For computing textual similarity using *title*, experimentation is done on 200 edge lists (50 for each centrality measure) from bibliography. First, 50 edge lists (10 iteration per edge list shown in Table ??) from *Betweeness* are picked. Then *titles* of nodes in the edge lists are extracted. After that, similarity of *titles* using *jaccard* and *cosine* similarity is calculated (as shown in Figure 4.1). In Figure 4.1, threshold on 5 different set of edge lists are shown on x-axis and on the y-axis, percentage of accurate identified citation links is shown. The same pattern is followed in all the figures. The resultant thresholds shown that the threshold *0.02* achieved the highest results with *36.8%* citation links. Same behaviour for threshold *0.05* in the all edge lists shows well identification of citation links. The main thing which can be seen in this figure is the *cosine* similarity. In case of all the edge lists, *cosine* similarity performed well with respect to *jaccard* similarity.

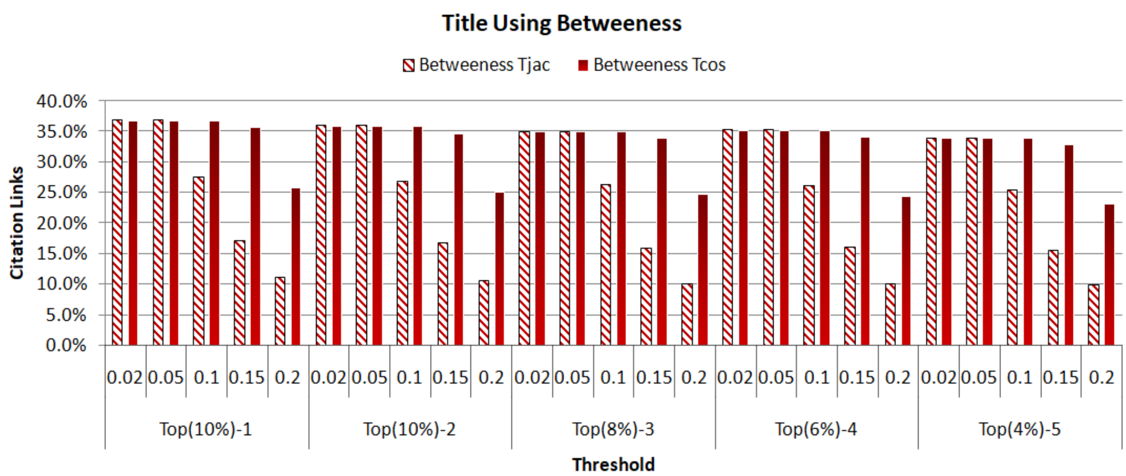


FIGURE 4.1: *Jaccard* similarity and *Cosine* similarity on top nodes using *Betweeness*

Results of *Closeness* are shown in Figure 4.2. The threshold values *0.02* and *0.05* almost achieved the same results by identifying *31.2%* citation links from all the edge lists. Out of all the edge lists, *Top10%-1* and *Top10%-2* are contributing well on all the thresholds. The *cosine* similarity again obtained good results than *jaccard* similarity. In all edge lists, threshold *0.15* and *0.2* presenting the big difference between *cosine* similarity and *jaccard* similarity.

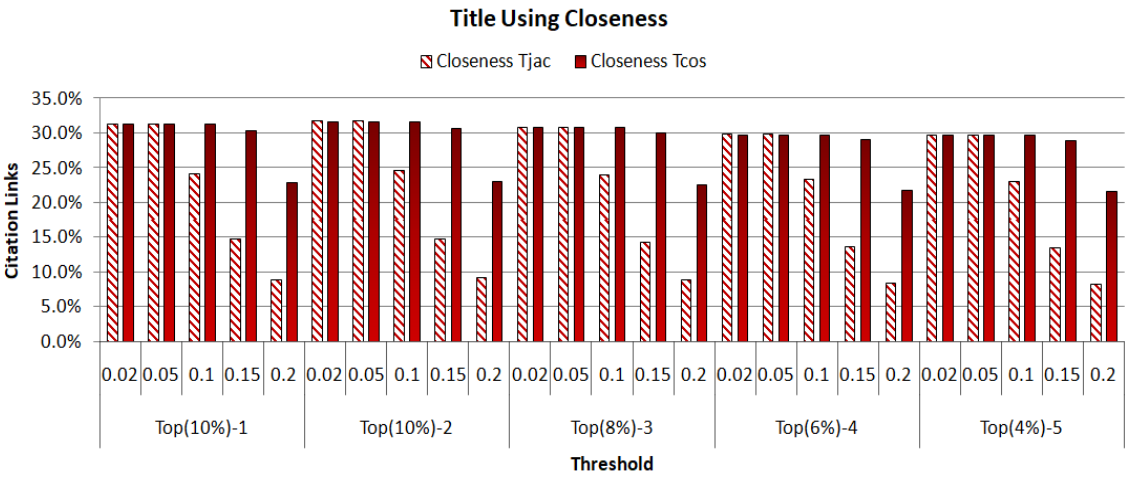


FIGURE 4.2: *Jaccard* similarity and *Cosine* similarity on top nodes using *Closeness*

Results of *Degree* are presenting in Figure 4.3. In this Figure 4.3, *jaccard* and *cosine* similarity obtained highest results by getting *35.8%* for the first two edge lists (*Top10%-1* and *Top10%-2*). The threshold values *0.15* and *0.2* shows that as the threshold increased, *jaccard* similarity decreased. At threshold *0.2* in the edge list *Top10%-2*, *cosine* similarity obtained *25.1%* and *jaccard* similarity only *9.9%*.

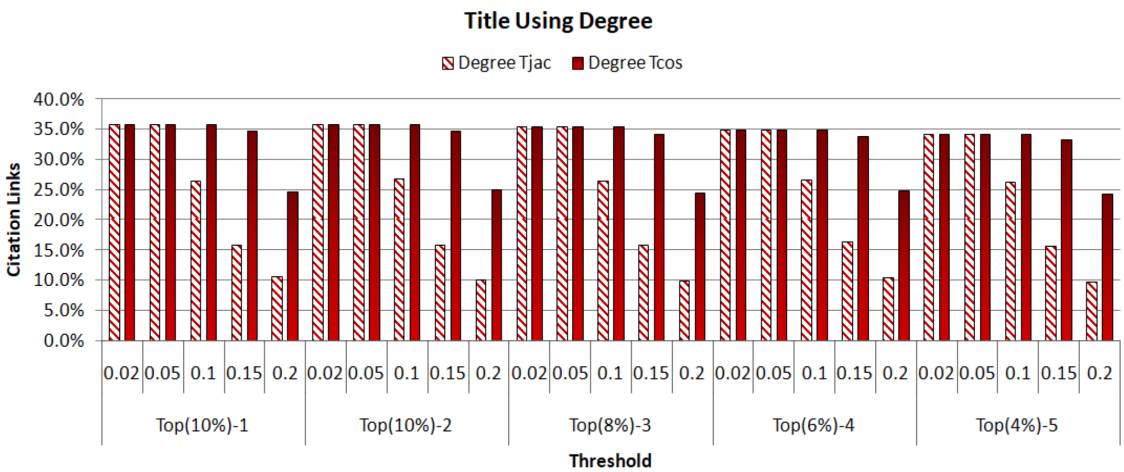


FIGURE 4.3: *Jaccard* similarity and *Cosine* similarity on top nodes using *Degree*

The Figure 4.4 presenting the results of *Pagerank*. In this Figure 4.4, same thresholds *0.02* and *0.05* obtained highest results. These thresholds in all edge lists, almost identify *40%* citation links. For the remaining thresholds, a big difference can be seen here between *cosine* and *jaccard* similarity. For all the edge lists, *cosine* similarity obtained good results, as almost *40%* on the threshold 0.1.

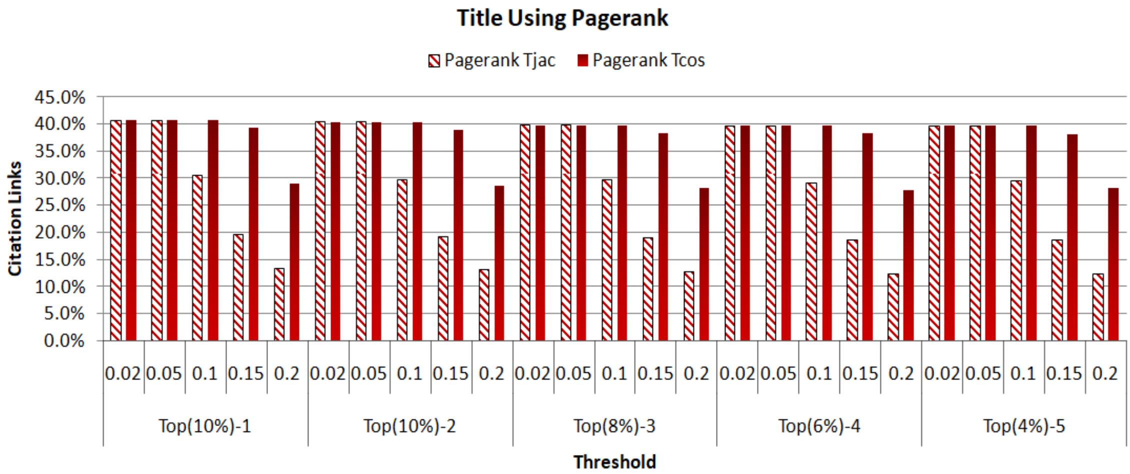


FIGURE 4.4: *Jaccard* similarity and *Cosine* similarity on top nodes using *Pagerank*

The Figure 4.5 is the combination of all the centrality measures. In this Figure 4.5, x-axis represents the average threshold from all the edge lists with respect to their centrality measure. For example, *Betweeness Tjac* and *Betweeness Tcos*, the first two bars at threshold *0.02* are the averages of thresholds *0.02* from all the edge lists from *Betweeness* (shown in Figure 4.1). The Figure 4.5 shows that at threshold values *0.02* and *0.05*, *Pagerank Tjac* and *Pagerank Tcos* achieved overall good results by correctly identifying *40%* citation links. Out of two similarity measures (*jaccard* and *cosine*), at threshold *0.15*, *Pagerank Tcos* (*cosine* similarity) obtained *39%* while *Pagerank Tjac* (*jaccard* similarity) identify *19%* citation links.

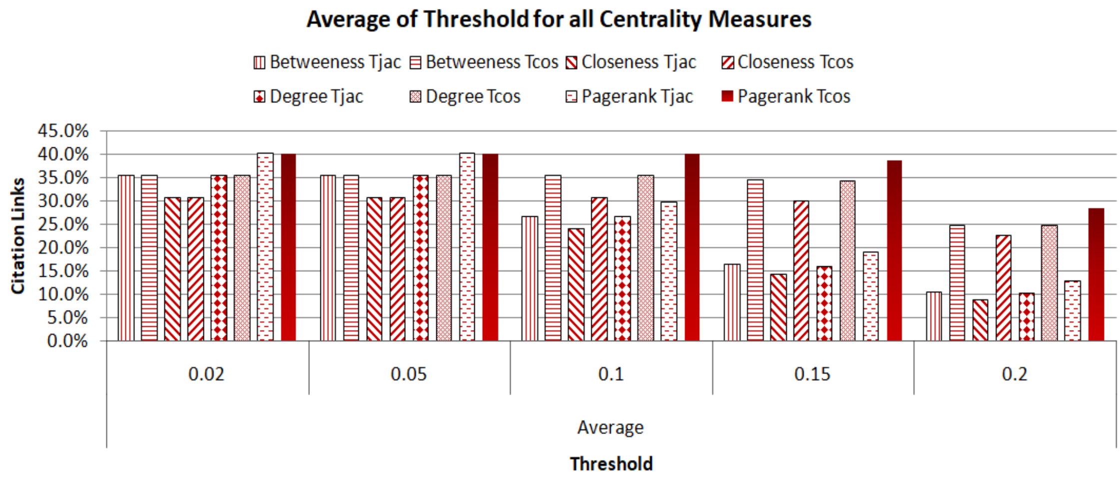


FIGURE 4.5: Average Title Similarity

Abstract Similarity

For computing textual similarity using *abstract*, 200 edge lists (50 for each centrality measure) from bibliography are used for the experimentation. First of all, abstracts of nodes in the edge lists are extracted. After that, similarity of papers using abstracts is calculated through *jaccard* and *cosine* similarity.

The results of *Betweeness* are shown in Figure 4.6. Figure 4.6 clearly shows that textual similarity using *abstract* produced better results than using *title*. The previous statement is further justified on the threshold *0.02*, there are almost *96.1%* citation links are identified by *Betweeness Acos*. Another interesting fact which can be seen is the *Betweeness Acos* (*cosine* similarity), which is competing the *Betweeness Ajac* (*jaccard* similarity) by achieving almost *49.6%* citation links on the threshold *0.15*. On the other hand, for the same threshold *0.15*, *Betweeness Ajac* (*jaccard* similarity) degrades its results by getting *2.5%* citation links.

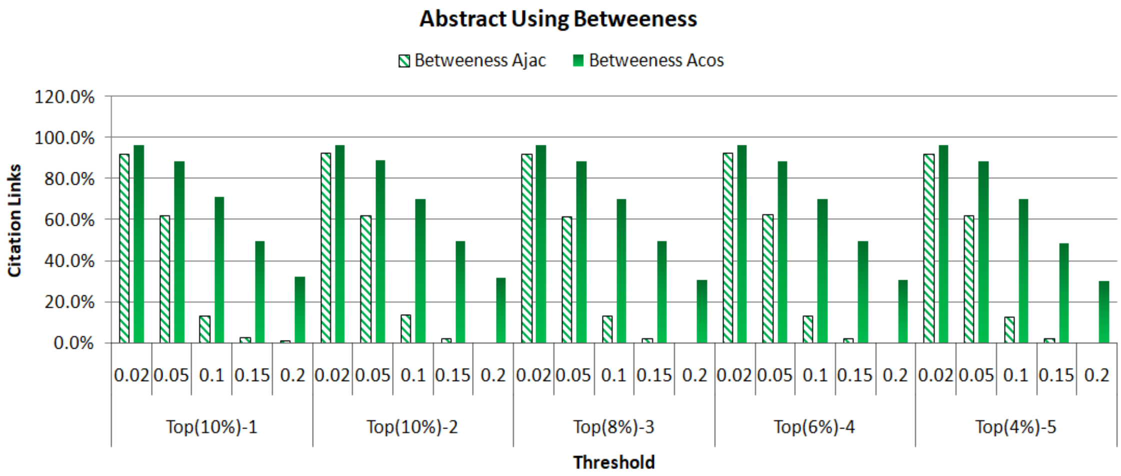


FIGURE 4.6: *Jaccard* similarity and *Cosine* similarity on top nodes using *Betweeness*

The results of *Closeness* are shown in Figure 4.7. There is a slight difference between results of *Closeness* and *Betweeness*. *Jaccard* is the only measure, which produced slight different results in the this Figure 4.7 compared to Figure 4.6.However, *cosine* produced the same results as produced in Figure 4.6.

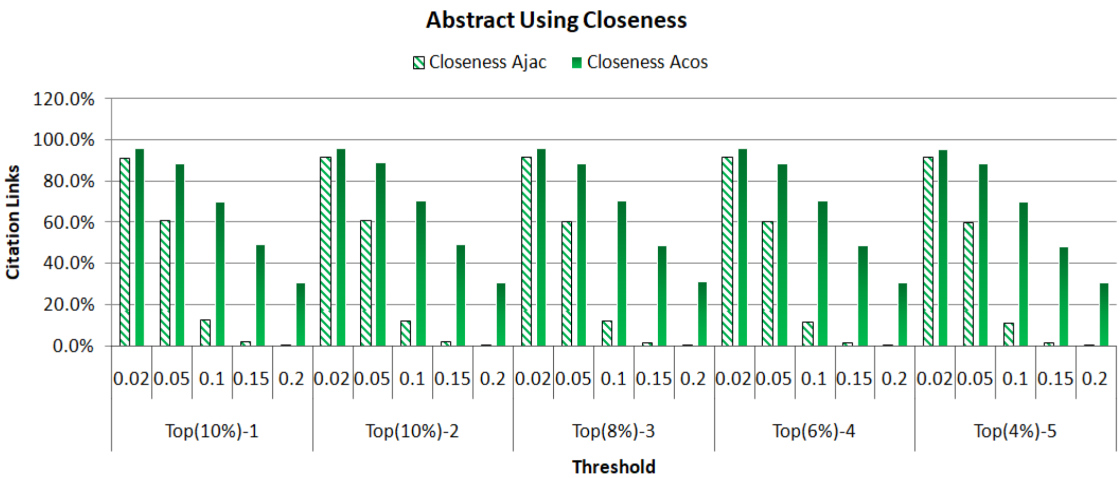


FIGURE 4.7: *Jaccard* similarity and *Cosine* similarity on top nodes using *Closeness*

The Figure 4.8 is presenting the results of *Degree*. For threshold *0.02*, *Degree Acos* is contributed in identification of *96.6%* citation links, while *Degree Ajac* only *92.4%*. In the first edge list *Top10%-1* for the threshold *0.2*, *Degree Acos* achieved *33.7%* and *Degree Ajac* only *0.4%*.

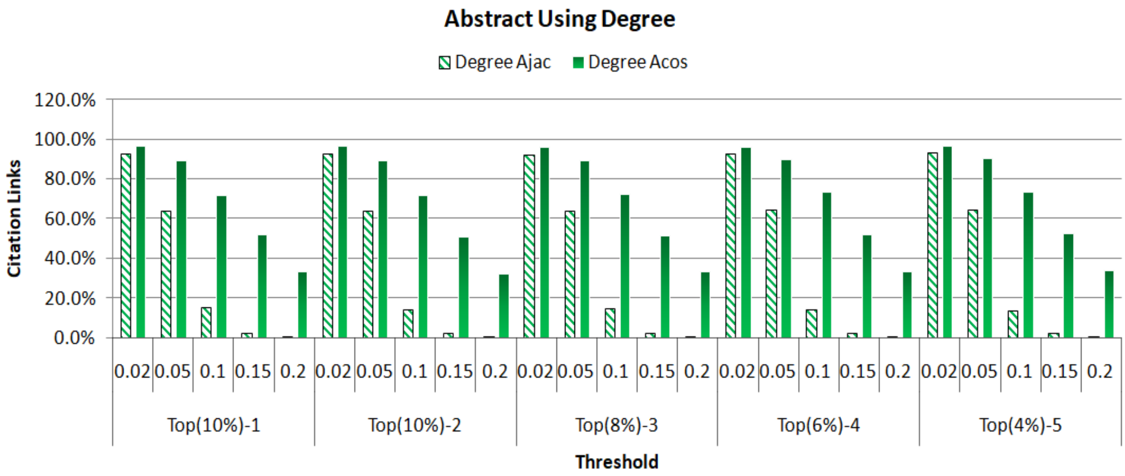


FIGURE 4.8: *Jaccard* similarity and *Cosine* similarity on top nodes using *Degree*

The results of *Pagerank* are shown in Figure 4.9. In this Figure 4.9, at threshold *0.02*, *Pagerank Acos* obtained *96%* score, while *Pagerank Ajac* achieved *92.5%*. For the remaining thresholds *0.1*,*0.15* and *0.2*, *Pagerank Acos* performed well against *Pagernk Ajac*.

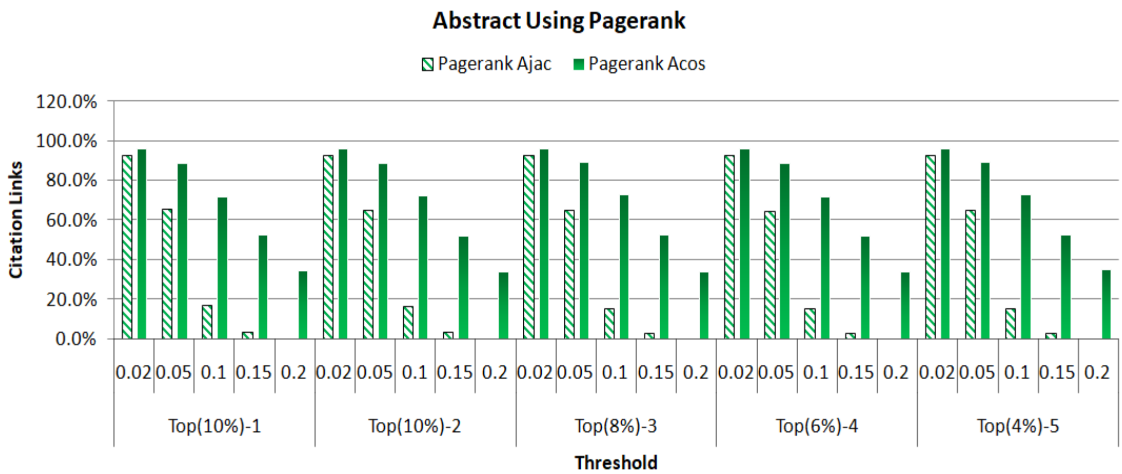


FIGURE 4.9: *Jaccard* similarity and *Cosine* similarity on top nodes using *Pagerank*

The Figure 4.10 is the combination of all the centrality measures. In this figure 4.10, x-axis represents the average threshold from all the edge lists with respect to their centrality measure. For example, *Betweeness Tjac* and *Betweeness Tcos*, the first two bars at threshold *0.02* are the averages of thresholds *0.02* from all the edge lists from *Betweeness* (shown in Figure 4.6). The Figure 4.10 shows that at the threshold *0.02*, all the centrality measures produced equally good results. But when moved towards threshold *0.2*, all the centrality measures degrade results. For all the thresholds, *Degree Acos* achieved highest results than others. Over all, *Cosine* (i.e., *Betweenss Acos, Closeness Acos, Degree Acos* and *Pagerank Acos*) similarity outperformed than *jaccard* (i.e., *Betweeness Ajac, Closeness Ajac, Degree Ajac* and *Pagerank Ajac*) similarity.

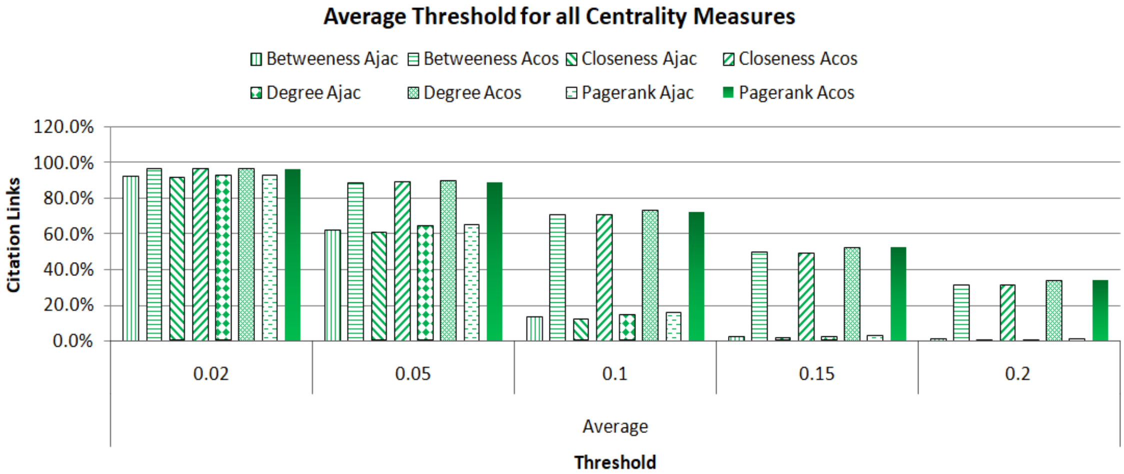


Figure 4.10: Average Abstract Similarity

4.3.2 Topological Similarity

For topological similarity, we have used citation network. Experimentation in topological similarity considered one parameter, which is neighbors of the paper. Here also, 200 edge lists (50 for each centrality measure) from bibliography are used. First, we picked these edge lists one by one, then remove these edges from the original graph and made another graph. In order to infer these removed edges, cosine and jaccard similarity measures are used. After applying these similarity measures, different thresholds are applied. After that, using formula 3.7, got some accuracy score for each edge list. This accuracy score represents the percentage of accurate identified citation links. The Figure 4.11 is presenting the results of *Betweeness*. At threshold *0.02* in first edge list *Top10%-1*, *Betweeenss Topcos* obtained *99.9%* citation links, while *Betweeness Topjac* achieved *97.1%*. When the threshold was *0.2* in edge list *Top4%-5*, *Betweeness Topjac* obtained *20.2%* citation links. For the same threshold *0.2*, Topological similarity outperformed than textual similarity(results shown from Figure 4.1 to Figure 4.10) using *title* and *abstract*.

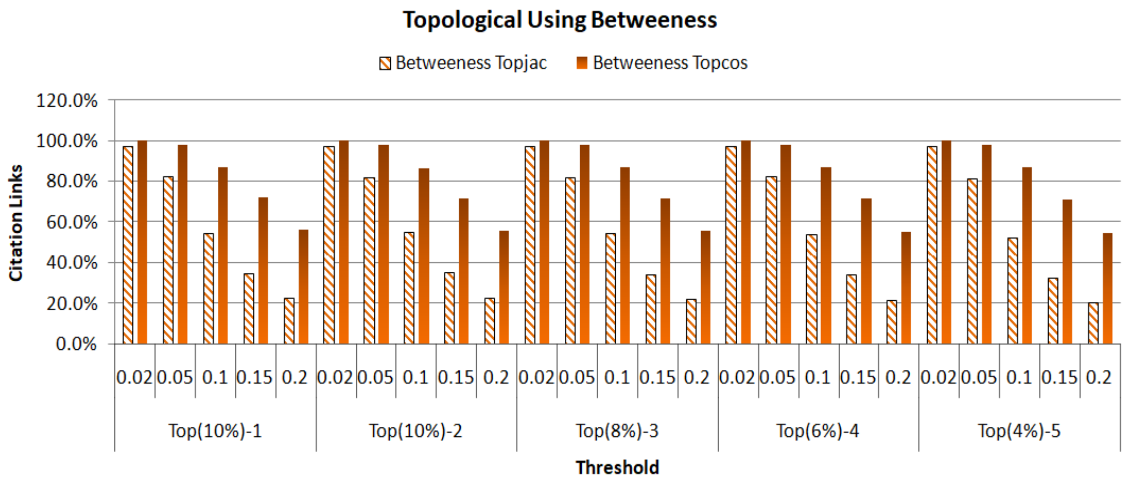


FIGURE 4.11: *Jaccard* similarity and *Cosine* similarity on top nodes using *Betweeness*

The results of *Closeness* are shown in Figure 4.12. Here in this Figure 4.12, *Closeness Topcos* obtained *99.9%* citation links and *Closeness Topjac* achieved *95.2%*. Out of all the edge lists, *Top10%-1*, *Top10%-2* and *Top8%-3* are contributing equally on all the thresholds. In the edge list *Top4%-5*, *Closeness Topcos* and *Closeness Topjac* produced not good results as they produced in *Betweeness Topcos* and *Betweeness Topjac* (shown in Figure 4.11).

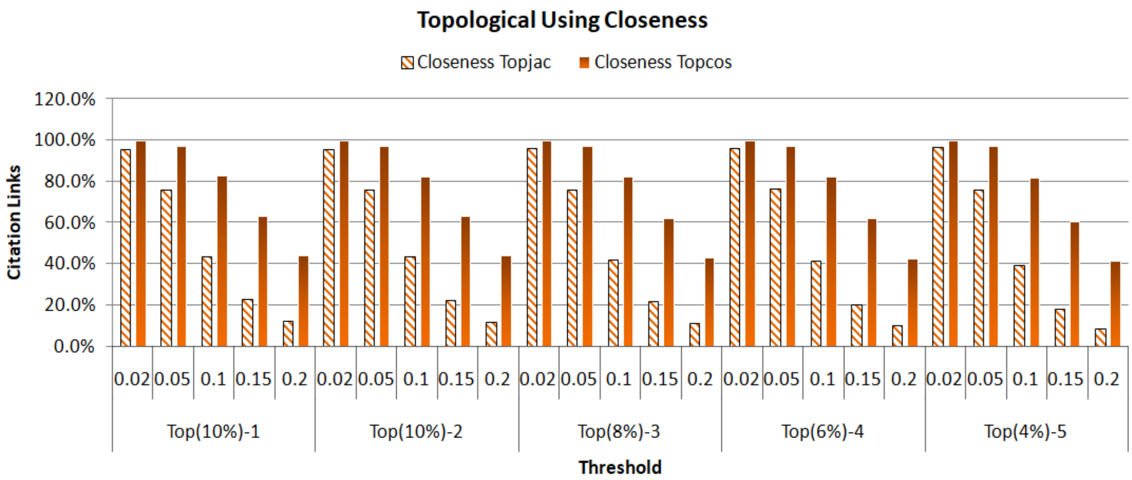


FIGURE 4.12: *Jaccard* similarity and *Cosine* similarity on top nodes using *Closeness*

In Figure 4.13, *Degree Topcos* and *Degree Topjac* followed almost the same pattern as followed by *Betweeness Topcos* and *Betweeness Topjac* in Figure 4.11. The results of *Degree* are presenting in this Figure 4.13. At threshold *0.02* in edge list *Top10%-1*, *Degree Topcos* contributed in identifying of *99.9%* citation links, while *Degree Topjac* obtained *97.7%*. When the threshold was *0.2* in edge list *Top10%-1*, *Degree Topcos* succeeds in getting *55.5%* and *Degree Topjac 20.6%*.

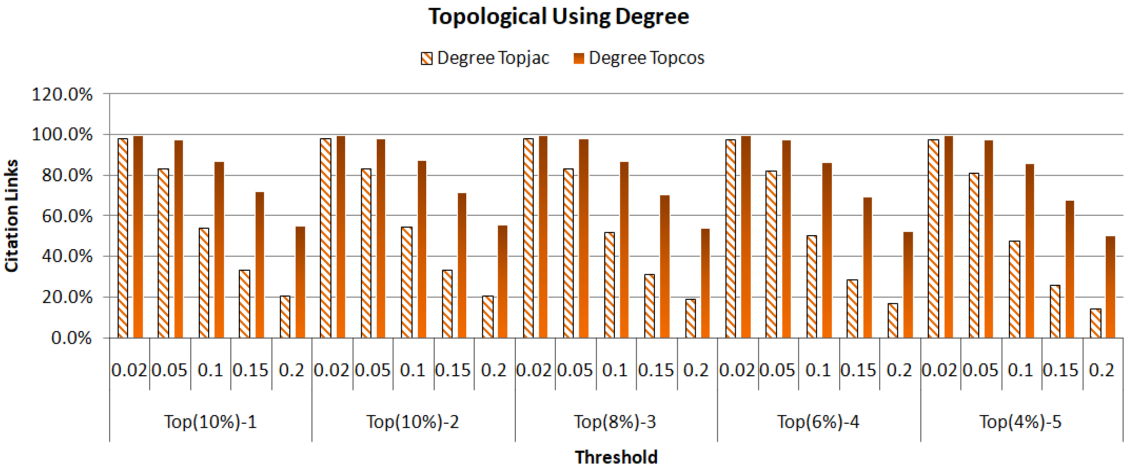


FIGURE 4.13: *Jaccard* similarity and *Cosine* similarity on top nodes using *Degree*

The Figure 4.14 is presenting the results of *Pagerank*. At threshold *0.02* in edge list *Top10%-2*, *Pagerank Topcos* obtained *99.8%* and *Pagerank Topjac 97.7%*. On the other hand, in edge list *Top4%-5*, *Pagerank Topcos* achieved *48.4%* and *Pagerank Topjac* succeeds in getting *13.4%* citation links.

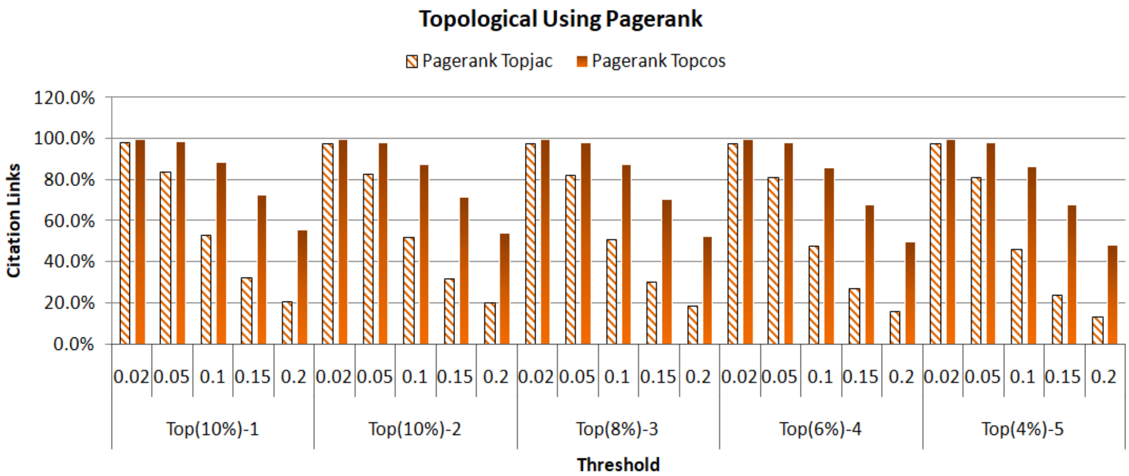


FIGURE 4.14: *Jaccard* similarity and *Cosine* similarity on top nodes using *Pagerank*

In the following Figure 4.15, all centrality measures (Shown in Figures 4.11,4.12,4.13 and 4.14) are combined by taking average of their thresholds from all the edge lists. In this Figure 4.15, at threshold *0.02*, all centrality measures perform well by identifying *99%* citation links. In case of threshold *0.2*, *Betweeness Topcos* obtained *55.4%* which is better than *Closeness Topcos*, *Degree Topcos* and *Pagerank Topcos*. For the same threshold *0.2*, *Jaccard* similarity (*Betweeness Topjac, Closeness Topjac, Degree Topjac* and *Pagerank Topjac*) failed in producing good results. Out of all the centrality measures, *Betweeness* (*Betweeness Topcos* and *Betweeeness Topjac*) performed well in identifying citation links.

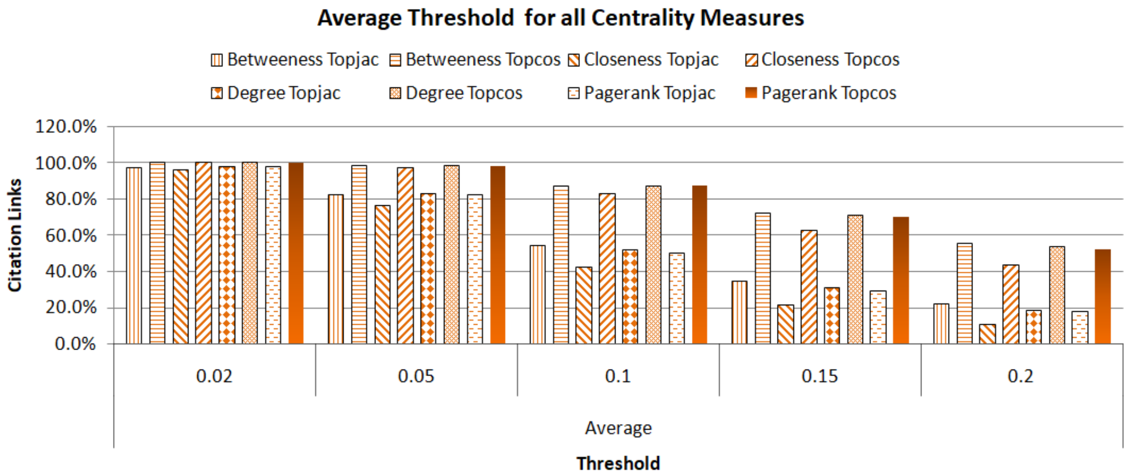


FIGURE 4.15: Average Topological Similarity

4.3.3 Centrality Matrices

In the following Figures (i.e., 4.16,4.17,4.18 and 4.19), results of previous two sections (textual and topological similarity) are combined by centrality measures. The Figure 4.16 is presenting the results of *Betweeness*. At threshold *0.02*, the last two bars (*Betweeness Topcos* and *Betweeness Topjac*) from topological similarity are the competing the textual similarity, where *Betweeness Topcos* succeeds in getting *100%* citation links and *Betweeness Topjac* achieved *97.1%*. Till the last threshold *0.2*, *Betweeness Topcos* retained success strike. Another thing which can be seen at threshold *0.2*, Textual and topological similarity measures are not performed well. In this Figure 4.16, at threshold *0.2*, the lowest result is obtained by *Betweeness Ajac* with *0.6%* citation links. Overall for all the average thresholds, *Betweeness Tjac* succeeds in getting *24.8%* citation links, *Betweeness Tcos 33%*, *Betweeeness Ajac 34%*, *Betweeeness Acos 67.1%*, *Betweeness Topjac 57.8%* and *Betweeenss Topcos 82.4%*.

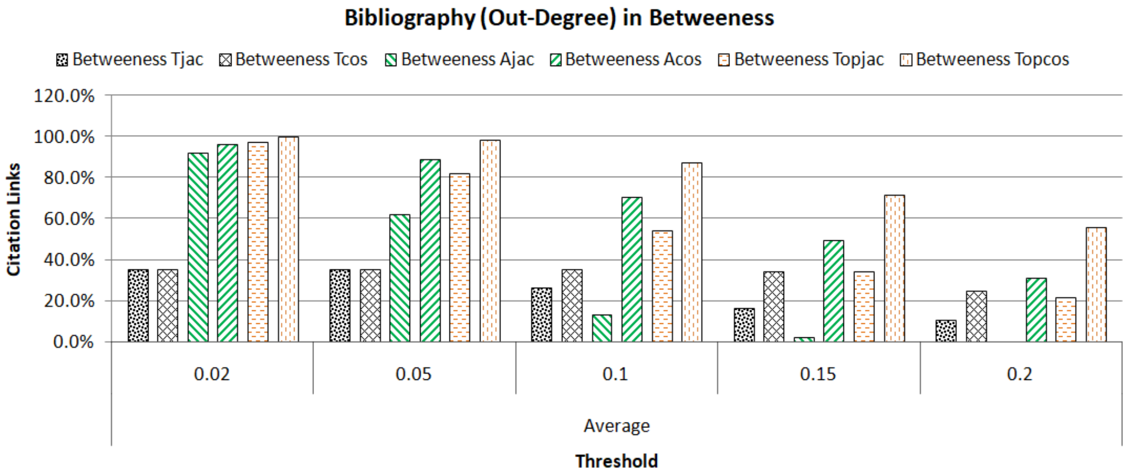


FIGURE 4.16: Textual similarity and Topological similarity on Bibliography(Outdegree Edges) using Betweeness list

The results of *Closeness* are shown in Figure 4.17, which clearly shows that Topological (*Closeness Topcos*) similarity obtained better results than textual (*Closeness Tcos* and *Closeness Acos*) similarity. In case of *jaccard* and *cosine* similarity within toplogical similarity, *Closeness Topcos* competing *Closeness Topjac* for all thresholds. In case of textual similarity using *title* and *abstract*, *Closeness Acos* obtained highest results than *Closeness Tcos*. Maximum citation links at threshold *0.02* achieved by *Closeness Tcos* (using *title*) are *30.7%*, obtained by *Closeness Acos* (using *abstract*) are *96%*, and *achieved* by *Closeness Topcos* (using *topological*) are *99.9%*. Overall for all the average thresholds, *Closeness Tjac* obtained *21.6%* citation links, *Closeness Tcos 28.9%*, *Closeness Ajac 33.2%*, *Closeness Acos 67.1%*, *Closeness Topjac 49%* and *Closeness Topcos 77%*.

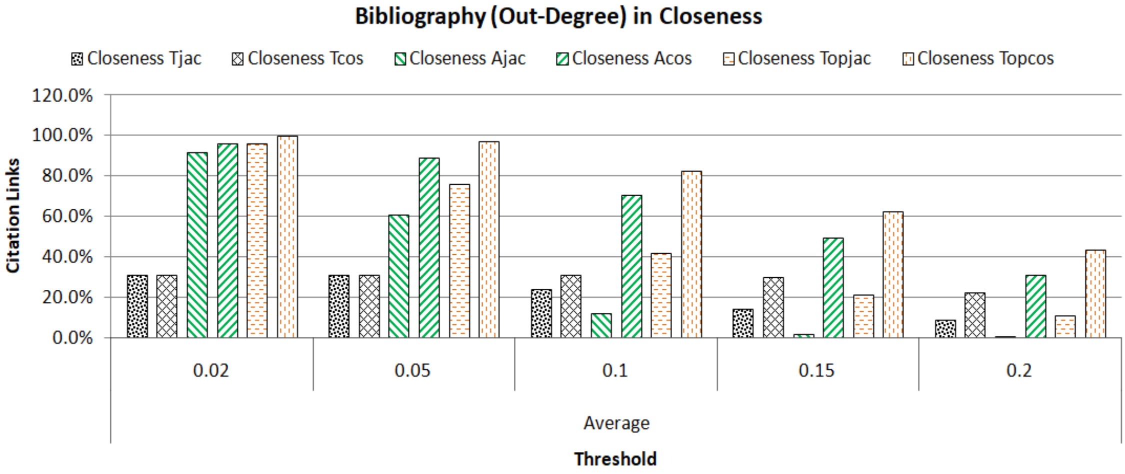


FIGURE 4.17: Textual similarity and Topological similarity on Bibliography(Outdegree Edges) using Closeness list

The Figure 4.18 presenting the results of *Degree*. In case of textual similarity using *title* and *abstract*, at threshold *0.02*, *abstract* (*Degree Acos*) obtained *96.6%* citation links while *title* (*Degree Tcos*) obtained *35.2%*. At the threshold *0.2*, *cosine* (i.e., *Degree Tcos, Degree Acos* and *Degree Topcos*) succeeds in getting *24.7%, 33.5%* and *53.7%*. For the same threshold *0.2*, *Jaccard* (i.e., *Degree Tjac, Degree Ajac* and *Degree Topjac*) achieved *10.1%, 0.4%* and *18.3%*. Toplogical (*Degree Topcos*) similarity outperformed all others at all the thresholds. Overall on all the average thresholds, *Degree Tjac* obtained *24.6%* citation links, *Degree Tcos* achieved *32.9%*, *Degree Ajac* scored *34.7%*, *Degree Acos* succeeds in *68.9%*, *Degree Topjac* obtained *56.1%* and *Degree Topcos* fetched *81.8%*.

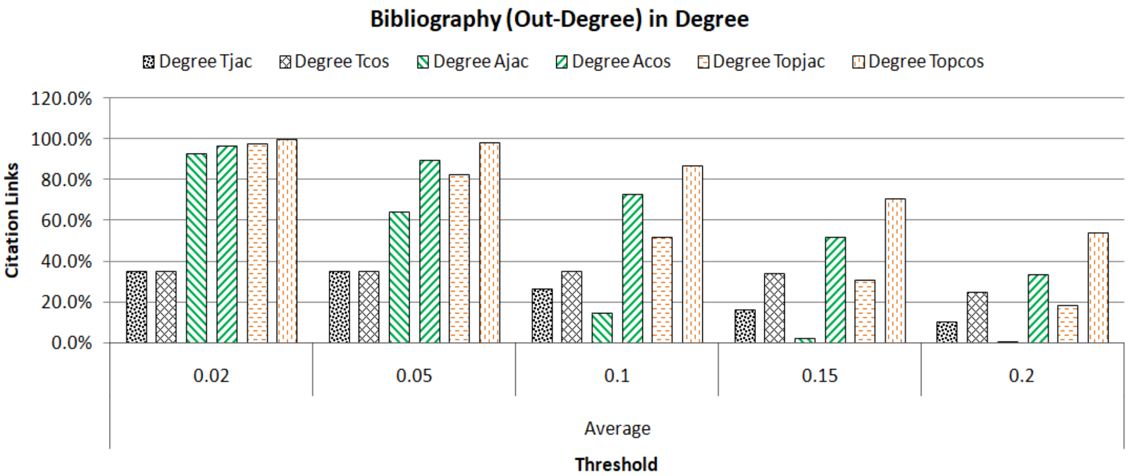


FIGURE 4.18: Textual similarity and Topological similarity on Bibliography(Outdegree Edges) using Degree list

The results of *Pagerank* are shown in Figure 4.19. At the threshold *0.02*, *Pagerank Topcos* and *Pagerank Topjac* from topological similarity are the competing the textual similarity, where *Pagerank Topcos* succeeds in getting *99.9%* citation links and *Pagerank Topjac* achieved *97.7%*. Uptill threshold *0.2*, *Pagerank Topcos* retained success strike. Another thing which can be seen at threshold *0.2*, *jaccard* (i.e., *Pagerank Tjac, Pagerank Ajac* and *Pagerank Topjac*) similarity did not perform well. In this Figure 4.19, at threshold *0.2*, the lowest result is obtained by *Pagerank Ajac* by getting only *0.6%* citation links. Overall for all the average thresholds, *Pagerank Tjac* succeeds in getting *28.3%* citation links, *Pagerank Tcos 37.4%*, *Pagerank Ajac 35.4%*, *Pagerank Acos 68.9%*, *Pagerank Topjac 55.3%* and *Pagerank Topcos 81.6%*.

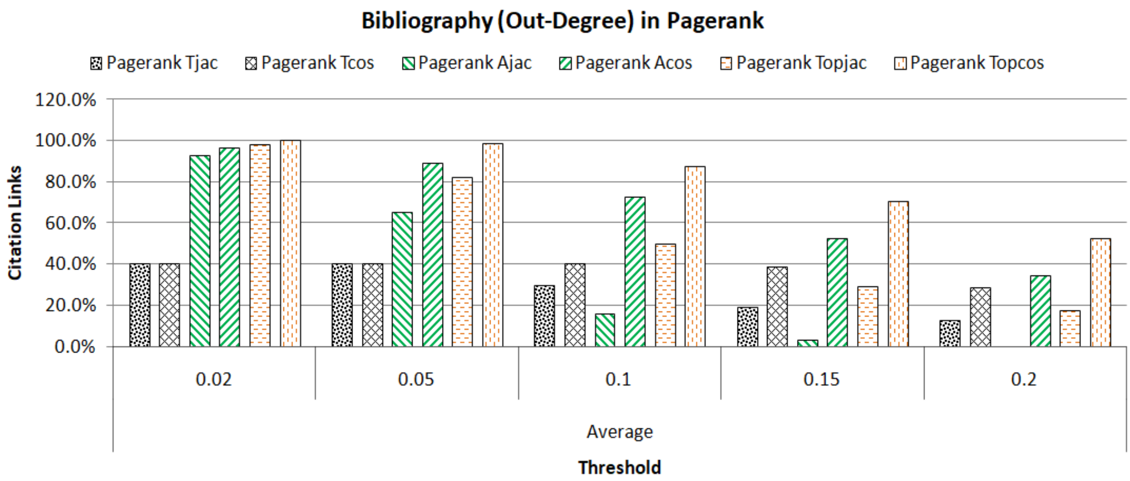


FIGURE 4.19: Textual similarity and Topological similarity on Bibliography(Outdegree Edges) using Pagerank list

4.4 Citation-Based Similarity Computation

As discussed in section 4.2, 200 edge lists (50 for each centrality measure) are computed using *indegree* edges. These *indegree* edges are the citations of the papers in citation graph. In this section, citation-based similarity is computed and results are presented. Moreover, textual and topological similarities are computed and results are presented. First, textual similarity is calculated using *title* and *abstract* of the paper. Then, topological similarity is calculated using neighbor nodes of the paper in citation graph. In the end, both (textual and topological) similarities are evaluated in order to identify the correct citation links.

4.4.1 Textual Similarity

Experiments for the textual similarity are done on *title* and *abstract*, which are used in Section 4.3.1. For the textual similarity, 200 edge lists are used. First of all, textual similarity using *title* is computed and results are presented. Then, *abstract* is used in order to compute textual similarity. In the end, both *title* and *abstract* are evaluated for their performance. Likewise, both similarity measures, *jaccard* and *cosine*, are evaluated.

Title Similarity

For computing textual similarity using *title*, 200 edge lists are used in experiments, where each centrality measure contained 50 edge lists. First of all, *titles* of nodes in edge lists are extracted. After that, *jaccard* and *cosine* similarity measures are performed on *titles* for computing similarity score. After calculating similarity of *titles* using edge lists from *Betweeness*, results are shown in Figure 4.20. For threshold values *0.02* and *0.5* in edge list *Top10%-2*, both *Betweeness Tjac* and *Betweeness Tcos* obtained *39.1%* citation links. At threshold *0.2* in edge list *Top4%-5*, the lowest results achieved by *Betweeness Tjac* and *Betweeness Tcos* are *10.6%* and *24.9%* respectively. Overall, at thresholds (i.e., *0.1*,*0.15* and *0.2*), a big difference between *jaccard* (*Betweeness Tjac*) and *cosine* (*Betweeness Tcos*) similarity can be seen.

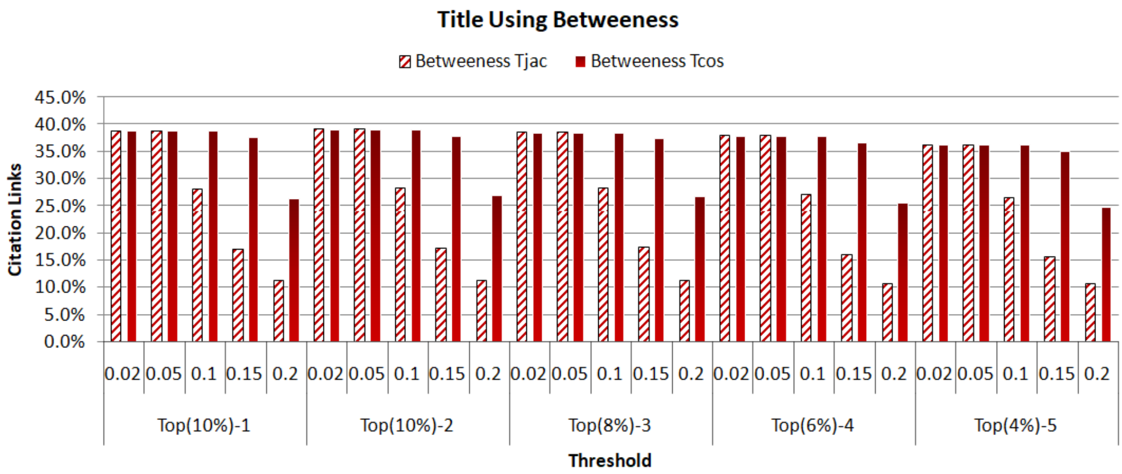


FIGURE 4.20: *Jaccard* similarity and *Cosine* similarity on top nodes using *Betweeness*

The Figure 4.21 shows the results of *Closeness*. In this Figure 4.21, highest result obtained by both *Closeness Tjac* and *Closeness Tcos* is *35.5%*. Likewise, *8%* is the lowest result, which is achieved by *Closeness Tjac* in edge list *Top6%-4*. The main thing which can be seen here is the *cosine* (*Closeness Tcos*) similarity, which outperformed the *jaccard* (*Closeness Tjac*) on different thresholds (i.e., *0.1,0.15* and *0.2*).

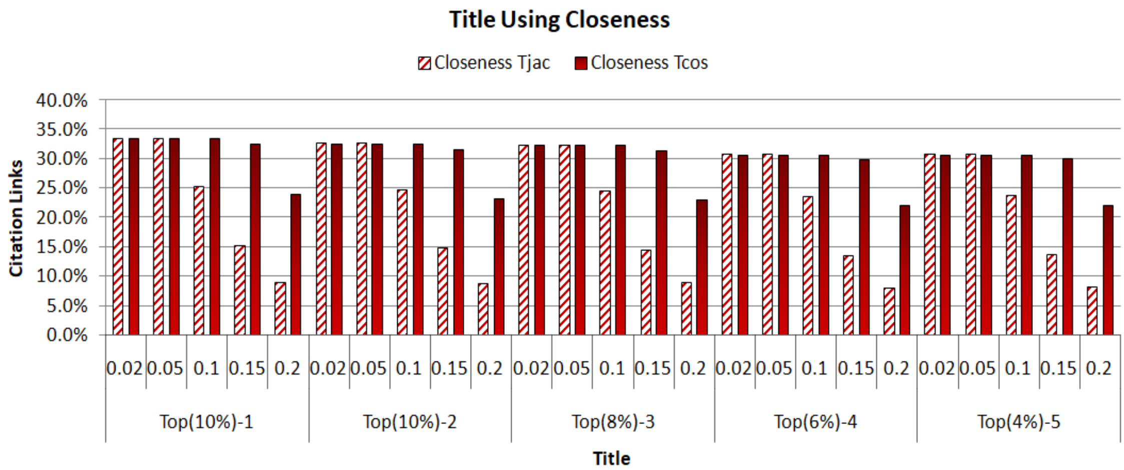


FIGURE 4.21: *Jaccard* similarity and *Cosine* similarity on top nodes using *Closeness*

The results of *Degree* are shown in Figure 4.22. The resultant thresholds shown that both *Degree Tjac* and *Degree Tcos* at threshold *0.02* achieved the highest results with *37.6%*. Same behaviour at threshold *0.05* in the all edge lists shows well identification of citation links. In case of all the edge lists, *cosine* (*Degree Tcos*) similarity performed well with respect to *jaccard* (*Degree Tjac*) similarity.

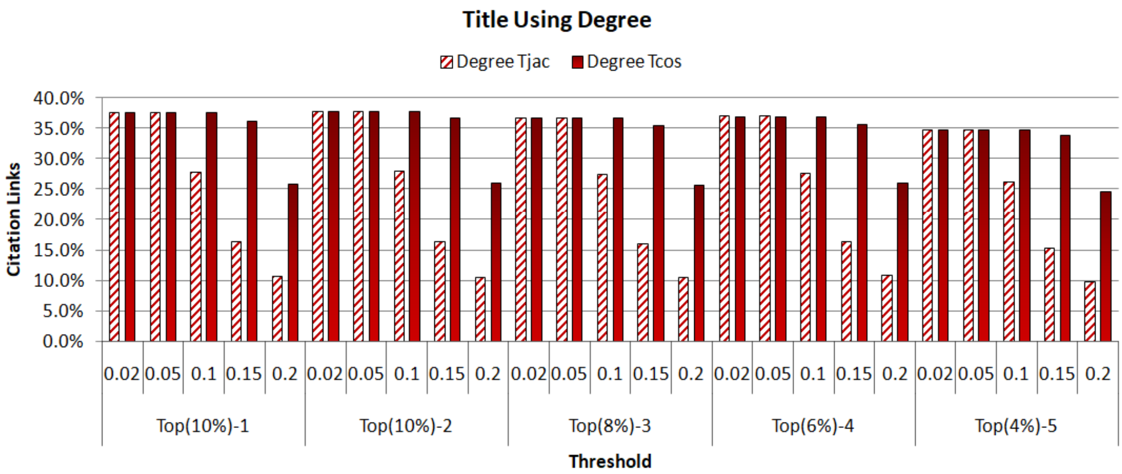


FIGURE 4.22: *Jaccard* similarity and *Cosine* similarity on top nodes using *Degree*

The Figure 4.23 is presenting the results of *Pagerank*. For thresholds *0.02* and *0.05*, both *Pagerank Tjac* and *Pagerank Tcos*, achieved the same results by identifying *38.6%* citation links within edge list *Top10%-1*. Considering increased thresholds values (*0.1*,*0.15* and *0.2*), *jaccard* (*Pagerank Tjac*) similarity decreased. When threshold was *0.2* in edge list *Top4%-5*, *Pagerank Tcos* obtained *26.3%* citation links, and jaccard similarity achieved only *11.1%*.

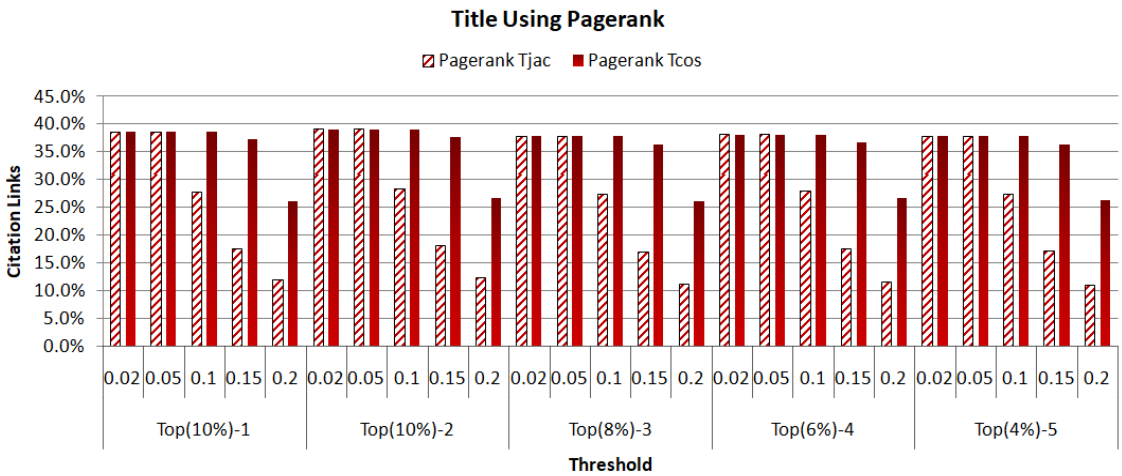


FIGURE 4.23: *Jaccard* similarity and *Cosine* similarity on top nodes using *Pagerank*

In Figure 4.24, all centrality measures results (which are shown in Figures 4.20, 4.21, 4.22 and 4.23) have been combined by taking average of their thresholds from all the edge lists. In this Figure 4.24, at threshold *0.02*, *Pagerank* (*Pagerank Tjac* and *Pagerank Tcos*) obtained highest results with *38.3%*. For the same threshold *0.02*, *Betweeness* (*Betweeness Tjac* and *Betweeness Tcos*) achieved second highest results with *38.1%*. In case of *cosine* (i.e., *Betweeness Tcos, Closeness Tcos, Degree Tcos* and *Pagerank Tcos*) and *jaccard* (i.e., *Betweeness Tjac, Closeness Tjac, Degree Tjac* and *Pagernk Tjac*) similarity, *cosine* similarity outperformed the *jaccard* similarity at thresholds *0.1, 0.15* and *0.2*. Overall for all average thresholds, *Betweeness* (i.e., *Betweeness Tjac* and *Betweeness Tcos*) obtained *26.3%* and *35.5%*, *Closeness* (i.e., *Closeness Tjac* and *Closeness Tcos*) obtained *22.2%* and *30%*, *Degree* (i.e., *Degree Tjac* and *Degree Tcos*) obtained *25.5%* and *34.3%*, and *Pagerank* (i.e., *Pagerank Tjac* and *Pagerank Tcos*) obtained *26.7%* and *35.6%*.

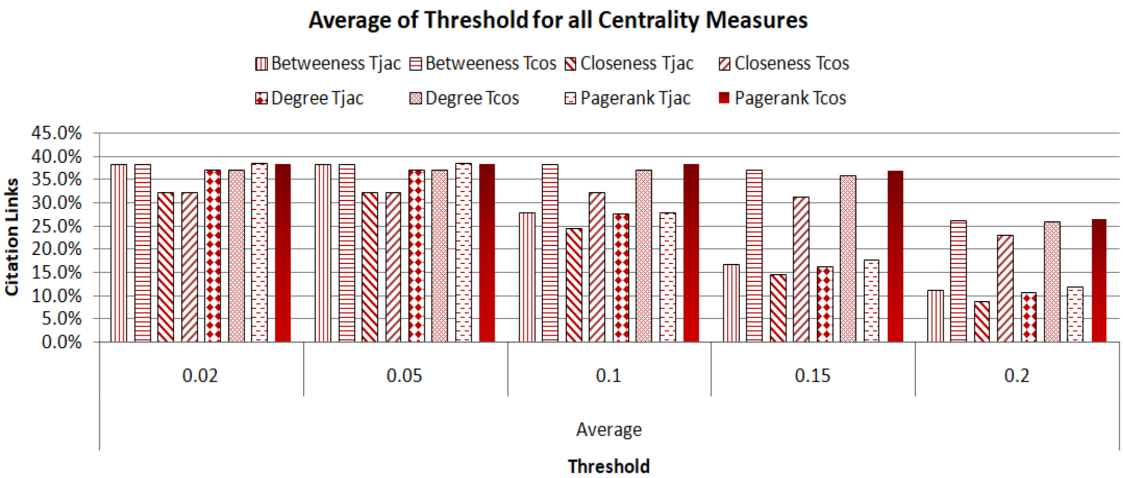


FIGURE 4.24: Average Title Similarity

Abstract Similarity

Abstract similarity is calculated between papers in the edge list. For this purpose, 200 edge lists from citation are used for the experiments. First of all, *abstracts* of nodes in edge list are extracted. After that, for computing similarity, two similarity measures (i.e., *cosine* and *jaccard*) are used. The results of *Betweeness* are shown in Figure 4.25. In this Figure 4.25, textual similarity using *abstract* produced better results than using *title*. At threshold *0.02* in edge list *Top6%-4*, there are almost *97.2%* citation links are identified by *Betweeness Acos*. Here, at threshold *0.1* in edge list *Top10%-1*, *Cosine* (*Betweeness Acos*) similarity present a big difference with *Jaccard* (*Betweeness Ajac*) similarity, where *Betweeness Acos* obtained *72.7%* and *Betweeness Ajac* only *16.6%*.

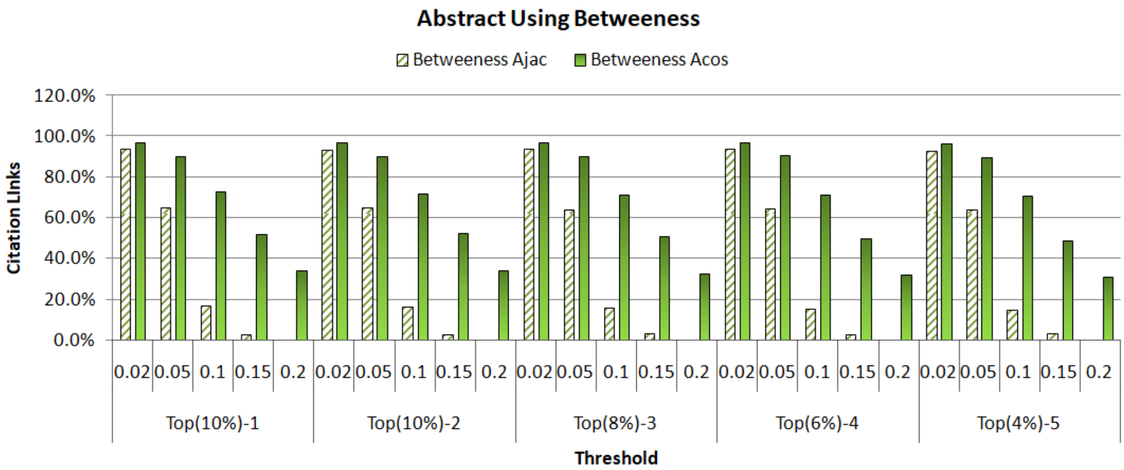


FIGURE 4.25: *Jaccard* similarity and *Cosine* similarity on top nodes usig *Betweeness*

The Figure 4.26 is presenting the results of *Closeness*. In this Figure 4.26, at threshold *0.02* in edge list *Top10%-1*, *Closeness Ajac* obtained *92.7%* and *Closeness Acos* achieved *96.7%*. Likewise, *Closeness Acos* outperformed the *Closeness Ajac* at all the thresholds. At threshold *0.15* in edge list *Top10%-1*, *Closeness Acos* obtained *50.7%* and *Closeness Ajac* achieved *2.3%*.

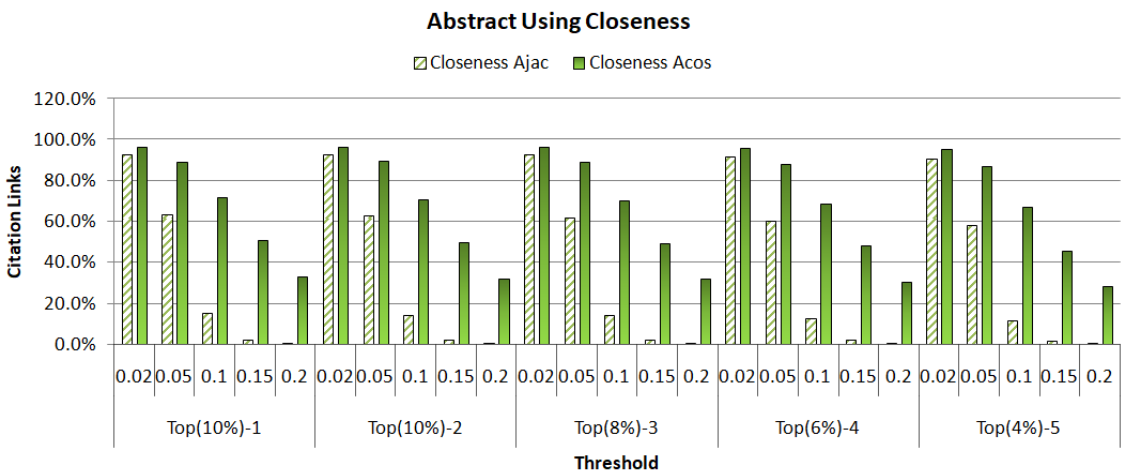


FIGURE 4.26: *Jaccard* similarity and *Cosine* similarity on top nodes using *Closeness*

The results of *Degree* are presented in Figure 4.27. At the threshold *0.02* in edge list *Top10%-1*, *Degree Acos* succeeds in getting *96.7%* citation links and *Degree Ajac* obtained *92.9%*. At threshold *0.1, 0.15* and *0.2*, *Degree Ajac* did not perform well. Overall, *Degree Acos* outperformed the *Degree Ajac*.

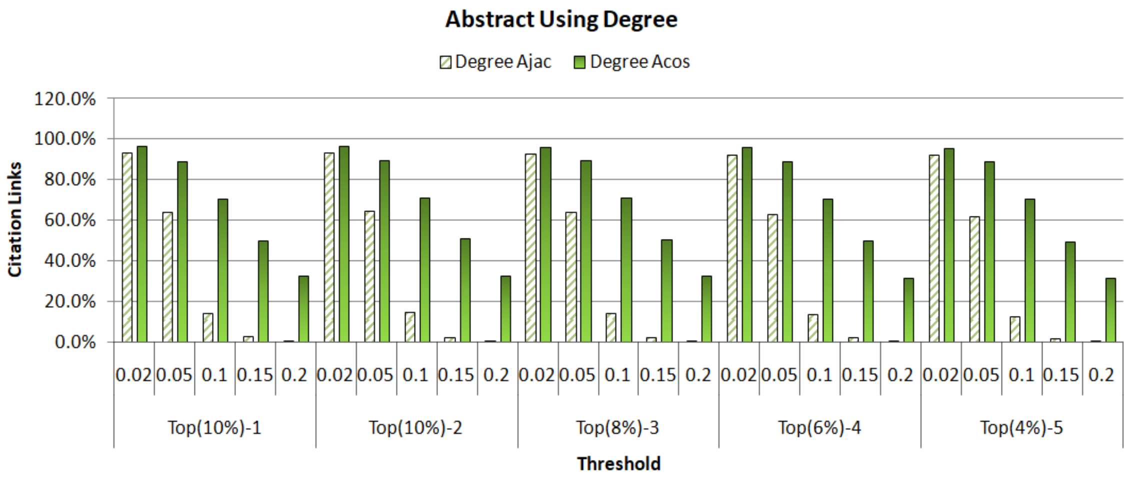


FIGURE 4.27: *Jaccard* similarity and *Cosine* similarity on top nodes using *Degree*

The Figure 4.28 is presenting the results of *Pagerank*. At the threshold *0.02*, *Pagerank Acos* achieved highest result by identifying *95.7%* citation links within edge list *Top10%-1*. At thresholds *0.1,0.15* and *0.2*, as threshold increased, *jaccard* (*Pagerank Ajac*) similarity decreased. when threshold was *0.2* in edge list *Top4%-5*, *Pagerank Acos* obtained *30.6%* citation links, and *Pagerank Ajac* achieved only *0.3%*.

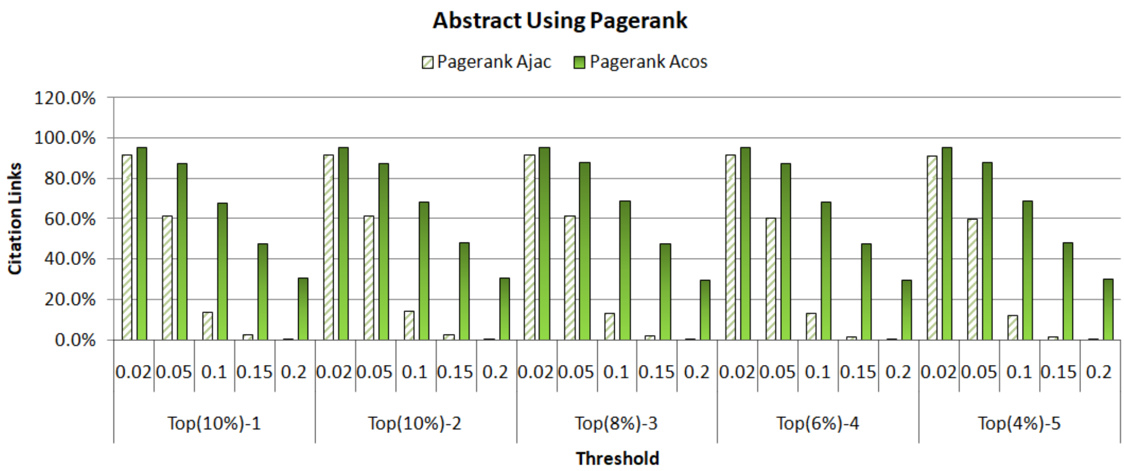


FIGURE 4.28: *Jaccard* similarity and *Cosine* similarity on top nodes using *Pagerank*

The Figure 4.29 is the combination of all the centrality measures. In this Figure 4.29, x-axis represents the average threshold from all the edge lists with respect their centrality measure. This Figure 4.29 shows that at the threshold *0.02*, all the centrality measures produced equally good results. However, moving towards threshold *0.2*, all the centrality measures degrade their results. For all the thresholds, *Betweeness Acos* achieved highest results than others. Over all, *Cosine* (i.e., *Betweenss Acos, Closeness Acos, Degree Acos* and *Pagerank Acos*) similarity outperformed the *jaccard* (i.e., *Betweeness Ajac, Closeness Ajac, Degree Ajac* and *Pagerank Ajac*) similarity. Overall for all the average thresholds, *Betweeness* (i.e., *Betweeness Ajac* and *Betweeness Acos*) obtained *35.3%* and *68.4%*, *Closeness* (i.e., *Closeness Ajac* and *Closeness Acos*) obtained *33.8%* and *66.9%*, *Degree* (i.e., *Degree Ajac* and *Degree Acos*) obtained *34.4%* and *67.8%*, and *Pagerank* (i.e., *Pagerank Ajac* and *Pagerank Acos*) obtained *33.7%* and *66.1%*.

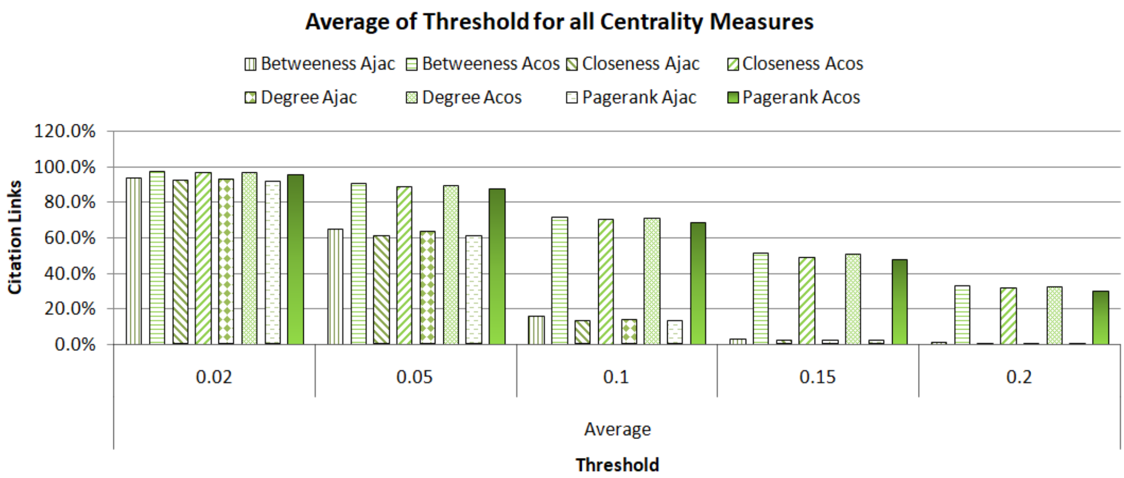


FIGURE 4.29: Average Abstract Similarity

4.4.2 Topological Similarity

For topological similarity, experiments have been performed with one parameter, which is neighbors of the paper. For this purpose, we have used citation graph. In this section, 200 edge lists are picked from citation, where 50 edge lists from each centrality measures. First of all, we picked these edge lists one by one. Then, remove these edges from original graph and made another graph. To infer these removed edges, *cosine* and *jaccard* similarity measures are used. Then, different thresholds are applied on similarity scores. In the end, we have find accuracy score for each edge list.

The Figure 4.30, presenting the results of *Betweeness*. At the threshold *0.02* in edge list *Top10%-1*, *Betweeness Topcos* obtained 100% citation links, while *Betweeness Topjac* achieved 98.4%. When the threshold was *0.2* in edge list *Top4%-5*, *Betweeness Topjac* obtained 23.5% citation links and *Betweeness Topcos* succeeds in getting *58.3%*. For the same threshold *0.2*, topological similarity outperformed the textual similarity (shown in Figures: 4.20,4.21,....,4.29) using *title* and *abstract*.

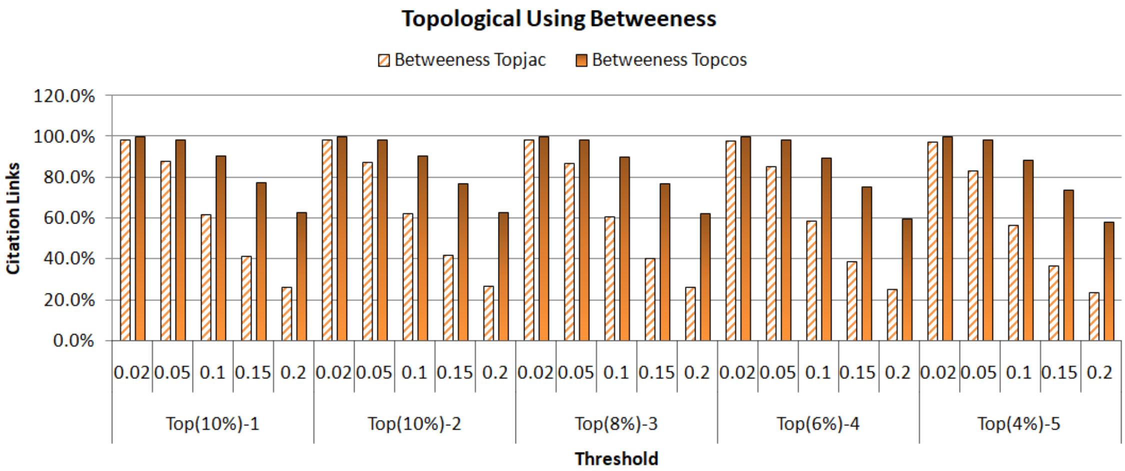


FIGURE 4.30: *Jaccard* similarity and *Cosine* similarity on top nodes using *Betweeness*

In Figure 4.31, results of *Closeness* are shown. In this Figure 4.31, at threshold *0.02* in edge list *Top10%-1*, *Closeness Topjac* obtained *97.6%* citation links and *Closeness Topcos* achieved *100%*. Out of all the edge lists, *Top10%-1 Top10%-2* and *Top8%-3* contributing equally at all the thresholds. In the edge list *Top4%-5*, *Closeness Topjac* and *Closeness Topcos* did not produced good results as they produced in *Betweeness Topjac* and *Betweeness Topcos* (shown in Figure 4.30).

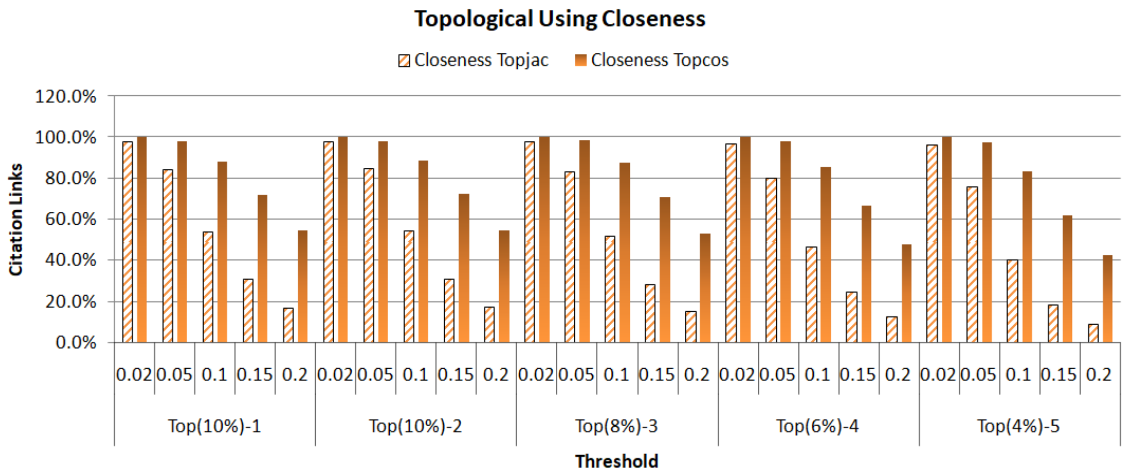


FIGURE 4.31: *Jaccard* similarity and *Cosine* similarity on top nodes using *Closeness*

The results of *Degree* are presenting in Figure 4.32. At the threshold *0.02* in edge list *Top10%-1*, *Degree Topcos* contributed in identifying of *100%* citation links, while *Degree Topjac* obtained *98.3%*. When the threshold was *0.2* in edge list *Top10%-1*, *Degree Topcos* succeed in getting *61%* and *Degree Topjac* obtained only *23.2%*.

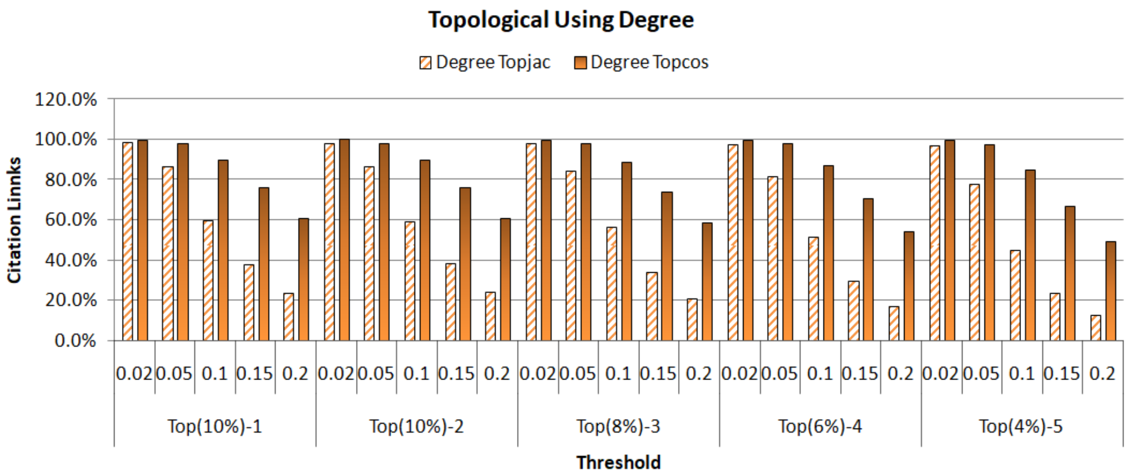


FIGURE 4.32: *Jaccard* similarity and *Cosine* similarity on top nodes using *Degree*

The Figure 4.33 is presenting the results of *Pagerank*. At the threshold *0.02* in edge list *Top10%-2*, *Pagerank Topcos* obtained *100%* and *Pagerank Topjac 98.1%*. On the other hand, at threshold *0.2* in edge list *Top4%-5*, *Pagerank Topcos* achieved *45%* and *Pagerank Topjac* succeeds in getting *10.4%* citation links.

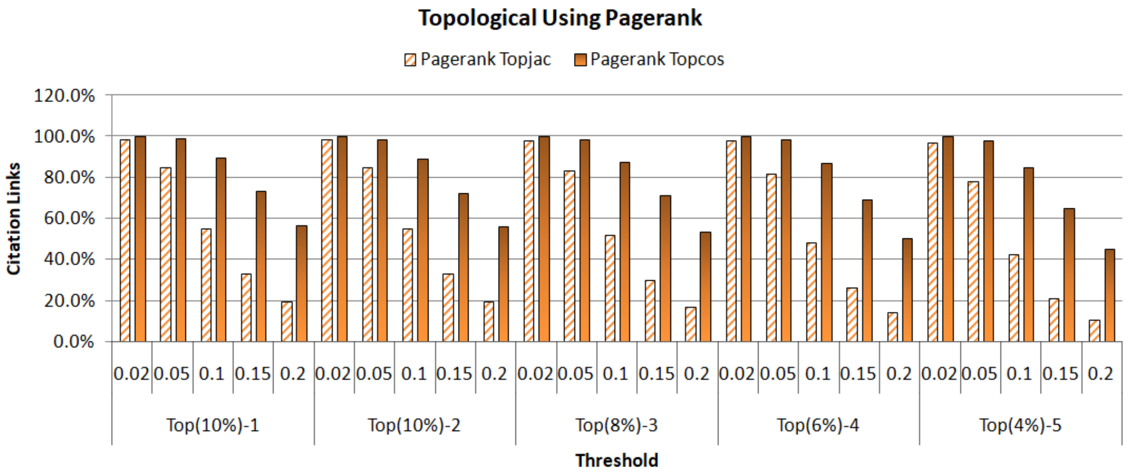


FIGURE 4.33: *Jaccard* similarity and *Cosine* similarity on top nodes using *Pagerank*

In the following Figure 4.34, all centrality measures (shown in Figures 4.30,4.31,4.32 and 4.33) are combined by taking average of their thresholds from all the edge lists. In this Figure 4.34, at threshold *0.02*, all centrality measures perform well by identifying *100%* citation links. In case of threshold *0.2*, *Betweeness Topcos* obtained *61.2%* which is better than *Closeness Topcos*, *Degree Topcos* and *Pagerank Topcos*. For the same threshold *0.2*, *Jaccard* similarity (i.e., *Betweeness Topjac, Closeness Topjac, Degree Topjac* and *Pagerank Topjac*) failed in producing good results. Out of all the centrality measures, *Betweeness* (i.e., *Betweeness Topcos* and *Betweeeness Topjac*) performed well in identifying citation links. Overall for all the average thresholds, *Betweeness* (i.e., *Betweeness Topjac* and *Betweeness Topcos*) obtained *61.9%* and *85.2%*, *Closeness* (i.e., *Closeness Topjac* and *Closeness Topcos*) obtained *53.8%* and *80.7%*, *Degree* (i.e., *Degree Topjac* and *Degree Topcos*) obtained *57.5%* and *83.2%*, and *Pagerank* (i.e., *Pagerank Topjac* and *Pagerank Topcos*) obtained *55%* and *81.7%*.

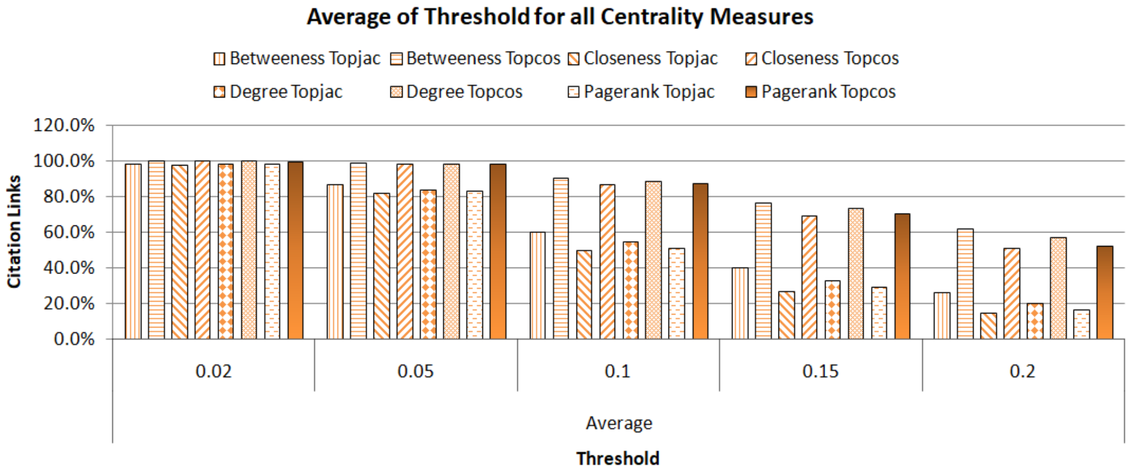


FIGURE 4.34: Average Topological Similarity

4.4.3 Centrality Metrices

In Figures (i.e., 4.35, 4.36, 4.37 and 4.38), results of previous two sections (textual and topological similarity) are combined by centrality measures. The Figure 4.35 contains the results of *Betweeness*. At threshold *0.02*, the last two bars (*Betweeness Topcos* and *Betweeness Topjac*) from topological similarity are competing the textual similarity, where *Betweeness Topcos* obtained *100%* citation links and *Betweeness Topjac* achieved *98%*. Till the threshold *0.2*, *Betweeness Topcos* maintained its success strike. Another thing which can be seen at threshold *0.2*, *Jaccard* similarity (*Betweeness Ajac*) on *abstract* did not perform well. Overall for all the average thresholds, *Betweeness Tjac* obtained *26.3%*, *Betweeness Tcos* achieved *35.5%*, *Betweeness Ajac* obtained *35.3%*, *Betweeness Acos* fetched *68.4%*, *Betweeness Topjac* obtained *61.9%* and *Betweeness Topcos* succeeds in getting *85.2%* citation links.

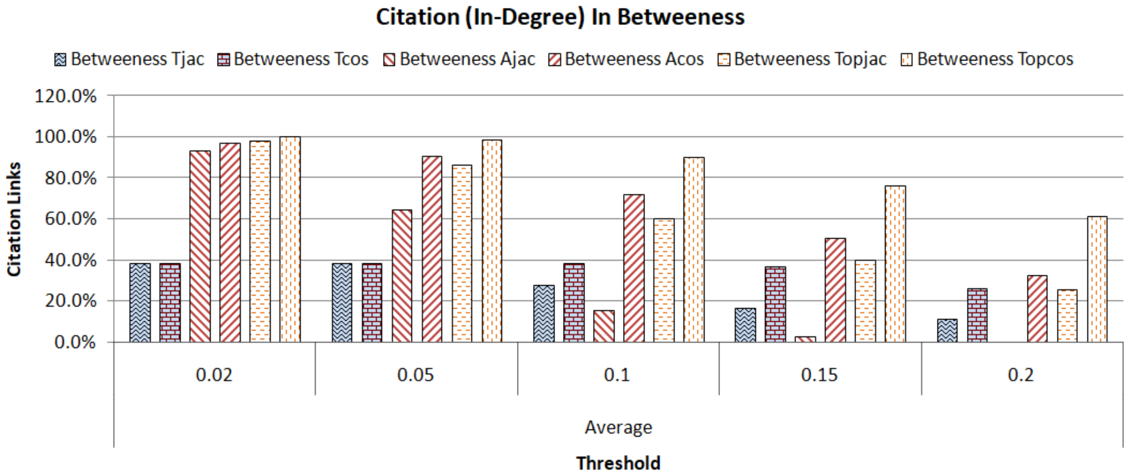


FIGURE 4.35: Textual similarity and Topological similarity on Citation(Indegree Edges) using Betweeness list

The Figure 4.36 is presenting the results of *Closeness*. In this Figure 4.36, it clearly shows that topological similarity (*Closeness Topcos*) obtained better results than textual similarity (*Closeness Tcos* and *Closeness Acos*). In case of *jaccard* and *cosine* within topological similarity, *cosine* (*Closeness Topcos*) produced better results than *jaccard* (*Closeness Topjac*). In case of textual similarity using *title* and *abstract*, *abstract* (*Closeness Acos*) obtained highest results than *title* (*Closeness Tcos*). Maximum number of citation links at threshold *0.02*, obtained by *Closeness Tcos* (using *title*) are *32%*, achieved by *Closeness Acos* (using *abstract*) are *96.2%* and obtained by *Closeness Topcos* (using *topological*) are *100%*. Overall for all the average thresholds, *Betweeness Tjac* succeeds in *22.2%*, *Betweeness Tcos* obtained *30%*, *Betweeness Ajac* achieved *33.8%*, *Betweeness Acos* obtained *66.9%*, *Betweeness Topjac* obtained *53.8%* and *Betweeness Topcos* achieved *80.7%*.

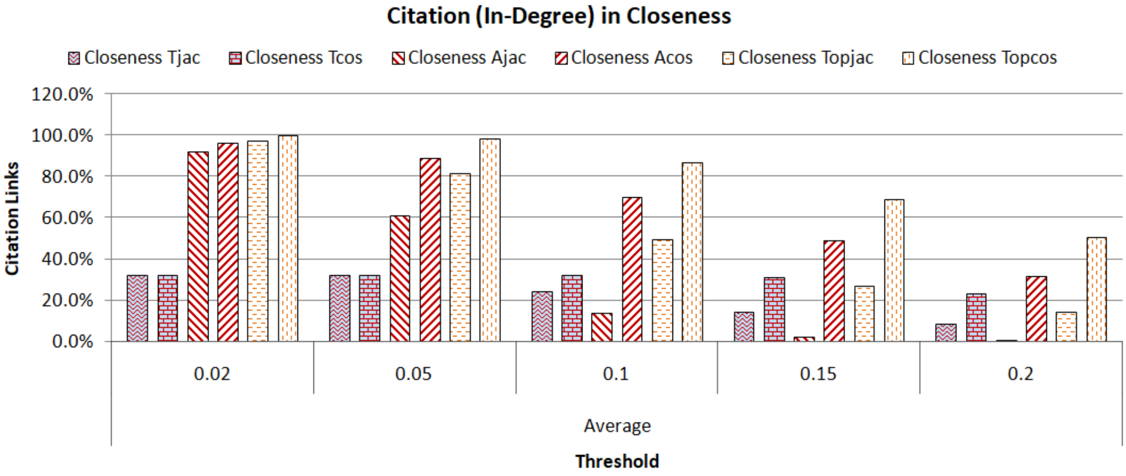


FIGURE 4.36: Textual similarity and Topological similarity on Citation(Indegree Edges) using Closeness list

The Figure 4.37 presenting the results of *Degree*. In case of textual similarity using *title* and *abstract*, at threshold *0.02*, *abstract* (*Degree Acos*) obtained *96.3%* citation links while *title* (*Degree Tcos*) obtained *36.8%*. At the threshold *0.2*, *cosine* (i.e., *Degree Tcos, Degree Acos* and *Degree Topcos*) succeeds in getting *25.7%, 32.3%* and *56.9%*. For the same threshold *0.2*, *Jaccard* (i.e., *Degree Tjac, Degree Ajac* and *Degree Topjac*) achieved *10.5%, 0.4%* and *19.5%*. Toplogical (*Degree Topcos*) similarity outperformed all others at all the thresholds. Overall for all the average thresholds, *Degree Tjac* obtained *25.5%* citation links, *Degree Tcos* achieved *34.3%*, *Degree Ajac* scored *34.4%*, *Degree Acos* succeeds in *67.8%*, *Degree Topjac* obtained *57.5%* and *Degree Topcos* fetched *83.2%*.

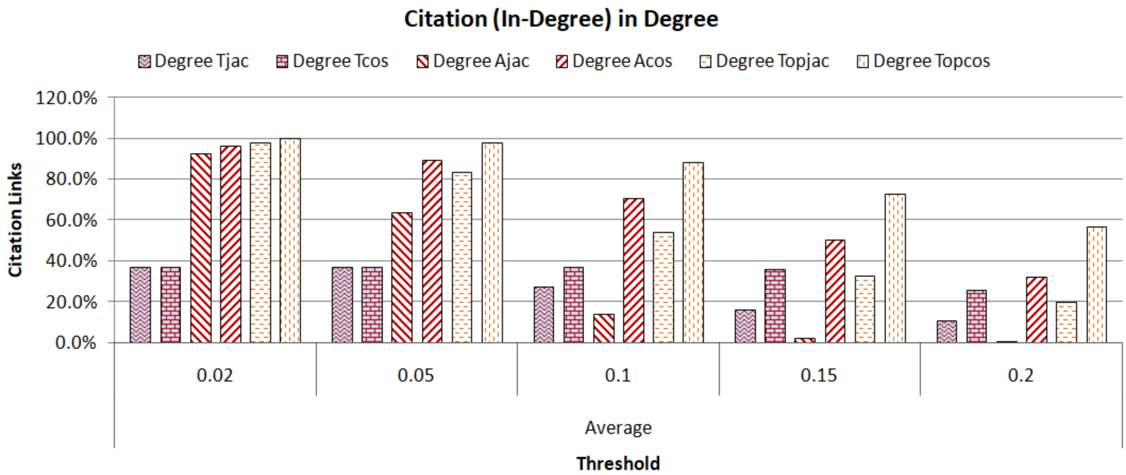


FIGURE 4.37: Textual similarity and Topological similarity on Citation(Indegree Edges) using Degree list

The results of *Pagerank* are shown in Figure 4.38. At the threshold *0.02*, *Pagerank Topcos* and *Pagerank Topjac* from topological similarity are competing the textual similarity, where *Pagerank Topcos* succeeds in getting *100%* citation links and *Pagerank Topjac* achieved *97.7%*. Till threshold *0.2*, *Pagerank Topcos* retained success strike. Another thing which can be seen at threshold *0.2*, *jaccard* (*Pagerank Tjac*, *Pagerank Ajac* and *Pagerank Topjac*) similarity did not perform well. In Figure 4.38, at threshold *0.2*, the lowest result is obtained by *Pagerank Ajac* by getting only *0.4%* citation links. Overall for all the average thresholds, *Pagerank Tjac* succeeds in getting *26.7%* citation links, *Pagerank Tcos 35.6%*, *Pagerank Ajac 33.7%*, *Pagerank Acos 66.1%*, *Pagerank Topjac 55%* and *Pagerank Topcos 81.7%*.

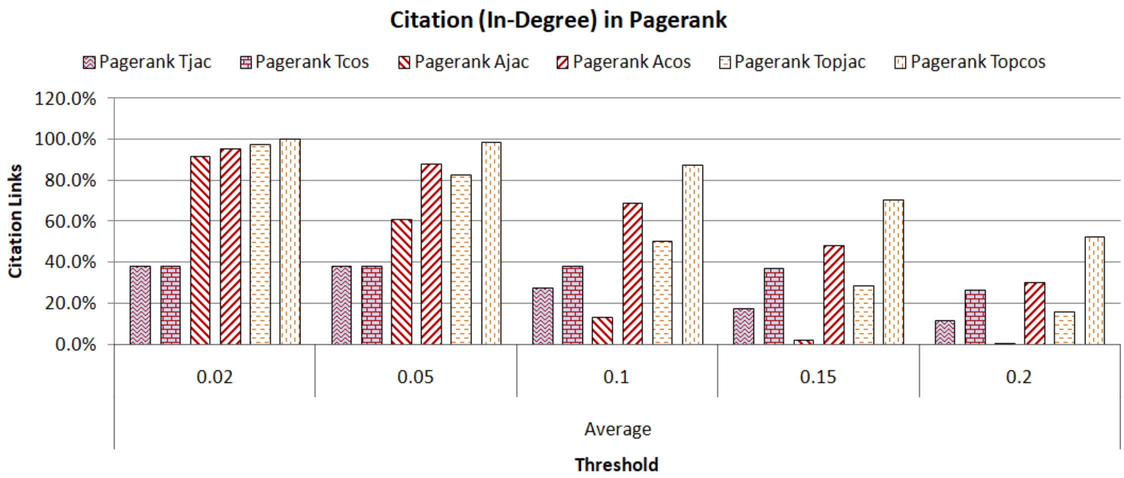


FIGURE 4.38: Textual similarity and Topological similarity on Citation(Indegree Edges) using Pagerank list

4.5 Evaluation

4.5.1 Bibliography vs Citation

In this thesis, experimentation is done on 400 edge lists of 5 different kinds, where 200 edge lists belongs to citation and 200 are of bibliography. In this section, performance of citation and bibliography are evaluated by giving answer of the following questions.

Q: Which aspect of citation analysis (Citation and Bibliography) is more suitable in identification of citation links ?

The answer of this question is results are shown in Figures 4.39, 4.40, 4.41 and 4.42.

- **Textual similarity(using *title*):** In case of bibliography, *Tcos* succeeds in getting 35.6% citation links, while *Tjac* obatined 26.7% (shown in Figure 4.42). On the other hand, in case of citation , highest results achieved by *Tcos* are 37.4%, and obtained by *Tjac* are 28.3%(shown in Figure 4.42). In case of textual similarity using *title*, bibliography is better option than citation.

- **Textual similarity (using *abstract*):** In case of bibliography, *Acos* achieved maximum of 68.4% citation links, while *Ajac* obtained 35.3% (shown in Figure 4.39). Likewise, in case of citation, *Acos* obtained 68.9% citation links, and *Ajac* achieved 35.4% (shown in Figure 4.42). In case of textual similarity using *abstract*, citation produced better results than bibliography. Overall, textual similarity produced better results through bibliography.

- **Topological Similarity:** In all the Figures (i.e., 4.39, 4.40, 4.41 and 4.42), *Topcos* and *Topjac* performed well through bibliography. The highest results obtained by *Topcos*, through bibliography are 85.2%, and through citation are 82.4% (shown in Figure 4.39).

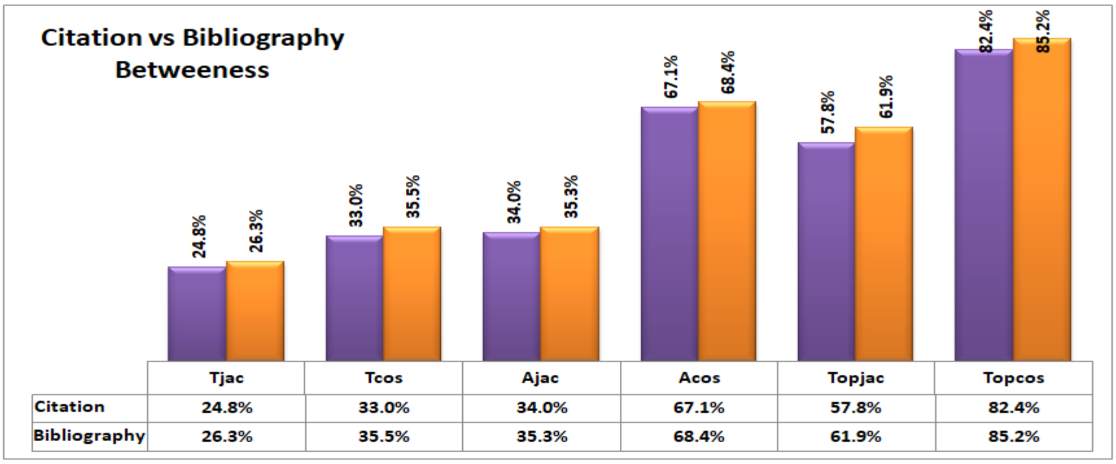


FIGURE 4.39: Comparison Between Citation and Bibliography Through Betweeness

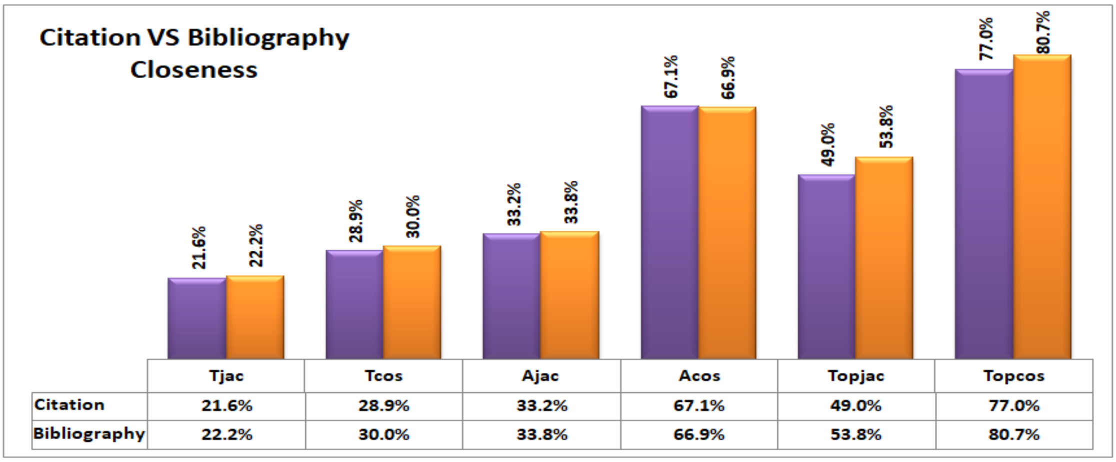


FIGURE 4.40: Comparison Between Citation and Bibliography Through Closeness

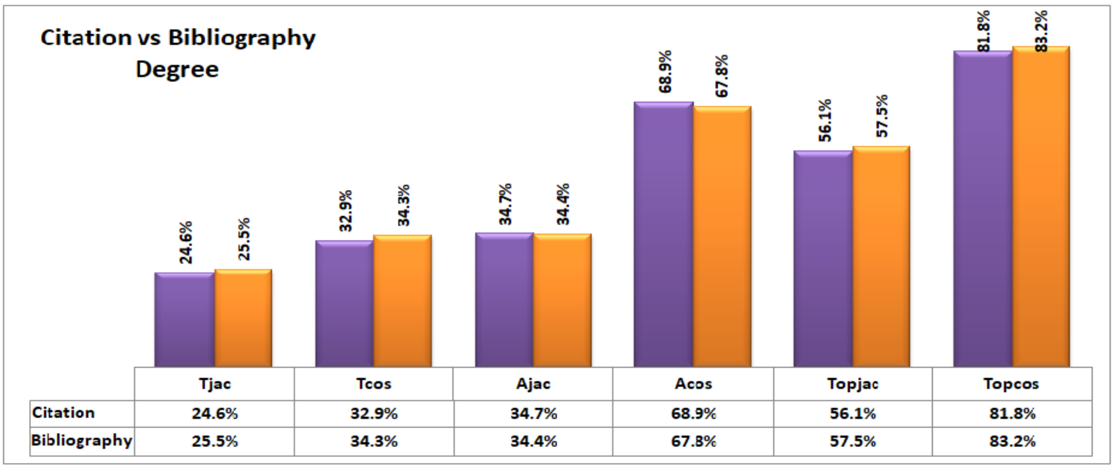
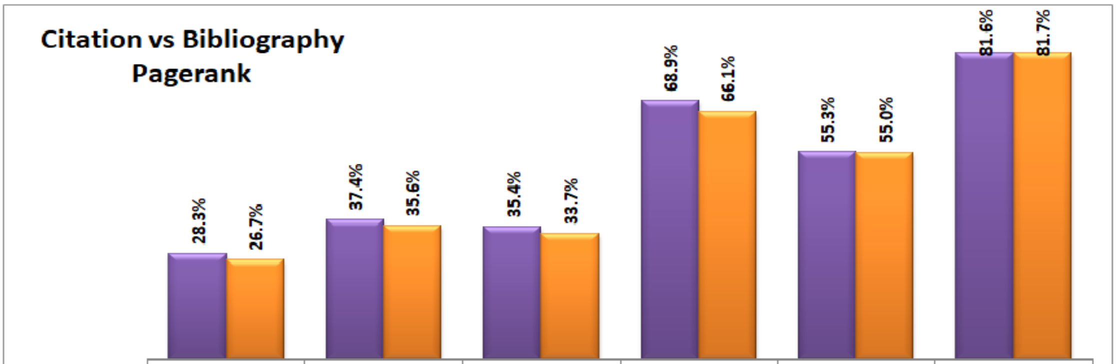


FIGURE 4.41: Comparison Between Citation and Bibliography Through Degree



	Tjac	Tcos	Ajac	Acos	Topjac	Topcos
Citation	28.3%	37.4%	35.4%	68.9%	55.3%	81.6%
Bibliography	26.7%	35.6%	33.7%	66.1%	55.0%	81.7%

FIGURE 4.42: Comparison Between Citation and Bibliography Through Pagerank

4.5.2 Textual Similarity vs Topological Similarity

Q: Which aspect (*Title*,*Abstract*) accurately identifies citation links for textual similarity?

In Figures 4.39, 4.40, 4.41 and 4.42, It clearly shows that textual similarity using *abstract* (*Acos* and *Ajac*) outperformed the textual similarity using *title* (*Tcos* and *Tjac*). The maximum result obtained by *Acos* is *68.9%* (Figure 4.41), and achvied by *Ajac* is *35.4%* (Figure 4.42). Likewise, *Tcos* succeeds in getting *37.4%* (Figure 4.42), and *Tjac* obtained *28.3%* (Figure 4.42). It clearly shows that textual similarity using *abstract* produced better results than textual similarity using *title*.

Q: Are topological similarity measures better than textual similarity measures to predict a citation link ?

- **Topological Similarity:** *Topcos* produced better results than *Tcos* and *Acos* by obtaining *85.2%* (shown in Figure 4.39). Likewise, *Topjac* competing with *Tjac* and *Ajac* by scoring *61.9%* (see Figure 4.39). In this way, topological similarity measures performed better than textual similarity measures.

- **Textual Similarity:** *Tjac* and *Ajac* failed in getting highest results than *Topjac* by getting *28.3%* and *35.4%* (see Figure 4.42). Likewise, *Tcos* and

Acos also did not perform well, *Tcos* obtained *37.4%* and *Acos* achieved *68.9%* (Figure 4.42).

The main point which can be seen here is the big difference between textual and topological similarity measures. In case of *jaccard*, *Tjac* and *Ajac* produced low results than *Topjac*. While, in case of *cosine*, *Topcos* outperformed than *Tcos* and *Acos*.

4.5.3 Cosine Similarity vs Jaccard Similarity

Q: How accurate are textual similarity measures (*Jaccard*, *Cosine*) for correct identification of citation link ?

- **Textual Similarity(using title):** *Cosine* (Tcos) similarity perform better than *Jaccard* (*Tjac*) similarity by obtaining 37.4% citation links, while *Jaccard* (*Tjac*) obtained 28.3%(shown in Figure 4.42).
- **Textual Similarity (using abstract):** *Cosine* (*Acos*) similarity obtained 68.9%, while *Jaccard* (*Ajac*) similarity achieved 35.4% (see Figure 4.42).

In this way, *Cosine* similarity outperformed than *Jaccard* similarity.
Q: How accurate are topological similarity measures (*Jaccard*, *Cosine*) for correct identification of citation link ?

- **Topological Similarity:** In case of topological similarity, *Cosine* (*Topcos*) similarity performed better than *jaccard* (*Topjac*) similarity. The maximum result obtained by *Topcos* is 85.2%, while achieved by *Topjac* is 61.9%(shown in Figure4.39).

It is clearly show that, *Cosine* similarity produced better results than *Jaccard*.

4.5.4 Betweeness vs Closeness vs Degree vs Pagerank

Q: Which centrality measure (Betweeness, Closeness, Degree and Pagerank) is more accurate in identification of citation links ?

- **Textual similarity (using title):** The highest results using *title* are obtained through Pagerank, where *Tcos* obtained 37.4% and *Tjac* obtained 28.3% (see Figure 4.42). Likewise, lowest results are obtained through Closeness, where *Tcos* obtained 28.9% and *Tjac* obtained 21.6% (shown in Figure 4.40). Therefore, textual similarity using title produced better results through Pagerank than other centrality measures.

- **Textual similarity (using abstract):** In case of textual similarity using *abstract*, Pagerank outperformed the other centrality measures. In Figure 4.42 of Pagerank, *Acos* succeeds in getting 68.9% citation links, and *Ajac* obtained 35.4%. Again, Closeness did not perform well in case of abstract.

- **Topological similarity:** Here, in case of topological similarity, Betweeness produced better results than other centrality measures. Through Betweeness, *Topcos* succeeds in getting 85.2% citation links, and *Topjac* obtained 61.9%. It is clear that Betweeness centrality is better option for topological similarity than other centrality measures.

4.6 Comparisons

In this section, comparisons are performed with *Bo et.al* [29]. They have proposed technique to recommend citations for non-profile users using cosine similarity on short queries and long queries. They have considered titles of papers as short queries and abstracts as long queries. For comparisons 8 different edge lists are used, where 4 from In-Degree edges and 4 from Out-Degree. Moreover, citation links are identified. Our comparisons results are shown in Figures(4.43,4.44,...4.50). In these Figures, Y-axis represents the percentage of identified citation links while X-axis shows the threshold. In case of in-degree edges, our approach using *Topological* and *Abstract* similarity performed well than *Bo et.al*. In Figure 4.43, at threshold *0.02*, *Topological* obtained *100%* citation links, *Abstract* achieved *97.6%* and *Bo et.al* succeed in getting *93.3%*. Although, *Title* similarity did not perform well, but overall proposed approach perform well. The same behaviour is shown in Figure 4.44, again proposed approach is succeed in getting high citation links. Considering thresholds *0.1,0.15 and 0.2*, clearly show that *Topological* and *Abstract* similarity identified high citation links than Bo et.al. At threshold *0.1*, *Topological* similarity obtained *87.8%*, *Abstract* similarity achieved *71.3%* and *Bo et.al* obtained only *50.5%*. Although, *Title* similarity did not succeed in getting high citation links, but competed well with *Bo et.al* at thresholds *0.15 and 0.02*. Furthermore, Figures 4.45 and 4.46 showing the same behaviours, where proposed approach performed well than *Bo et.al*.

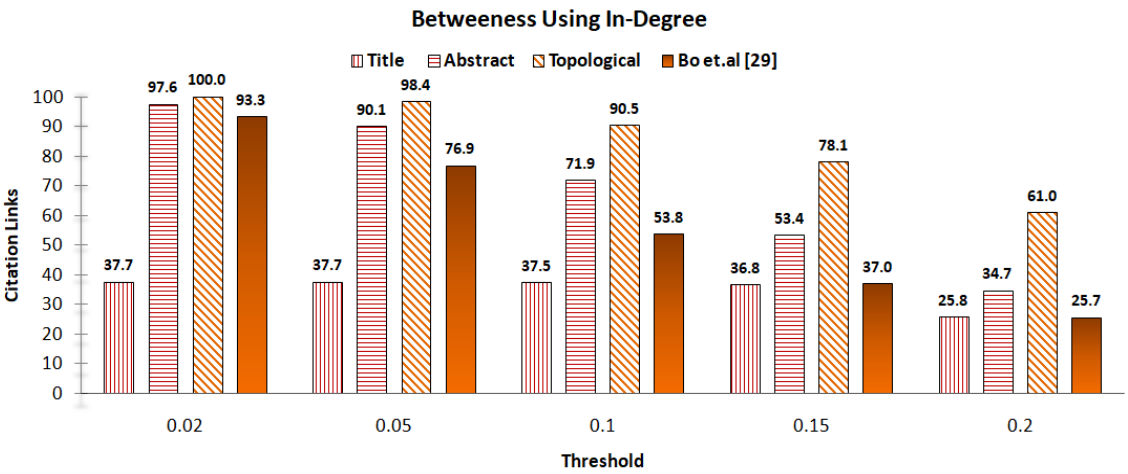


FIGURE 4.43: Comparisons with Bo et.al using Betweeness

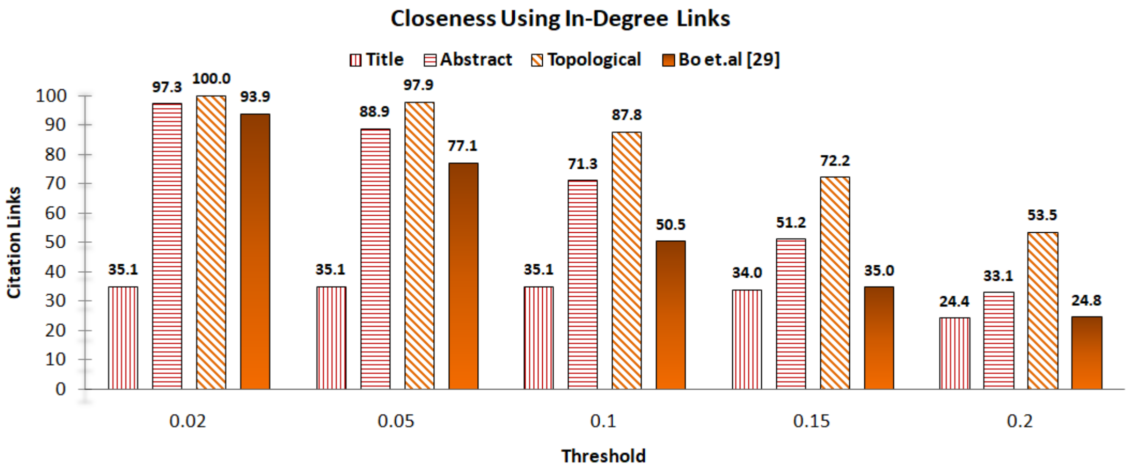


FIGURE 4.44: Comparisons with Bo et.al using Closeness

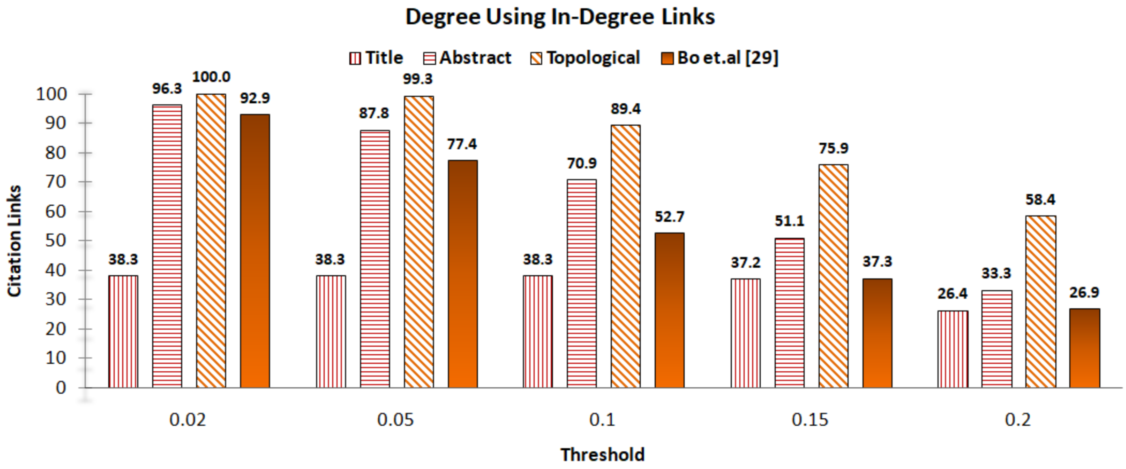


FIGURE 4.45: Comparison with Bo et.al using Degree

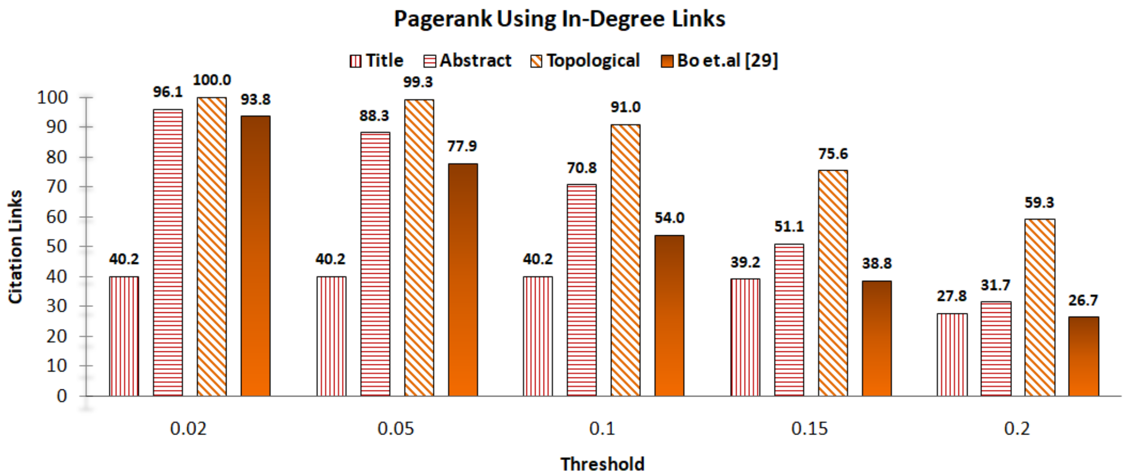


FIGURE 4.46: Comparison with Bo et.al using Pagerank

In case of out-degree edges, proposed approach again performed well than Bo et.al. Overall, *Topological* similarity at threshold *0.1*, in Figure 4.47 obtained *87.3%*, in Figure 4.48 achieved *83.9%*, in Figure 4.49 obtained *86.6%* and in Figure 4.50 obtained *87.8%* citation links. Similarly, *Abstract* obtained *73.3%,69.9%,72.1% and 72%* citation links. On the other hand, at threshold *0.1*, *Bo et.al* obtained *52.7%,48.8%,50.2% and 55.6%* citation links.

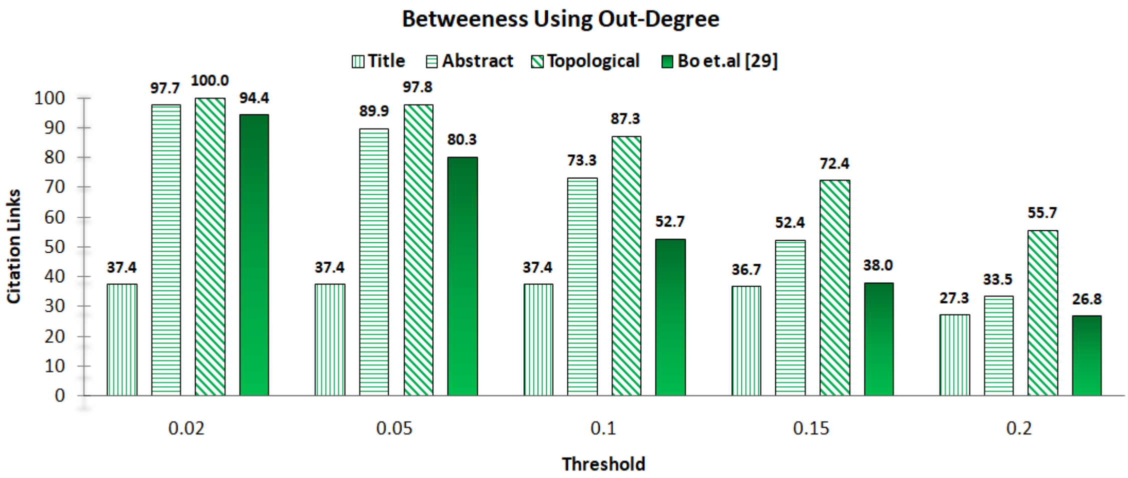


FIGURE 4.47: Comparison with Bo et.al using Betweeness

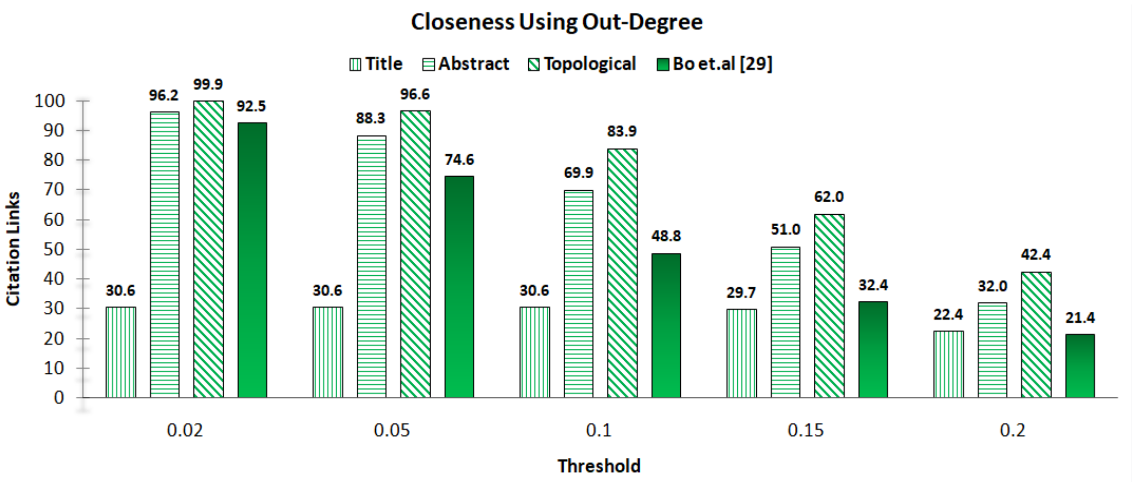


FIGURE 4.48: Comparison with Bo et.al using Closeness

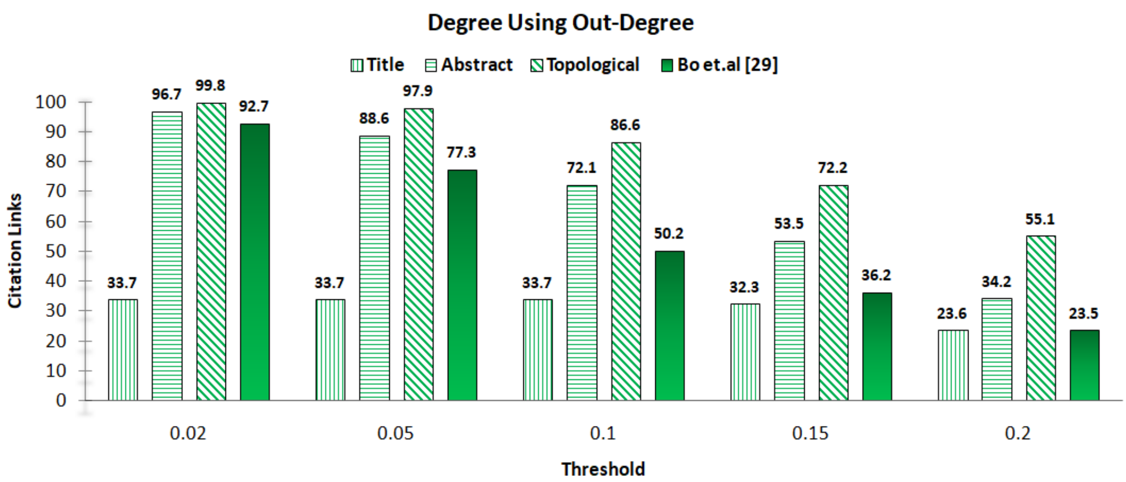


FIGURE 4.49: Comparison with Bo et.al using Degree

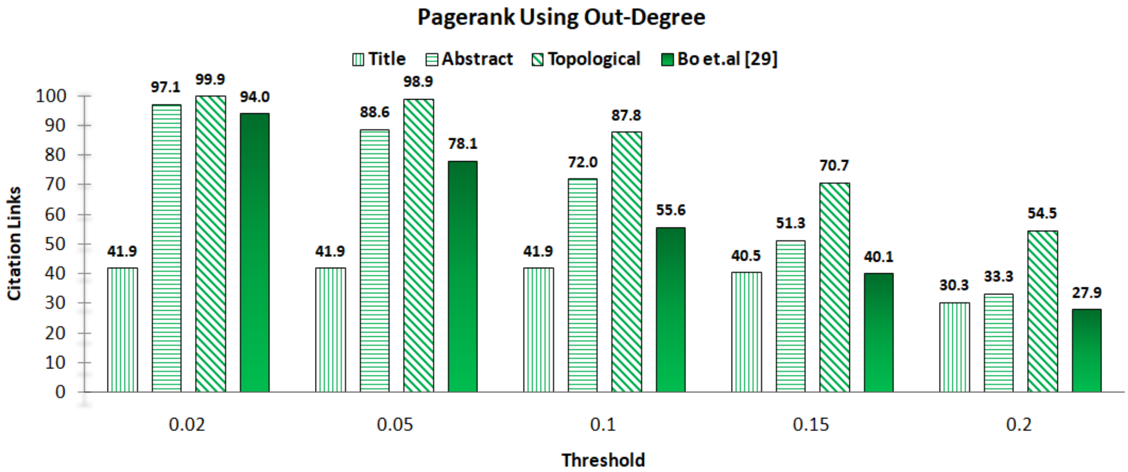


FIGURE 4.50: Comparison with Bo et.al using Pagerank

Chapter 5

Conclusion and Future Work

The number of research publications are increased and created a problem to search the required and relevant research papers. Moreover, it became difficult for the researchers to keep up-to-date with the new research ideas and the previous research. Recommender systems facilitated researchers to keep in touch with the current as well as previous research. Furthermore, such systems provided help to researchers in finding the topic of their interest. Research paper recommendation systems takes input (i.e. research articles, words, and sentences) from the researchers and processed to provide related documents. The recommendation systems worked with set of approaches which are used to match the researcher query with documents. Most of the citation recommender systems recommended similar papers on textual basis.

This thesis have evaluated textual and topological-based similarity measures for citation recommendation. Moreover, centrality metrics are used to find the influential papers for recommendation. The experimentation setup contains dataset of 8179 (with two textual parameter *title* and *abstract*) papers with citation graph (contain 1,43,906 edges). Graph centrality measures are applied on citation graph to choose the top (i.e., *top10%, top8%,top6%* and *top4%*) research papers.

First, we applied textual and topological similarity measures and analyzed that topological based similarity outperformed the textual-based similarity. The experimental results shows that for the citation recommendation, topological-based similarity is better as compared to textual-based similarity. Where *Topcos* (Toplogcial *cosine*) obtained *85.2%* citation links and *Topjac* (Topological *jaccard*) obtained *61.9%*. Likewise, *Tcos* (textual *cosine*) obtained *37.4%* and *Tjac* (textual *jaccard*) achieved *28.3%*. Secondly, the results of *cosine* and *jaccard* similarity are analyzed, where *cosine* competed *jaccard* similarity with highest score. Third, evaluate the centrality measures to check which centrality measures is best to find the influential papers. In case of textual-based similarity, the highest results were obtained through *Pagerank*, while for topological similarity *Betweeness* is the better options. Finally, results from *citation* (indegree) and *bibliography* (outdegree) are analyzed. In case of textual-based similarity using *title*, similarity measures performed best on *bibliography* (outdegree). In case of textual based similarity using *abstract*, similarity measures achieved best results through *citation* (indegree). However, in case of topological-based similarity, results from *bibliography* were best. The overall finding of this thesis is that, Topological-based similarity is better option for finding and recommending similarity papers and on the other hand, importance of paper should be considered in citation recommendation.

5.1 Future Work

In this study, two similarity measures are used, which are cosine and jaccard. Both similarity measures are used the "**Symmetric**" relationship of papers for finding the similarity of two papers. In some environments, such as social network, one sided similarity should be computed by using "**Asymmetric**" relationship instead of "**Symmetric**". This could be the best thing for identification of link in social network. Second future direction could be the use of "**Multi-Attribute Decision Making(MADM)**" to find the top ranked influential paper for citation. Where, ranked papers should be categorized in most important, important, less important and not important.

Bibliography

[1] S. Bethard and D. Jurafsky, “Who should i cite: learning literature search models from citation behavior,” in *Proceedings of the 19th ACM international conference on Information and knowledge management.* ACM, 2010, pp. 609–618.

[2] P. Lops, M. De Gemmis, and G. Semeraro, “Content-based recommender systems: State of the art and trends,” in *Recommender systems handbook.* Springer, 2011, pp. 73–105.

[3] M. Deshpande and G. Karypis, “Item-based top-n recommendation algorithms,” *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 143–177, 2004.

[4] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, “Recommending and evaluating choices in a virtual community of use,” in *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM Press/Addison-Wesley Publishing Co., 1995, pp. 194–201.

[5] H. Small, “Co-citation in the scientific literature: A new measure of the relationship between two documents,” *Journal of the American Society for information Science*, vol. 24, no. 4, pp. 265–269, 1973.

[6] E. Rich, “User modeling via stereotypes,” *Cognitive science*, vol. 3, no. 4, pp. 329–354, 1979.

[7] K. D. Bollacker, S. Lawrence, and C. L. Giles, “Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications,”

in *Proceedings of the second international conference on Autonomous agents.* ACM, 1998, pp. 116–123.

[8] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2016.

[9] J. Wang, H. Mo, F. Wang, and F. Jin, "Exploring the network structure and nodal centrality of chinas air transport network: A complex network approach," *Journal of Transport Geography*, vol. 19, no. 4, pp. 712–721, 2011.

[10] D. Bertsimas, E. Brynjolfsson, S. Reichman, and J. Silberholz, "Moneyball for academics: Network analysis for predicting research impact," 2014.

[11] M. M. Kessler, "Bibliographic coupling between scientific papers," *American documentation*, vol. 14, no. 1, pp. 10–25, 1963.

[12] T. Strohman, W. B. Croft, and D. Jensen, "Recommending citations for academic papers," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2007, pp. 705–706.

[13] E. Meij and M. De Rijke, "Using prior information derived from citations in literature search," in *Large scale semantic access to content (text, image, video, and sound)*, 2007, pp. 665–670.

[14] L. Egghe and R. Rousseau, *Introduction to informetrics: Quantitative methods in library, documentation and information science.* Elsevier Science Publishers, 1990.

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

[16] W. Huang, Z. Wu, P. Mitra, and C. L. Giles, "Refseer: A citation recommendation system," in *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on.* IEEE, 2014, pp. 371–374.

[17] C. Scholz, M. Atzmueller, and G. Stumme, "On the predictability of human contacts: Influence factors and the strength of stronger ties," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom).* IEEE, 2012, pp. 312–321.

[18] F. Isinkaye, Y. Folajimi, and B. Ojokoh, "Recommendation systems: Principles, methods and evaluation," *Egyptian Informatics Journal*, vol. 16, no. 3, pp. 261–273, 2015.

[19] J. Beel, S. Langer, M. Genzmehr, and A. Nürnberger, "Introducing docear's research paper recommender system," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries.* ACM, 2013, pp. 459–460.

[20] F. Ferrara, N. Pudota, and C. Tasso, "A keyphrase-based paper recommender system," in *Italian Research Conference on Digital Libraries.* Springer, 2011, pp. 14–25.

[21] S. M. McNee, N. Kapoor, and J. A. Konstan, "Don't look stupid: avoiding pitfalls when recommending research papers," in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work.* ACM, 2006, pp. 171–180.

[22] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles, "Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach," in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence.* Morgan Kaufmann Publishers Inc., 2000, pp. 473–480.

[23] C. Yang, B. Wei, J. Wu, Y. Zhang, and L. Zhang, "Cares: a ranking-oriented cadal recommender system," in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries.* ACM, 2009, pp. 203–212.

[24] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in *The adaptive web.* Springer, 2007, pp. 291–324.

[25] A. Adams, K. Milland, S. Savage, C. Callison-Burch, J. Bigham *et al.*, "A data-driven analysis of workers' earnings on amazon mechanical turk," 2018.

[26] J. Beel, "Towards effective research-paper recommender systems and user modeling based on mind maps," *arXiv preprint arXiv:1703.09109*, 2017.

[27] J. Beel, B. Gipp, and E. Wilde, "Academic search engine optimization (aseo) optimizing scholarly literature for google scholar & co." *Journal of scholarly publishing*, vol. 41, no. 2, pp. 176–190, 2009.

[28] L. Steinert and H. U. Hoppe, "A comparative analysis of network-based similarity measures for scientific paper recommendations," in *Network Intelligence Conference (ENIC), 2016 Third European.* IEEE, 2016, pp. 17–24.

[29] D. Hanyurwimfura, L. Bo, V. Havyarimana, D. Njagi, and F. Kagorora, "An effective academic research papers recommendation for non-profiled users," *International Journal of Hybrid Information Technology*, vol. 8, no. 3, pp. 255–272, 2015.

[30] H. Xue, J. Guo, Y. Lan, and L. Cao, "Personalized paper recommendation in online social scholar system," in *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* IEEE Press, 2014, pp. 612–619.

[31] S. Philip, P. Shola, and A. Ovye, "Application of content-based approach in research paper recommendation system for a digital library," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 10, 2014.

[32] T. Huynh, K. Hoang, L. Do, H. Tran, H. Luong, and S. Gauch, "Scientific publication recommendations based on collaborative citation networks," in *2012 International Conference on Collaboration Technologies and Systems (CTS).* IEEE, 2012, pp. 316–321.

[33] N. Shibata, Y. Kajikawa, and I. Sakata, "Link prediction in citation networks," *Journal of the American society for information science and technology*, vol. 63, no. 1, pp. 78–85, 2012.

[34] R. Dong, L. Tokarchuk, and A. Ma, "Digging friendship: paper recommendation in social network," in *Proceedings of Networking & Electronic Commerce Research Conference (NAEC 2009)*, 2009, pp. 21–28.

[35] B. Bulut, B. Kaya, R. Alhajj, and M. Kaya, "A paper recommendation system based on user's research interests," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 911–915.

[36] H. Sahijwani and S. Dasgupta, "User profile based research paper recommendation," *arXiv preprint arXiv:1704.07757*, 2017.

[37] L. Yang, Y. Zheng, X. Cai, H. Dai, D. Mu, L. Guo, and T. Dai, "A lstm based model for personalized context-aware citation recommendation," *IEEE Access*, vol. 6, pp. 59 618–59 627, 2018.

[38] Y. Zhang, L. Yang, X. Cai, and H. Dai, "A novel personalized citation recommendation approach based on gan," in *International Symposium on Methodologies for Intelligent Systems*. Springer, 2018, pp. 268–278.

[39] X. Cai, J. Han, W. Li, R. Zhang, S. Pan, and L. Yang, "A three-layered mutually reinforced model for personalized citation recommendation," *IEEE transactions on neural networks and learning systems*, no. 99, pp. 1–12, 2018.

[40] J. Son and S. B. Kim, "Academic paper recommender system using multilevel simultaneous citation networks," *Decision Support Systems*, vol. 105, pp. 24–33, 2018.

[41] X. Cai, Y. Zheng, L. Yang, T. Dai, and L. Guo, "Bibliographic network representation based personalized citation recommendation," *IEEE Access*, vol. 7, pp. 457–467, 2019.

[42] L. Yang, Z. Zhang, X. Cai, and L. Guo, "Citation recommendation as edge prediction in heterogeneous bibliographic network: A network representation approach," *IEEE Access,vol.8,pp.382-395*, 2019.

[43] X. Cai, J. Han, and L. Yang, "Generative adversarial network based heterogeneous bibliographic network representation for personalized citation recommendation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[44] F. Ayala-Gómez, B. Daróczy, A. Benczúr, M. Mathioudakis, and A. Gionis, "Global citation recommendation using knowledge graphs," *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 5, pp. 3089–3100, 2018.

[45] H. Jia and E. Saule, "Graph embedding for citation recommendation," *arXiv preprint arXiv:1812.03835*, 2018.

[46] L. Yang, Y. Zheng, X. Cai, S. Pan, and T. Dai, "Query-oriented citation recommendation based on network correlation," *Journal of Intelligent & Fuzzy Systems*, no. Preprint, pp. vol.4,1–8, 2018.

[47] T. Dai, L. Zhu, Y. Wang, H. Zhang, X. Cai, and Y. Zheng, "Joint model feature regression and topic learning for global citation recommendation," *IEEE Access*, vol. 7, pp. 1706–1720, 2019.

[48] D. Mu, L. Guo, X. Cai, and F. Hao, "Query-focused personalized citation recommendation with mutually reinforced ranking," *IEEE Access*, vol. 6, pp. 3107–3119, 2018.

[49] J. D. West, I. Wesley-Smith, and C. T. Bergstrom, "A recommendation system based on hierarchical clustering of an article-level citation network," *IEEE Transactions on Big Data*, vol. 2, no. 2, pp. 113–123, 2016.

[50] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 177–187.

[51] J. Gehrke, P. Ginsparg, and J. Kleinberg, "Overview of the 2003 kdd cup," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 149–151, 2003.

[52] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal, Complex Systems*, vol. 1695, no. 5, pp. 1–9, 2006.

[53] L. C. Freeman, S. P. Borgatti, and D. R. White, "Centrality in valued graphs: A measure of betweenness based on network flow," 1991.

[54] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13–18, 2013.

[55] A. Samad, M. A. Islam, M. A. Iqbal, M. Aleem, and J. U. Arshed, "Evaluation of features for social contact prediction," in *2017 13th International Conference on Emerging Technologies (ICET)*. IEEE, 2017, pp. vol.6,1–6.

[56] H. Calvo, O. Méndez, and M. A. Moreno-Armendáriz, "Integrated concept blending with vector space models," *Computer Speech & Language*, vol. 40, pp. 79–96, 2016.