CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Impact of Field of Study Trend on Citation Count of Scientific Articles and Authors

by

Lubna Zafar

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

in the

Faculty of Computing

Department of Computer Science

2023

Impact of Field of Study Trend on Citation Count of Scientific Articles and Authors

By

Lubna Zafar

(DCS161004)

Dr. Muhammad Younis, Reader
Oxford Brookes University, Oxford, UK
(Foreign Evaluator 1)

Dr. Suhuai Luo, Associate Professor
The University of Newcastle, Australia
(Foreign Evaluator 2)

Dr. Nayyer Masood
(Dissertation Supervisor)

Dr. Abdul Basit Siddiqui
(Head, Department of Computer Science)

Dr. Muhammad Abdul Qadir
(Dean, Faculty of Computing)

DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
ISLAMABAD
2023

Dedicated to my parents & siblings.

CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD

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

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the dissertation, entitled "**Impact of Field of Study Trend on Citation Count of Scientific Articles and Authors**" was conducted under the supervision of **Dr. Nayyer Masood**. No part of this dissertation has been submitted anywhere else for any other degree. This dissertation is submitted to the **Department of Computer Science, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Computer Science.** The open defence of the dissertation was conducted on **July 17, 2023**.

Student Name : Lubna Zafar (DCS161004)

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

Examination Committee :

(a) External Examiner 1: Dr. Manzoor Ilahi Tamimy,
Professor
COMSATS University, Islamabad

(b) External Examiner 2: Dr. Rabeeh Ayaz Abbasi,
Associate Professor
QAU, Islamabad

(c) Internal Examiner : Dr. Umair Rafique,
Assistant Professor
CUST, Islamabad

Supervisor Name : Dr. Nayyer Masood
Professor
CUST, Islamabad

Name of HoD : Dr. Abdul Basit Siddiqui
Associate Professor
CUST, Islamabad

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

AUTHOR'S DECLARATION

I, **Lubna Zafar (Registration No. DCS161004)**, hereby state that my dissertation titled, '**Impact of Field of Study Trend on Citation Count of Scientific Articles and Authors**' is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

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

(**Lubna Zafar**)

Dated: 17 July, 2023

Registration No: DCS161004

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the dissertation titled "**Impact of Field of Study Trend on Citation Count of Scientific Articles and Authors**" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete dissertation 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 dissertation declare that no portion of my dissertation 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 dissertation even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD degree and that HEC and the University have the right to publish my name on the HEC/ University Website on which names of students are placed who submitted plagiarized dissertation.

(**Lubna Zafar**)

Dated: 17 July, 2023

Registration No: DCS161004

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 dissertation:-

1. **L. Zafar** and N. Masood, ”Impact of Field of Study Trend on Scientific Articles,” IEEE Access, vol. 8, p. 128295-128307, 2020.
2. **L. Zafar**, N. Masood, and S. Ayaz, ”Impact of Field of Study (FoS) on Authors’ Citation Trend,” Scientometrics, 2022.

(Lubna Zafar)

Registration No: DCS161004

Acknowledgement

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

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

I must also acknowledge the help that I got from **Dr. Arshad Islam** who was always prepared to guide me in achieving the knowledge that was required to complete this project. I have to offer my gratitude to **Dr. Muhammad Abdul Qadir** and all my teachers who have always been a guiding light for me in achieving the knowledge and completing the project. I also want to thank Ms. Fakeeha Jafari, Ms. Samreen Ayaz, Ms. Rabia Mumtaz and Mr. Abdul Samad who were always helpful to me.

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

(Lubna Zafar)

Abstract

Millions of new scientific articles are published each year. Researchers work and publish in their respective fields of interest. A major portion of the scientific community publishing in the same Field of Study (FoS) forms a trend of that field. A novice researcher chooses his field of interest based upon its popularity. This may have a positive impact on the acceptance of a study or high count of citations in future. There are multiple studies in literature that focus on FoS trend detection and analysis, birth and establishment of an FoS trend, number of publications and researchers in an FoS trend, communities of researchers being formed around an FoS trend, author's FoS switching, vanishing of an FoS trend, trends in different disciplines etc. However, the previous work contains a gap, that is, there is no work on impact of following an FoS on citation trend of scientific articles and authors.

This study identifies how significant it is to follow an FoS trend and the impact of the FoS trend on research paper citations and on authors citation count. For this purpose, we have chosen the field of Computer Science and Microsoft Academic Graph (MAG) dataset from the 1950-2018 time period. We extracted publications of different FoS of Computer Science and also citation counts for these publications. First, we established similarity between citation trends of papers belonging to same FoS using rand index and correlation. Then we proposed a technique to identify trend setters and trend followers that would help to identify influential authors in a particular FoS. Finally, we established the impact of FoS on the citation patterns of authors by achieving a consistent R2 values of papers belonging to same FoS. The results depict that if papers belong to the same FoS, then there are 69% of the chances of having a similar citation pattern and that they have the same citation trend as they also have achieved a high correlation value. Experimental results show that there is a similarity between citation trend of authors that belong to the same FoS as compared to different FoS and achieved consistent R2 value. FoS trend following has a certain impact on the citation count of authors. The result also shows that if an author publishes in a particular FoS, then the citation trend of this author's work resembles more to the overall citation trend of that particular

FoS than that of some other FoS. This proves that FoS has a certain impact on the citation count of a paper and researchers should contemplate on the FoS trend before selecting a particular research area

Contents

List of Figures

List of Tables

Abbreviations

ACM	Association of Computational Machinery
ANN	Artificial Neural Network
FoM	Field of Study Multigraph
FoS	Field of Study
IDT	Information of Diffusion Theories
MAG	Microsoft Academic Graph
MLR	Multiple Linear Regression
RDF	Resource Description Framework
RI	Rand Index
SS	Semantic Search
SVM	Support Vector Machine

Chapter 1

Introduction

The volume and diversity of scientific literature are escalating each day. According to study, almost 2.5 million new scientific papers are published every year [1]. The research environment changes and evolves continuously and as a result new research fields emerge while some other fade out. The new research fields that emerge generally form a trend in research. A trend is the general direction or field of study (FoS) a research area is taking during a specified period of time. An FoS is defined as an area that is increasing in importance and effectiveness over time [2]. For example, it can be said that the Artificial Intelligence is an FoS that currently has trend in Computer Science research area. Trend in a research area grabs the attention of large number of researchers. Moreover, it generally has applicability in different domains. For example, p2p network and semantic search are two FoS that emerged around 2002, however, semantic search is still a trendy FoS as we see many publications in this FoS, on the other side, p2p network is not that trendy and number of publications and applications are quite less relatively.

Keeping up with the state of research is one way is to stay informed with the trendy FoS. Not only for a new researcher but also for established or seasoned ones, having knowledge of prior, recent, and emerging FoS patterns is important. For instance, a researcher might choose to conduct research in a FoS that hasn't received much attention. It can also be helpful to a businessman who is attempting to weigh the risks of funding a start-up.

Identifying FoS trends is important because it allows researchers to choose their areas of interest based on their success or impact. Similar to this, it is crucial to identify researchers who were active during the early stages of a FoS trend because this will reveal the key individuals who contributed to or kicked off the trend's rise to fame. The ability to recognize FoS trends is noteworthy for anyone involved in the research environment, including researchers, academic publishers, journal editors, institutional funding bodies and other relevant stakeholders.

Is this really necessary to follow FoS trend? Or researchers work in traditional way of research and stick with that? Sticking to very traditional means of research can be very damaging as other researchers can take the lead. Think about FoS trends this way; if researchers follow them, they can start seeing a pattern grow. They will be able to smartly guess at what could be coming ahead of the research fields and will help to make positive and insight decisions for the future of their research fields. Similarly, researchers will aptly be able to anticipate the new FoS trends coming, because they will have already have a good idea of what is already coming.

Various techniques have been used in literature for FoS trend detection and analysis. Trend analysis is gathering data and trying to spot a pattern from collected information. Trend detection is commonly used to discover topic areas which are evolving in interest over time [3]. A very significant problem for trend detection is to catch research trends in a pool of scientific articles. A manual evaluation of the articles in a specific field would be a time consuming process. The automatic detection of research trends can help researchers swiftly recognize the occurrence of the trend of a scientific field and discover the most recent correlated fields in their research domain.

A number of approaches for detecting FoS trends have been proposed in literature. Initially, it was observed as the specialty and responsibility of domain experts. Domain experts were asked to determine the trendy FoS based on their knowledge and experience. Even to this time, expert consultation process is quiet extensively used in the practice of science and technology policy making [4]. The manual detection of research trends is an intensive job and a time-consuming process. Likewise, the continuous rise in research corpus available each year, builds the method created by specialists a fewer practicable. Therefore, it is significant to plan and improve instinctive and accessible approaches capable to execute the job in an automatic way [5].

There are three main stages for an FoS to become a trend:(a) embryonic, (b) early and (c) recognised. In embryonic stage of an FoS, a concept or an idea or concept did not emerge, yet. An FoS has not yet been clearly labelled and recognized by research community in this stage, though it is already taking shape. As the researchers from various fields are publishing and starting new collaborations to discuss the issues and the paradigms linked with the emerging new FoS. In early stage of an FoS, now it has been recently emerged and few researchers starting publications and will agree on certain concepts. Afterward, an FoS becomes mature and enters in its recognized stage and several researchers actively publishing their results [4].

For example, Figure 1.1 shows the embryonic, early and recognized stage of an FoS. The embryonic stage of "Semantic Search" FoS is 2003, it still was a concept in which a number of of researchers from, World Wide Web, Information Retrieval, Semantic Web and Search engine were linking their forces. After 2003, the FoS emerges, getting its identity, and enters in the early stage, and a group of researchers started publishing in this FoS. After few years, the FoS reaches its recognized stage with an increasing number of publications per year.

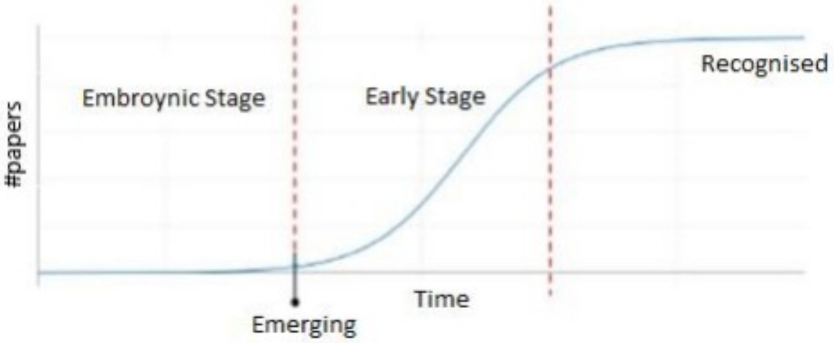


Figure 1.1: "Semantic Search" FoS lifecycle representation [4].

The dynamic increase in the research plethora has made it difficult for the scientific community to discover hidden patterns from a particular field of study. The FoS determines the area of focus of a particular scientific article. For instance, a paper focusing on comparison between various machine learning algorithms such as Support Vector Machine (SVM) and Naive Bayes, etc. will belong to the FoS, "Machine Learning" or "Artificial Intelligence" [3]. Typically, the inclination of the scientific community towards certain field of study (FoS) is more among other fields due to emerging of trends in the field. Publications by a large group of researchers in the same FoS may form a trend, resulting in increased popularity of the FoS among other fields. For example, figure 1.2 shows the trend of some selected FoS from year 1990-2015. The X-axis shows FoS and Y-axis indicates the score which is a metric proposed in [3] to represent the level of importance of a specific FoS. The figure shows importance or significance of different FoS at individual level and relevant to each other. Web search engine is leading till 2012, when social relation and recommender system arise in fame.

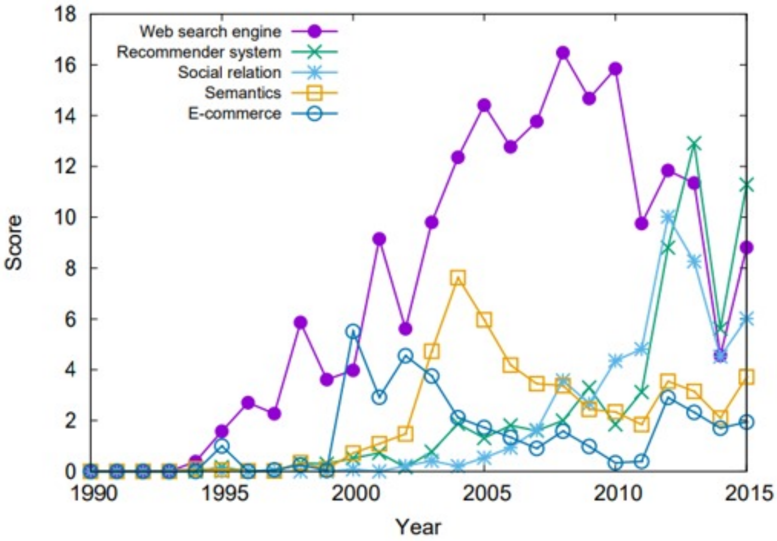


Figure 1.2: "Trend of selected fields of study in World Wide Web [3].

A pioneering researcher typically opts for a field that is more popular or its trends are being followed by the wider scientific community. This is done based on an assumption that contemplation of these aspects may increase the acceptance probability of the piece of work done in the trendy FoS, and further lead towards the rapid gain of citations in future. In scientific literature, following the research trends and dynamics can hold noteworthy benefits and this is significant to specify the interest of researchers [6].

The world of research does not stance still for long, it fluctuates with the passage of time new research fields and areas emerge whereas some others diminish. Such changes is precisely challenging. The capability to identify significant innovative research trends and forecasting the upcoming effect is important for recognizable members, such as researchers, academic publishers and official finance organizations. This may have direct impact on the acceptability, response, productivity and usefulness of a piece of work. One way to supplement this argument is through google N-gram viewer [7] that shows the frequency of use of different terms in published literature. For some terms we see the number rising and for some others it shows declining numbers. Figure.1.3 presents the frequencies of the terms "Remote Health Monitoring", "Smart Power Grids" and "P2P Network". Same impression is obtained from Google scholar [7], where the number of publications in recent years from these three fields is 395000, 51000 and 21900 respectively.

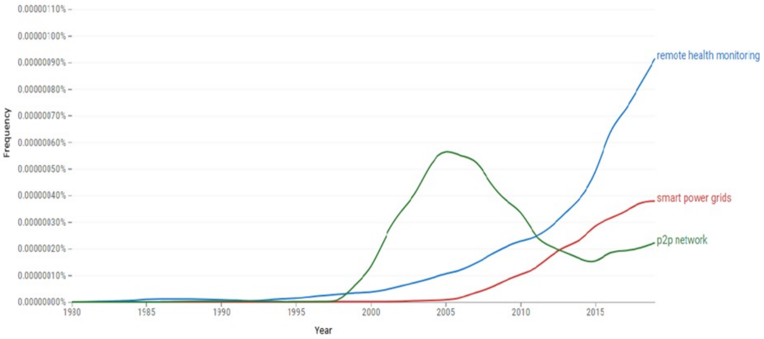


FIGURE 1.3: Trend of remote health monitoring, smart power grids and p2p network FoS [7].

The scientific community has presented various studies on FoS trend detection and analysis. These studies focus on multiple issues like, (i) birth and establishment of an FoS trend, (ii) number of publications and researchers in an FoS trend, (iii) communities of researchers being formed around an FoS trend, (iv) measuring strength of an FoS trend, (v) vanishing of an FoS trend, (vi) lifespan of an FoS trend, (vii) grouping of different FoS trends, (viii) trends in different disciplines, etc. In this work, we are going to discuss the (i) nature of response received to work done in a particular FoS trend, (ii) analyze relationships between different FoS trends by using graph centrality measures, (iii) researchers who are involved at the early stage of an FoS, and (iv) the impact to follow FoS trend on research paper citations and on authors citation count.

One major difficulty in addressing these issues is the availability of relevant data, that is, a worthy source of dataset is required. Since bibliographic datasets having features like title, authors, conference, and journal information in the field of Computer Science are not so hard to acquire as DBLP [3] freely provides this metadata in a structured manner. However, features like citations, keywords, and FoS are harder to acquire as they are not available in the form as DBLP provides other features. Discovering the FoS of a research paper is itself a research problem.

Microsoft Academic Graph (MAG) provides a rich source of dataset making it easier to acquire such information [8]. Precisely, MAG has a study that depicts a relationship between research papers and their corresponding FoS in a hierarchical manner. In MAG, each paper contains a list of FoS and these FoS are organized into four different levels; level-0 to level-3 with level-0 being the most general FoS, e.g., Computer Science and level-3 being the most specific e.g., cluster analysis (MAG has been explained in detail in section 3.2).

The focus of this thesis is to study the impact of FoS trend following on research paper citations and author citation count. This research uses the scientific articles published from 1950-2015 time period in the domain of Computer Science from MAG dataset. The research identifies that (i) how significant is to follow an FoS trend in Computer Science field? (ii) can we use any measure other than citation count to detect the trend of an FoS? (iii) is there any relationship between different fields of study? (iv) which are the trendy FoS in Computer Science field? (v) who are the individuals involved at the early stage of an FoS trend? (vi) what is the effect of following FoS trend on research paper citations and on authors citation count.

1.1 Scope

Following the FoS trends and dynamics in scientific literature can have significant advantages for researchers, editors, and funding agencies who want to keep themselves informed about the advancements made in various study fields in the past, present, and future. Being trendy in the field of study doesn't necessarily mean sticking to old-fashioned methods or always seeking perfect symmetry. Instead, it's about staying open to new ideas and being aware of innovative ways of thinking. People who keep themselves updated with the latest research and ideas are considered trendy compared to those who don't follow current trends in the field. Following the latest FoS trends also provides researchers a good advantage in the fast-paced world and to be able to connect well professionally and also it creates a high impact on their paper citations and their careers.

Hence, it is important to follow FoS trends in scientific research. It is believed that editors recognizing innovative evolving FoS in advance is critical for presenting the utmost and motivating contents e.g. the editor being the first person to identify the significance of FoS trend and publish about the issue. It is also believed that funding agencies known the effect of FoS trends on scientific literature may get sight and learn the antiquity of a field by depict the structure of its scientific inventions, in order to do tactically design and define the research significances in the field.

1.2 Problem Statement

Scientific researchers work and publish in their respective areas of interest. There is a paradox while adopting a particular research direction is that it must be very useful in the future, simultaneously it must be very popular and trendy at present. A plethora of work has been carried out on detecting Field of Study (FoS) trends, however, the impact of following an FoS trend on citation count of scientific articles and authors has not been studied up till now. The study of FoS trends can have an immense impact on the acceptance or citation of scientific work and on the researchers citation count. In this context, this work aims to explore the impact of following an FoS trend on acceptance of a piece of work and on researchers citation count.

1.3 Research Questions

This research uses the scientific articles published in Computer Science and analysis following research questions. The detail of these questions is given in methodology chapter. The research questions investigated in this thesis are;

RQ1. How similar is the citation trend of papers belonging to the same FoS?

This research question describes the impact of following a particular FoS on research papers. To address this question, we plan to study the citation trends of papers belonging to different FoS. If we could find a reasonable level of similarity in the citation pattern of papers belonging to same FoS and a level of dissimilarity in the citation pattern of paper belonging to different FoS, this will provide a basis to claim that FoS has impact on the citation trend or acceptance of a piece of work. Section 3.3 describes proposed approach of this research question.

RQ2. How can we differentiate between trend setters and trend followers?

This research question differentiates between researchers who were involved at the early stage of an FoS trend and the authors who followed it afterwards. Classifying authors into these two categories will help researchers to identify the influential authors in a specific FoS. Studying work of trend-setters of an FoS guides a researcher that how an FoS was originally conceived and proposed, the later review on that FoS will guide the stages it has gone through. For example, E.F Codd's work on Relational Data Model or Tim Berners Lee's work on Semantic Web gives real insight into these areas. That is why their work is still being cited heavily even today. In section 3.5, we presented our work on this research question.

RQ3. What is the impact of FoS trend on authors citation count?

This research question describes the effect of FoS trend following on scientific authors citation count. A significant aspect in scientific research is evaluating the impact of scientific contribution of a particular author. As the scientific publication data increased massively per day, the issue of Field of Study (FoS) selection by authors started to be studied quantitatively in recent years. Therefore, this is significant to measure the impact of FoS trend following on authors scientific growth. In section 3.6, we will address and discuss the proposed approach of this research question.

1.4 Aim and Objectives

FoS trend following is significant for Computer Science researchers as they will be able to smartly guess at what could be coming ahead of the research fields in Computer Science and will help to make positive and insight decisions for the future of their research fields in this area. Similarly, researchers will aptly be able to anticipate the new FoS trends coming in Computer Science field, because they will have already have a good idea of what is already coming.

The aim of this thesis is to discover the significance of FoS trend following in the field of Computer Science from 1950-2015 time period. The main objectives are: (i) how significant is to follow an FoS trend in Computer Science field? (ii)can we use any measure other than citation count to detect the trend of an FoS? (iii)Is there any relationship between different fields of study? (iv)who are the trendy FoS in Computer Science field? (v)who are the individuals involved at the early stage of an FoS trend? (vi)what is the effect of following FoS trend on research paper citations and on the careers of scientific authors.

This research could be beneficial for researchers and subject experts as well as policy makers. The researchers and subject experts at a glimpse can see which FoS trend tailed in their discipline by their peers, and which areas have been less attended. The fallouts of such study would help the policy makers in the distribution of research aid to specific FoS and subject fields with more assurance.

1.5 Research Contributions

Contributions. The main contributions of this thesis are as follows;

1. A novel technique Field of Study Multigraph (FoM) is proposed, by using centrality measures: degree, closeness and betweenness to explore the trendy FoS in Computer Science field and the relationship between different FoS.

2. A technique is proposed to identify trend setters and followers of an FoS trend. We have proposed an approach to detect influential researchers who

were involved at the early stage of an FoS trend known as trend setters and the authors who followed it afterwards known as trend followers. The influential authors (trend setters) achieved high citation count and significance in a particular FoS.

3. A method is proposed to detect the FoS trend that an individual author is involved in his/her career years and the impact of FoS trend following on authors careers.

1.6 Thesis Organization

The chapters of the thesis are described as follows:

Chapter 2 describes the literature review of trend detection and analysis techniques like probabilistic, co-citation network, keywords based and hybrid techniques. The chapter also presents topic evolution and author topic switching techniques. Further, describes a comprehensive analysis of these techniques in scientific trends, highlight the issues and research gap.

Chapter 3 describes the dataset in detail and the proposed methodology for each research question. First, we present FoS trend, their impact on research paper citations. We propose a novel technique Field of Study Multigraph (FoM), formed by using different centrality measures to examine the FoS trend, citation trend, and the relationship between different research areas in Computer Science papers. Then, we analyze the FoS - debut year, publication count, author count and FoS trend for the identification of trend setters and followers. Finally, we detect an individual author's - FoS, publication count, citation count, citation trend and author FoS trend to analyze the impact of FoS trend on authors citation count.

Chapter 4 discusses the experimental setup, results, evaluation and comparison to the existing state-of-the-art approaches.

Chapter 5 closes the thesis, deliberating the limitations, and outlining the future work.

Chapter 2

Literature Review

The importance of detecting and analyzing FoS trends for a piece of work as well as for researchers is discussed in this chapter. This chapter discusses various approaches of trend detection and analysis in research fields, the emergence of FoS, the detection of researchers at various stages of FoS, researchers' FoS interest, the challenges and issues that represent the importance of this research, In section 2.1, we look at various strategies for detecting and analysing trends in research papers, journals, conferences, and keywords. The evolution of FoS and FoS detection techniques are discussed in section 2.2. The strategies recommended for detecting FoS in scientific authors'careers are also presented in section 2.2. Finally, in section 2.3, the limitations and research gap of these studies are explored.

2.1 FoS Trend Detection & Analysis Techniques

FoS trend detection's purpose is to classify and organise all of the articles according to their FoS. In a seminal study published in 1983, Callon et al. [9] developed a method called co-word analysis, which was one of the first attempts to assess the content of texts. Their main claim is that keywords are adequate to describe the content of an article. The keywords used to define a publication's content can be thought of as the essential building blocks for describing science's structure in this approach. This method selects keywords from the papers and counts the number of times they appear. The purpose of every phrase in a document is to assist in its placement in the proper network area. The program then divides the network of keywords into groups depending on how frequently they occur together. The scientific literature's concepts or themes will be represented by the clusters generated.

Many other research groups built on Callon and his colleagues' work as computing power improved, focusing in Scientometrics field [10], and technological research [11], and analysing various fields like Informatics [12], Software Engineering [13], and Geographic Information [14]. The fundamental issue is that such a technique is incapable of dealing with polysemy and synonymy concerns because it focuses on the recurrence of specific words and ignores context. Defense Advanced Research Projects Agency (DARPA) created Topic Detection and Tracking (TDT) programme, which laid the framework for this field [15]. Despite its focus on broadcast news, this computer recognised the problem and provided useful suggestions for automatically recognising subjects and organising a stream of papers. They use cluster analysis to organise objects by their relatedness into bins that represent subjects in order to locate topics. Allan and his colleagues experimented with a variety of clustering techniques.

Eichmann et al. reference [16] created a system designed to identify the most crucial phrases in news articles and then utilized term frequency-inverse document frequency (tf-idf) to compute cosine similarity between each article. This process allowed them to form clusters of related news pieces. On the other hand, Kullback-Leibler [15] employed a different approach to select relevant news articles, utilizing distance as a metric.One of the significant contributions of this work was the formalization of the problem of TDT (Text Data Mining), which played a crucial role in advancing the field. In reality, based on one of Allan et al's formalized tasks, there are several approaches in the literature. The issue was divided into four research tasks by Allan and his colleagues: Topic Detection, Topic Tracking, First Story Detection, Story Segmentation and Story Link Detection. The clustering task specifically, aids in discovery of the corpus's thematic structure, and it essentially comprises of technologies that summarise document content and categorise it according to the subjects it covers.

The data types provided in the literature can be used to classify all existing techniques. The most essential pieces of information employed by techniques for detecting themes in collections of scientific publications are titles, keywords, abstracts, citations and complete text. To improve clustering performance, few techniques focus on just keywords information whereas others use the combination of abstracts and keywords. It's worth noting, however, that the corpus may impose limitations on the proposed technique on occasion. Certain approaches, for example, are unable to use the majority of a paper's content due to a lack of data. The next sections go through the most prevalent topic detection methods, organised by the data types they use (abstracts or full text, citations, keywords and taxonomies, and hybrid solutions that combine different domains).

In systems that rely on abstracts or full content of publications or both, the text of scientific papers is frequently synthesised first. This dimensionality reduction is achieved by eliminating specific relevant keywords known as bag of words, these words precisely reflect the content. Previously, we discussed the widely used tf-idf technique for extracting relevant terms from document content. This method [17] [18] involves creating an index for each word, considering its significance in the context of the entire document collection. Afterward, only the words with the highest values are retained. The text pieces are subsequently sorted using cosine similarity, which measures the similarity between documents based on the shared terms they contain.

In molecular biology Roche et al. [19] choosed a similar method by using $tf - idf$ and may find discriminatory word groupings in texts, it ignores text structure, meaning, and co-occurrences across documents. Furthermore, the reduction in dimensionality is restricted, and it does not provide sufficient information about the statistical organisation of texts between or within them. Topic modelling has become an integral part of the next generation of topic recognition systems. One of the most significant techniques in this field is Latent Dirichlet Analysis (LDA) developed by Blei et al. [20] . In LDA, each paper is represented by a set of topics, and each topic is characterized by a multinomial word distribution, indicating the probability of each word occurring within that particular topic. The primary objective of LDA is to uncover hidden patterns in the word structure of documents, revealing the underlying subject and topic of each document. This involves calculating the hidden variables associated with topics and words.

On the other hand, the Tf-idf (Term frequency-inverse document frequency) approach operates on a lexical level. Unlike LDA, Tf-idf considers the specific characteristics of words in documents. For instance, it would treat a network-related text differently from a book about graphs, taking into account the word frequency and its significance in the context of the entire document collection.

LDA has an advantage when dealing with related topics that often co-occur, like networks and graphs appearing together in articles. Since LDA can identify such closely related topics, it provides a comparable representation for articles discussing these subjects.

Apart from LDA and Tf-idf, other notable topic modelling techniques include the Correlated Topic Model (CTM), Latent Semantic Analysis (LSA), and the Probabilistic Latent Semantic Analysis (pLSA). Each of these methods brings its unique approach to extracting and organizing topics from a collection of documents.

One of the earlier methods used for addressing the limitations of the $tf - idf$ approach is called Latent Semantic Analysis (LSA). LSA is particularly effective in handling large collections of text data due to its ability to significantly reduce dimensionality. The process involves calculating the frequency of each term in the text while preserving the similarity structure between columns. Subsequently, LSA applies a technique called Singular Value Decomposition (SVD) to reduce the number of rows in the data.

By leveraging this approach, LSA can group together documents that share similar keywords, as well as identify common terms among a group of documents [21]. This enables LSA to capture two important linguistic features: synonymy (words with similar meanings) and polysemy (words with multiple meanings). Through these patterns, LSA detects the latent components present in a batch of documents.

Hofmann created the pLSA [22] as a substitute for the LSA. Text words are analysed using the pLSA as samples from a mixed model, then extracts topics as multinomial random variables, and hence mixture components using the Expectation-Maximization technique. Blei and his colleagues created the LDA to address some of the flaws in the pLSA, such as the fact that it only learns topic mixes in the case of papers seen in the training period. Logistic normal distribution used the Correlated Topic Model [23] to address the fact that LDA fails to capture the link among topics.

Because it is realistic to predict that a fraction of the underlying hidden themes will be closely connected. For example, data on health and sickness is likely to be included in a scholarly study on genetics.LDA has been improved and expanded for a wide range of applications since its inception. Some of the methods employed include supervised topic models, latent Dirichlet co-clustering, author-topic analysis, temporal text mining, and LDA-based bioinformatics [24] [25]. Two more LDAs developments are the LDA's hierarchical structure [26], which arranges subjects in a hierarchy, and the topic model based on relationships [27], which combines the topic and network models for groups of papers that are linked together.

However, one of the major flaws of the LDA, as well as many other topic modelling methods, is the lack of subject labels. These approaches communicate a topic by utilising several phrases that are also descriptive of the circumstance, but choosing one as a label is challenging, at least automatically. Morinaga and Yamanishi [28], for example, use the entire text to organise the subjects in a collection of papers. A Model is used to locate subjects, and changes in the extracted components can be used to keep an eye on the advent of new ones (word clusters) by using [29] Kleinberg's approach, that will be described in further detail in the subject evolution section. Morinaga and his colleagues characterise the topic structure using a finite mixture after defining it as a single component of a mixture. Because this technique was only tested on an email corpus, it is unknown how well it will work on scientific papers.

Chavalarias and Cointet [30] came up with a strategy for rebuilding scientific representations on their own. They searched the Thomson Web of Science corpus of over 200000 entries for anything connected to embryology research. They created an n-grams list of 2000 that reflected the majority of important sentences using the CorText Manager application. The clique identification method was then employed to perform clustering analysis on the co-occurrence matrix for the purpose of reveal scientific evolution tendencies. Sayyadi and Raschid [31] provide another network-based technique. Sayyadi and a coworker came up with a way to extract all significant terms from a document's text. The publications are then utilised to form a network of keyword co-occurrences. They then use a community discovery technique to find term clusters that match to network topics using a community discovery technique.

The majority of citation network approaches are based on Small's concept of clustering scientific papers via co-citation analysis [32]. Citations have been used to locate subjects in a variety of ways, with some systems combining keywords and abstracts are examples of citations with other entities. Small [33] is credited with inventing one of the first approaches for determining subjects from citations. Small discovered "hot domains" by accumulating and linking highly co-cited publications over time. Boyack and Klavans [34] and Small [35] have provided more current analyses of this subject, claiming that the fundamental technique remains applicable despite changes in thresholds and normalizations throughout time. Similar studies have been done by Small et al., as well as Upham and Small [36]. The ISI corpus (now Web of Science) was used between 1999 and 2004 to find 20 developing subjects that formed co-citation clusters. Small et al. [37] on the other hand, use Scopus data to do co-citation analysis, allowing them to undertake a more comprehensive worldwide analysis rather than a more specific subject analysis.

CiteSpace, a programme created by Chen [38] uses a combination to find unique emergent patterns using co-citation analysis and burst detection. They specialise in network analysis that is progressive which entails combining network slices and concentrating on key nodes that influence the evolution of the network through time. In the fields of Regenerative Medicine [39], Mass-extinction and Terrorism [38], Peptic Ulcer, Gene Targeting and String Theory[40] CiteSpace has been used to characterise creative emergent ideas. According to their findings, the increase in citations and the relevance of their articles' betweenness are two of the most important factors in the formation of new topics. This is when two or more current themes are combined to form new subjects or clusters. Combining phrase distributions like n-grams considering the citation graph distribution linked with publications that contain that phrase, Jo et al. [41] established a method for recognising subjects. The word would have been the name for that topic in their previous estimate, but they altered their method to recognise themes as a relationship between a number of terms.

The fundamental disadvantage of systems According to co-citation analysis, each document may only be assigned to one topic. Rarely is a document monothematic. As a result, these strategies may result in fractured document clusters. Keywords are another way to recognise subjects in an anthology of papers. Scholars carefully make use of keywords also known as subject terms to describe the research field of their work [42] [43]. They highlight the paper's uniqueness while expressing the essential arguments and ideas. Such representation is critical for search engines when it comes to returning relevant publications in response to a query. In the literature, many keyword-based strategies for discovering subjects have been discussed. Duvvuru et al. [44] [45] investigated how link weights varied over time in keyword co-occurrence networks. They intend to detect research patterns as well as emerging subject fields using this strategy. Keywords as a theme alternative, on the other hand, has a number of disadvantages. Words like "case study," according to Osborne and Motta [46], are ambiguous and may not always imply research topics.

They also struggle with synonymy and polysemy, which arise when many phrases refer to the same item or when a single phrase conveys multiple ideas. Yi and Choi [47] proposed a method for dealing with some of Duvvuru et al's issues. The authors created a method for cleansing the keyword set before generating the keyword network. When two keywords are considered similar, such as "agent" and "agents, "Agent" and "agents" are examples of words that have been combined into a single form. If a term has two different keywords, the phrase "efficiency and effectiveness" is broken down into "efficiency" and "effectiveness." However, there are just a few guidelines on how to implement this method. Decker [48] devised a method for analysing topic trends over time by using abstract keywords and phrases to create paper-subject correlations. This technique, in particular, links a group of scientific papers to a well-chosen taxonomy of subjects. This taxonomy was created by hand using topics from the proceedings, which were then supplemented using phrases and keywords from the abstracts.

They can then look for research trends by looking at the number of publications on particular themes has changed. Erten et al. [48] followed research trends by tracking the evolution of subject graphs using the ACM Computing taxonomy. These two approaches, which use subject taxonomies rather than keywords, could be regarded an improvement over the strategy used by Duvvuru et al. [44]. The primary problem The number of publications on particular themes has changed. Erten et al. [48] followed research trends by tracking the evolution of subject graphs using the ACM Computing taxonomy are gradually becoming obsolete. The ACM Computing taxonomy was last updated in 2012 (six years ago), replacing an earlier version from 1998. As a result, fresh emergent themes are not given the opportunity to be included in the taxonomy, resulting in inaccurate classification. To classify physics papers, Herrera et al. [49] suggested a concept network, with each node representing a code for a particular physics problem. If the two relevant codes appear in at least one article, the two nodes are linked. The Palla et al. [50] used the Clique Percolation Technique to connect together comparable communities over time to analyse the evolution of various sectors.

To undertake a comparable analysis in medicine, Ohniwa et al. [51] employed the Medical Subject Heading (MeSH) 16. The National Library of Medicine in the United States maintains the MeSH taxonomy, which is similar to PhySH in that it is updated on a regular basis. Another tool in this area is Osborne and Motta's [46] Klink-2, a method for automatically constructing the Computer Science Ontology, an ontology of study disciplines in computing. Klink-2 uses semantic technology, machine learning, as well as knowledge gleaned from outside sources to reveal relationships and develop DBpedia.

Klink-2 employs keywords in the same way that Erten et al. [48], Duvvuru et al. [44] and Decker [52],: relatedEquivalent, contributes To, and broader Generic. Compared to human-created ontologies like this computer-generated ontology has two key advantages. For starters, because the ontology is automatically constructed, it may could be quickly up to date accommodate new emergent themes by reinstalling Klink-2 on a fresh batch of texts. The second advantage of Klink-2 is that it can connect more ideas than other taxonomies. Rexplore [53], Smart Topic Miner [54], Smart Book Recommender [55], and the detection of topic-based research [56] are just a few of the applications that have profited from this semantic characterization of research topics.

Finally, in literature, there are numerous techniques for selecting study subjects using keywords. As we've seen, some systems rely solely on keywords, which can lead to polysemy, synonymy, and other difficulties. Other algorithms, on the other hand, infer from taxonomies of topics rather than keywords to provide more accurate results. Methods that use automatically generated taxonomies, on the other hand, are more complete and up to date. So far, we've looked at methods for selecting subjects based on scientific journal abstracts or full texts, as well as citations and keywords. We'll look at hybrid systems This section contains documents that employ a variety of metadata, such as venues, titles, confereneces, journals, and authors. Some systems rely solely on one of these components, while others employ a combination. The writers are the fundamental agents in scientific endeavours. They do scientific study, publish their findings, and present their discoveries to the public. Knowing about them can be valuable and reveal surprise conclusions when analysing a corpus's theme organisation. Author networks are employed, with links representing co-authorship ties to analyse the research environment in a variety of ways [57].

Other approaches, such as the Author-Topic Model (ATM) [58], build upon the probabilistic themes model mentioned in the previous section to characterize writers' patterns. The main objective of ATM is to integrate authorship information into Latent Dirichlet Allocation (LDA). In LDA, text is represented as a distribution of themes, and ATM expands this concept by associating themes not only with words but also with authors. By doing so, ATM aims to identify which topics are prevalent among different authors based on the entire corpus.

To improve the results of ATM, the Probabilistic Author-Topic Model [59] was introduced by the same research group. In this model, the text is divided into various themes, where themes are treated as probability distributions over words, and authors are represented as probability distributions over topics. Instead of relying on LDA, they enhance Hofmann's Probabilistic Latent Semantic Analysis [22] for their approach.

However, the probabilistic approaches have their limitations. They tend to oversimplify the representation of the data, overlooking factors such as topic correlation and author interactions. To address these issues, a generative model called S-ATM was developed by [60]. S-ATM utilizes the temporal ordering of documents to detect topic evolution over time and evaluates the importance of relevant terms in texts based on citations.

Scientific advancement and subject identification can also be found in conferences and periodicals. Indeed, the organisation of a research topic may change over time as a result of many publication channels. Currently, there are various approaches to leverage publication venues, which can be broadly categorized into two groups: methods that expand thematic models and approaches based on network analysis. For instance, a notable contribution in this field is the Author-Conference-Topic (ACT) model, introduced by reference [61]. This model extends the conventional Author-Topic model [62] by incorporating data from conferences and publications. The authors propose three distinct implementations of the ACT model, each offering a unique perspective on the connections between authors, subject distribution, and conference information.

Yan and his colleagues [62] used the ACT model to explain how scholarly communities and research subjects are linked and evolve together rather than as two distinct entities. According to [63], the Author-Conference-Topic technique has a flaw in that it maps subjects from a corpus of works to research fields provided in conference "calls for papers." Because the latent themes derived using the that the LDA will not always match to the conference subjects, this technique is not always feasible. Wang and his colleagues improve this functionality, they developed Author Conference Topic Connection (ACTC) model, which combines conference subject information with information about latent mapping between subjects and themes. [64] [65] and [66] are instances of techniques that rely on networks to gather data from venues. [44] and [49] utilised a similar approach, constructing a collection of entities that can be used to analyse and identity scientific subjects [64]. The method developed by Sun and her colleagues, on the other hand, is unique in that it employs author co-occurrences to determine conference similarity. They establish a network for each year, each node represents a current conference for that year and their degree of connectedness is determined by the total number of authors who have published in both conferences. They use the Louvain technique to detect communities within the network once it has been built in order to locate research areas [67].

Boyack et al. [65] developed a new strategy for locating participants that they want to expand and use in the future. Using journal citation data, the authors created a map depicting the structure of all research. They came up with eight classification criteria for Journals are classified according to how they cite one another. Despite the fact that this method is used the authors use visualisation to highlight that each cluster of journals is linked to a distinct research topic. Infact, to explore the evolution of Computer Science topics, [66] used a similar method to visualise the at different moments in time, the knowledge network has changed. Their method integrates data from DBLP and CiteSeerX on venues and citations, respectively. They started by integrating them with citations from papers and conferences to create a venue knowledge network. The network is then clustered and the evolution of the network is tracked to uncover computer science sub-areas. Other research groups, such as [68] [69], have looked into journal maps.

From more than 36,000 big data articles across all academic disciplines between 2012 and 2017, authors used topic modelling and word co-occurrence analysis approaches to identify relevant topics [70]. The Several topics related to the storage, gathering, and analysis of huge datasets were exposed by the results; The majority of the papers were in the computational sciences. Other known studies themes demonstrate how big data techniques and procedures are used outside of computer science like business, health, and medical sciences. In actuality, the predominance of these topics has grown throughout time. In contrast, some topics like big data analytics, parallel computing and network modeling have lost favour in recent years. These outcomes most likely demonstrate the development of key big data subjects and spotlight thriving new research trends relevant to large data in new fields, especially in the social sciences, health, and medicine.

A large-scale knowledge network that classifies research papers in accordance with the research themes from the Computer Science Ontology was used by the authors to propose a framework for detecting, analysing, and forecasting research topics (CSO) [71]. They first provided an example of how to add a set of research topics from a domain ontology to a scientific knowledge network describing research publications and their metadata.They presented many approaches for analysing research from various angles that build on this knowledge graph. The benefits of a solution built on a formal description of themes were presented, and they provided an account of how it was put to use to create bibliometric studies and cutting-edge tools for analysing and forecasting research dynamics.

Big data refers to enormous databases that make analysis using conventional data processing methods difficult [72]. Researchers who want to work on this fascinating topic will benefit by recognising and grouping developing topics in this field. Algorithms for text mining and social network analysis are used to spot the newest trends in the big data field. In this study, authors first gathered all of the papers that are pertinent to the big data field, and based on the extracted keywords, a word co-occurrence network was built. The association rules technique was used to determine the relationships between the keywords after the best clusters had been found.

The authors created worldwide maps of research based on journal-journal citation links using data from the Social Sciences Citation Index (SSCI) and the Journal Citation Reports (JCR) in these investigations. There are a number of other possibilities that are fairly similar to the ones we've just looked at. [73], [74], [75] devised a method for creating computer science maps using DBLP paper titles. They start by extracting common phrases and words, then utilising title co-occurrence to calculate their similarity. The words are then clustered according to their similarity ratings, resulting in a depiction of topic space in graphic form.

The corpus is the key reason why authors only look at titles in the DBLP, which is restricted to titles, authors and locations. However, there are drawbacks, such as the fact that the title of a paper may not always reflect all of the topics addressed in the research.

As we can see, there are a plethora of cutting-edge methods for locating topics within a collection of articles. The extent to which the corpus is comprehensive, as well as the availability of specific entities, may have an impact on the creation and acceptability of research. To that purpose, we organised our research by grouping strategies that use a comparable collection of things together, then including methods that combine many entities at the end.

2.2 FoS Evolution and Author's FoS Detection

The static component of a FoS, i.e., its identification within a collection of documents, is the focus of FoS detection. Evolution of the FoS, on the other hand, is concerned with the dynamic nature of FoS, or how they change over time. [15] focused on two objectives in particular: First Story Tracking and Detection. The First Story Detection (FSD) task is used to detect the first story find previously unnoticed rising ideas. This task, in particular, keeps an eye on the incoming document flow to see if any new subjects have arrived. A competent FSD system, for example, should be able to recognise early Semantic Web papers from 2001, as well as Deep Learning and Cloud Computing papers from the mid-decade (2000-2009). This task, on the other hand, can only recognise people after they've already appeared, rather than anticipating or forecasting them.

Tracking, on the other hand, searches for fresh articles that address issues that have already been covered. When analysing incoming articles, the analysing system should be aware of the topics covered in the document collection and be able to categorise them appropriately. This implies it will extract the themes from each new document and arrange them with similar documents in the collection. The system may now do a statistical analysis to track the current state and evolution of each issue. Twenty years ago, some of the technology employed to aid in this endeavour was rather advanced. Cutting-edge algorithms for extracting topics from texts, such as PLSA [22], are now available, as well as a variety of similarity metrics for grouping articles based on their topics. These two tasks can be used by users to keep track of and analyse concerns as they arise. Although these are two distinct tasks, some solutions combine them to assess the existing state of each topic while simultaneously discovering new ones and organising a large number of incoming articles.

Di Caro et al. [76] devised a system for tracking the progression of topics inside a corpus through time, as we'll see later.They may even detect the emergence of new motifs that have never been seen before in history if they use their method. There are a number of approaches in the literature that attempt to track the progression of themes as well as their origins in general. Others perform extensive research [77], [78], [79] on some custom metrics based on the total number of papers related to the issue [80], [81], or the number of authors [82]. Others may employ co-word analysis [83], [84], hybrid studies [77] or citation analysis [78], [79], while still others may employ citation analysis to discover document citation trends. Finally, a third approach creates science maps using overlay mapping techniques and relies on human experts to analyse new topics [69], [74].

The burst detection approach for detecting emerging subjects recognises rapid changes in word usage, according to [37]. Kleinberg [29] is credited with inventing the burst detection approach for spotting emergent subjects, which detects rapid changes in word usage. This technique has stirred a good number of approaches for identifying research trends [52], [78], [85], [86], [87], [88]. Citespace II [89], Sci2 and Network Workbench [90], and all include burst detection as part of their bigger tool sets.

Augur [91] is a revolutionary way for identifying study volunteers early on. Augur looks at the diachronic linkages across fields of study and can spot clusters of subjects with dynamics linked to the formation of new disciplines. A novel community discovery algorithm, the Advanced Clique Percolation Method (ACPM), was devised expressly for this objective, is also featured. From 2000 through 2011, Augur was compared to a gold standard of $1,408$ new themes. Kleinberg's method involves analysing a stream of documents for bits that behave “bursty,” that is, when they occur in a rapid burst of activity. This method uses probabilistic automation, with numerous phases dependent on how frequently each term is used. There are as many automata as there are words, and they switch states when the frequency of their associated word varies dramatically, such as at the start or end of the burst period.

This technique can be used to discover FoS and concepts that have gained traction and have sparked heated debate for some time. Because it does burst analysis for each word, including stop words, it must be put into a pipeline that first preprocesses the texts. Jo et al. [78] devised a method that incorporated phrase distributions including n-grams and the citation graph distribution for publications containing the term in question. If a term is related to a topic, for example, the authors expect that documents that contain that term will have a stronger relationship than documents chosen at random. The approach is successful, according to their findings, and can even detect new developing subjects. However, because the citation network of a phrase takes time to build to become firmly connected, their approach has a temporal lag.

By looking at how citation patterns have changed throughout time, Morris et al., [92], Small and Upham, [93], Shibata et al., [79],Morris, [94], Takeda and Kajikawa, [95], Astrom, [96], can discover the birth of a new region. These methods are based on the premise that bringing two previously unconnected or poorly connected locations together will result in a better outcome could signal the formation of a new subject that can build on earlier Takes. Because the authors focused on a small area of optics, manual analysis was possible; nevertheless, we could argue that if the domain was increased, such a strategy would not be scalable. Morris et al. [92] use co-citation networks in their research to group bibliographically related texts or to share a list of publications that have been cited. This strategy, however, has the same faults as Takeda and Kajikawa [95]. Furthermore, no statistical metric for analysing the introduction of a new topic is provided because these approaches are evaluated by a human expert.

Shibata et al. [79] on the other hand, used topological methodologies to identify the rise of new topics without the help of expert specialists. The citation network was separated into clusters by the authors and assigned the most representative word to each cluster. By looking at the age of the cluster, the method detects impending topics. This strategy, according to the authors, has a time lag, which we agree with. This is a problem that any citation and co-citation network-based strategy encounters. They are under represented in such networks because new articles can take up to two years to be mentioned. Shibata et al. [79] urge that these algorithms be supplemented with data from other sources, such as venues, to detect the introduction of new subjects. Clarivate Analytics employs a different approach, relying on citations rather than topological network analysis. Clarivate Analytics has published a report called Research Fronts since 2013, which highlights a variety of important research fronts, including emerging and hot ones. According to "Research Fronts 2017", "A research front is made up of a core of highly cited articles that are linked to the citing journals that often co-reference the core," this report lists 100 hot research fronts as well as 43 new ones.

They grouped the total number of research fronts (9,690) into ten macro-areas to discover the most promising research fronts. The top ten research fronts for each of these ten organisations are then chosen based on their highly mentioned publications' average year i.e., core publications. Following that, the identified core articles, associated nations, and institutions are shown. Instead, they seek for research fronts that are increasing in fields where in the last two years, notable publications have been published (2015 to onwards). Human experts next analyse and interpret the evolving research fronts in order to catch recent trends and estimate their importance. This strategy has two significant flaws. This method, like others according to citation analysis, there is a time lag between the emergence of a research topic (emergent research fronts) and its identification. We can detect the emergence of a new issue two years later in the worst-case scenario, even with a two-year time lag. The second issue is the method's potential for low recall. Despite the lack of statistical data, in this report precision and recall were used to identify hot and developing research fronts, the method can be used to detect a problem like this. Because their primary papers garnered inadequate citations in the past two years, many fascinating subjects may go undiscovered.

The number of authors and co-authorship networks have been used to investigate how the number of writers influences the birth of new topics. Guo et al. [97] proposed a model that incorporates three different emergence signals.They came to their conclusions based on the frequency of keywords, the expanding number of authors, and the interdisciplinarity of the sources mentioned. The Rao-Stirling diversity index, which is calculated on a year-to-year basis [98],[99] is used to calculate the final indicator. Bursts of keywords appear before the introduction of new themes, followed by rapid increases in the number of authors cited, as well as the interdisciplinarity of the references cited,according to the researchers.Bettencourt et al. [77] looked analysed co-authorship networks to determine if there were any trends that could be linked to the formation of new research fields. Three main patterns were discovered: (i) the average number of nodes grows, showing that the network that surrounds such nodes is growing denser; (ii) the average path length ins two nodes stays the same or shrinks, suggesting that the network's width is changing; and (iii) the largest component has a growing number of edges. These all developments point to a tightening of the co-authorship network. As a result, forming a new research group is seen as a precursor to the development of new research topics.

The authors developed a method for determining the genesis of themes by studying the expansion of conference networks. They started by creating a progressive conference network using co-word analysis, with nodes representing links and conferences signifying connections indicating proximity based on keywords extracted from published papers. They then look for conferences that are becoming more and more similar to one another, collapsing over one another as a sign of new topics emerging. Di Caro et al. [100] devised a mechanism for monitoring subjects' progress over time. The approach takes two successive slices of the corpus after splitting it into discrete time windows, using LDA, extract the topics, and then examines how these subjects changed over time. The primary concept is that by comparing topics created over a short period of time, one can discern how subjects develop and how their birth and death are captured. Morinaga and Yamanishi [101] developed a method for forecasting the birth of a new subject using a probabilistic model called Finite Mixture Model. Using this method, the authors dynamically learned the structure of subjects from the papers in each year. Researchers then looked at the differences in the extracted components to see if any new subjects had emerged. Their research, on the other hand, has never been put to the test. Using scholarly publications to forecast the emergence of new study areas is difficult.

One of the first overlay mapping methodologies was developed by Boyack et al. [65], who mapped the "backbone of science". They started by classifying the data into temporal frames, then looking for phrase clusters and linking them to research areas for each window, such as year.By monitoring the clusters for two years in a row, they were able to match similar themes across time and discover new clusters connected to new topics. Similarly, Leydesdorff et al. [69] developed soverlay maps to assist policymakers in locating research bodies that cross traditional academic boundaries. These overlay mapping technologies are fascinating because they enable users to visually analyse locations in a global research environment where the number of publications is rapidly increasing. They can only provide a coarse-grained perspective since they neglect intricate linkages between research subjects. According to Rafols et al. [102] they should be used in conjunction with other maps that provide more detailed viewpoints.

Different approaches and techniques are employed in literature to identify an author's research interests and track their development throughout their career, culminating in the discovery of new knowledge [103–107]. With the increasing availability of large datasets in research, there is a unique opportunity to utilize cutting-edge computational and mathematical tools to uncover the dynamic patterns of scientific publications [108–110]. Apart from conventional studies [79–81] that assess an author's scientific impact using metrics like citation count and H-index, recent methodologies have shifted towards analyzing authors' careers by quantifying and modeling the growth of research originality [111–117]. The increase in an author's output, measured by the number of publications, follows a steady evolutionary process over time, often referred to as the Matthew effect [118].

Authors' publications and their impact, as measured by citation counts, appear to follow a random pattern, with one study highlighting the emergence of an author's most important work among their various publications [112]. Additionally, research has shown that authors may experience a peak period in their careers when their performance surpasses typical levels [113]. Throughout an author's career, factors such as output, movement, reputation, and social connections have been extensively studied [114–119]. Moreover, the evolution of authors' research interests is influenced by their shifts between different topics over time [120]. Sociologists have been investigating the reasons behind authors' topic choices and have discovered a potential trade-off between conventional productivity and risky innovation [121].

Authors in the literature have employed various strategies, and sociologists have developed models to categorize these strategies [122]. In recent years, the increased accessibility of scientific papers has allowed for statistical examination of topic/field selection. Several approaches have been proposed to determine authors' research domains in language-based themes [123, 124]. Additionally, scientific funding has been suggested as a potential priority in these research areas [125].

Research interests of individual physicists may evolve over the course of their careers, as indicated by a study that introduced physics categorization codes [126]. One recent study focused on using co-citing networks of papers to reveal community structures, with each main community representing a distinct research topic. This study investigated the variation of topic/field switching from the beginning to the end of an author's career [127]. Moreover, researchers analyzed the publishing records of individual scientists to quantify their subject switching dynamics and the resulting impact. Various methods exist in the state of the art for detecting and evaluating research trend formation based on the type of analysis and entities used. These methods can be classified similarly to detection strategies and, despite their limitations, they contribute to advancing the state of knowledge. However, they each have their drawbacks. For instance, citation-based approaches suffer from a temporal lag, making it challenging to quickly identify new topics. On the other hand, co-word analysis approaches focus on previously recognized topics that are already associated with specific labels or collections of terms. There are diverse approaches for studying the evolution and development of Fields of Study (FoS) over time. Some methods primarily concentrate on well-established FoS that have a substantial publication count, while others explore the embryonic stage of FoS. It is evident from the literature that researchers tend to work and publish within their areas of expertise and interest.

When embarking on research, scientists often consider the popularity or trend of a particular field. This can significantly impact their careers as scientific authors. Choosing an FoS based on its popularity might have its advantages, but it also carries the risk of potentially jeopardizing a researcher's career if the field loses its prominence or support in the future. Based on the critical analysis of the literature review, there is not any ample solution on the analysis and evaluation of impact of FoS trend following and how significant is to follow an FoS trend in Computer Science field?. Can we use any measure other than citation count to detect the trend of an FoS? Is there any relationship between different fields of study? Which are the trendy FoS in Computer Science field? Who are the individuals involved at the early stage of an FoS trend? What is the effect of following FoS trend on research paper citations and on authors citation count [2, 73, 74, 94, 128]? Still, these are inspiring and wide-open research questions. Table 2.1 displays most related studies summary:

2.3 Research Gap and Analysis of Existing Approaches

As we have observed that in previous state-of-art approaches like keyword-based, graph-based, bibliometric and hybrid approaches are used for FoS trend detection and analysis in scientific articles.

Table 2.1: Summary of some related studies

Reference, Year	**Technique**	**Outcome**	**Trend Focus**	**Detect FoS Trend**	**Detect FoS Trend Following**
Ehsan et al. [2022]	Topic Modeling ad word co-occurrence.	Detect big data core topics and new research trends pertinent to big data in new domains, especially in social sciences, health, and medicine	All academic disciplines publications	yes	no
Angelo et al. [2021]	Knowledge based graph	Analyze research trends and predicting their impact on academia and industry	Computer Science Ontology publications	yes	no
Seyed et al. [2020]	Text mining and social network analysis algorithms	Identify emerging trends	Big data domain publications	yes	no
Chengyao et al. [2019]	Network based-model	Detectrising trending topics and the most influential peer	Computer Science conferences	yes	no
Salatino et al. [2018]	Augur	Detect researchtopics in embryonic stage	Publications	yes	no
Binling et al. [2018]	Text mining approaches, including bibliometric analysis	Reveal research trends and evolution of research areas	Publications	yes	no
Effendy et al. [2017]	Field of Study Score	Detect research trends	Computer Science Conferences	yes	no
Cano et al. [2017]	Keyword based-model	Topic detection	Keywords	yes	no
Osborne et al. [2016]	Semantic topic model	Detect trends in new research areas	Publications	yes	no
Motta et al. [2016]	Graph-based approach	Detect trends	Publications	yes	no

Trend detection and analysis techniques are used to detect embryonic, current, and imminent FoS trends from the scientific data. Researchers have also proposed techniques for the detection, evolution and the development of FoS over time. Their focus is on embryonic, well established and recognized FoS, where a few and an active number of authors are involved with a number of publications. We have also observed that various techniques have been studied in authors careers such as the growth of authors' productivity, reputation, social ties, and mobility. Previous literature describes that an important aspect of scientific research is the growth of authors' research interest, which is represented in the switching of authors among diverse FoS over time.

The detailed analysis shows that the scientific community has presented various studies on;

1. FoS trend detection and analysis
2. birth and establishment of an FoS trend
3. number of publications and researchers in an FoS trend
4. communities of researchers being formed around an FoS trend
5. author's FoS switching
6. measuring strength of an FoS trend
7. vanishing of an FoS trend
8. lifespan of an FoS trend
9. grouping of different FoS trends
10. trends in different disciplinesetc.

On the other hand, the state-of-the-art approaches is still missing a comprehensive analysis on;

1. nature of response received to work done in a particular FoS trend
2. significance of following an FoS trend in Computer Science field
3. impact of following FoS trend on research paper citations in Computer Science articles
4. can we use any measure other than citation count to detect the trend of an FoS?

5. analyze relationships between different FoS trends by using graph centrality measures
6. detect trendy FoS in Computer Science field
7. researchers who are involved at the early stage of an FoS
8. detect FoS switching in scientific author's career years
9. impact of following FoS trend on the careers of scientific authors

In the following chapter, we will elaborate on how to close these research gaps by explaining the proposed approaches in detail.

Chapter 3

Proposed Methodology

3.1 Overview

The previous chapter presents a detailed discussion on the literature related to different aspects related to FoS trends. The research gap in the literature review forms the basis of our problem statement which then leads to the research questions. This chapter describes the proposed methodology that we have adopted to answer the research questions. In this thesis, we have used Microsoft Academic Graph (MAG) dataset; a well-known dataset that has already been used in many studies [8]. We have used MAG dataset to analyze FoS trend and their impact on research paper citations and authors citation count.

This chapter describes the proposed methodologies for research questions 1-3. To identify the significance of FoS trend on research paper citations (RQ-1), we have performed clustering on FoS and citations pattern separately. We presented a novel method of Field of Study Multigraph (FoM), formed by using centrality measures; degree, betweenness and closeness to analyze the FoS trend, citation trend and the relation between research areas in Computer Science scientific articles. The frequency of FoS in papers is also calculated to detect FoS trend. Rand Index value is computed to find the similarity between two data clustering's to analyze the impact of FoS on citation count. Finally, we have used the correlation coefficient to find the nature of a relation between FoS and citation patterns.

We have proposed an approach to detect researchers who are involved at the early stage of an FoS trend (RQ-2). First, we calculated the debut year of an FoS. Then, we have computed the FoS publication count, its author count and FoS trend by using FoM with degree centrality measure. Afterwards, we applied Rogers [129] for the detection of trend setters and followers. Lastly, we have compared our list of researchers (trend setters) with two existing lists that contain highly recognized Computer Science scientists. The lists are as follows; (i) top 10 influential authors identified by [91] and (ii) an existing list of Computer Science scientists with H-index of 40 or higher (www.cs.ucla.edu/ palsberg/h-number.html)

Finally, to detect the impact of FoS on authors citation trend (RQ-3), we have proposed an approach that detect the FoS trend of an individual author in his/her career years by characterizing the relations between his/her publications and citations. We detected the FoS of authors, then we selected those authors who follow the maximum trend of an FoS. Further, we calculated the citation count of authors and we computed the citation trend of authors. We used the citation trend of authors as input and predict the next year citation trend as output by using Multiple Linear Regression (MLR) and Artificial Neural Network (ANN). We detected the FoS trend of authors by using the field of study multigraph (FoM), formed by using degree centrality measure. We have also applied MLR and Artificial Neural Network on FoS degree values to predict the citation count.

The structure of the chapter is as follows: Dataset description is discussed in section 3.2. Proposed Methodology for RQ-1 is presented in section 3.3-3.4.3. RQ-2 overview and its proposed methodology is discussed in section 3.5-3.5.4. RQ-3 introduction and proposed methodology is presented in 3.6-3.6.3.

3.2 Dataset Description

The dataset employed for this study is taken from Microsoft academic [8] and is known as Microsoft Academic Graph (MAG) dataset which contains information about different academic articles, fields of study and the association between academic articles. The academic articles include conference papers, journal papers and books. The data about these articles include id, title, authors.name, venue, year, keywords, FoS, n_citation, references, doc_type, publisher, doi, and abstract as shown in Table 3.1 below.

TABLE 3.1: MAG articles schema.

Field Name	**Description**	**Example**
Id	MAG ID	00000707-26a7-491e-85b2-31063816253a
Title	paper title	The Research and Application of Resource Dissemination Based on Credibility and UCON
authors.name	author name	Fengying Wang, Fei Wang
Venue	paper venue	Computational intelligence and security
Year	published year	2007
Keywords	keywords	['mirrors', 'technological innovation', 'certificate authority', 'image databases', 'computational intelligence', 'trust management', 'contracts', 'fuzzy set theory', 'usage control', 'access control authorization fuzzy set theory image databases mirrors contracts computational intelligence security fuzzy systems technological innovation', 'access control models', 'membership function', 'authorization', 'access control', 'security', 'fuzzy systems', 'digital right management ']
FoS	fields of study	['Membership function', 'Computer Science', 'Knowledge management', 'Artificial intelligence', 'Information security', 'Access control', 'Computational intelligence', 'Data mining', 'Authorization', 'Fuzzy set', 'Computer security', 'Certificate authority ']
n_citation	number of citation	50
References	citing papers' ID	['2d2bcea7-33f9-4f58-81dc-34eacc8d5945','61a8529d b737-49e2-9c78-69040374bc8f','6e01112a-8c85-4252 ac08-f7977ae449a2','6ea51a04-50da-4621-b7d7-f89bee33ac26','727312a4-6798-4195-98d63d3b84965c5f', '93b8b7b4-6d81-4d59-bb3b-fdfa494eff41','b04b629f19ef-447e-806d-90644a78d670','b28a681c-35e9-4c2d-aab6-d8208b1cb55a','c13e496c-b997-4b83-ae464fe86768f891', 'cae74a58-fd52-45af-94fd-957939759810','e07bd0fa690d-4ea5-a693-72b49b11254e ']
doc_type	paper type	Conference
Publisher	publisher	IEEE
Doi	Doi	doi:10.1109/cis.2007.47
Abstract	abstract	Based on the concept of credibility

The academic articles in MAG are from multiple disciplines such as Physics, Computer Science, Engineering, Chemistry, and many others. The statistics about overall data and data specific to Computer Science are thus, can be separated by the FoS of each paper without analyzing the paper content or abstract of the paper.

Table 3.2 shows the MAG dataset statistics about multidiscipline and Computer Science.

Table 3.2: MAG dataset count of multidiscipline and Computer Science entities.

Entity	Total Count	Computer Science Count
Papers	228,956,810	1,354,603
Authors	231.969,837	2,324,591
Conferences	4,414	1,277
Fields of Study(FoS)	50,007	9,800

MAG investigated the communication activities within the academics, representing them as a diverse graph containing six distinct types of entities [8]. These entities comprise fields of study, authors, institutions (author affiliations), papers, venues (such as journals and conference and events.

The relationships among these entities are quite straightforward and intuitive. For example, the association between papers and venues is evident since papers are published in journals/conferences, which justifies the connection between paper and venue nodes in the graph.

Authors gather information about paper and author entities from two main sources: (1) feeds provided by reputable publishers such as ACM and IEEE, and (2) webpages that are indexed by Bing. While the bulk of our data originates from the indexed pages, it's worth noting that the data obtained from the feeds of established publishers generally exhibits higher quality and reliability.

Authors label the field of study (FoS) entities in their in-house knowledge base where their type is currently missing. To achieve this, authors employed a seeding process using two sources. The first source consists of entities already labeled as FoS in the knowledge base. The second source involves identifying potential FoS entities by matching their names with the keyword attributes in paper entities.

With these ”seed” FoS entities, then leverage the relationships present in in-house knowledge base. These relationships are calculated based on entity contents, hyperlinks, and web-click signals. The goal is to identify new FoS candidates that may not be explicitly labeled as such but are highly related to existing FoS entities.

To do this, authors assess the entities that are related to known FoS entities but lack any explicit type labeling. These entities are considered candidates for being potential FoS entities. Then, use a classification approach based on the ratio of the number of entities of the same FoS type in their top N related entities to N. This allows us to create a final list of FoS entities, significantly expanding the size of FoS entity collection.

Based on initial testing, this process has shown promising results. Then, identified new FoS entities with a significant increase in the overall FoS entity count, expanding it by twenty times. Moreover, sample results indicate that the accuracy of the identified new entities is above 98 percent.

Conference-related entities are gathered from several semi-structured websites, which are indexed by Bing and act as central platforms for conference organizers to announce their latest calls for papers. These websites primarily contain information about individual conference instances, such as ”WWW 2015,” but they may occasionally feature notices for journal special issues as well. To consolidate the data and establish relationships, authors merge similar conference instances from different websites and identify the corresponding conference series (venue). This process involves using various signals, including acronyms, full names, years, locations, and other relevant data extracted from the semi-structured content [8].

In MAG, FoS determines the area of focus of a particular scientific article. For instance, a paper focusing on comparison between different machine learning algorithms like, Support Vector Machine SVM and Naïve Bayes etc. will belong to the FoS, ”Machine Learning” or ”Artificial Intelligence” [2]. Every paper in MAG has a unique ID and is mapped to one or more associated FoS in the multiple levels of MAG hierarchy i.e. level-0 to level-3 as presented in Figure 3.1.

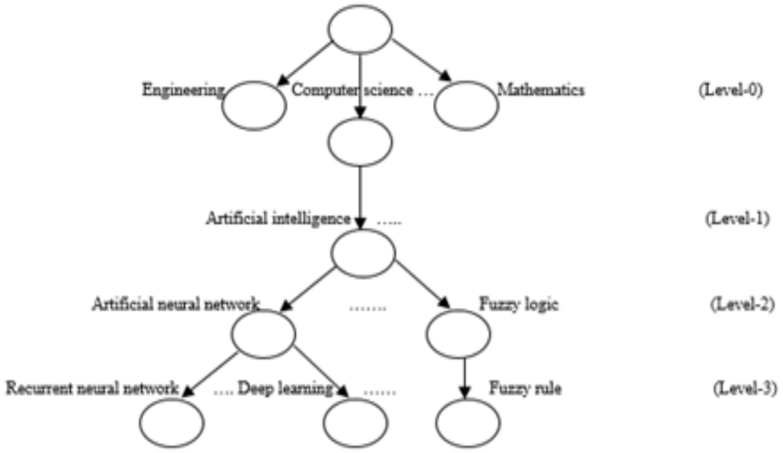


FIGURE 3.1: MAG different levels.

Figure above shows a snippet of the MAG hierarchy from level-0 to level-3. Computer Science field lies at level-0 and there are total 35 FoS at level-1 of Computer Science. Level-0 contains FoS at a more generic level, like Engineering, Computer Science, etc. The lower levels contain more specific FoS as shown in the Figure 3.2.

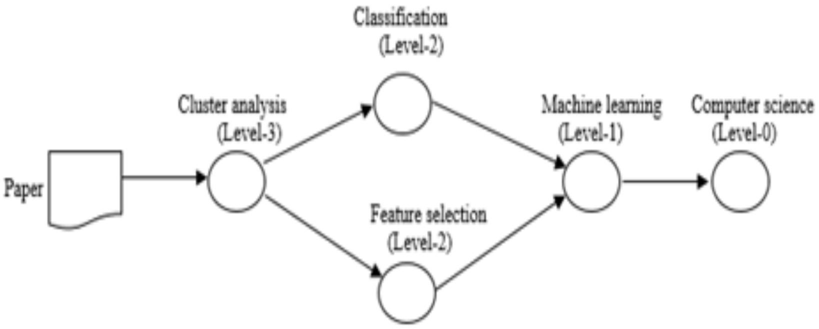


FIGURE 3.2: An example of Computer Science FoS levels.

An example of mapping is shown in figure where a paper from the domain of Computer Science is mapped to different FoS from level-3 to level-0. In general, the structure of the FoS in MAG is in the form of a directed acyclic graph, i.e., an FoS may have more than one parent FoS. For example, Cluster analysis (level-3), belongs to Feature selection (level-2) and Classification (level-2) which belongs to Machine learning (level-1) and Computer science (level-0).

3.2.1 Dataset Preprocessing

As explained earlier, the MAG dataset contains articles from different domains. For this study, we have selected the research papers from the field of Computer Science published during 1950-2018. Even though the MAG contains the papers that are published in journals and conferences. In this research, we have considered studies which are published in journal and conferences as shown in Table 3.3. Dataset preprocessing of research questions 1-3 are discussed in detail in their relevant sections.

Table 3.3: Level-1 FoS of sampled papers.

Paper id	Year	Type	Publisher	Paper Title	Author	Level-0 FoS	Level-1 FoS
103729	1988	Journal	International Journal of Pattern Recognition and Artificial Intelligence	On automatic feature selection	W Siedlecki, J Sklansky	Feature Selection, Pattern Classifier, Pattern Recognition, Decision rule, Computer Science, Simulation	Pattern Recognition, Simulation
105642	1987	Journal	International Journal of Computer Applications	Survey of Expert Systems and the Cognitive Approaches towards an Effective Tutoring System	DK Chaturvedi, AP Prajapati	Expert Tutoring Systems, Pedagogy, Decision Making, Computer Science, Artificial Intelligence, Expert System	Artificial Intelligence
109827	1999	Journal	Computer Networks	Finding related pages in the World Wide Web.	Dean, J. and Henzinger, M.R	World Wide Web, Search Engine, Computer Science, Algorithm, Pattern Recognition, Web Page, Feature Selection.	World Wide Web, Algorithm, Pattern Recognition
100981	1962	Conference	Work	Standardised Clinical Datasets-Pre-Requisites to Successful Data Mining	AIMIS, Dean White MSc	Speech Recognition, Computer Science, Pattern Recognition, Data Mining	Speech Recognition, Pattern Recognition, Data Mining
100231	1970	Conference	In Space Optics	Computerized image dynamic analysis	F.B. Brown, K.W.Hering	Computer Vision, Simulation, Computer Science, Computer Graphics	Computer Vision, Simulation, Computer Graphics

Paper id	Year	Type	Publisher	Paper Title	Author	Level-0 FoS	Level-1 FoS
101432	1970	Journal	The Journal of Internet Banking and Commerce	Quality of Web-based information systems	Worwa, Kazimierz and Stanik, Jerzy	Web-based information systems, Web engineering, computer science, Web-based software, world wide web, software quality modeling	Web-based information systems, Web engineering, Web-based software, world wide web, software quality modeling

3.3 Proposed methodology-RQ-1

RQ-1: How similar is the citation trend of papers belonging to the same FoS?

We proposed an approach to identify the significance of FoS trend on research paper citations. The answer to this question will lay the foundation for our next questions where we try to establish nature of association between the FoS and citation trend of author working in that FoS. This research question will establish the similarity between the citation trends of authors belonging to same FoS. To answer this question, we collected papers from MAG dataset belonging to different FoS, and listed the citation patterns of all papers for five years (section 3.3.1). We then performed clustering on FoS and citations trend of papers separately (section 3.3.2-3.3.3). We compared the similarity between these two clusters.

Our proposal is that if there is reasonable level of similarity between these two clusters then it means there is an association between FoS and the citation pattern of papers. Rand Index (RI) has been used to compare the similarity between the two clusters (detailed in section 4.1.1). We performed another experiment for the same RQ with a novelty that we proposed an FoS Multigraph (FoM) from where we computed different centrality measures. Then, we used these centrality measures in the same experiment with the objective to find a better metric to establish similarity in the trend of papers belonging to same FoS. Once again we used RI for the purpose (section 4.1.1).

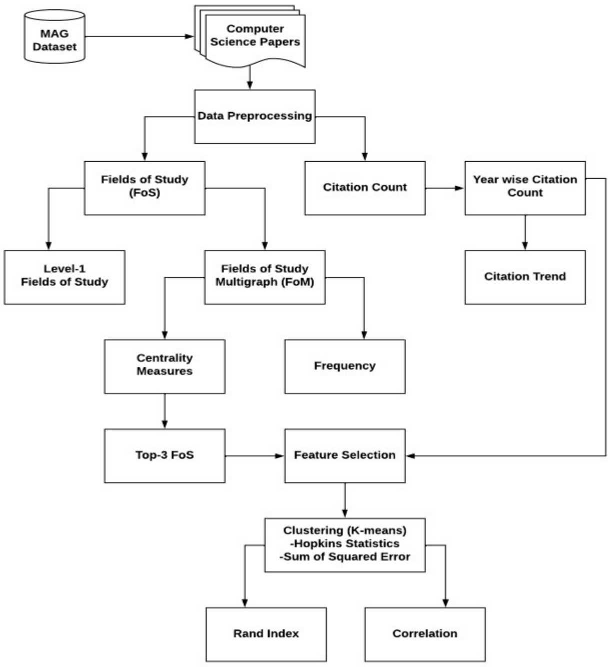


FIGURE 3.3: The proposed methodology.

Figure 3.3 is a graphical representation of modules of the proposed methodology of RQ-1. The proposed methodology describes the FoS extraction process, clustering technique, FoS clusters, relationship between different FoS, FoM graph construction by using graph centrality measures: degree, closeness and betweenness. Further, presents rand index and correlation as evaluation metrices.

3.3.1 Data Collection

This approach works on FoS of level-1 because it is the earliest and most generic distribution of FoS of a particular domain of knowledge [2]. The FoS in MAG becomes more specific when we move down in the hierarchy. After getting the level-1 FoS of Computer Science papers, we have stored the paper id, year, title, FoS, level-0, and level-1 FoS associated with the paper in a separate file named as FoS dataset as shown in Table 3.4.

Table 3.4: FoS of a sampled paper.

Paper ID	**Year**	**Title**	**FoS**	**Leve-0 (FoS)**	**Level-1 (FoS)**
24c7ef8ab2398h455217b64b	2007	Joint optimization of relay strategies and resource allocations in cooperative cellular networks	Cellular network, tele communications, computer science, base station, resource management, operating system, wireless network, relay channel, computer network	Computer science	Tele communication, operating system, computer network

To find out the association between the citation trend of papers and their corresponding FoS, we need to process our dataset to collect the yearly citation count of each paper and the number of publications for each FoS over the years. The MAG dataset does not contain the year-wise count of citations. For this purpose, we selected those papers that were published between 2007 and 2011, and calculated the yearly citation count of each paper for the next five years, as shown in Table 3.5 below.

Table 3.5: Yearly citation count of five sampled papers.

Paper Year	Level-1 FoS	Publication Year	yearly citation count				
			PY+1	PY+2	PY+3	PY+4	PY+5
p^1_{2007}	telecommunications, operating system, computer networks	2007	8	76	104	120	112
p^2_{2008}	World Wide Web, Computer Security, Computer Networks	2008	1	19	21	22	21

p^3_{2009}	Machine Learning, Data Mining, Artificial Intelligence, Simulation	2009	0	1	2	0	0
p^4_{2010}	Computer Vision, Simulation, Artificial Intelligence, Machine Learning	2010	0	4	9	12	15
p^5_{2011}	Data Mining, Database, Machine Learning, Information Retrieval	2011	4	10	20	25	16

In the above table, the first column shows the paper number and its publication year, the second column illustrates the level-1 FoS associated with the paper. The third column contains the publishing year, the next five columns contain the citation count of papers over the next five years. After calculating the citations pattern of an individual paper, we have calculated the citations pattern of each of 34 level-1 FoS of Computer Science. For this purpose, we have summed the citation count of papers belonging to different FoS. Table 3.6 shows the citations pattern of some of FoS over five years.

Table 3.6: FoS citation count.

Level-1 FoS	yearly citation count of Different Level-1 FoS				
	2007	2008	2009	2010	2011
Machine Learning	1283	844	1214	1412	1733
Data Mining	979	1039	1775	1836	1144
Computer vision	970	550	1023	1131	1108
Artificial Intelligence	919	887	1084	1084	1554
Operating System	885	534	663	992	748
Theoretical Computer Science	820	433	551	992	644

Finally, we replaced the FoS associated with each paper with the citation count of FoS for the publication year of the paper. Out of those citation counts, we picked the top three ones. The example of pre-processed data used to perform experiments is shown in Table 3.7 below.

Table 3.7: Papers with their citation counts and those of associated FoS.

$Paper_{Year}$	Yearly Citations of FoS			Yearly Citations of Papers				
	Top1	Top2	Top3	2007	2008	2009	2010	2011
p^1_{2007}	885	696	530	1	1	2	2	4
p^2_{2007}	884	854	696	5	7	8	12	14
p^3_{2007}	1283	979	919	4	1	3	5	5
p^4_{2007}	1283	970	919	1	3	1	1	3
p^5_{2007}	1283	979	745	3	7	12	15	20

The example of pre-processed data used to perform experiments is shown in Table 3.7 below. In this table, five papers published in the year 2007, the citation count of the top three associated FoS for 2007, and the citation count of each paper for the next five years, are shown as an example. The prepared data set contains the papers published from 2007-2011. In the next section, we have presented our approach to investigate the similarity between FoS and citations pattern.

3.3.2 Clustering

We have applied the clustering technique to analyze the impact of FoS on citation count of papers. Clustering is a method of grouping similar objects (commonly signified as a vector of measurements) into different clusters based on similarity. Clustering analysis is one of the key analytical methods in data mining. The clustering technique is mainly appropriate for the studies focusing on capturing inter-relationships amongst the data items [100]. This study forms two different sets of clusters to address the RQ-1.

In one set of clusters, a 5-year count of citations of papers is considered as the feature set and in the other set, we used the citation count of top three level-1 FoS associated with papers. Thereafter, similarity between two sets of clusters is calculated using rand index and correlation. Before applying clustering, we first analyzed the clustering tendency of our dataset. For this purpose, Hopkins Statistic H was chosen. This is a spatial statistic that tests the spatial randomness of a variable as distributed in a space [126]. The equation is as given below.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} xi + \sum_{i=1}^{n} yi} \tag{3.1}$$

where: H is the Hopkins statistic value ranging from 0-1, n is the number of points in dataset, yi is distance from the the i-th data point, xi is the distance from the i-th randomly generated point.In the test, if data is uniformly distributed $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i$ would be near to each other, and therefore H would be about 0.5. This test is conducted iteratively using 0.5 as a threshold. If the value of H is less than 0.5, it means that data does not have statistically significant clusters. If the value of H is close to 1, this means that the data can significantly form clusters. We have computed H for our dataset separately on the citation pattern of papers and also the citation count of FoS. This has been computed year-wise for all the papers. All the values of H were more than 0.5 suggesting that our dataset tends to form meaningful clusters. Table 3.8 shows the value of H calculations.

Table 3.8: Hopkins statistic values for two feature sets.

Year	Citation Count	
	FoS	Papers
2007	0.7	0.6
2008	0.7	0.6
2009	0.7	0.6
2010	0.7	0.7
2011	0.7	0.7

As indicated by the values of H, our dataset has a reasonable tendency for clustering. We have applied k-means clustering on Computer Science papers for five different years with two different selected feature sets, which are yearly citation counts of corresponding FoS and papers' citation counts as shown in table (above). Afterwards, the similarity between the two sets of clusters is calculated for evaluation.

K-means clustering [128] is a partition-based cluster analysis method. According to this algorithm, first, we randomly selected k data values as initial cluster centers or centroids, then calculated a proximity metric (generally Euclidean distance) between each data value and each centroid and assigned it to the closest cluster, after that we updated the averages of all clusters, repeated this process until the criterion is not matched. K-means clustering aims to partition data into k clusters in which each data value belongs to the cluster with the nearest mean. The equation used for Euclidean distance is:

$$d = \sum_{k=1}^{k} \sum_{i=1}^{n} ||x_i - u_k||^2 \tag{3.2}$$

where k signifies k cluster centers, u_k signifies the k^th center, and x_i represents the i^th the point in the dataset. The value of k, in K-means, is set by evaluating Sum of Squared Error (SSE) with different values of k generally starting from 2 and moving onwards. For our experiments, the graph between the value of k and corresponding SSE is shown in Figure 3.4.

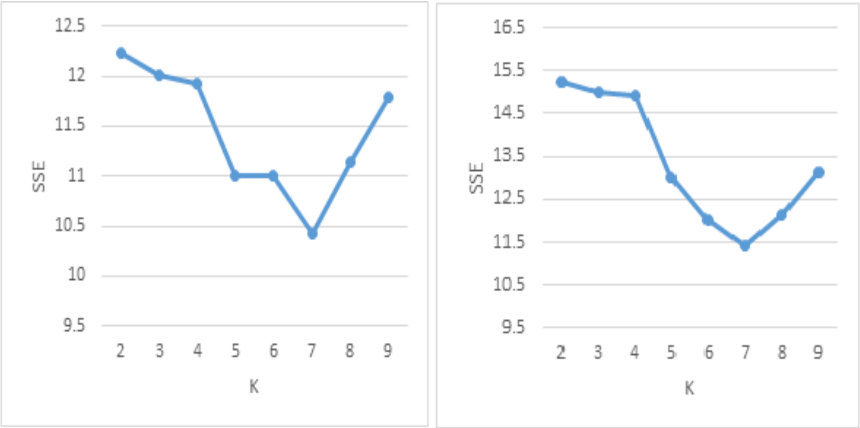


Figure 3.4: The relationship between SSE and the value of k for citation count(left), FoS(right).

As per this diagram, the value of SSE falls with an increase in the value of k and it rises at 8. Therefore, we set the value k as 7. After applying K-means clustering on citation counts of FoS with k equals to 7, a total of seven clusters were formed.

3.3.3 Field of Study Clusters

The clustering results show the interaction of certain FoS with each other. We can see this with the interaction between co-appearance of FoS in a research paper and similar citation trends as they are clustered in the same group. We can see this with the interaction such as co-appearance of FoS in a research paper as shown in Table 3.9. In particular, in research fields interdisciplinary interactions such as Machine Learning, Data Mining, Data Science, FoS may co-exist within one article, and the relationship between FoS may be important information. Therefore, it is essential to analyze the FoS that has a great influence on other FoS, such as the relationship between FoS, and the FoS that co-exists in articles.

Table 3.9: Grouping of different FoS based on the similarity of their citation count patterns.

Clusters	Field of Study (FoS)
Cluster0	Distributed Computing, Real-time Computing, Operating System, Parallel Computing.
Cluster1	Artificial Intelligence, Machine Learning, Computer Vision, Simulation.
Cluster2	Computer Security, Computer Networks, World Wide Web, Telecommunications.
Cluster3	Data Mining, Data Science, Database, Machine Learning.
Cluster4	Theoretical Computer Science, Algorithm, Computer Vision, Computer Graphics.
Cluster5	Operating System, Telecommunications, Computer Networks.
Cluster6	Machine Learning, Data Mining, Database, Information Retrieval.

As it can be seen from the above table that cluster0 comprises following FoS of level-1: ”Distributed Computing, Real-time Computing, Parallel Computing, Operating System”. These combinations look very natural, e.g., there is a possible relationship between the Distributed Computing, Real-time Computing, and Parallel Computing. These FoS usually occur together in the majority of research publications and both FoS seem to be more equal in terms of influence on each other. We can also observe that similar FoS shows similar citation trends of papers as they are clustered in the same group. Cluster1 comprises these FoS: ”Computer Networks, Real-Time Computing, Operating System, Telecommunications”. We have also generated 7 clusters based on the citations pattern of the papers as shown in Table 3.9. We discussed results in the next chapter section 4.1.2.

3.4 Field of study trend and relation between research areas

In this thesis, we have used a multigraph with centrality measures to measure an FoS trend other than citation count. Since most of the papers in our dataset correspond to more than one FoS, which establish a link or relation between them. One possible approach to explore the significance or trend of an FoS other than the citation count could be the co-occurrence of an FoS with other FoS. More an FoS co-occurs with other FoS, more significant or trendy it is. The graph is a natural representation of such links between objects providing different centrality measures to measure the significance of objects within the graph.

For this purpose, we propose to construct an FoS multigraph (FoM) from the articles. Next, the trend of each FoS can be determined using graph centrality measures. In this study, we have applied three classic centrality measures (degree centrality, closeness centrality, and betweenness centrality). These centrality measures have been evaluated in the context of FoS. Lastly, these metrics are considered as FoS trend metrics and compared with the results obtained for the citation count (Table 3.6).

3.4.1 Field of study multigraph (FoM) construction

A field of study multigraph (FoM) is built from the FoS of Computer Science papers. A multigraph is permitted to have multiple edges (also called parallel edges) between two nodes. Thus, two vertices (nodes) may be connected by more than one edge. A multigraph is a set of vertices, V which shows FoS, a set of edges, E which shows relation ship between different FoS, and a function $f : E \longrightarrow u, v : u, v \epsilon V and u \neq v$ The significance of every FoS is then resolute

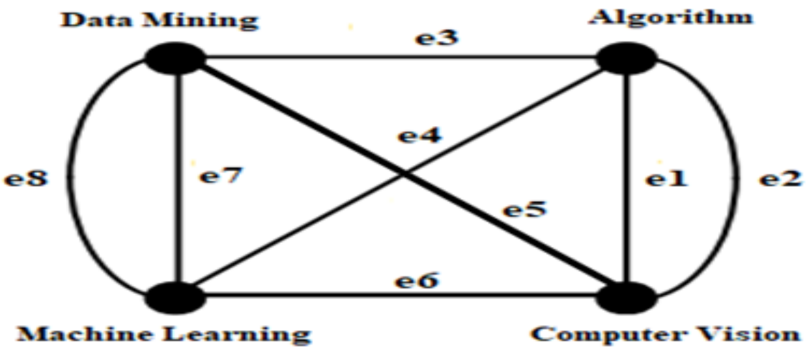


Figure 3.5: FoM for three example papers.

using graph centrality measures and papers are categorized based on the FoS they comprise. The construction of the FoM graph is principally based on the FoS which are enclosed in a research paper and their vicinity. Each FoS that is enclosed within the research paper is signified by a system of a labeled node. The edges are focused to grab the structure of the FoS as they occur inside the research papers (relationship of FoS in the paper) as illustrated in Figure 3.5. The nearness between the FoS is signified by the edges that join the nodes and is defined using an explicit extensive diversity of FoS. As an example, let us suppose three papers with their corresponding FoS, as given below.

Paper1 FoS: Algorithm, Computer Vision.

Paper2 FoS: Algorithm, Computer Vision, Data Mining, Machine Learning.

Paper3 FoS: Data Mining, Machine Learning.

The FoM for the above papers is shown in Figure 3.5. The FoM shows that Algorithm is connected to Computer Vision, Machine Learning and Data Mining. Similarly, Computer Vision is connected to Algorithm, Machine Learning and Data Mining, and Data Mining is connected to Algorithm, Computer Vision, and Machine Learning. Likewise, Machine Learning is connected to Algorithm, Data Mining, and Computer Vision. Algorithm and Computer Vision have parallel edges $(e1, e2)$ as these FoS have appeared in paper 1 and paper 2. Similarly, Data Mining and Machine Learning have parallel edges $(e7, e8)$ as they appeared in paper 2 and paper 3. As soon as the FoM graph is constructed, centrality measures including degree, betweenness, and closeness are computed for each node by using the formulas shown in equation 3.3, 3.4 and 3.5 in the following sections.

Once the FoM is constructed, centrality measures are calculated to assign a value to each node. Let $G = (V, E, f)$ be a multigraph with a set of vertices (FoS) V, a set of edges E and f mapping edges between nodes. Starting with degree centrality, this section describes all the centrality measures employed in this study.

1. **Degree centrality:** is defined as the number of edges incident upon a node. Applied to FoM, the degree of a node v_i represents the number of FoS that co-occur with the FoS equivalent to v_i. Let $C_D(v_i)$ be the degree centrality of a node v_i is given by [130]:

   $$C_d(v_i) = deg(v_i) \tag{3.3}$$

   Generally, vertices with a higher degree or more connections tend to have a greater capacity to influence others. In the context of FoM, the value of degree centrality indicates the co-occurrence of a node (FoS) with other FoS in different papers which may be considered as influence or trend of that FoS.

2. **Closeness centrality:** measure the node centrality in a connected graph and calculated as the sum of the length of the shortest paths between the node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes in the network. Let distance (v_i, v_j) be the shortest distance between nodes v_i and v_j. The closeness centrality of a node v_i is [130]:

   $$C_c(v_i) = \frac{1}{\sum_y distance(v_j, v_i)} \tag{3.4}$$

   The degree centrality signifies the importance of a node (FoS) based on its direct connections with other nodes (FoS), whereas the closeness centrality covers both direct and indirect connections of an FoS showing how central a node in the FoM is.

3. **Betweenness centrality:** is a measure of centrality in a graph based on the shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through. This computes the number of times an FoS (node) behaves as a bridge alongside the shortest path between two other FoS (nodes). Here, $\sigma(s_t)$ is the total number of shortest paths from node s to node t and $\sigma(s_t, v)$ is the number of those paths that pass through v [130].

$$C_B(v) = \sum s \neq v \neq \frac{\sigma(s_t, v)}{\sigma(s_t)} \tag{3.5}$$

Being between means that a node can act as a bridge to provide flow of knowledge between most of the nodes in a network. FoS with high betweenness are the pivots in the network knowledge flowing. The nodes with the highest betweenness also result in the largest increase in typical distance between others when they are removed.

Table 3.10: Top-10 centrality measures.

FoS	Centrality Measures for year 2007		
	Degree	Betweenness	Closeness
Algorithm	2150	0.9393939	0.01053
Artificial Intelligence	3200	0.9487532	0.0194119
Computer Networks	2925	0.9393939	0.0172436
Computer Vision	2680	0.9257143	0.0183707
Data Mining	2755	0.9211111	0.0182324
Database	2435	0.9193939	0.0100324
Machine Learning	3064	0.9117647	0.0194327
Operating System	2720	0.9293939	0.0105444
Theoretical Computer Science	2387	0.9387543	0.0128119
World Wide Web	2545	0.9193939	0.0191463

After constructing FoM, we calculated the degree centrality measures for all nodes of FoM (representing FoS) starting from the year 2007 till 2011. Table 3.10 shows the values of centrality measures of top-ten trendy FoS for the year 2007.

3.4.2 Citation Trend of Trendy FoS

Bibliometric analysis is used to identify citation trends from various aspects. Citation analysis is a bibliometric method used to reveal different patterns of the scientific community. Researchers can measure the significance of their publications with the help of citation analysis. They may gain facts about that paper's effect on its field by calculating the number of times it has been cited in research publications. Additionally, the citation trend is a good measure to analyze the impact of a research publication as high count of citation specifies usefulness and effectiveness.

A citation trend **ct** is the group of citation sequences sharing a common pattern of evolution of citation count. Citation sequences of various citation trends show different evolutions of citation count [6]. A citation-sequence of a research paper **p**, indicated as, $s_{\Delta t}(p) = [c_1(p), c_2(p), \ldots c_{\Delta t}(p)]$ is a sequence of citation count $c_i(p)$ over a period of time $1, 2, 3, \ldots t$, where c_i is the citation count of the i^{th} year after **p** gets published. For a collection of research papers, given a paper $p \epsilon P$, its citation count $c(p)$ is the number of papers that cite **p**, denoted by, $c(p) = |p' \epsilon P : p' cites p|$.

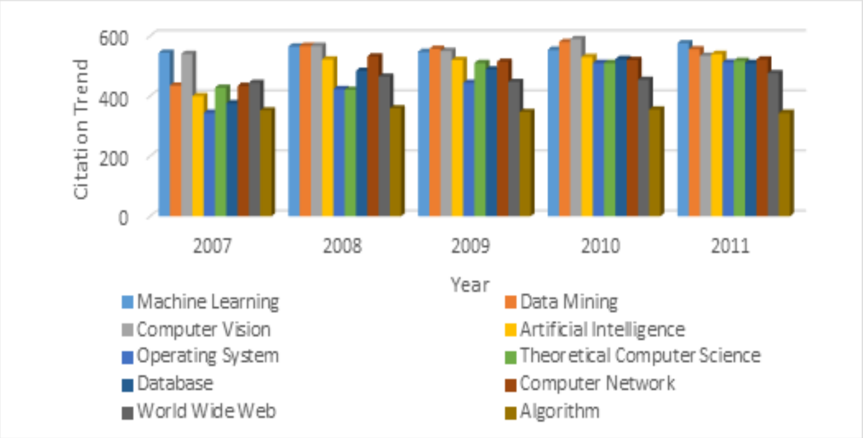


FIGURE 3.6: Top-10 trendy FoS citation count.

An FoS receiving comparatively high citation count is considered the more influential FoS as it is being followed by more researchers in more papers. This can be used to establish an order among FoS and the ones at top most levels can be called as trendy or popular FoS [6]. We derive our definition of Trend of an FoS over time t from the work of [6] as follows: Let F be set of all FoS defined in MAG, F = f1, f2, ..., fn, the FoS of a paper F(p) is defined as;

$$F(p) = f_i(p)|pcontainsf_i \tag{3.6}$$

Similarly, we define papers of an FoS as;

$$P(f) = p_1, p_2, ...p_n(p)|p_icontainsf \tag{3.7}$$

Then, trend of an FoS f is defined as scintometric value of (c) of that FoS for the next t years of its publication years as;

$$s_{\Delta t}(f) = [c1(f), c2(f), \dots c_{\Delta t}(f)] \tag{3.8}$$

where scintometric value of a particular year for an FoS is sum of values of all papers belonging to that FoS as given in the last equation below:

$$c(f) = \sum(c(\pi) : \pi \epsilon P(f)) \tag{3.9}$$

$$c(p) = |p' \epsilon P : p' citesp| \tag{3.10}$$

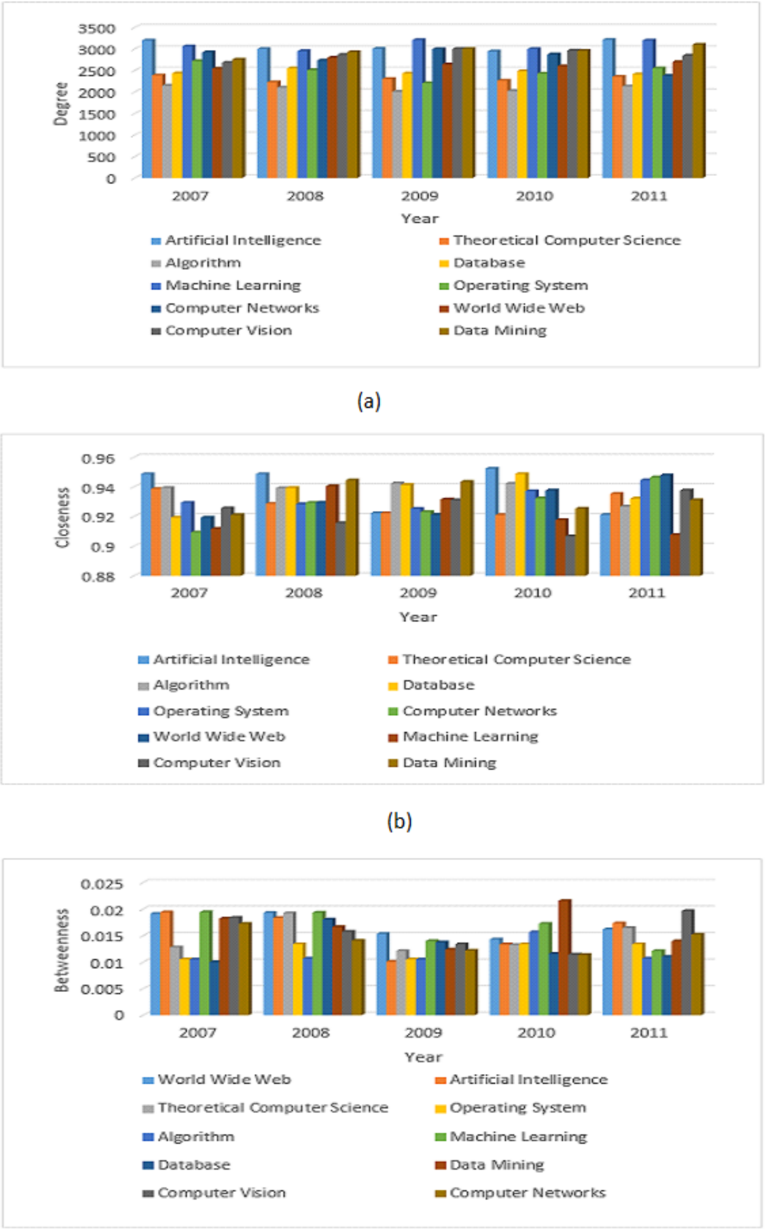
Here, the goal of bibliometric classification is to evaluate the citation trend of top-10 FoS in the Computer Science area. Counting the number of citations for each paper (where top FoS appears) and then calculate total citations of top-10 FoS give the FoS citation count, as shown in Figure 3.6. This exposes the impact and worth of the scientific research field. Machine Learning, Data Mining, and Computer Vision have the highest citation count in 2007, 2008, and 2010. Machine Learning, Artificial Intelligence, and Computer Networks have achieved maximum citation count in 2009. Whereas, Machine Learning, Data Mining, and Artificial Intelligence have the highest citation count in 2011.

3.4.3 Trendy FoS

As mentioned above, the previous work on Trendy FoS is based on citation counts of the papers, however, the centrality measures computed from FoM can also be used as the basis of Trendy FoS. The reason being that these measures demonstrate the link or interaction of and FoS with other FoS, that can be used to determine the trend. By analyzing the constructed FoM (section), we found the FoS with the highest degree, closeness, and betweenness to understand the trends of FoS over the time. Figure 3.7 (a-c) shows the top-ten trendy FoS with repect to different centrality measures, that is, degree, closeness and betweenness centrality.

As we can see in Figure 3.7(a), Artificial Intelligence, Machine Learning, and Computer Networks have a maximum degree in 2007. Artificial Intelligence, Machine Learning, and Data Mining have achieved a high degree in 2008. Artificial Intelligence, Computer Vision and Data Mining have a high degree in 2009. Machine Learning, Computer Vision and Data Mining have a maximum high degree in 2010. However, Artificial Intelligence, Machine Learning, and Data Mining attained a high degree in 2011. Closeness centrality shows the top-10 trendy FoS as shown in Figure 3.7(b). Artificial Intelligence, Theoretical Computer Science, and Algorithm has a maximum value in 2007. Artificial Intelligence, Machine Learning, and Data Mining have a maximum value in 2008. Algorithm, Database, and Data Mining in 2009 has revealed the high value. Artificial Intelligence, Algorithm, and Database have achieved a maximum value in 2010. Whereas, Operating System, Computer Networks, and the World Wide Web has a maximum value in 2011.

Betweenness centrality shows the top-10 trendy FoS as shown in Figure 3.7(c). Betweenness means that a node can act as a bridge to provide flow of information between most of the nodes in a network. FoS with the high betweenness are the most influential nodes in the network. Artificial Intelligence and Machine Learning, World Wide Web has maximum betweenness value in 2007. World Wide Web, Theoretical Computer Science and Machine Learning have the highest value in 2008. World Wide Web, Machine Learning, and Database has a maximum value in 2009. Data Mining, Machine Learning, and World Wide Web achieved high value in 2010. Computer Vision, Artificial Intelligence, and Theoretical Computer



(a)

(b)

(c)

FIGURE 3.7: Top-10 Trendy FoS(a)degree, (b)closeness, (c)betweenness.

Science have a maximum value in 2011. Table 3.11 below shows the ordering of top-10 FoS in year 2007 (as an example) with respect to different metrices.

TABLE 3.11: Top-10 FoS order in 2007.

Level-1 FoS	Order of Top-10 FoS w.r.t Different Metrics in 2007			
	Frequency	Degree	Betweenness	Closeness
Machine Learning	1	2	9	4
Computer vision	2	6	5	7

Level-1 FoS	Order of Top-10 FoS w.r.t Different Metrics in 2007			
	Frequency	Degree	Betweenness	Closeness
Operating System	3	5	4	3
Database	4	8	7	5
World Wide Web	5	7	8	1
Data Mining	6	4	6	6
Artificial Intelligence	7	1	1	9
Theoretical Computer Science	8	9	3	2
Computer Networks	9	3	10	8
Algorithm	10	10	2	10

An interesting fact that can be noticed while analyzing the values of different metrics is that the top-10 FoS across multiple metrics are the same, however, their order among the top-10 values is different.

In this section, we discussed about the significance of FoS trend on research paper citations. We have presented the FoS trends from 2006-2011 using different metrices. The comparison of these metrices has been given in next chapter (section 4.1.2). In the next section, we describe the process of identification of scientific researchers at the early Stage of an FoS trend.

3.5 Identification of Scientific Researchers at the Early Stage of FoS Trends

After identification of FoS trends, it is possible to determine the researchers who were involved at the early stage of an FoS trend known as “trend setters” and the authors who followed it afterwards known as “trend followers”. Classifying authors into these two categories will help researchers to identify the influential authors in a specific FoS. Studying work of “trend-setters” of an FoS guides a researcher that how an FoS was originally conceived and proposed, the later review on that FoS will guide the stages it has gone through. For example, E.F Codd’s work [131] on Relational Data Model (2601 citations upto 2022) or Tim Berners Lee’s work [132] on Semantic Web (2190 citations upto 2022) gives real insight into these areas. That is why their work is still being cited heavily even today.

Classifying researchers at the early phase of an FoS is of importance as it will define who the noteworthy authors that started were or in growth the popularity of a particular FoS trend. For example, the Association of Computational Machinery (ACM) program distinguishes and regards researchers for their accomplishments in the Computer Science and Information Technology fields. The detection of researchers that are recognized as “trend setters” might assistance in defining the researchers to cogitate for such honors.

According to state-of-art approaches, as soon as an innovative scientific area of research emerges, it drives over two key stages. In the preliminary phase, a group of researchers come to an agreement on few elementary concepts, construct a theoretical background and instigate to form a new scientific community. Subsequently, the research area moves in an acknowledged phase, where ample number of researchers start working on it, creating and publishing results [133].

An approach [91] also highlighted the earlier phase, known as an embryonic phase, in which an FoS has not yet been clearly known and labelled by a research community. However it is now taking shape, as proved by the evidence that researchers from diverse fields are making new collaborations and creating new research, initially to explain the paradigms and issues related with the emergence/early phase of new FoS.

New FoS emergence at the early stage can bring noteworthy benefits to anyone involved in the research community. Academic editors and publishers can use this information and suggest the most recent and motivating contents. Researchers may not only be involved in new FoS trends associated to their fields however, they may also find it very beneficial to be notified about the progresses of important new research fields. Companies and institutional funding agencies also required to be frequently updated on how the research landscape is evolving, therefore that they can make initial choices about their important funds.

The goal of our study is to develop an approach that will detect ”trend setters” and ”trend followers” by identifying; (i) FoS debut year, (ii) FoS trend in papers and (ii) FoS trend of authors, at early stage by constructing a multigraph using degree centrality measure. We also focus on determining who were the researchers that published at the early stage of an FoS. This approach builds on the work of [91] where they identify the influential authors of an FoS in the embryonic stage, whereas our approach detects trend setters at the early stage of an FoS after its birth?

3.5.1 Proposed Methodology-RQ-2

RQ-2: How can we differentiate between trend setters and trend followers?

Classifying researchers at the early phase of an FoS is of importance as it will define who the significant authors that started were or in growth the popularity of a particular FoS trend. We have proposed an approach to detect influential researchers who were involved at the early stage of an FoS trend known as trend setters and the authors who followed it afterwards known as trend followers. The influential authors (trend setters) achieved high citation count and significance in a particular FoS. In our proposed approach, firstly, we have considered the debut year of an FoS as per approach of [91]. We selected the “Semantic Search” FoS because it has been discussed in our reference paper. From the debut year, our approach determines the trend setters through following steps:

1. We selected all the authors who published in the birth year of selected FoS. In this work, we have selected “Semantic Search” with birth year 2003 as it has been discussed in our base paper [91] (section 3.5.2).

2. Then, we computed the publication count of these authors for next five years in Semantic Search, their citation count for the papers on Semantic Search and the degree centrality of FoM for their papers on Semantic Search for next five years. We sorted three lists in descending order (section 3.5.3).

3. Afterwards, we applied Rogers Information of Diffusion Theories (IDT) [129] on three lists generated above. As per the Rogers IDT, top 2.5% authors are taken as trend setters and rest as different types of trend followers.

4. Lastly, we have compared our lists of researchers (trend setters) with two existing lists that contain highly recognized Computer Science scientists. The lists are as follows; (i) top 10 influential authors identified by [91] and (ii) an existing list of Computer Science scientists with H-index of 40 or higher (www.cs.ucla.edu/ palsberg/h-number.html).

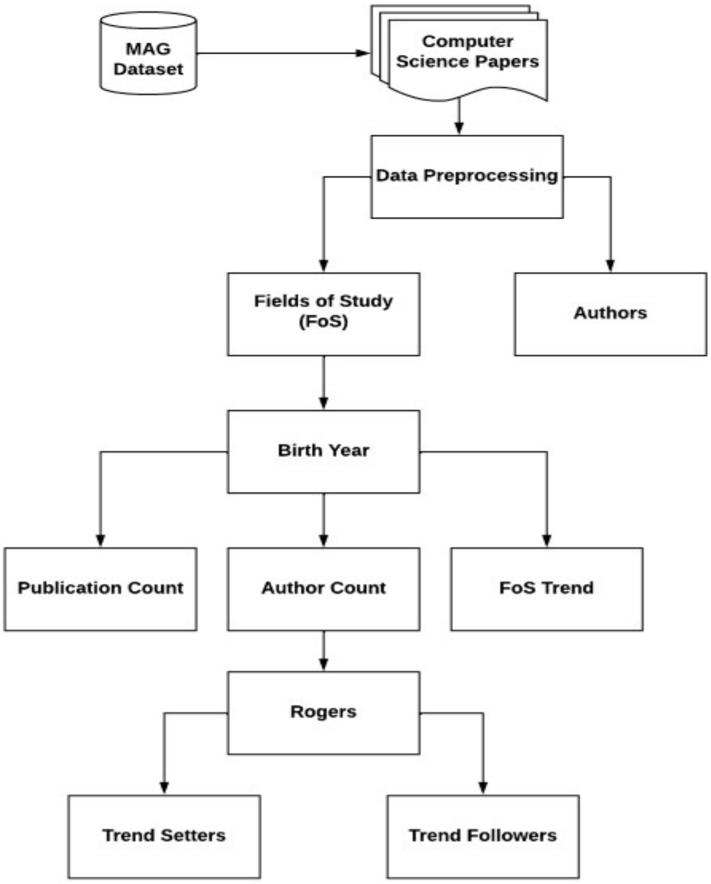


FIGURE 3.8: The proposed approach.

This section describes the proposed methodology of RQ-2. Dataset description is already discussed in section 3.2 in detail. In proposed approach FoS debut year is calculated first and discussed in section 3.5.2. Different counts for selected FoS is presented in section 3.5.3. In section 3.5.4, we discussed the emerging FoS rate of adoption, trend setters and followers. Figure 3.8 describes the modules of our proposed approach.

3.5.2 FoS Debut

Authors in [91] identified FoS belonging to the Computer Science, which emerged in the period from 2000-2009 as shown in Table 3.12 and all these FoS lies in level-2 in MAG dataset. The simplest way to detect the debut of an FoS is to consider the year in which the label of the FoS was used for the first time as keyword in a paper. For example, the FoS ”cloud computing”, first appeared in the year 2006. However, considering only the year in which its label firstly appeared as the year of debut can be misleading. An FoS label can in fact be discussed in few articles with some meaning and then become popular in later years with a totally changed meaning.

It is the case of ”linked data”, that initially was used in the context of databases to refer to pieces of data linked to each other before being adopted by the semantic web as a specific method for publishing data using the Resource Description Framework (RDF) format [91]. This label misuse can create significant noise. To handle this issue, authors choose as debut year of an FoS the first year in which it reaches at least 5 publications. At the same time, they named the previous five years of debut year as embryonic duration and from this duration they identified influential authors; those whose work ultimately gave birth to this new FoS in the debut year. In this way, they can be more certain that a new label is already recognized by multiple researchers.

Table 3.12: Selected debutant FoS and year [91].

Fos	**Year of Debut**
Service Discovery	2000
Ontology Engineering	2000
Ontology Alignment	2005
Service-oriented architecture	2003
Smart power grids	2005
Sentiment analysis	2005

Fos	**Year of Debut**
Semantic search	2003
Linked data	2004
Semantic web technology	2001
Vehicular ad-hoc networks	2004
Mobile ad-hoc networks	2001
P2P Networks	2002
Location-based services	2001
Service-oriented computing	2003
Ambient intelligence	2002
Social tagging	2006
Community detection	2006
Cloud computing	2006
User-generated content	2006
Information retrieval technology	2008
Web 2.0	2006
Ambient assisted living	2006
Internet of Things	2009

3.5.3 Different Counts for Selected FoS

As mentioned above, we are working on the Semantic Search (SS) FoS, following our base paper. We selected all those authors who worked in SS in the debut year, that is 2003. Table 3.13 below shows the number of authors and number of publications in SS in five years starting from birth year:

Table 3.13: Semantic search publication count and author count from 2003-2007.

Fos	Year	Publication Count	Author Count
Sematic Search	2003	232	545
Sematic Search	2004	313	756
Sematic Search	2005	421	959
Sematic Search	2006	482	1151
Sematic Search	2007	609	1413

As shown in the table above, there were 545 authors who worked in SS in 2003 (the birth year) and our proposal is that the trend setters for the SS are among these 545 authors. We selected all papers of these authors involving SS FoS for five years (2003-2007) and also the citation counts of those papers. Table 3.14 shows some of the authors for SS with their respective paper count and citation count:

Table 3.14: Semantic search authors publication count and citation count from 2003-2007.

Sr. No.	Researcher	Publication count	Citation Count
1.	Dieter Fensel	97	375
2.	Dan Suciu	62	323
3.	Justin Zobel	44	164
4.	James Allan	42	111
5.	W. Bruce Croft	29	109
6.	Dragomir R. Radev	28	67
7.	Alon Halevy	27	74
8.	Katia Sycara	26	60
9.	James Hendler	25	54
10.	Clement Yu	24	76
11.	Wolfgang Nejdl	23	48
12.	Victor Vianu	22	62
13.	Amit Sheth	20	45
14.	Andre Esteva	20	35
15.	Tom Gillespie	19	37
16.	Richard Christie	18	30
17.	Wenpeng Yin	17	32
18.	William W. Cohen	15	32
19.	Yuanzhang Li	15	34
20.	Berthier Ribeiro-Neto	13	22

After having these two lists, we computed the third list and that is the degree centrality measure of FoM constructed for each author against his work on SS for the years 2003-2007. For this purpose, we collected the papers of individual authors working in SS FoS during 2003-2007. For each author, we prepared co-occurrence data for the SS FoS.

Table 3.15 shows the SS co-occurrences and its degree with other FoS during a specific time period. This data has been compiled for one author's publications during 2003-2007.

Table 3.15: Semantic search co-occurrences with other FoS and its degree from 2003-2007.

Level-1 FoS	Semantic Search co-occurrence with other FoS	FoS Trend-Degree				
		2003	2004	2005	2016	2007
Semantic Search	Content-Based Retrieval	4	-	-	-	-
Semantic Search	Computational Semantics	3	2	4	4	3
Semantic Search	Semantic Equivalence	3	2	-	3	3
Semantic Search	Social Semantic Web	3	5	3	3	4
Semantic Search	Semantic Computing	-	3	-	5	5
Semantic Search	Digital Libraries	-	4	4	4	3
Semantic Search	Intelligent agents	-	-	6	-	5
Semantic Search	Explicit Semantic Analysis	-	-	3	3	4
Semantic Search	Similarity Heuristic	-	-	4	5	4
Semantic Search	Support Vector Machine	-	-	-	3	6
Semantic Search	Information Retrieval System	-	-	-	4	5

As can be seen from the table that SS appeared with content-based retrieval, computational semantics, semantic equivalence and social semantic web in 2003. Likewise, with computational semantics, semantic equivalence, social semantic web, semantic computing and with digital libraries in 2004 and other years. From this data we prepared FoS multigraph (FoM) for a particular author as discussed in section 3.4.1. From this FoM we computed the degree centrality of the SS FoS. In this way, we prepared the degree centrality for all authors working in SS FoS. Table 3.16 3.15 below shows the values of degree centrality of authors working in SS FoS.

Table 3.16: Authors SS degree from 2003-2007.

Sr. No.	Researcher	FoS Degree	Sr. No.	Researcher	FoS Degree
1.	Dieter Fensel	323	11.	Wolfgang Nejdl	63
2.	Dan Suciu	276	12.	Victor Vianu	54
3.	Justin Zobel	113	13.	Amit Sheth	61
4.	James Allan	101	14.	Andre Esteva	55
5.	W. Bruce Croft	82	15.	Tom Gillespie	51
6.	Dragomir R. Radev	50	16.	Richard Christie	49
7.	Alon Halevy	97	17.	Wenpeng Yin	39

Sr. No.	Researcher	FoS Degree	Sr. No.	Researcher	FoS Degree
8.	Katia Sycara	71	18.	William W. Cohen	35
9.	James Hendler	87	19.	Yuanzhang Li	31
10.	Clement Yu	88	20.	Berthier Ribeiro-Neto	29

The Table 3.16 contains degree centrality of the authors working in the SS FoS in year 2003 (debut year) and we compiled this table for their work between year 2003 and 2007.

So far, we have prepared three lists containing publication count and citation count (Table 3.14) and degree centrality of 545 authors working in SS FoS in year 2003. We are going to use these lists to find out the trend setters for the field of Semantic Search as explained in the next section.

3.5.4 Emerging FoS and Rate of Adoption

After the detection of FoS debut year, its publication count, author count and FoS trend. Now, it is possible to identify the researchers involved at the early stage of FoS trend and who followed FoS trend afterwards. An FoS in its debut year seems appears only in the papers of this time period and not in papers (back years), it can be now the early stage of an FoS. The authors involved at this stage of FoS are the trend setters or innovators and others are trend followers. We use Rogers IDT [129] to detect the trend setters and followers from the early stage of an FoS trend.

The above figure 3.9 shows the trend setters and followers categories as depicted in [129]. We applied Rogers IDT on the three lists that we prepared for 545 authors who worked on SS FoS in the debut year, that is, 2003. We have presented the trend setters, the five adopter classes and the estimated fraction of authors encompassed to each are positioned on the adopter dispersal. The part to the leftward of the mean time **x** of adoption minus two standard deviations **2sd** comprises the initial 2.5 percentage of the researchers intricate in the emergence of a trend the innovators or trend setters. The subsequent 13.5 percentage of researchers who

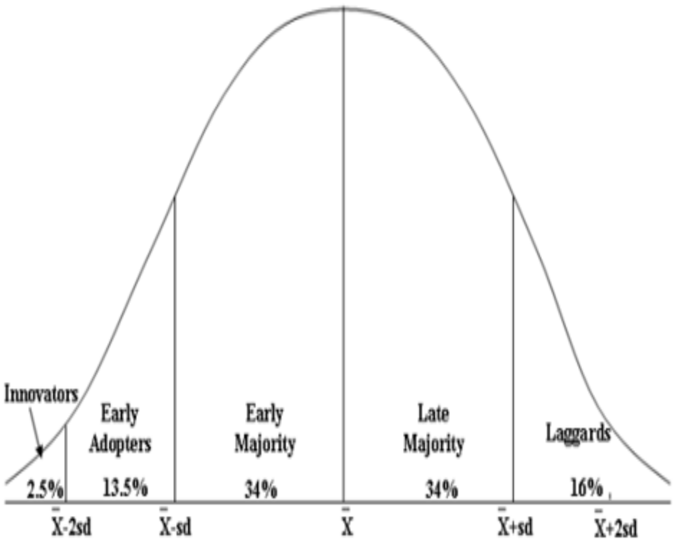


Figure 3.9: Rogers Innovation Diffusion Theory adopter categories [129].

adopt/accept the new trend are encompassed in the part among the mean minus one standard deviation **sd** and the mean minus two standard deviations; they are labeled early adopters.

The succeeding 34 percentage of the researchers are adopters, known as early majority, are comprised in the part among the mean time of adoption and minus **1sd**. Amongst the mean and **1sd** to the right of the mean are positioned the subsequent 34 percentage of authors to accept/adopt the trend, the late majority. The preceding 16 percentage of authors are known as laggards [129].

1. **Trend Setters/Innovators** According to Rogers [129] innovativeness is the value to which a person or a rate of adoption is comparatively earlier in accepting novel concepts than other fellows of an organization. Innovativeness guided in understanding the chosen and important behavior in the innovation-decision process. Therefore, he classifies the adopters on the basis of innovativeness. Trend setters/innovators are willing to experience new concepts and ideas. They act like the gatekeepers bringing the new concepts in from the outside of a system. Trend setters are capable to handle with advanced stages of ambiguity about a novelty than followers. Such as, they are the earliest to expose a new idea/concept in their method, they cannot be influenced by upon the particular assessments of the innovation from

other followers of their organization. Trend setters accept a novel idea there is nearly no one in the organization/ structure who has familiarity with the innovation. Trend setters who adopt an innovation as the first 2.5% of the individuals in an organization.

2. **Trend Follower Categories** Trend followers place their stamp of approval on a new idea by accepting it. They feel that it is safe to adopt now. These are the members of a system who wait till most of their peers adopt the innovation. There are also some individuals in a system who most need the benefits of a new idea are generally the last to adopt an innovation. Because of the limited resources and the lack of awareness-knowledge of innovations, they first want to make sure that an innovation works before they adopt. Therefore, they tend to decide after looking at whether the innovation is successfully adopted by other members of the system in the past. Due to all these characteristics, some followers innovation-decision period is relatively long [129]. Here, we are considering early adopters, early majority, late majority and laggards in trend follower categories.

   We have detected trend setters who are involved at the early stage of an FoS (only in debut year) as 2.5% of authors and trend followers who followed the FoS trend after debut year by applying Rogers [129]. As Table 3.17 shows trend setters and followers distributions.

   After the detection of researchers as trend setters and followers, we have calculated the trend setters publication count, FoS publication count and author FoS trend by using FoM with degree centrality measure.

Table 3.17: Semantic search trend setters and followers in 2003.

FoS	Debut Year	#Papers	#Authors	Trend Setters	Trend Followers
Semantic Search	2003	232	545	13.625%	86.375%

3.6 Quantifying the impact of FoS on author's citation trend

A significant aspect in scientific research is evaluating the impact of scientific contribution of a particular author. In literature, various metrices have been proposed for this purpose such as; h-index, citation, expertise, venue, versatility, social and network features such as; betweenness, closeness and PageRank. Authors select a particular Field of Study (FoS) to work in based on their interest and seeing the currently hot topics. One aspect that may encourage/support in the selection of their research area is the future impact of their work.

There are multiple studies in literature that focus on FoS trend detection and analysis; measuring strength of an FoS trend and of vanishing of an FoS trend etc. However, the previous work contains a gap of working on effect of following an FoS on citation trend of scientific authors. We have proposed an approach that detects the FoS trend of an individual author in his/her career years by characterizing the relations between his/her publications and citations. Hence our research question number three is:

RQ-3: What is the impact of FoS on authors citation trend?

The main focus in this question is that what is the effect of following FoS trend on the careers of scientific authors? Following the latest FoS trends provides researchers a good advantage in the fast-paced world and to be able to connect well professionally and also it creates a high impact on their paper citations and their careers. FoS trend following is significant for Computer Science researchers as they will be able to smartly decide at what could be coming ahead of the research fields in Computer Science and will help to make positive and insight decisions for the future of their research fields in this area. Similarly, researchers will aptly be able to anticipate the new FoS trends coming in Computer Science field, because they will already have a good idea of what is already coming. The significance of identifying FoS trends, a researcher could determine fields of interest with respect to its success or impact. The ability to recognize FoS trends is noteworthy for anyone involved in the research environment, including researchers, academic publishers, journal editors, institutional funding bodies and other relevant stakeholders.

This section encompasses details about RQ-03. Section 3.6.1 describes the proposed approach. Authors FoS extraction, publication count, citation count, citation trend and FoS degree trend discusses in section. 3.6.2,3.6.3.

3.6.1 Proposed Methodology-RQ-3

We have proposed an approach that compares the citation trends of papers belonging to same FoS and with different FoS. Our claim is that if there is more association between the citation trends of papers belonging to same FoS, then it establishes the fact that an author working in a particular FoS will receive the similar citation trend as that of the FoS. Our methodology comprises of following major steps:

- From MAG dataset in the field of Computer Science, we identified the FoS of authors and combined the citation counts of authors belonging to same FoS forming different groups (section 3.6.2).
- We fitted the predictive model on these different groups using Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) (section 3.6.3).
- We tested the performance of each of these models with test data from same group and from each of the different groups using R^2 as performance metric. Finally, we analyzed the results of R^2 to find any support (chapter 4 section 4.3)

This research uses the scientific articles published from 1950-2017 time period in the domain of Computer Science from Microsoft Academic Graph (MAG) dataset [3].

This section describes the proposed methodology of RQ-3. As we already discussed Dataset description in section 3.2 and dataset preprocessing is described in section

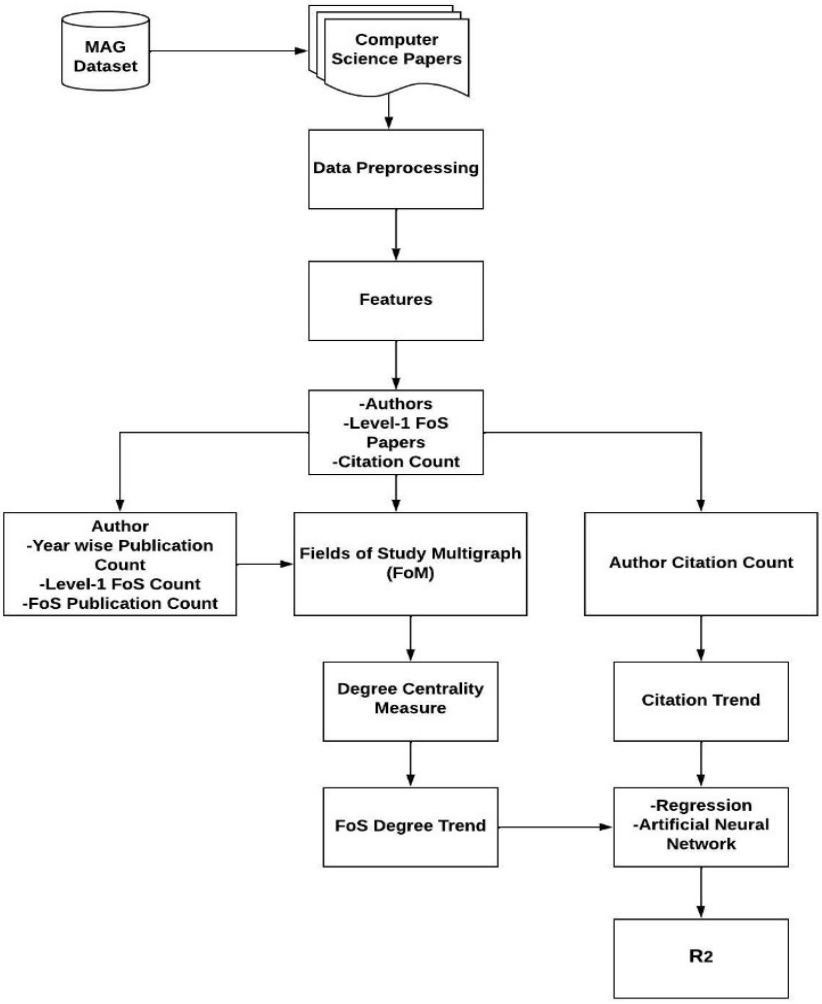


Figure 3.10: The proposed approach.

3.6.2 in detail. Finally, in section 3.6.3, we discussed the proposed model. Figure 3.10 presents the proposed approach of RQ-3.

3.6.2 Data Preprocessing

As explained earlier, the MAG dataset contains articles from different domains. For this study, we have selected the research papers from the field of Computer Science published during 1950-2017. Even though the MAG contains the papers that are published in journals and conferences. In this research, we have considered conference and journal papers both as shown in appendix A Table **??**. We chose three different time periods to perform our experiments, such as, 1970-1975, 1990-1995 and 2008-2013. These durations represent the earlier, mid and last part of our dataset, so they become an effective representation of the entire data. We extracted all publications of these time periods from MAG dataset. Then, we stored paper id, publication year, paper title, author, FoS, level-0, and level-1 FoS in a separate file named as author FoS dataset as shown in Table 3.18.

TABLE 3.18: FoS of three sampled papers.

Paper id	Year	Paper Title	Author	Level-0 FoS	Level-1 FoS
100981	1962	Standardised Clinical Datasets-Pre-Requisites to Successful Data Mining	AIMIS, Dean White MSc	Speech Recognition, Computer Science, Pattern Recognition, Data Mining.	Speech Recognition, Pattern Recognition, Data Mining.
101432	1970	Quality of Web-based information systems	Worwa, Kazimierz and Stanik, Jerzy	Web-based information systems, Web engineering, computer science, Web-based software, world wide web, software quality modeling	Web-based information systems, Web engineering, Web-based software, world wide web, software quality modeling.
100231	1970	Computerized image dynamic analysis.	F.B. Brown, K.W.Hering	Computer Vision, Simulation, Computer Science, Computer Graphics	Computer Vision, Simulation, Computer Graphics.

Table 3.18 shows the paper id, publication year, paper title, author name, level-0 and level-1 FoS of three sample papers from 1962 and 1970. The level-1 FoS of these papers in MAG dataset is: Speech Recognition, Pattern Recognition, Data Mining, Web-based Information Systems, Web Engineering, Web-based Software, World Wide Web, Software Quality Modeling, Computer Vision, Simulation, Computer Graphics.

From this data, we identified the main FoS for an author, that is, the FoS in which he/she has maximum publications. This data is computed for authors in six selected FoS for five years in aforementioned three durations, that is, 1970-1976,1990-1996 and 2008-2014 (each year separately). The FoS selected are Data Mining, Computer Network, World Wide Web, Computer Vision, Library Science and Computer Engineering. However, Google scholar and Google Ngram Viewer also exposed that Data Mining, Computer Network, World Wide Web, Computer Vision, Library Science and Computer Engineering FoS appeared before 1970 as shown in Figure 3.11.

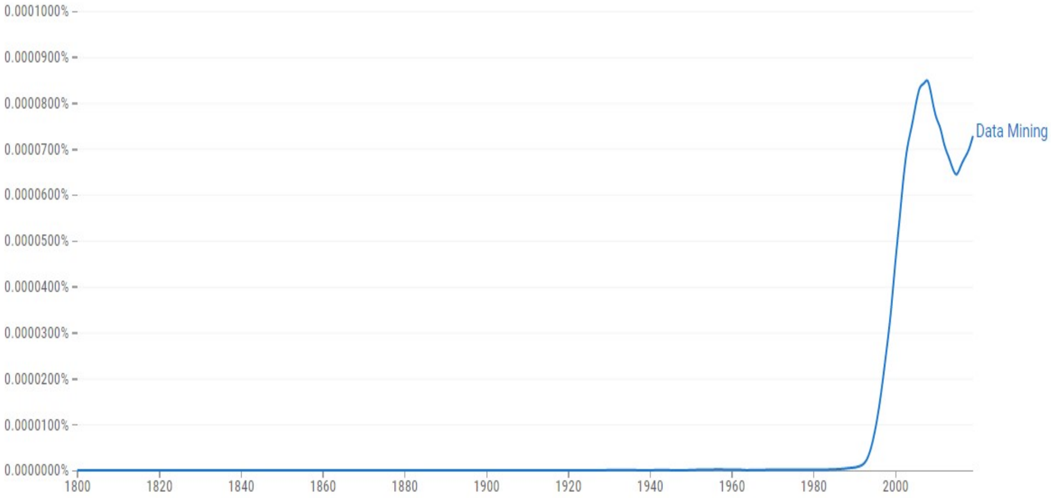


FIGURE 3.11: Trend of Data Mining FoS [7].

Now, we extracted authors who followed the maximum trend of a specific FoS in their publications in a specific year. For example, an author1 has 3 publications p1, p2, and p3 in year 1970. The FoS in p1 = Data Mining, Machine Learning, in p2 = Artificial Intelligence, Machine Learning and in p3 = Data Science, Data Mining, Machine Learning. Here, author1 uses 4 different FoS in 3 papers i.e; Data Mining, Machine Learning, Artificial Intelligence and Data Science.

Afterwards, we have calculated the author's publication count $p(c)$, citation count $c(p)$, and citation trend up-to 5 years of a specific FoS. As the Table 3.19 shows publication count and FoS count of an author A1 from 1970-2015. A1 have two publications in 1970 and papers covers following level-1 FoS: Data Mining, Machine Learning and Artificial Intelligence. A1 have three publications in 1971 and papers comprises of the following level-1 FoS: Data Mining, Data Science, Database, Machine Learning, Information Retrieval and Artificial Intelligence. A1 has one publication in 1972 and papers encompasses following level-1 FoS: Data Mining, Data Science and Database and upto so on.

Table 3.19: Publication count and FoS count of an author A1 from 1970-2015.

Author	Year	Publications	Level-1 FoS	p(c)	FoS (c)
A1	1970	P1	Data Mining, Machine Learning.	2	2
		P2	Data Mining, Artificial Intelligence.		2
	1971	P3	Data Mining, Database.	3	2
		P4	Data Mining, Data Science, Machine Learning,		3
		P5	Information Retrieval, Artificial Intelligence.		2
	1972	P6	Data Mining, Machine Learning, Data Science, Database.	1	4
	…..	…..	…..	…..	…..
	2015	P202	Data Science, Data Mining, Machine Learning.	8	3
		P203	Pattern Recognition, Speech Recognition, Machine Learning.		3
		P204	Information Retrieval, Data Mining, Machine Learning, Artificial Intelligence.		4
		P205	Data Mining, Artificial Intelligence, Data Science, Database, Algorithm.		5
		P206	Pattern Recognition, Speech Recognition, Machine Learning, Artificial Intelligence, Data Mining.		5
		P207	Computer Security, Database, Simulation.		3
		P208	Information Retrieval, Data Mining, Algorithm, Machine Learning, Artificial Intelligence.		5
		P209	Data Mining, Artificial Intelligence, Data Science, Database.		4

However, A1 have eight publications in 2015 and papers comprises of the following level-1 FoS: Data Science, Data Mining, Machine Learning, Pattern Recognition, Speech Recognition, Artificial Intelligence, Information Retrieval, Algorithm, Computer Security and Simulation. Now, we have the publication count and FoS count of an author. FoS publication count in a specific year is also calculated and shown in Table 3.20.

TABLE 3.20: Author A1 FoS Publication Count.

Year	Data Mining	Machine Learning	Artificial Intelligence	Database	Data Science	Information Retrieval	Algorithm	Pattern Recognition	Speech Recognition	Computer Security	Simulation
					FoS(p(c))						
1970	2	1	1	0	0	0	0	0	0	0	0
1971	2	1	1	1	1	1	0	0	0	0	0
1972	1	0	1	1	0	0	0	0	0	0	
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....		
2015	6	5	5	3	3	2	2	2	2	1	1

Next, we calculated the citation count of a paper of an author. For an archive of scientific publications P of an author A1, a paper $p\epsilon P$, its citation count $cc(p)$ is the number of papers that cite p [6]:

$$cc(p) = |p\prime\epsilon P : p\prime citesp| \tag{3.11}$$

For the papers published by a particular author during the time period of 1970-1975, we calculated their citation counts for next five years. Table 3.21 shows the citation count for author A1 for his papers published in the selected time period.

TABLE 3.21: Yearly Citation count of papers of an author starting from 1970.

Author	Publication Year = Y	Citation count in....					
		Y+1	Y+2	Y+3	Y+4	Y+5	Y+6
	P1970	1	1	2	1	2	2
	P1970	1	1	1	1	1	1
A1	P1971	1	2	2	3	2	3

Author	Publication Year = Y	Citation count in. . . .					
		Y+1	Y+2	Y+3	Y+4	Y+5	Y+6
	P1972	1	2	2	2	3	2

For the papers published by a particular author during the time period of 1970-1975, we calculated their citation counts for next five years. Table 3.21 shows the citation count for author A1 for his papers published in the selected time period. As shown in the table, author A1 published two papers in 1970, one paper in 1971 and one paper in 1972. We calculated the citation counts of papers from birth year up-to next 5 years. After calculating the yearly citation count of each author then we calculated the citation trend $ct_y(a)$ of an author (a) for the year (y) as the sum of citation counts $cc(p_i)$ of all papers of author (a) published in year y.

$$ct_y(a) = \sum_{i=y+1}^{y+5} cc_i(p) \tag{3.12}$$

In equation 3.12, p represents the papers of author (a) published in year y, whereas the right side of equation sums the citation counts of papers (p) for next five years. The table 3.22 below shows the citation trends of three different authors for the year 1970.

Table 3.22: Yearly Citation count of papers of an author starting from 1970.

Author	1971	1972	1973	1974	1975	1976
A1	2	2	3	2	3	3
A2	0	1	2	2	3	2
A3	1	1	1	2	2	2

In this research, we have selected six different FoS from level-1. We are considering Data Mining, Computer Network, World Wide Web, Computer Vision, Library Science and Computer Engineering. We have randomly selected these six FoS considering it a good representation of overall data. Moreover, we selected papers published in different eras of our dataset, one from initial times, one from middle and one relatively recent. The prepared dataset contains the papers published from 1970-1975, 1990-1995 and 2008-2013.

Next, we combined the authors having the same FoS forming six different groups of our selected areas. The Table 3.23 below shows the 5 years' data about different groups that we created:

Table 3.23: Dataset statictics of six FoS from 1970-1975.

	1970	1971	1972	1973	1974	1975	Total
No. of authors in Computer Science	39627	33924	36560	38955	41500	47137	237703
No. of papers in Computer Science	26845	23,643	25123	26516	28214	32032	162373
No. of authors in Data Mining	717	626	624	698	700	821	4468
No. of papers in Data Mining	406	370	366	397	417	496	2643
No. of authors who's max FoS trend is Data Mining	60	67	70	75	82	90	444
No. of papers of authors -max FoS trend -Data Mining	72	79	80	82	90	100	503
No. of authors in Computer Network	578	512	558	708	744	805	3914
No. of papers in Computer Network	353	323	353	403	439	500	2371
No. of authors who's max FoS trend is Computer Network	45	54	59	67	70	75	370
No. of papers of authors -max FoS trend -Computer Network	60	76	77	80	86	90	469
No. of authors in World Wide Web	2264	1773	1832	1809	1869	2067	11614
No. of papers in World Wide Web	1773	1414	1439	1421	1490	1616	9153
No. of authors who's max FoS trend is World Wide Web	1000	800	850	790	860	800	5100
No. of papers of authors -max FoS trend -World Wide Web	700	650	600	560	600	620	3730
No. of authors in Computer Vision	1929	1846	2064	2289	2441	2878	13447

	1970	1971	1972	1973	1974	1975	Total
No. of papers in Computer Vision	11431	1107	1226	1334	1405	1646	7861
No. of authors who's max FoS trend is Computer Vision	960	850	1000	1120	1130	1670	6730
No. of papers of authors -max FoS trend -Computer Vision	700	690	710	750	780	800	4430
No. of authors in Library Science	1097	1180	11581	1144	1252	1273	7104
No. of papers in Library Science	450	480	485	450	500	520	2885
No. of authors who's max FoS trend is Library Science	430	460	480	500	510	530	2910
No. of papers of authors -max FoS trend -Library Science	180	190	205	210	225	230	1240
No. of authors in Computer Engineering	260	298	211	251	260	275	1555
No. of papers in Computer Engineering	146	152	129	149	155	154	885
No. of authors who's max FoS trend is Computer Engineering	57	60	72	75	79	80	423
No. of papers of authors -max FoS trend -Computer Engineering	42	53	60	67	70	83	375

Table 3.23 shows summary of data set that we have created out of MAG data set. The first two rows show the overall number of authors and number of papers in the field of Computer Science for the duration of 1970-1975. The third row shows the total number of authors who published in Data Mining. The next row shows number of papers having Data Mining FoS. Next is the number of authors who's maximum FoS is Data Mining and then the papers of those authors. In this way, every set of four rows describes the summary of data about six selected FoS.

In the next section, we discuss our proposed model to investigate the similarity between authors FoS trend and citation patterns.

3.6.3 Proposed Model

As discussed earlier, we had citation trends of six different FoS for five years. We fitted Multiple Linear Regression (MLR) on each data set separately. Multiple linear regression is a statistical method used to analyze the relationship between a dependent variable and two or more independent variables. It is an extension of simple linear regression, where only one independent variable is considered. In multiple linear regression, the dependent variable is predicted using a linear combination of the independent variables.

The multiple linear regression model can be expressed as:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + ... + \epsilon \tag{3.13}$$

where Y is the dependent variable, $\beta 0$ is the intercept or constant term, $\beta 1$, $\beta 2$, $\beta 3$, $\ldots$, β_n are the coefficients of the independent variables $X1, X2, X3, \ldots, Xn$, respectively, and ϵ is the error term.

The main purpose of multiple linear regression is to determine the coefficients of the independent variables that best predict the dependent variable. This is done by minimizing the sum of the squared residuals between the predicted values and the actual values of the dependent variable.

For each data set we performed different experiments taking one year's citation count as input and predicting citation count for the second year. Then, we took two years' citation count as input and predicting for third year. In the same way, at the end we took five years' citation counts as input and predicted for the sixth year. We proposed a model to predict the citation trend of authors.

Table 3.24: Proposed Model for each six FoS.

Input (x)	Output (y)
Year 1	Year2
Year1-Year2	Year3
Year1-Year2-Year3	Year4

Input (x)	Output (y)
Year1-Year2-Year3-Year4	Year5
Year1-Year2-Year3-Year4-Year5	Year6

Table 3.23 above presents a summary of our proposed model. Reason for increasing input feature one by one is that, we wanted to study the behavior of FoS impact on citation trend of authors in different environments. That is, starting from a relatively weaker fit where we have just one feature to a relatively stronger model where we have five features. We performed these experiments for papers published in 1970 till 1975. For paper published in 1970, we calculated citation count from 1971-1976. For those published in 1971, we calculated citation count from 1972-1977. Likewise for papers published in 1975, we calculated citation court from 1976-1981. For every set, we performed experiments as shown in Table 3.24.

We fitted models in each of the six FoS as sketched in Table 3.24 above and used R^2 as the performance metric. We tested each of the models with the test data set from same FoS from which model was fitted and from all other five FoS and analyzed the R^2 value in each case. We also performed same experiments using Artificial Neural Network (ANN) to further strengthen our findings.

An Artificial Neural Network (ANN) is a computational model inspired by the structure and function of the biological neural networks of the human brain. An ANN consists of layers of interconnected nodes, called neurons, which are organized into input, output, and hidden layers. Each neuron in the network receives input signals from the previous layer, performs a computation on those inputs, and passes the output to the next layer.

The inputs to the network are typically feature vectors, and the outputs are the predictions made by the network for a given input. During training, the network learns to adjust the weights of the connections between neurons to minimize the difference between its predictions and the actual outputs.

As mentioned earlier, we planned to evaluate the results obtained by testing models using the test dataset of the same FoS versus those obtained by testing on other

FoS. The layout and strategy for experiments have been described in this section. We have presented and evaluated the results in the next chapter 4.3.

We performed our experiments for **RQ3** using an established metric, that is, citation count, and our results seem promising that we are going to discuss in next chapter. However, we performed same experiments using the degree centrality measures from FoS Multigraph created using the papers of authors from same FoS. The purpose of these experiments was to evaluate the performance of degree centrality in the prediction of response that authors may get for their work in a particular FoS. If results of degree centrality are comparable or better than those obtained through citation count then we may propose to use this metric as the representative of authors' performance. Table 3.25 shows the degree centrality values for some authors working in a particular FoS:

TABLE 3.25: Authors FoS degree trend starting from year 1970.

Author	1971	1972	1973	1974	1975	1976
A1	4	3	3	3	4	3
A2	3	3	2	2	3	2
A3	3	4	2	3	4	4

The degree centrality values of three example authors are shown in the Table 3.25.

These values have been used in the experiments mentioned in section 3.6.3. The results are discussed and compared with others in next chapter section 4.3.

This chapter describes the proposed methodologies for research questions 1-3. To identify the significance of FoS trend on research paper citations **(RQ-1)**, we have performed clustering on FoS and citations pattern separately. We also presented a novel method of Field of Study Multigraph (FoM), formed by using centrality measures; degree, betweenness and closeness to analyze the FoS trend, citation trend and the relation between research areas in Computer Science scientific articles. Further, we have proposed an approach to detect researchers who are involved at the early stage of an FoS trend **(RQ-2)** known as trend setters. We calculated the debut year of an FoS. Then, we have computed the FoS publication count, its author count, FoS degree and lastly, we applied Rogers [129] for the detection of trend setters and followers.

Finally, to detect the impact of FoS on authors citation count **(RQ-3)**, we have proposed an approach that detect the FoS trend of an individual author in his/her career years by characterizing the relations between his/her publications and citations. We detected the FoS of authors, the citation count of authors and then we computed the citation trend of authors. Finally, we used citation trend of authors to predict the citation count of authors. We also computed FoS degree values and used to predict the citation count of authors.

The next chapter describes the experimental results and detailed discussion about **RQ-1-3**.

Chapter 4

Results and Discussion

The previous chapter describes the proposed methodologies for the research questions that we established from the gap in the literature survey. The research questions focus on significance of FoS trend on research paper citations; detection of researchers who are involved at the early stage of an FoS trend; and establishing the impact of FoS on the citation trend of an individual author. Establishing research questions from the literature survey and proposing methodologies to answer those questions is important but equally important is the evaluation of those results in order to establish the authenticity and validity of the methodologies. This chapter describes the experiments, results and discussions about the proposed techniques presented in methodology chapter. Results are described with reference to each research question.

Structure of the chapter is as follows: Section 4.1,4.3 describes the experiments and results of **RQ-1**. **RQ-2** experiments and results discusses in section 4.2 and section 4.3 presents the experiments and results of **RQ-3**.

4.1 Experiments and Results (RQ-1)

This section presents the experiments dataset statistics for RQ-1. In RQ-1, we have selected those papers that were published in 2007-2011 time period. As mentioned in section 3.3.2,3.3.3 we performed two separate clustering; one on the

FoS associated with research papers, and second on the citation trend of those papers for five years. Our proposal is that if there is significant level of similarities between these two clusters then we can establish that there is similarity in the citation trend of paper belonging to same FoS. The dataset statistics are shown in Table 4.1. We have used Rand Index and correlation as evaluation metrics for experiments.

Table 4.1: Publication count from 2007-2011.

Year	Publication Count
2007	5863
2008	6599
2009	7159
2010	7070
2011	6315

4.1.1 Rand Index

To find out the similarity between two sets of formed clusters (FoS and citation pattern clusters), we used the Rand Index RI which is defined as a measure of the percentage of correct decisions made by the algorithm [134]. Rand Index gives a value between 0 and 1, where 1 means two clustering outcomes match identically. Rand Index can be calculated using the following formula 4.1;

$$RI = \frac{a+b}{a+b+c+d} \tag{4.1}$$

where, **a**: two similar documents to the same clusters, **b:** two dissimilar documents to different clusters, **c:** two similar documents to the different clusters, and **d:** two dissimilar documents to the same

Table 4.2: Similarity between FoS and citation clusters from 2007-2011.

Publication Year of Paper	Duration of Citation Pattern	Value of Rand Index
2007	2007-2011	0.67
2008	2008-2012	0.67

Publication Year of Paper	Duration of Citation Pattern	Value of Rand Index
2009	2009-2013	0.68
2010	2010-2014	0.67
2011	2011-2015	0.68

As can be seen from Table 4.2 that there is a reasonable level of similarity between the cluster formed independently on the bases of FoS and the citation pattern of papers. It proves that the papers belonging to same FoS have similar citation patter, or in other words FoS has certain level of impact on the citation trend of the papers following that FoS.

4.1.2 Similarity between Trendy FoS and Citation Clusters

We also performed same experiment, but using different metrics instead of citation count. As mentioned in section 3.4.3, we calculated the RI of FoS that are selected as trendy FoS by FoM method using graph centrality measures and frequency. The RI is used to compute the similarity between two data clustering i.e., FoS and citations clusters and compared the resulting values for each other. The RI values of four metrics have been illustrated in Table 4.3 below and also shown in the form of a graph in Figure 4.1.

Table 4.3: Similarity between FoS and citation clusters from 2007-2011.

FoS Year	Citation	Rand Index			
		Frequency	Degree	Betweenness	Closeness
2007	2007-2011	0.67	0.68	0.63	0.61
2008	2008-2012	0.67	0.68	0.63	0.62
2009	2009-2013	0.68	0.68	0.64	0.62
2010	2010-2014	0.67	0.69	0.63	0.61
2011	2011-2015	0.68	0.69	0.63	0.60

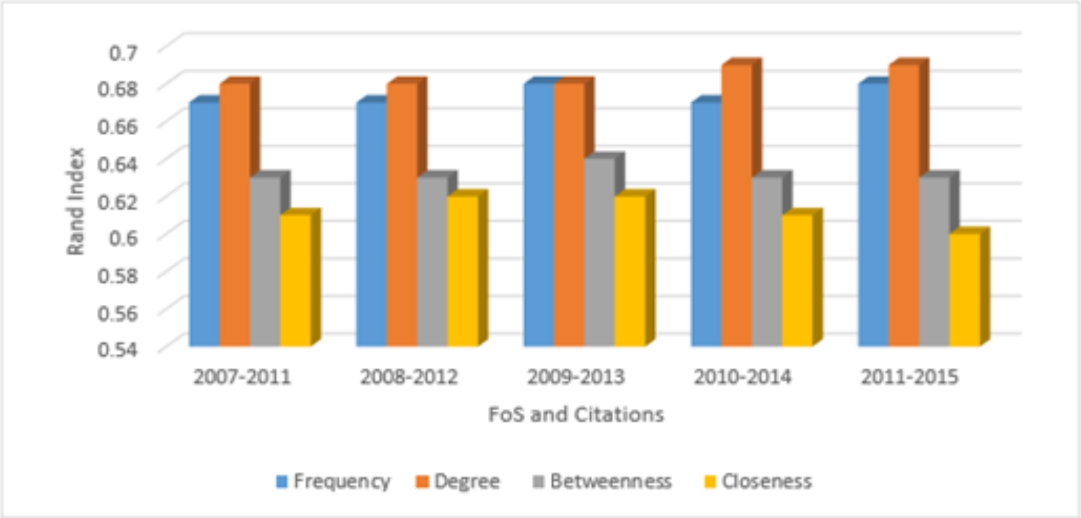


FIGURE 4.1: Rand index of frequency, degree, betweenness, and closeness.

The RI results show a similar level of similarity between clustering based on FoS and four different measures, i.e., frequency, degree, betweenness, closeness. Frequency and degree centrality have relatively higher values of RI as compared to the other two and out of these two- degree centralities have the highest RI values across multiple years. As the results indicate, the degree has achieved the highest RI value 0.69. The results indicate that if the papers belong to the same FoS, then there are 69% of chances, that they have the same citation trend. This proves that a field of study has a certain impact on citation count of a paper and researchers should also contemplate on the trend of a field of study while selecting a particular research area. Also, the degree centrality is a more suitable metric to measure the trend of an FoS than a simple citation count.

4.1.3 Correlation

We have also computed the correlation coefficient to examine the relationship between FoS citations pattern. Correlation is one of the most common and useful statistics to examine the nature of the relationship between data items [135]. A positive correlation indicates the extent to which two variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases. Equation is given as below;

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - \sum(x)^2][N\sum y^2 - \sum(y)^2]}} \tag{4.2}$$

where: r is the correlation coefficient

n is the number of observations or data points

$\sum xy$ is the sum of the product of each x and y value

$\sum x$ and $\sum y$ are the sums of the x and y values, respectively

$\sum(x)^2]$ and $\sum(y)^2]$ are the sums of the squares of the x and y values, respectively.

The correlation coefficient measures the strength and direction of the linear relationship between two variables. A positive correlation coefficient indicates that the variables are positively related, meaning that as one variable increases, the other variable also tends to increase. A negative correlation coefficient indicates that the variables are negatively related, meaning that as one variable increases, the other variable tends to decrease.

The correlation coefficient ranges from -1 to 1, with 0 indicating no correlation, -1 indicating a perfect negative correlation, and 1 indicating a perfect positive correlation. However, it is important to note that correlation does not necessarily imply causation, and further analysis is necessary to establish a causal relationship between variables.

For this experiment, we have considered 5 years citation count of papers belonging to a particular FoS cluster. Out of these papers, we took stratified random subset of 80% papers and used them as training data set and remaining 20% as a test set. In this way, 7 different training and test data sets were formed which comprise of five years average of citation count of papers belonging to the same cluster. These values are shown in Table 4.4 below.

TABLE 4.4: Average citation count of papers from 2007-2011.

Clusters	Yearly Average of Training Data Set				
	2007	2008	2009	2010	2011
cluster0	2.2	1.7	1.3	1.5	4.4

Clusters	Yearly Average of Training Data Set				
	2007	2008	2009	2010	2011
cluster1	2.2	1.5	1.4	1.5	4.4
cluster2	2.2	1.4	1.4	1.5	3.4
cluster3	2.3	2.5	1.2	1.5	4.4
cluster4	3.2	1.4	1.3	1.5	4.4
cluster5	2.2	1.5	1.3	1.5	4.4
cluster6	2.4	1.4	1.3	2.5	4.4
Clusters	Yearly Average of Test Data Set				
	2007	2008	2009	2010	2011
cluster0	2.2	1.4	1.2	1.5	4.4
cluster1	1.1	1.4	1.3	1.7	3.4
cluster2	1.1	1.4	1.3	1.3	2.5
cluster3	2.1	1.4	1.3	1.5	3.4
cluster4	2.1	1.4	1.3	1.5	4.4
cluster5	2.1	1.3	1.3	1.5	4.4
cluster6	1.1	1.4	1.3	1.5	3.4

The values illustrated in the above table reveal that average citation count across multiple FoS is approximately similar. Next, to find the level of similarity among papers belonging to the same FoS, we performed two steps: (1) we have calculated the correlation coefficient between training dataset of one year with test dataset of every other year and compared them.

(2) Then, we plotted the training dataset against the test dataset of the same year to graphically see the level of similarity between them. Table 4.5 below shows the correlation coefficient between different clusters training dataset with each of the other clusters test dataset. The highlighted values show that every cluster has the highest correlation with the test dataset of its cluster. This proves that the papers belonging to the same FoS have similar citation patterns and if we select a particular FoS to work in, then we can have an estimate of the citation pattern that we may receive on our work.

TABLE 4.5: Correlation matrix of FoS citations from 2007-2011.

Clusters	Correlation						
	test	test	test	test	test	test	test
Cluster0 (training)	**0.99**	0.65	0.51	0.46	0.73	0.22	0.54
Cluster1 (training)	0.65	**0.92**	0.49	0.54	0.7	0.22	0.37
Cluster2 (training)	0.51	0.49	**0.83**	0.29	0.6	0.35	0.26
Cluster3 (training)	0.46	0.4	0.29	**0.91**	0.71	0.23	0.53
Cluster4 (training)	0.73	0.56	0.6	0.41	**0.93**	0.23	0.13
Cluster5 (training)	0.22	0.22	0.35	0.23	0.23	**0.97**	0.65
Cluster6 (training)	0.54	0.37	0.36	0.53	0.43	0.69	**0.87**

Figure 4.2 shows the plots of training and test datasets of different clusters and citations pattern. The plots also show the similarity between the average citation trend of the same FoS. Moreover, the level of the correlation coefficient is also clear from the corresponding graph, for example, cluster0 has the maximum value of correlation coefficient which is also evident from the corresponding plot of Figure 4.2, where both lines are almost identical.

The correlation result shows the papers belonging to the same FoS and following the trend, have similar increasing or decreasing patterns of citations, as shown in Figure 4.2. The experimental results show that FoS has a certain impact on citation count. Furthermore, a high count of citation depicts that if a paper belongs to the same FoS, then it may have the same citation trend. This proves that a field of study has a certain impact on citation count of a paper and researchers should also consider the trend of a field of study while selecting a particular research area.

4.2 Experiments and Results (RQ-2)

This section presents the results of RQ-2, as we have detected the individuals involved at the early stage of an FoS in section 3.5.1. This is challenging to evaluate trend setters at the early stages of an FoS. We have compared our list of researchers (trend setters) with two existing lists that contain highly recognized Computer Science scientists. The lists are as follows; (i) top 10 influential authors identified by [91] and (ii) an existing list of Computer Science scientists with H-index of 40 or higher (www.cs.ucla.edu/ palsberg/h-number.html).

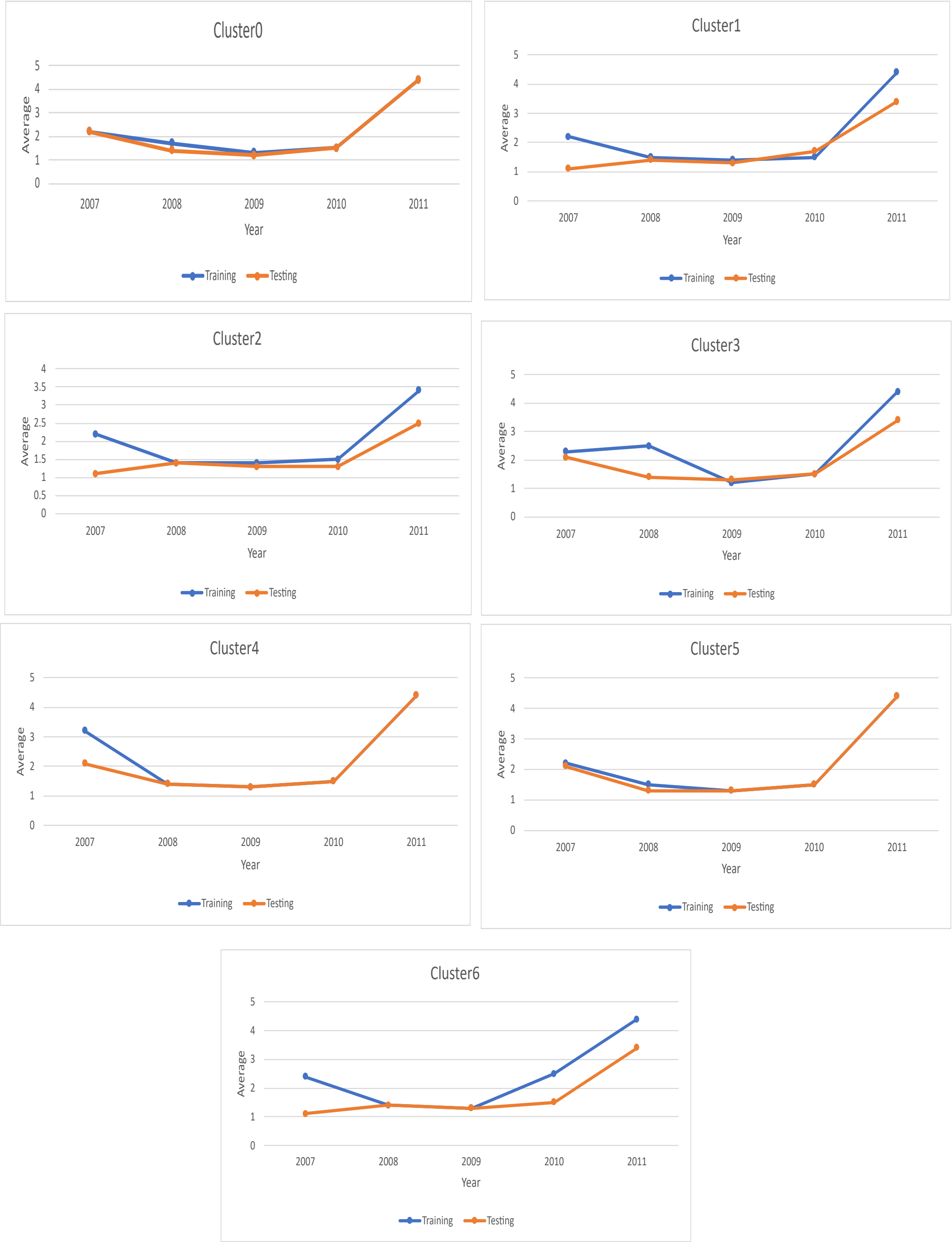


FIGURE 4.2: Training and test datasets of different clusters and citation patterns.

The H-index is defined as a measure to compute the scientific output of a researcher, where h is the number of publications with citation count higher or equal to h [136].

Table 4.6 shows a comparison of the ”Semantic Search” FoS trend setters at the early stage identified by our approach with top 10 influential authors identified by [91].The table shows that Dan Suciu, Justin Zobel, Dieter Fensel, W. Bruce Croft, Clement Yu, Dragomir R. Radev, James Allan and Victor Vianu have the exact match with influential authors [91]. These authors worked and published at the embryonic and early stage of ”Semantic Search” FoS.

Table 4.6: Top-left, we show trend setters and on top-right, the top 10 influential authors of the semantic search FoS.

Influential Authors	Ranking by [91]	Ranking by Proposed Approach
W. Bruce Croft	1	5
Dieter Fensel	2	1
Dan Suciu	3	2
William W. Cohen	4	18
Berthier Ribeiro-Neto	5	20
Clement T. Yu	6	10
James Allan	7	4
Justin Zobel	8	3
Dragomir R. Radev	9	6
Victor Vianu	10	12
Alon Halevy	–	7
Katia Sycara	–	8
James Hendler	–	9

Table 4.6 above shows comparison of influential authors in Semantic Search FoS identified in [91] and those established by our approach. The table highlights following aspects:

1. Seven out of top ten influential authors are common in both approaches, however, there is difference in the rankings of such authors as highlighted in the table

2. Three of the authors that are not in the top ten lie within top twenty trend setters as proposed by our approach

3. The strength of approach by [91] is that they identify the influential authors from the five years prior to the birth year of an FoS, whereas we identify the trend setters from the work done in next five years of the birth year of FoS. In spite of this difference, majority of authors are common in both approaches.

In order to evaluate that which of the two approaches identifies better trend setters, we evaluated the major authors working in the Semantic Search FoS from 2003-2007.

Table 4.7: Researchers appears in various lists, their publication count, citation count, FoS degree in semantic search FoS from 2003-2007.

Rank	Researcher	Publication count	Citation Count	FoS Degree	Influential Author	H-Index	Trend Setter
1.	Dieter Fensel	97	375	323	✓	✓	✓
2.	Dan Suciu	62	323	276	✓	✓	✓
3.	Justin Zobel	44	164	113	✓	✓	✓
4.	James Allan	42	111	101	✓	✓	✓
5.	W. Bruce Croft	29	109	82	✓	✓	✓
6.	Dragomir R. Radev	28	67	50	✓	✓	✓
7.	Alon Halevy	27	74	97	✗	✓	✓
8.	Katia Sycara	26	60	71	✗	✓	✓
9.	James Hendler	25	54	87	✗	✓	✓
10.	Clement Yu	24	76	88	✓	✓	✓
11.	Wolfgang Nejdl	23	48	63	✗	✓	✓
12.	Victor Vianu	22	62	54	✓	✓	✓
13.	Amit Sheth	20	45	61	✗	✓	✓
14.	Andre Esteva	20	35	55	✗	✓	✓
15.	Tom Gillespie	19	37	51	✗	✓	✓

Rank	Researcher	Publication count	Citation Count	FoS Degree	Influential Author	H-Index	Trend Setter
16.	Richard Christie	18	30	49	✗	✓	✓
17.	Wenpeng Yin	17	32	39	✗	✓	✓
18.	William W. Cohen	15	32	35	✓	✓	✓
19.	Yuanzhang Li	15	34	31	✗	✓	✓
20.	Berthier Ribeiro-Neto	13	22	29	✓	✓	✓

Table 4.7 shows researchers identified by our approach in the early stage of an FoS trend, that is, trend setters for the FoS of Semantic Search. As shown in the table, the authors selected by our approach have more work in the concerned FoS, whereas, the authors at serial 7, 8 and 9, had relatively less work in the later years as compared to other authors which have been identified by our approach. Moreover, all of the top twenty authors identified by our approach are in the list of authors having high h-index [136]. So in the nutshell, we can say that the approach of [91] identifies influential authors before the birth of an FoS, however, our approach identifies trend setters in the early years after the birth. Most of the authors are common in both lists, however, those identified by our approach proved more influential in the future. This is the edge that we have over our base approach.'

4.3 Experiments and Results (RQ-3)

This section presents the results of the experiments that we performed on six FoS data that we collected from MAG data set (section 3.2). In the first set of experiments, we applied our proposed model using MLR on citation trend from $1970-1975, 1990-1995, 2008-2013$ time periods. We used R^2 as the performance metric. Figures 4.3, 4.7 shows plots of the R^2 values of five MLR experiments where

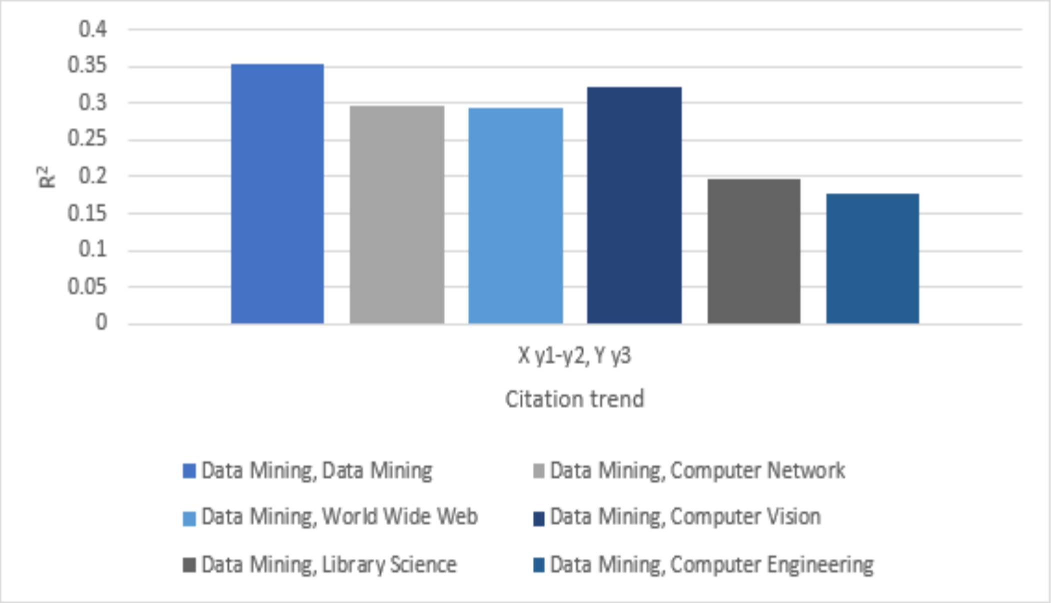


Figure 4.4: MLR models trained on Data Mining data set and tested with different FoS data sets.

we used Data Mining data set to train the model and for every trained model we used six different FoS data sets as test data for the time period of 1970 – 1975.

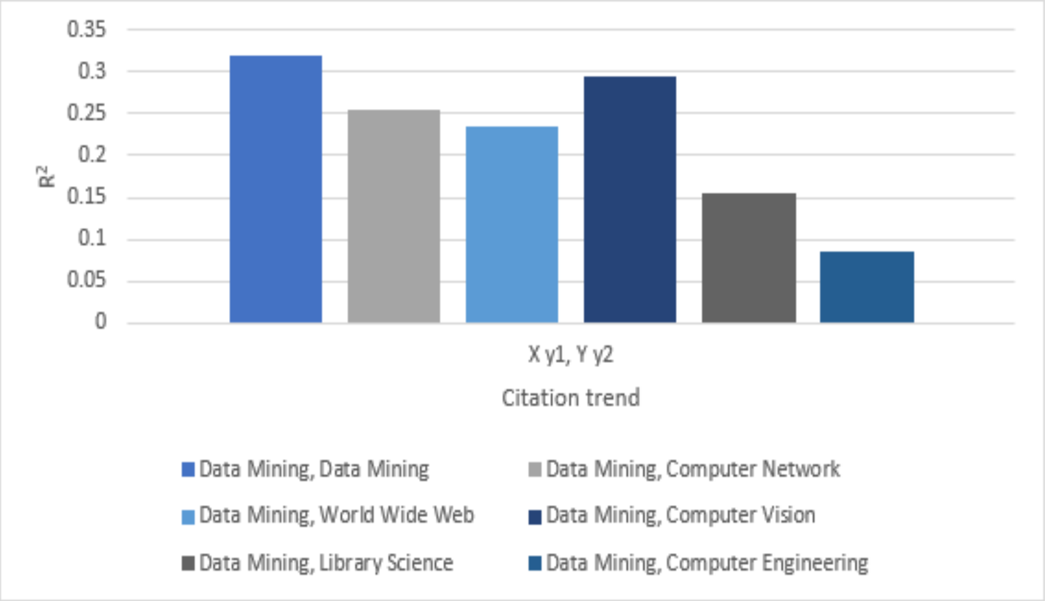


Figure 4.3: MLR models trained on Data Mining data set and tested with different FoS data sets.

The results shown in Figure 4.3 trained on Data Mining FoS first year data set and tested with different FoS second year data sets from 1970-1975 using MLR. The results shown in Figure 4.4 trained on Data Mining FoS two year data sets

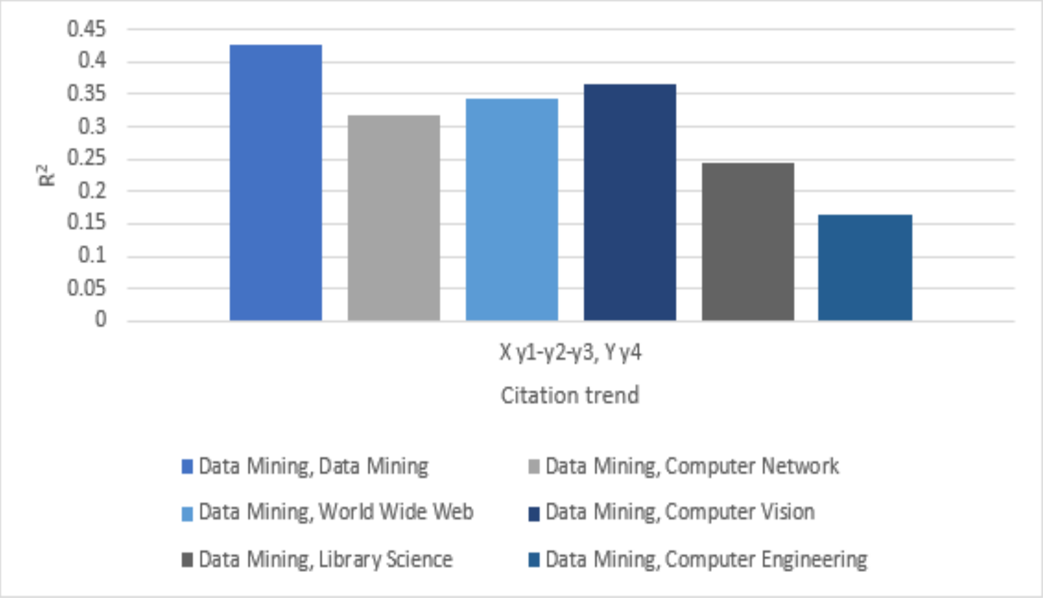


FIGURE 4.5: MLR models trained on Data Mining data set and tested with different FoS data sets.

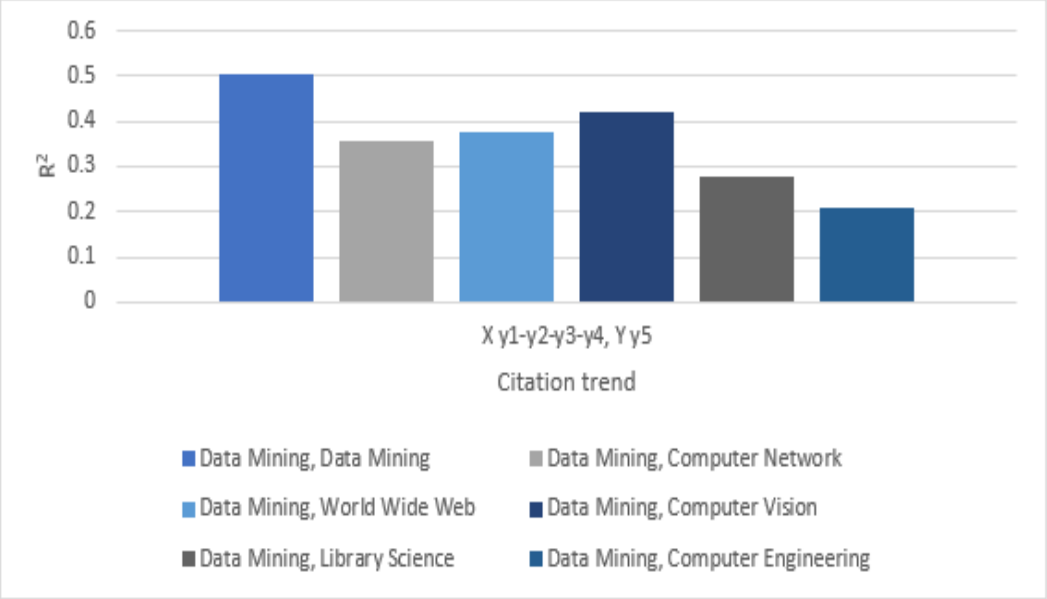


FIGURE 4.6: MLR models trained on Data Mining data set and tested with different FoS data sets.

and tested with different FoS third year data sets from 1970-1975 using MLR. The results shown in Figure 4.5 trained on Data Mining FoS three year data sets and tested with different FoS four year data sets from 1970-1975 using MLR. The results shown in Figure 4.6 trained on Data Mining FoS four year data sets and tested with different FoS five year data sets from 1970-1975 using MLR. The results shown in Figure 4.7 trained on Data Mining FoS five year data sets and tested

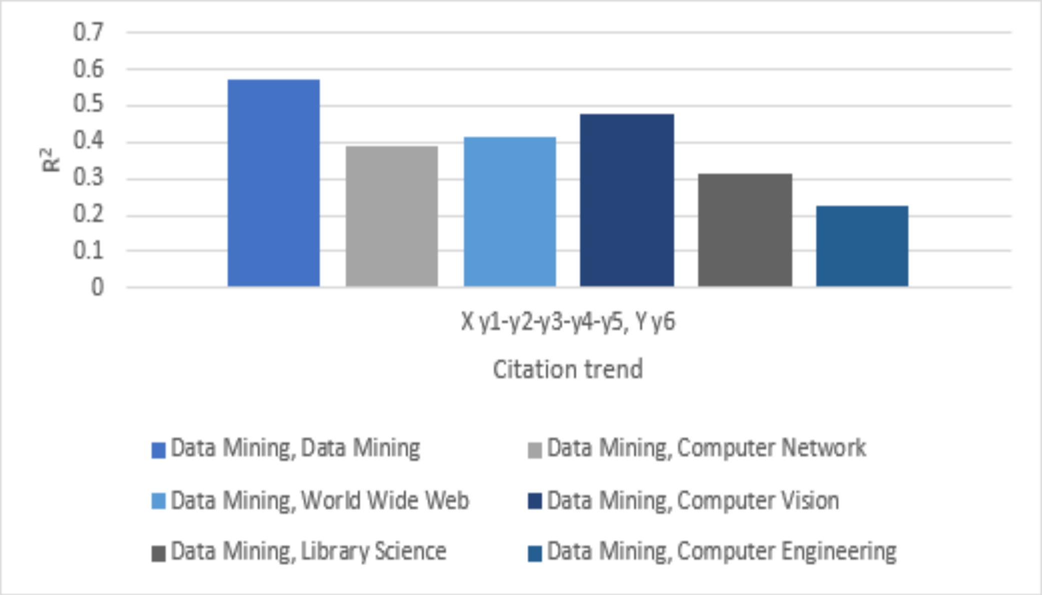


FIGURE 4.7: MLR models trained on Data Mining data set and tested with different FoS data sets.

with different FoS six year data sets from 1970-1975 using MLR.

Figure 4.3 above shows the R^2 values of the model that we trained using one value as feature, that is, citation counts of the authors for the year 1971 in the field of Data Mining and predicted the citation counts for next year, that is, 1972. The first bar shows the value of R^2 when model was tested using Data Mining data set. The next five bars show the R^2 values when same model was tested using test data from other five FoS. Figure 4.4 shows the R^2 values of the model that we trained using the citation trend of two years $y1 - y2$ from the field of Data Mining and predicted the citation trend of third year $y3$ as output Y.

Once again, the first bar shows R^2 value when model was tested using test data set from Data mining and remaining five bars show the values when model is tested using other FoS. In the similar fashion, we increased the number of features one by one taking them up to five features and predicting the citation count for the next year. Hence, Figure 4.7 shows the results of the models that was built using the citation counts of papers from the Data Mining FoS for the five years (1971-1975), and predicted the citation count for the year 1976. As before, we tested the model using all six FoS.

The study of the graphs in Figure 4.3 reveal the following facts:

1. Models get better as the number of features are increased which is generally true in such MLR models.

2. Among all the results, it is evident that the models perform better when they are tested using the test data set of the same FoS as compared to others. This finding leads to the answer of our research question, that is, what is the impact of FoS on the citation trend of the authors? Our analysis is that if a person publishes in a particular FoS, then the citation trend of this author's work resembles more to the overall citation trend of that particular FoS than that of some other FoS. The R2 value is though not very high, but it is higher in all the cases in the same FoS than the other ones. This gives enough evidence to believe the similarity of an author's citation trend with that of the particular FoS.

In order to further strengthen our finding, we performed same experiment with all six FoS, that is, training model in one FoS and testing it with data from all six FoS from 1970-1975, 1990-1995 and 2008-2013. Here, we do not show the results of 2008-2013 as they also depict the similar results.

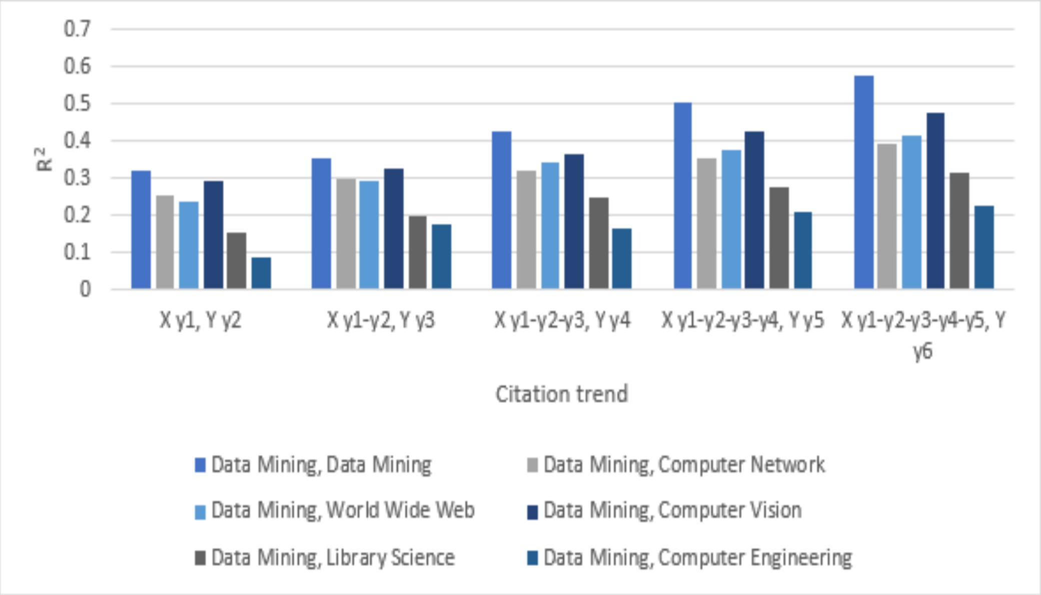


FIGURE 4.8: MLR models trained on Data Mining FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figures 4.8 trained on Data Mining FoS data set and tested with different FoS data sets from 1970-1975 using MLR.

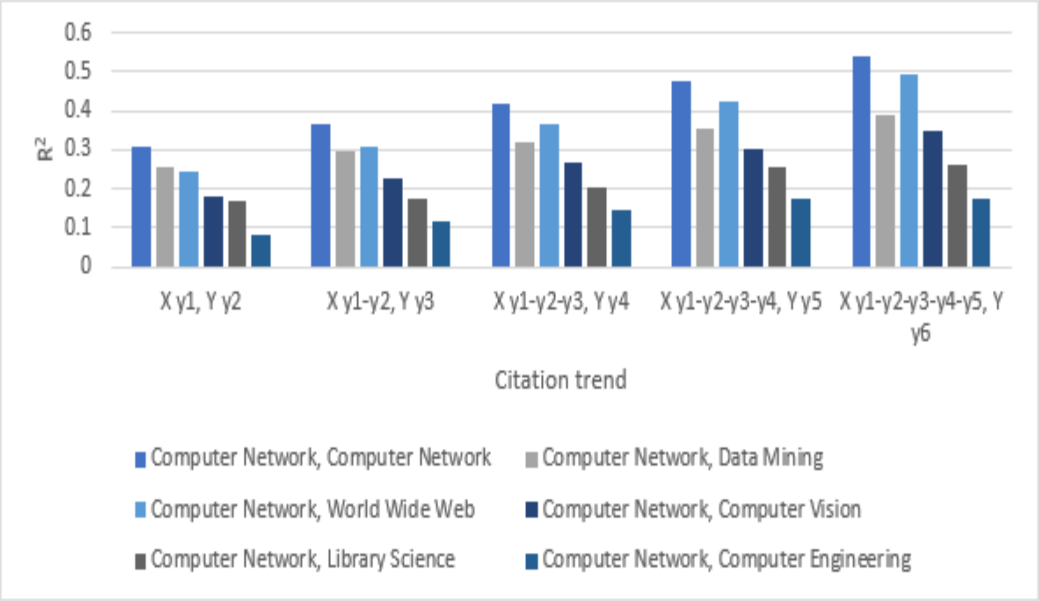


FIGURE 4.9: MLR models trained on Computer Network FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figures 4.9 trained on Computer Network FoS data set and tested with different FoS data sets from 1970-1975 using MLR.

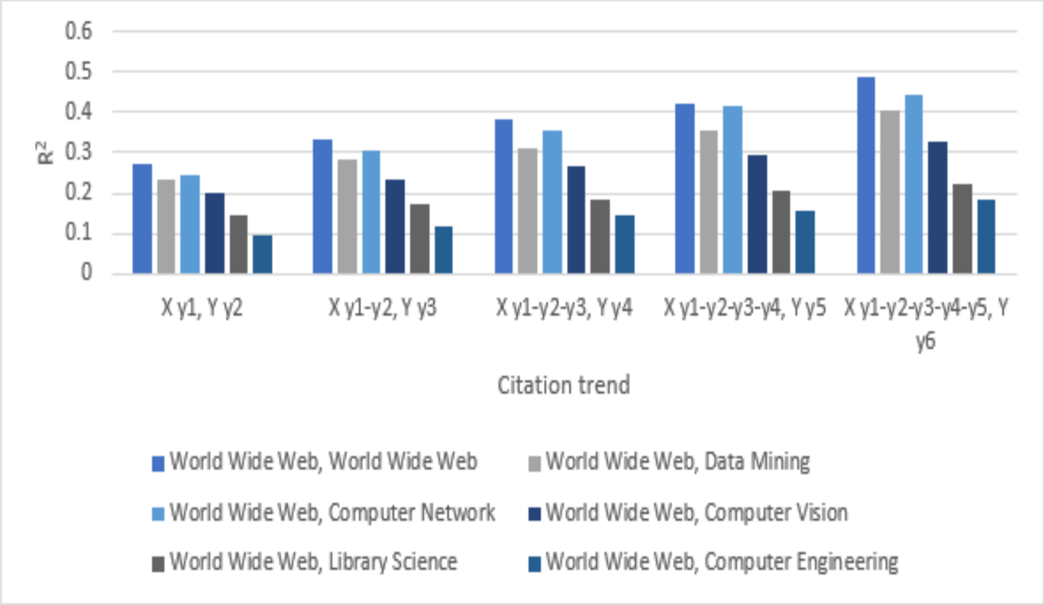


FIGURE 4.10: MLR models trained on World Wide Web FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in figures 4.10 trained on World Wide Web FoS data set and tested with different FoS data sets from 1970-1975using MLR.

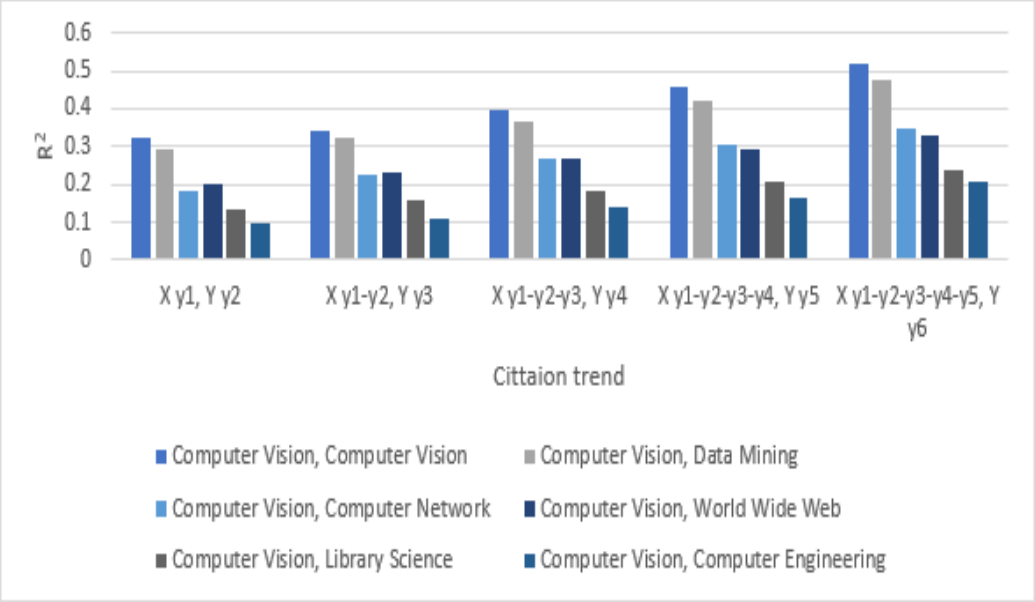


FIGURE 4.11: MLR models trained on Computer Vision FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in figures 4.11 trained on Computer Vision FoS data set and tested with different FoS data sets from 1970-1975 using MLR.

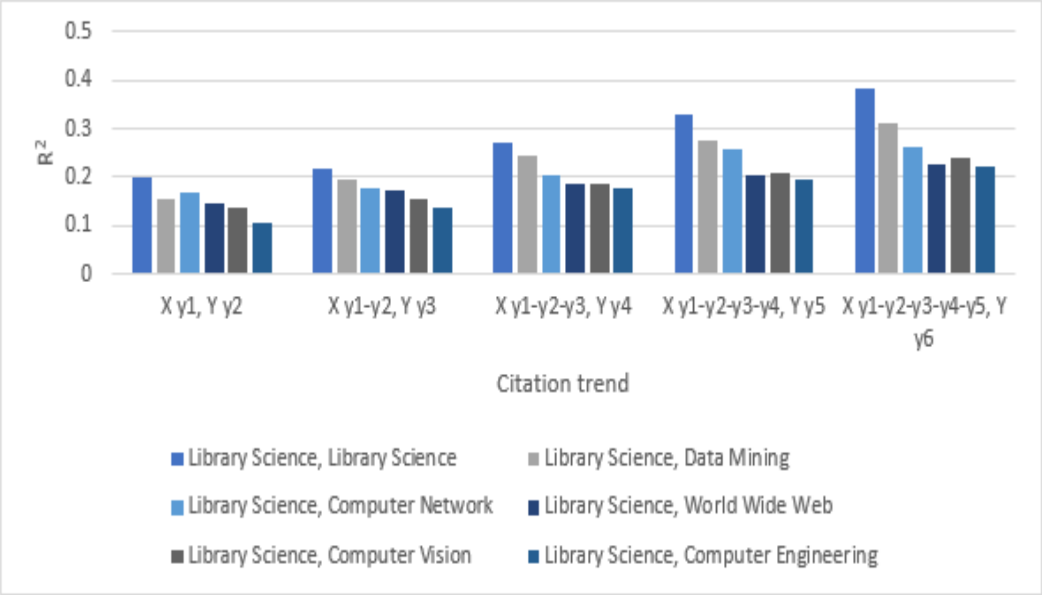


FIGURE 4.12: MLR models trained on Library Science FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in figures 4.12 trained on Library Science FoS data set and tested with different FoS data sets from 1970-1975 using MLR.

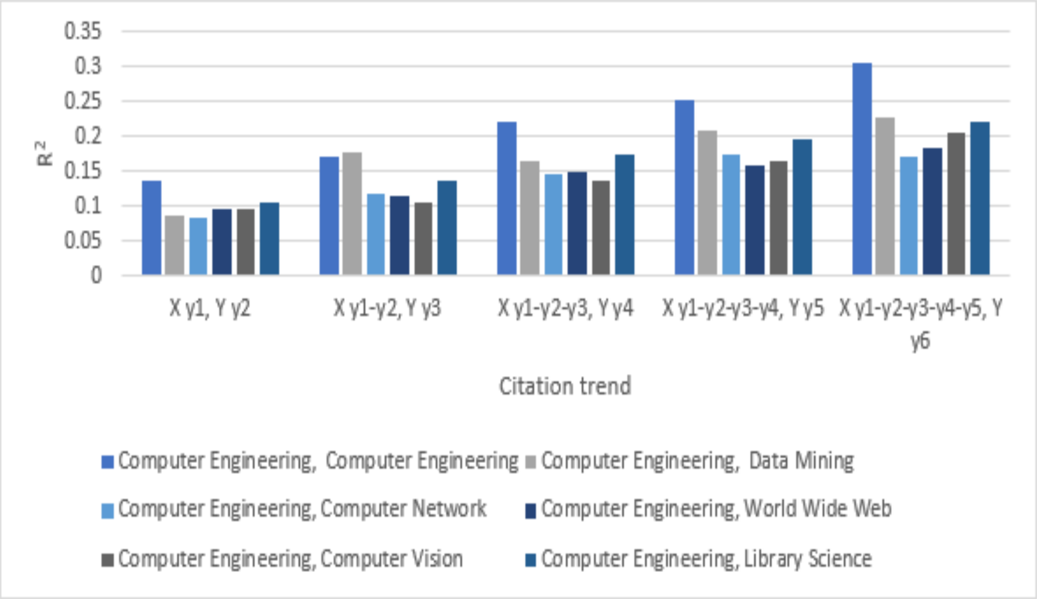


FIGURE 4.13: MLR models trained on Computer Engineering FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figures 4.13 trained on Computer Engineering FoS data set and tested with different FoS data sets from 1970-1975 using MLR.

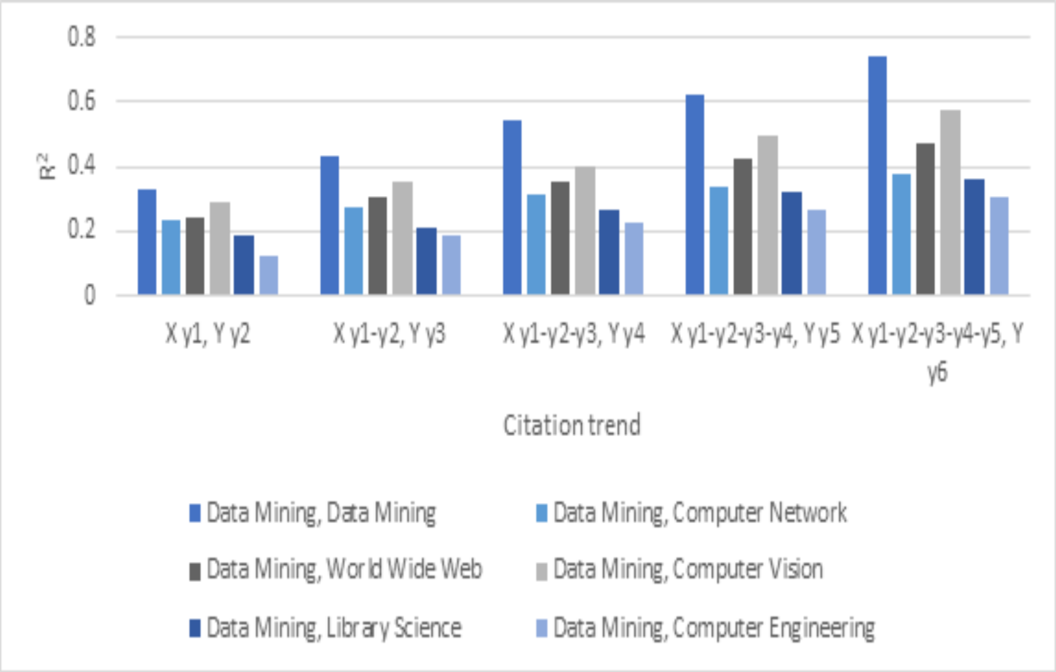


FIGURE 4.14: ANN models trained on Data Mining data set and tested with different FoS data sets from 1970-1975.

The results shown in Figures 4.14 trained on Data Mining FoS data set and tested with different FoS data sets from 1970-1975 using ANN.

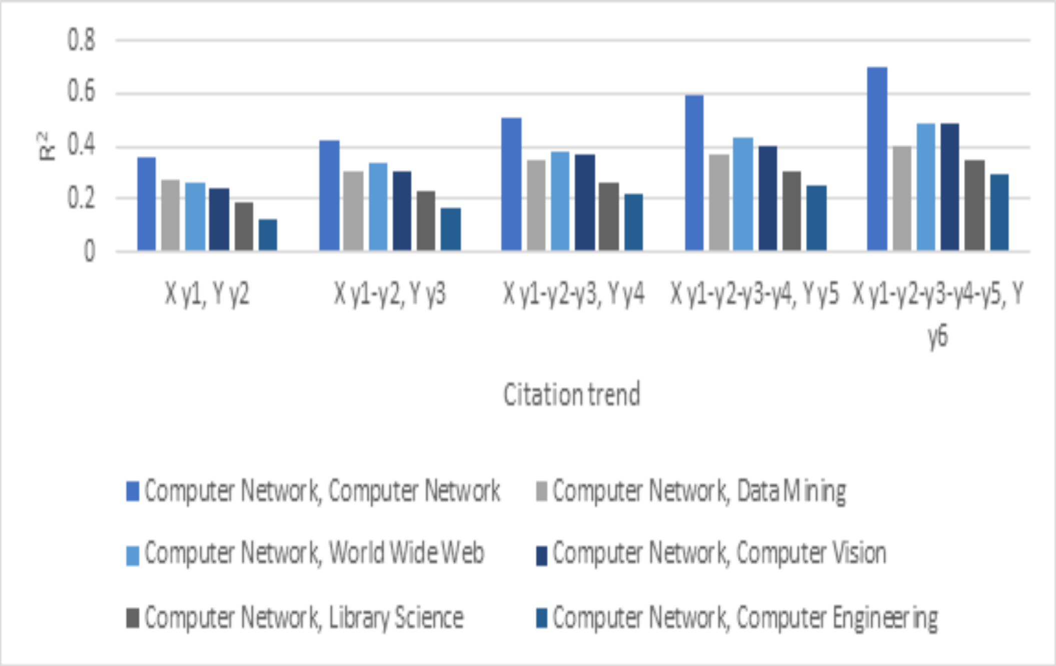


FIGURE 4.15: ANN models trained on Computer Network data set and tested with different FoS data sets from 1970-1975.

The results shown in Figures 4.15 trained on Computer Network FoS data set and tested with different FoS data sets from 1970-1975 using ANN.

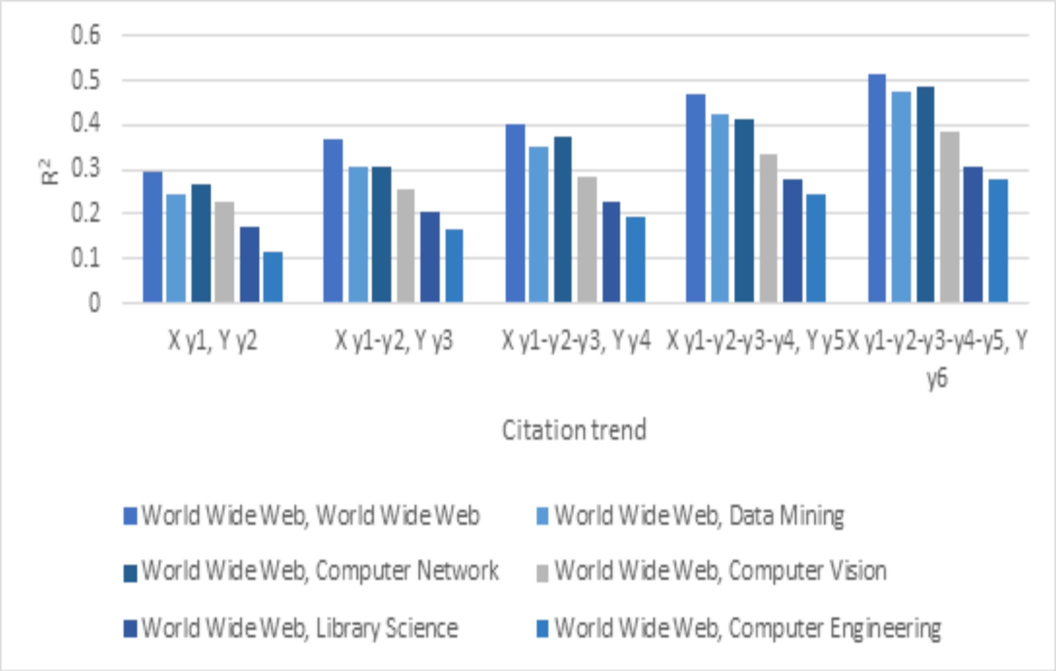


FIGURE 4.16: ANN models trained on World Wide Web data set and tested with different FoS data sets from 1970-1975.

The results shown in Figures 4.16 trained on World Wide Web FoS data set and tested with different FoS data sets from 1970-1975 using ANN.

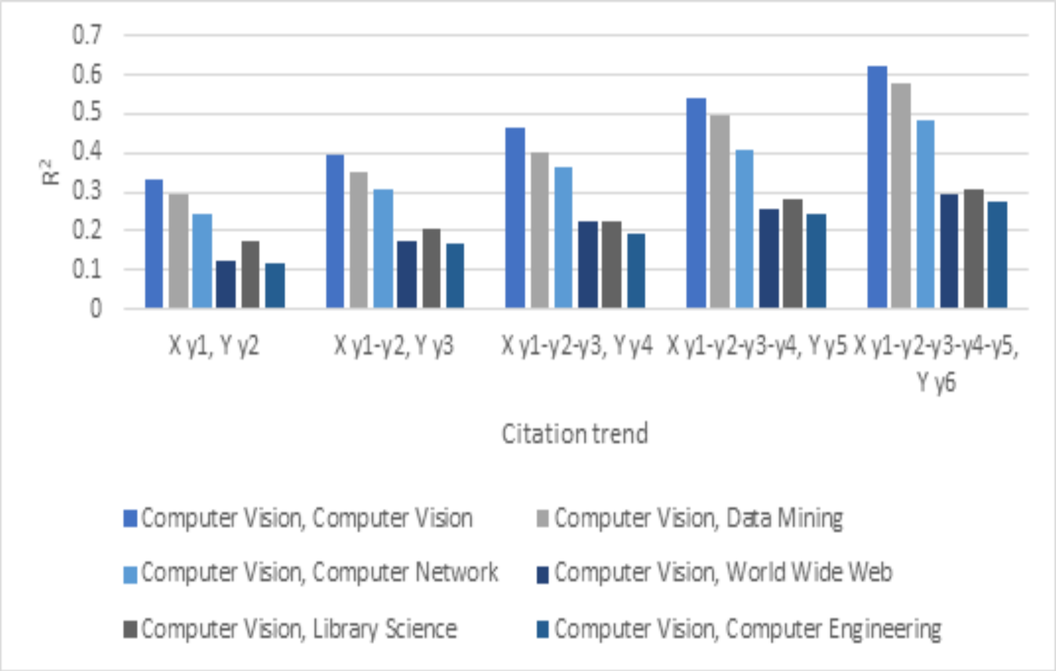


FIGURE 4.17: ANN models trained on Computer Vision data set and tested with different FoS data sets from 1970-1975.

The results shown in Figures 4.17 trained on Computer Vision FoS data set and tested with different FoS data sets from 1970-1975 using ANN.

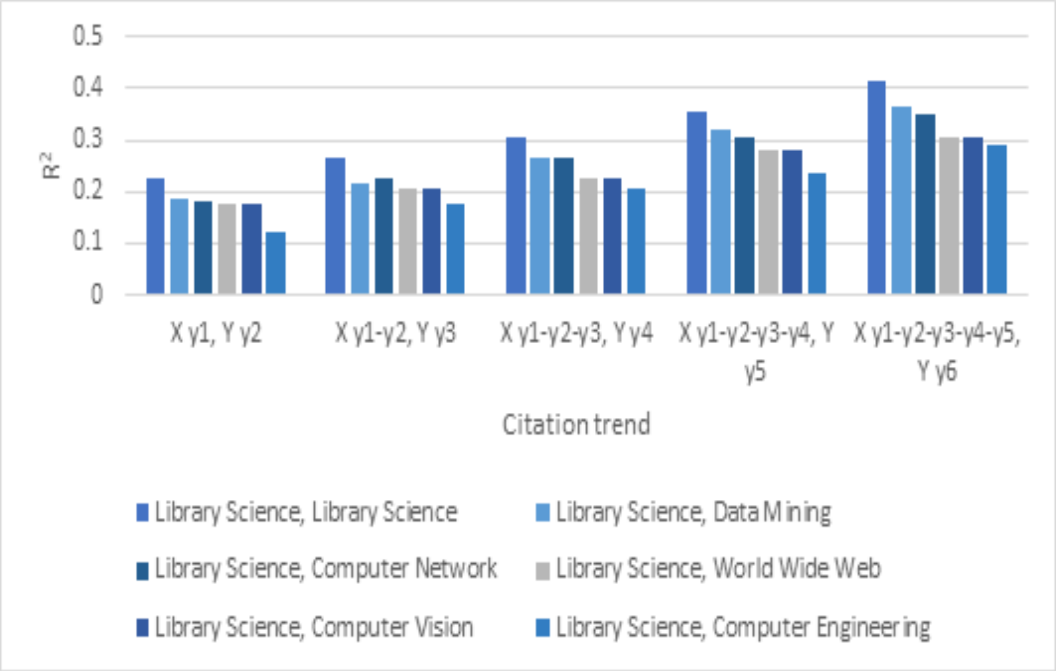


FIGURE 4.18: ANN models trained on Library Science data set and tested with different FoS data sets from 1970-1975.

The results shown in Figures 4.18 trained on Library Science FoS data set and tested with different FoS data sets from 1970-1975 using ANN.

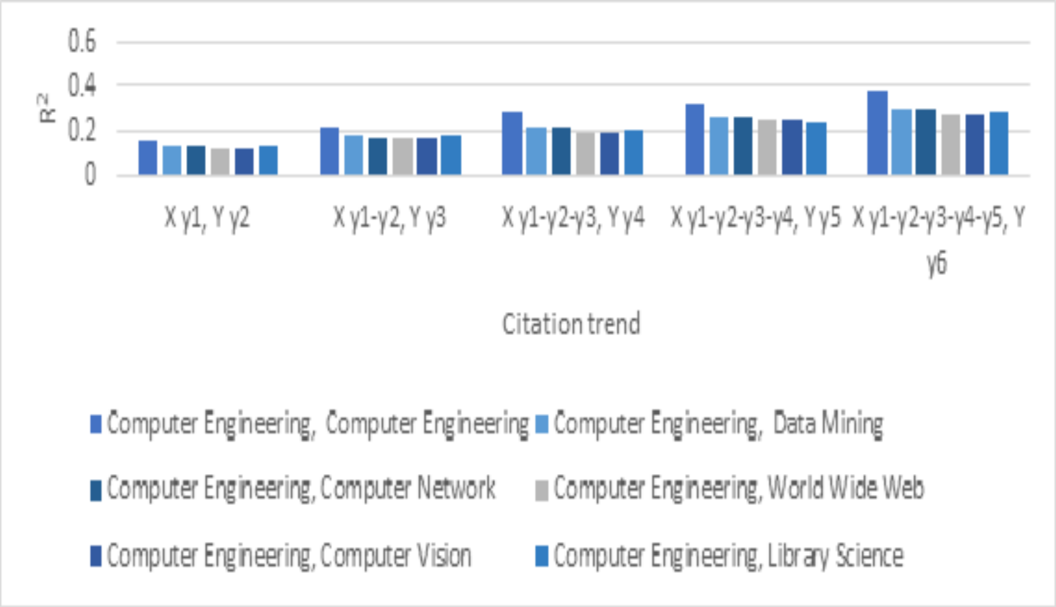


FIGURE 4.19: ANN models trained on Computer Engineering data set and tested with different FoS data sets from 1970-1975.

The results shown in Figures 4.19 trained on Computer Engineering FoS data set and tested with different FoS data sets from 1970-1975 using ANN.

TABLE 4.8: A comparative analysis of MLR and ANN models and clearly presents improvement in results.

Training FoS Dataset	Test FoS Dataset											
	Data Mining		Computer Networks		World Wide Web		Computer Vision		Library Science		Computer Engineering	
	MLR	ANN	MLR	ANN	MLR	ANN	MLR	ANN	MLR	ANN	MLR	ANN
Data Mining	**0.5736**	**0.7453**	0.3910	0.3765	0.4154	0.4749	0.4667	0.5765	0.3123	0.3628	0.2267	0.3048
Computer Networks	0.3910	0.3765	**0.5404**	**0.7004**	0.4945	0.4879	0.3478	0.4857	0.2622	0.3485	0.1723	0.2984
World Wide Web	0.4154	0.4749	0.4945	0.4879	**0.4904**	**0.5132**	0.3286	0.2943	0.2244	0.3046	0.1834	0.2756
Computer Vision	0.4667	0.5765	0.3478	0.4857	0.3286	0.2943	**0.4710**	**0.5091**	0.2384	0.3046	0.2054	0.2747
Library Science	0.3123	0.3628	0.2622	0.3485	0.2244	0.3046	0.2384	0.3046	**0.3841**	**0.4122**	0.2223	0.2885
Computer Engineering	0.2267	0.3048	0.1723	0.2984	0.1834	0.2756	0.2054	0.2747	0.2223	0.2885	**0.3037**	**0.3864**

Table 4.8 shows a comparative analysis of MLR and ANN models and clearly presents improvement in results. Figures 4.8-B.6 reproduces the results shown in

Figures 4.3-4.7 above plus the results of other five FoS with test data from all six FoS. Our finding presented above is supported for all FoS. Results show that same trend is found across these different time periods as well.

One concern that we had in these experiments is the low value of R2. In order to further improve the results, we applied ANN on the same data set and we found improvement across all models. As an example, we are presenting the results of ANN for five years' models for all six FoS in the table below:

In experiment 2, we used FoS degree trend values from 1970-1975,1990-1995, 2008-2013 time periods using MLR and ANN and we again used six different FoS "Data Mining", "Computer Network", "World Wide Web", "Computer Vision", "Library Science" and "Computer Engineering" for our experiments. Figures 4.20-4.24 shows MLR models trained on Data Mining dataset and tested with different FoS datasets.

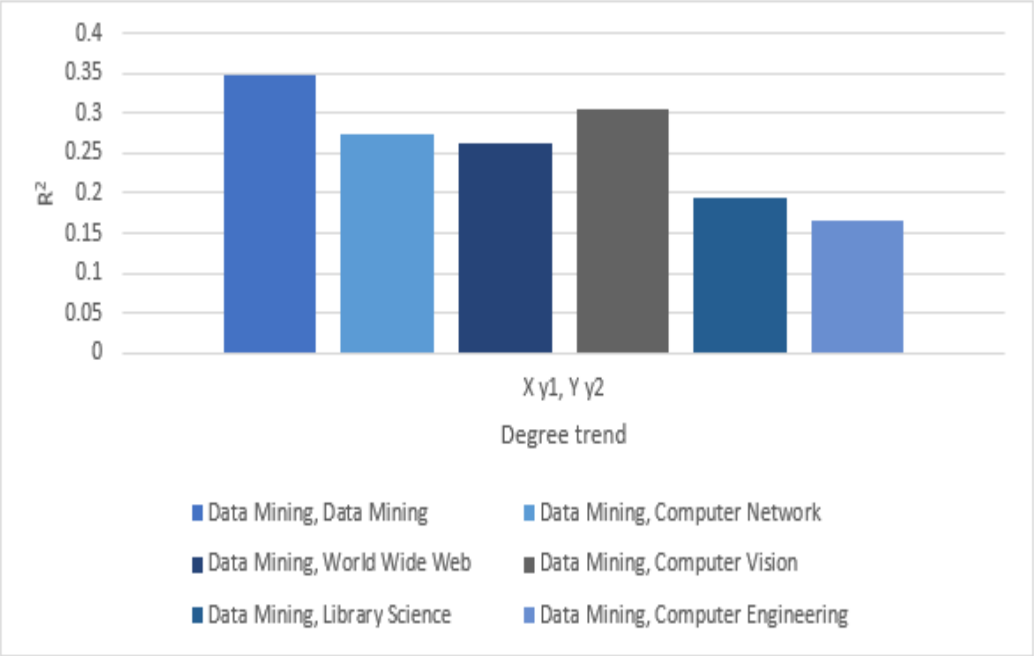


FIGURE 4.20: MLR models trained on Data Mining data set and tested with different FoS data sets.

The results shown in Figure 4.20 trained on Data Mining FoS first year data set and tested with different FoS second year data sets from 1970-1975 using MLR.

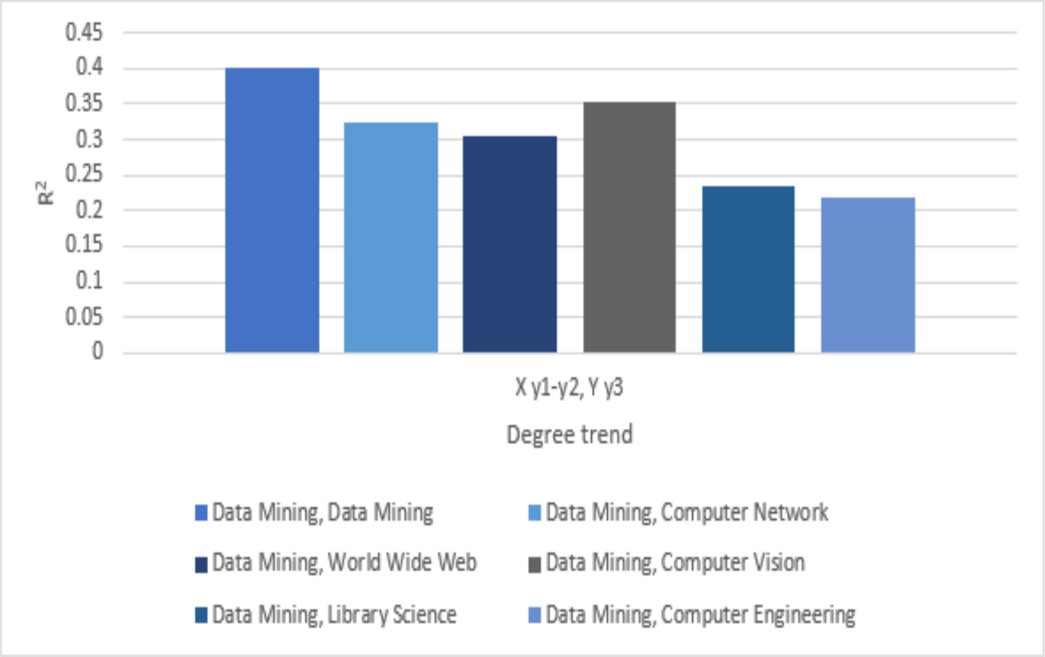


FIGURE 4.21: MLR models trained on Data Mining data set and tested with different FoS data sets.

The results shown in Figure 4.21 trained on Data Mining FoS two years data set and tested with different FoS third year data sets from 1970-1975 using MLR.

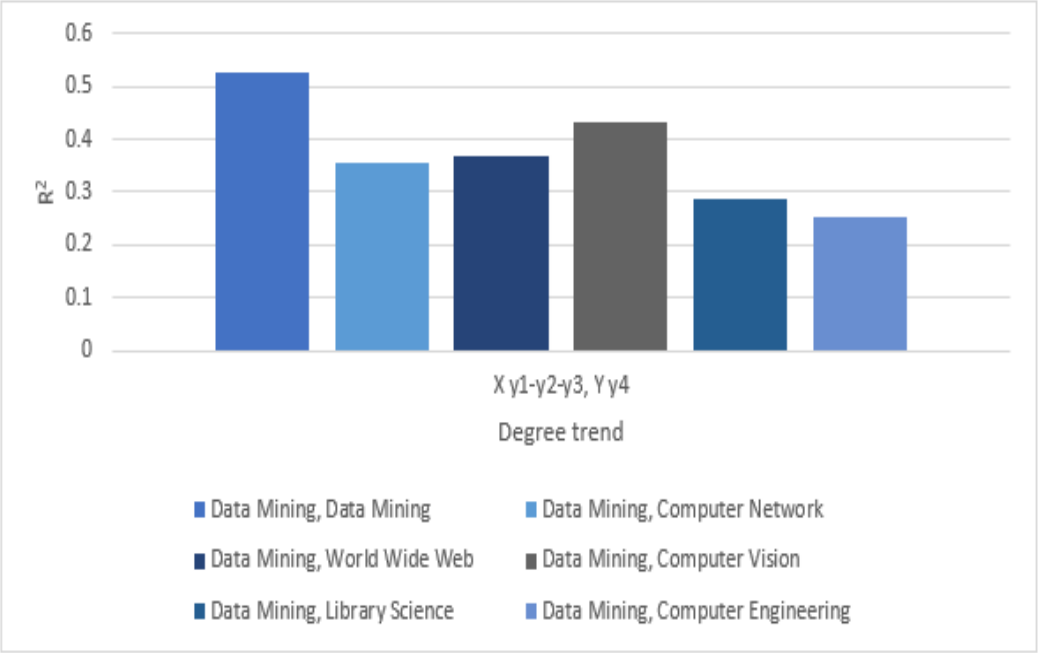


FIGURE 4.22: MLR models trained on Data Mining data set and tested with different FoS data sets.

The results shown in Figure 4.22 trained on Data Mining FoS three years data set and tested with different FoS four year data sets from 1970-1975 using MLR.

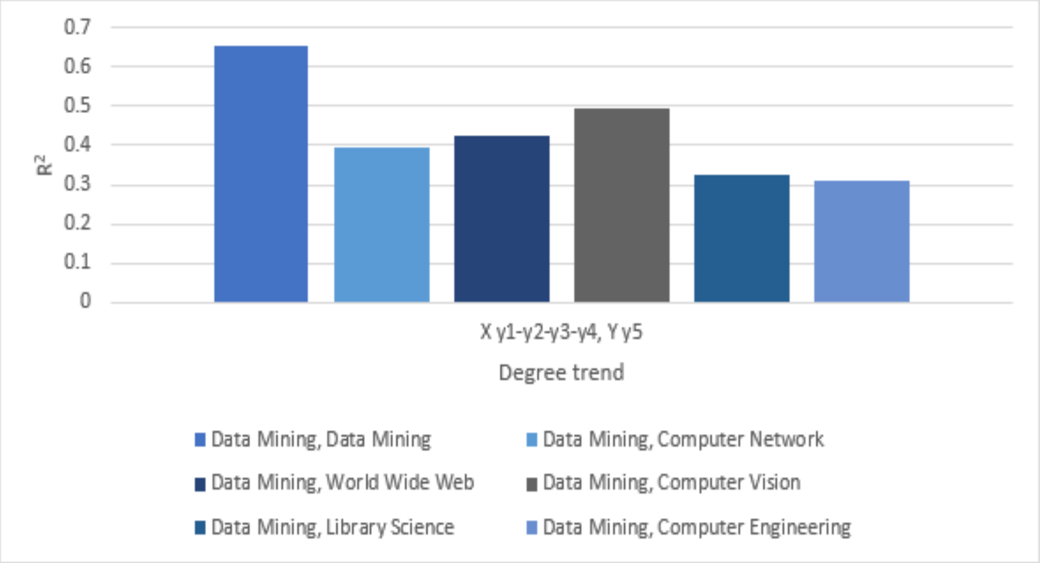


FIGURE 4.23: MLR models trained on Data Mining data set and tested with different FoS data sets.

The results shown in Figure 4.23 trained on Data Mining FoS four years data set and tested with different FoS five year data sets from 1970-1975 using MLR.

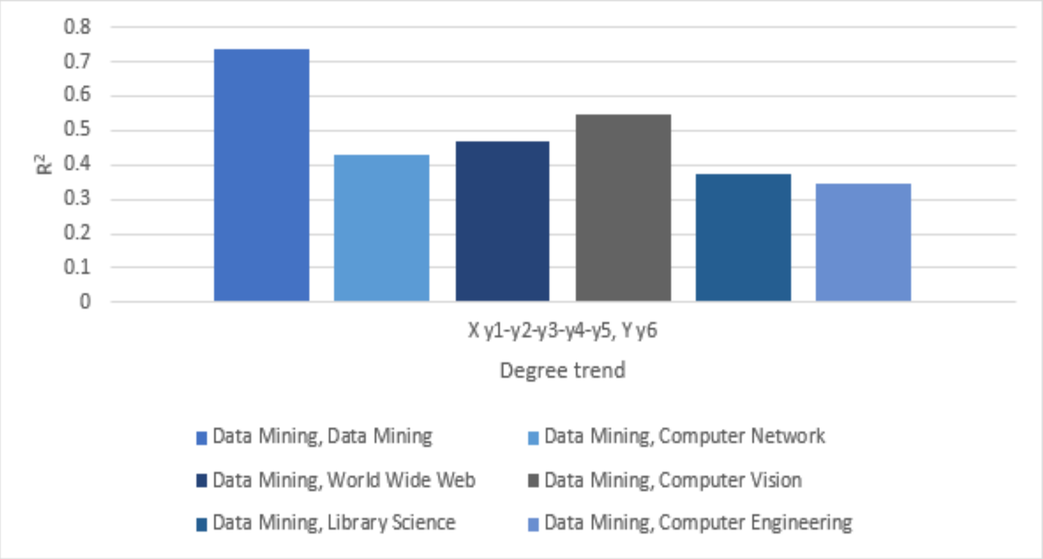


FIGURE 4.24: MLR models trained on Data Mining data set and tested with different FoS data sets.

The results shown in Figure 4.24 trained on Data Mining FoS five years data set and tested with different FoS six year data sets from 1970-1975 using MLR.

Figures 4.20-4.24 results show that same FoS attained similar citation trends as compared to different FoS. The results depict that FoS degree plays a significant role and this is an important measure in FoS trend following because R^2 values are much better as compared to citation trend. The results show that same FoS have similar R^2 values in different time periods and R^2 increased as number of input years are also increased.

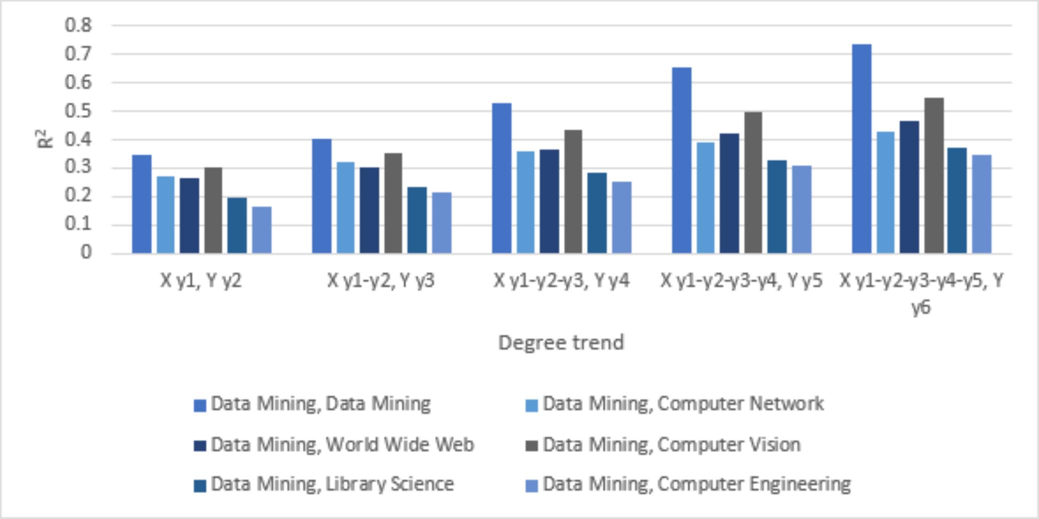


FIGURE 4.25: MLR models trained on Data Mining FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figure 4.25 trained on Data Mining FoS data set and tested with different FoS data sets from 1970-1975 using MLR.

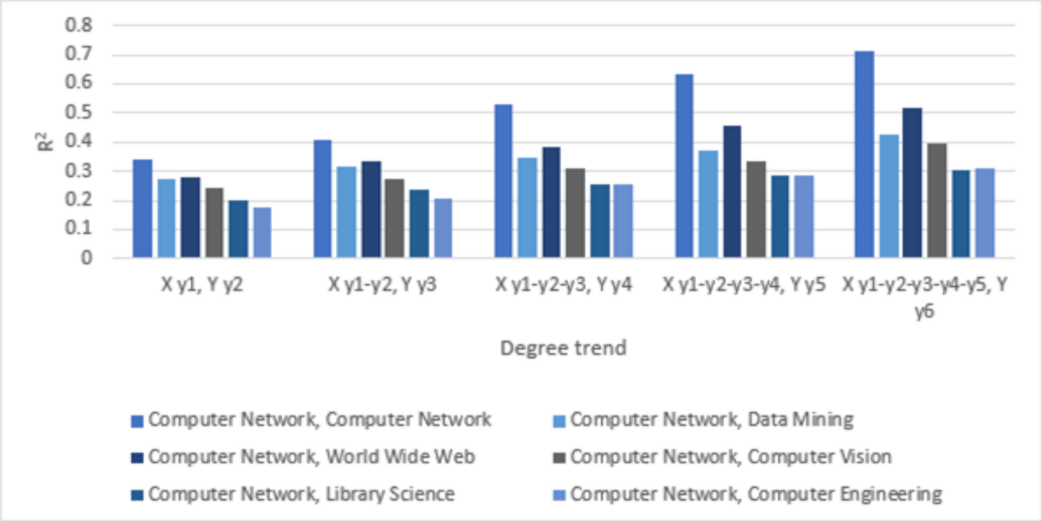


FIGURE 4.26: MLR models trained on Computer Network FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figure 4.26 trained on Computer Network FoS data set and tested with different FoS data sets from 1970-1975 using MLR.

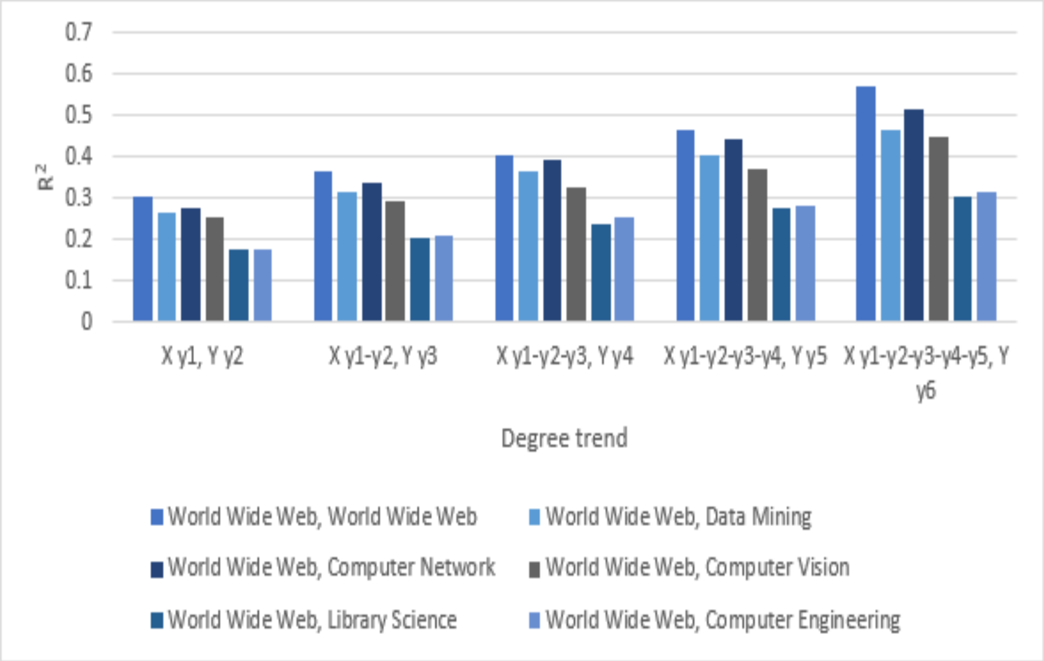


FIGURE 4.27: MLR models trained on World Wide Web FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figure 4.27 trained on World Wide Web FoS data set and tested with different FoS data sets from 1970-1975 using MLR.

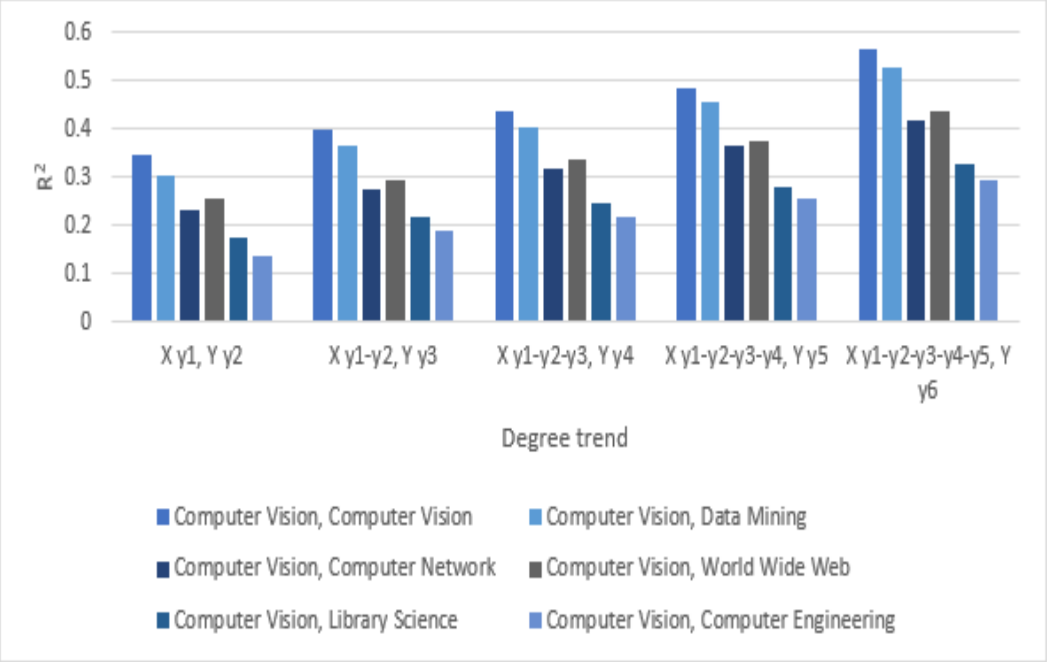


FIGURE 4.28: MLR models trained on Computer Vision FoS data set and tested with different FoS data sets from 1970-1975

The results shown in Figure 4.28 trained on Computer Vision FoS data set and tested with different FoS data sets from 1970-1975 using MLR.

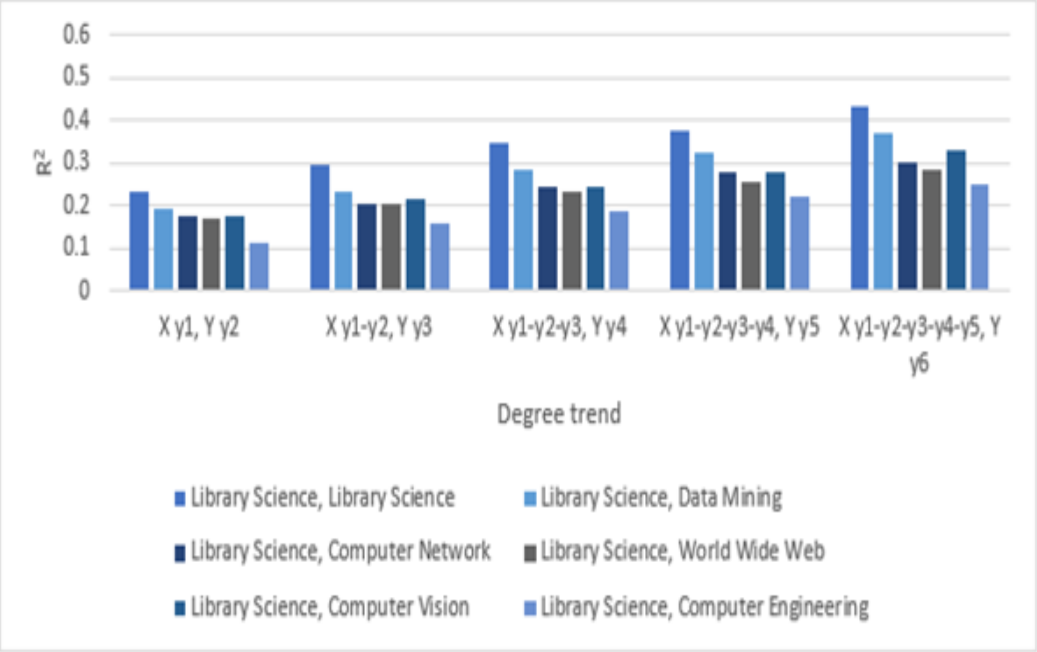


FIGURE 4.29: MLR models trained on Library Science FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figure 4.29 trained on Library Science FoS data set and tested with different FoS data sets from 1970-1975 using MLR.

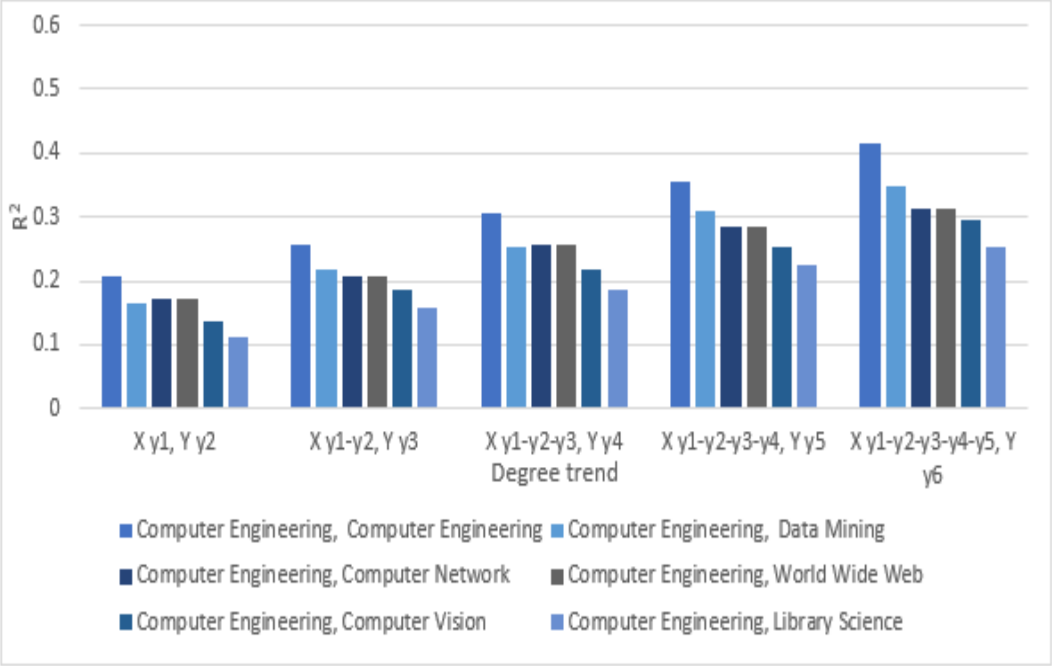


FIGURE 4.30: MLR models trained on Computer Engineering FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figure 4.30 trained on Computer Engineering FoS data set and tested with different FoS data sets from 1970-1975 using MLR.

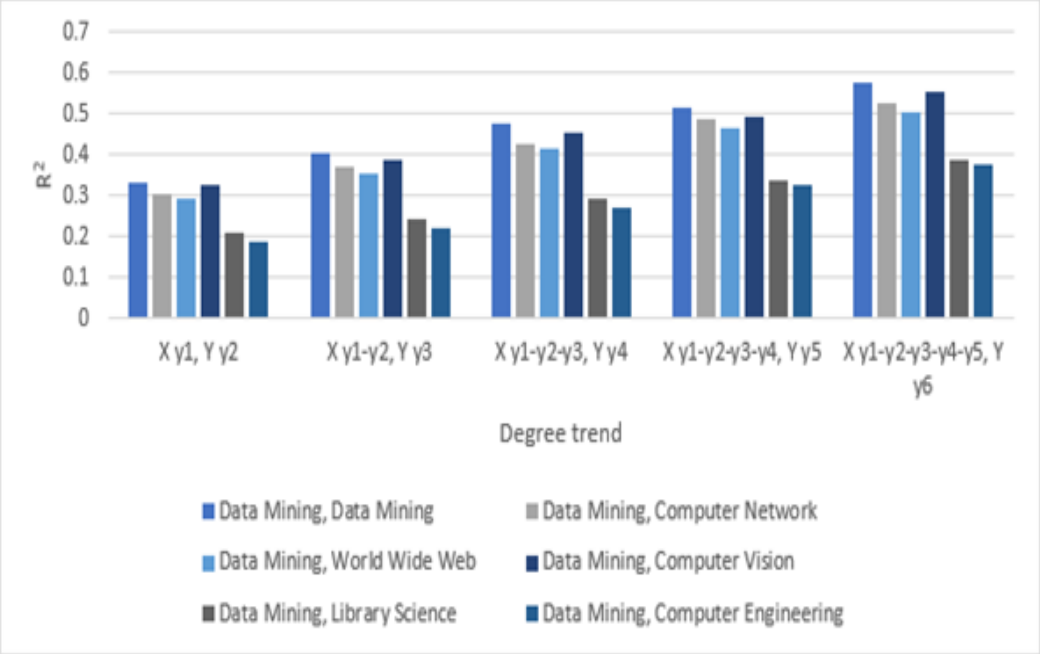


FIGURE 4.31: MLR models trained on Data Mining FoS data set and tested with different FoS data sets from 1990-1995.

The results shown in Figure 4.31 trained on Data Mining FoS data set and tested with different FoS data sets from 1990-1995 using MLR.

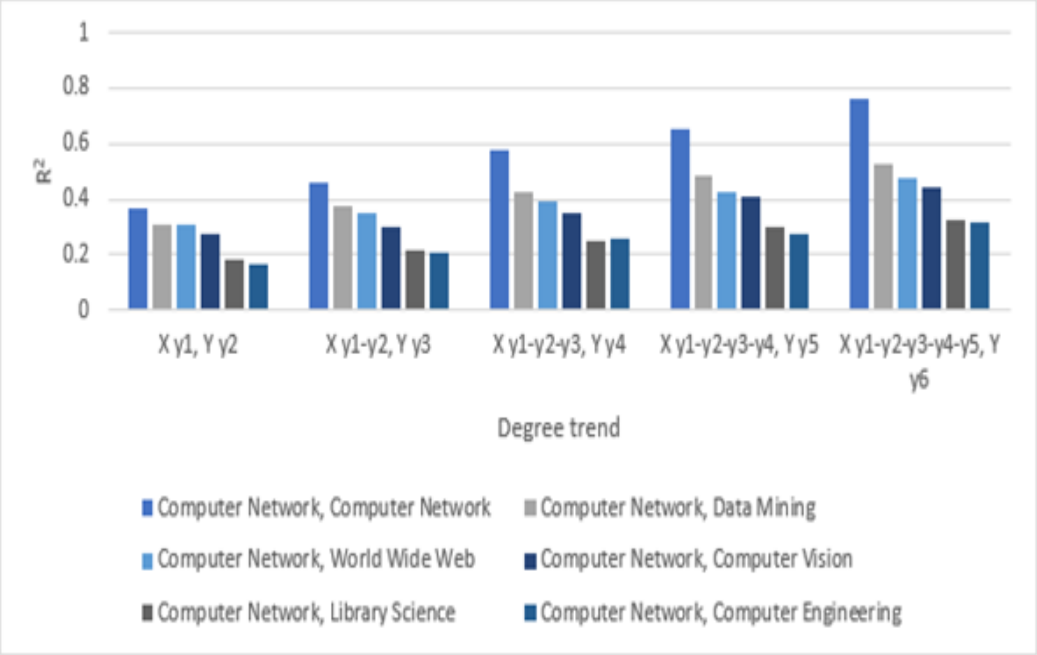


FIGURE 4.32: MLR models trained on Computer Network FoS data set and tested with different FoS data sets from 1990-1995.

The results shown in Figure 4.32 trained on Computer Network FoS data set and tested with different FoS data sets from 1990-1995 using MLR.

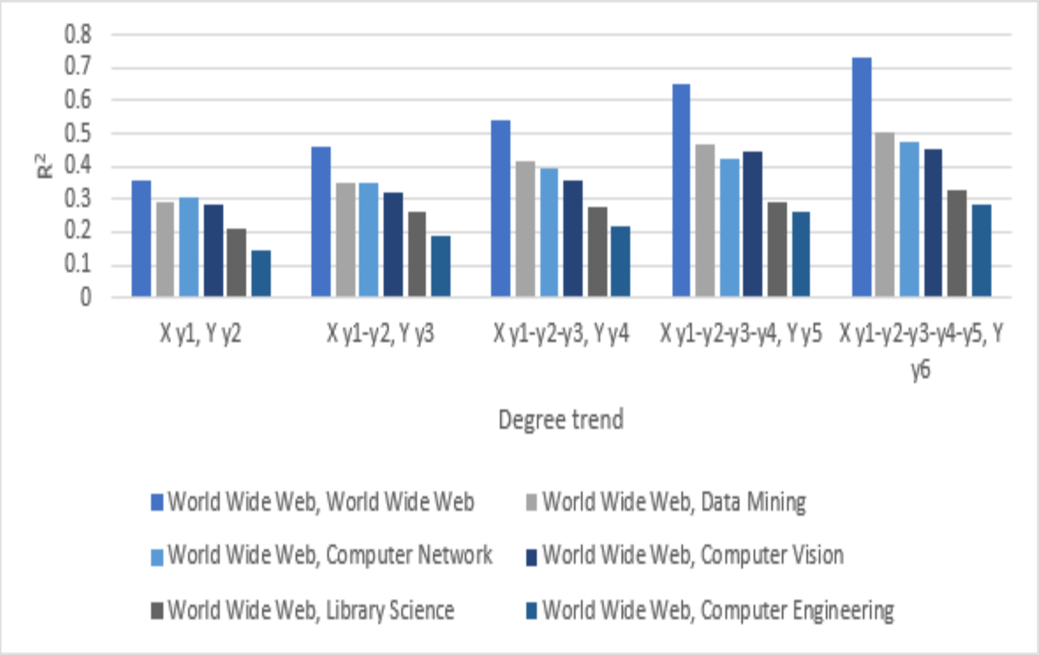


FIGURE 4.33: MLR models trained on World Wide Web FoS data set and tested with different FoS data sets from 1990-1995

The results shown in Figure 4.33 trained on World Wide Web FoS data set and tested with different FoS data sets from 1990-1995 using MLR.

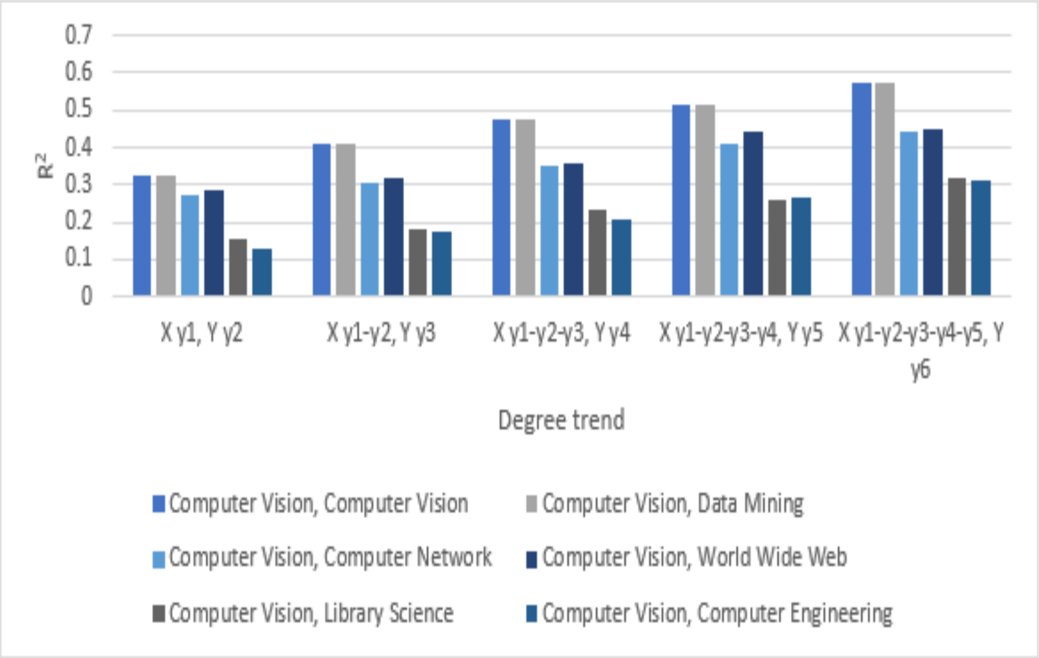


FIGURE 4.34: MLR models trained on Computer Vision FoS data set and tested with different FoS data sets from 1990-1995.

The results shown in Figure 4.34 trained on Computer Vision FoS data set and tested with different FoS data sets from 1990-1995 using MLR.

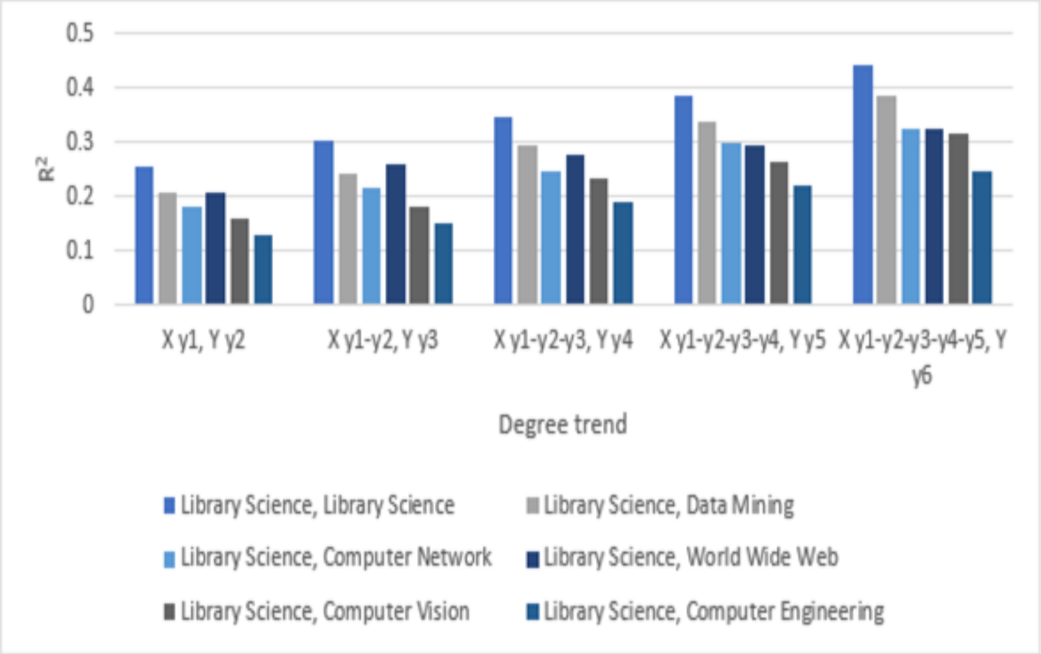


FIGURE 4.35: MLR models trained on Library Science FoS data set and tested with different FoS data sets from 1990-1995.

The results shown in Figure 4.35 trained on Library Science FoS data set and tested with different FoS data sets from 1990-1995 using MLR.

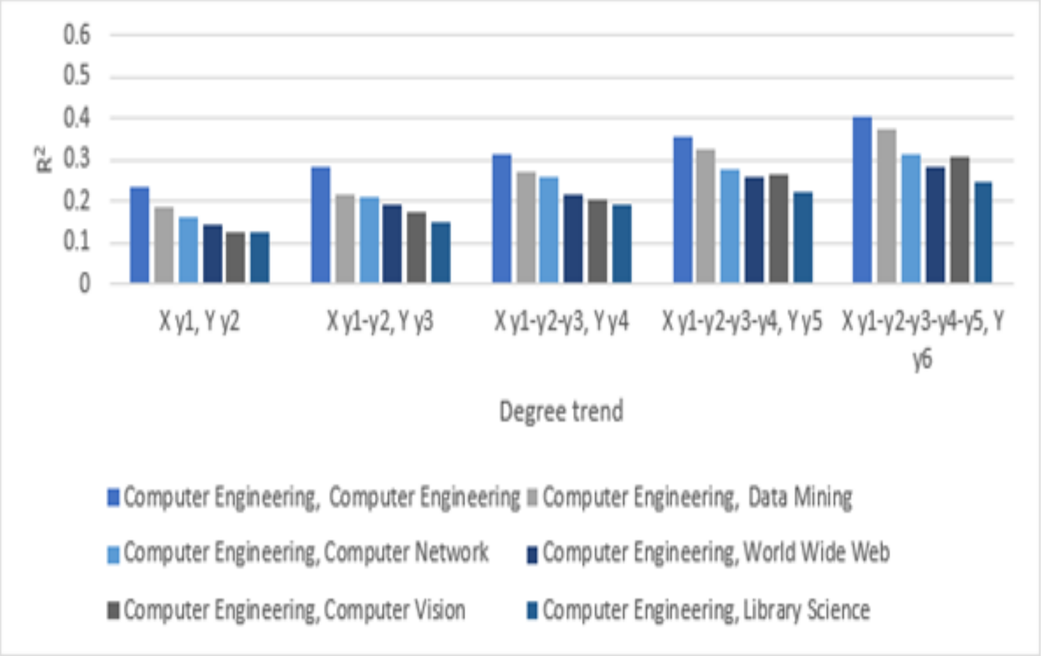


FIGURE 4.36: MLR models trained on Computer Engineering FoS data set and tested with different FoS data sets from 1990-1995.

The results shown in Figure 4.36 trained on Computer Engineering FoS data set and tested with different FoS data sets from 1990-1995 using MLR.

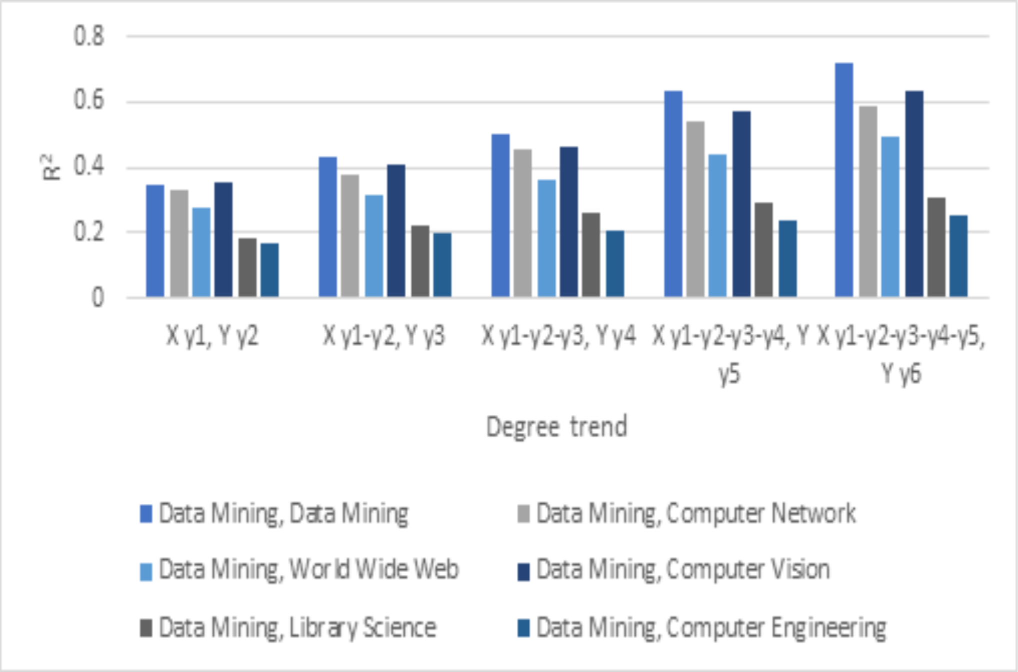


FIGURE 4.37: ANN models trained on Data Mining FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figure 4.37 trained on Data Mining FoS data set and tested with different FoS data sets from 1970-1975 using ANN.

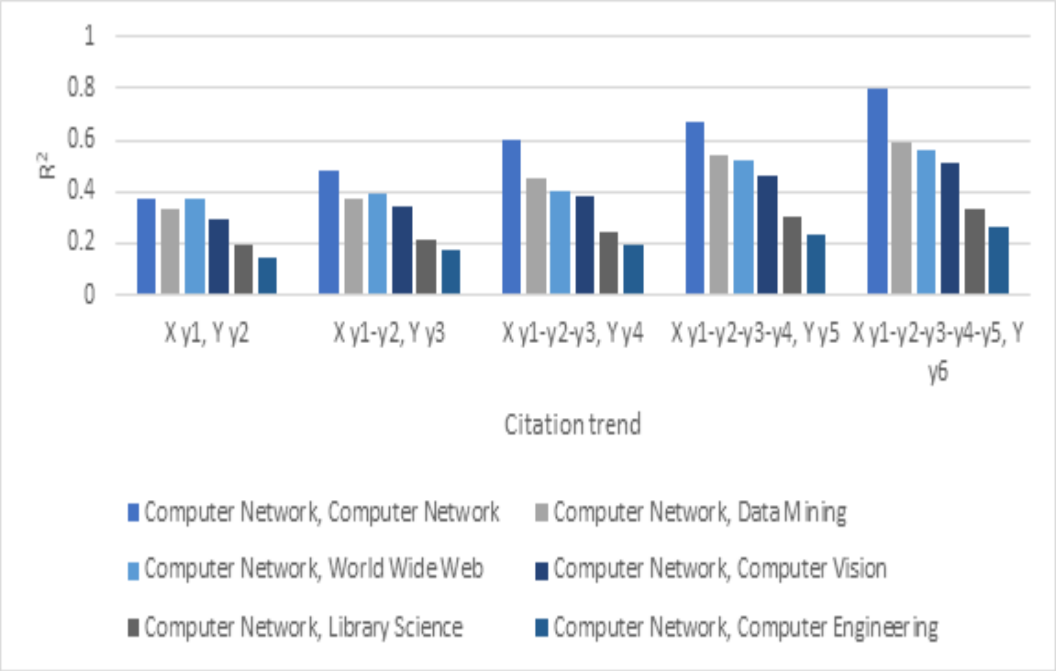


FIGURE 4.38: ANN models trained on Computer Network FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figure 4.38 trained on Computer Network FoS data set and tested with different FoS data sets from 1970-1975 using ANN.

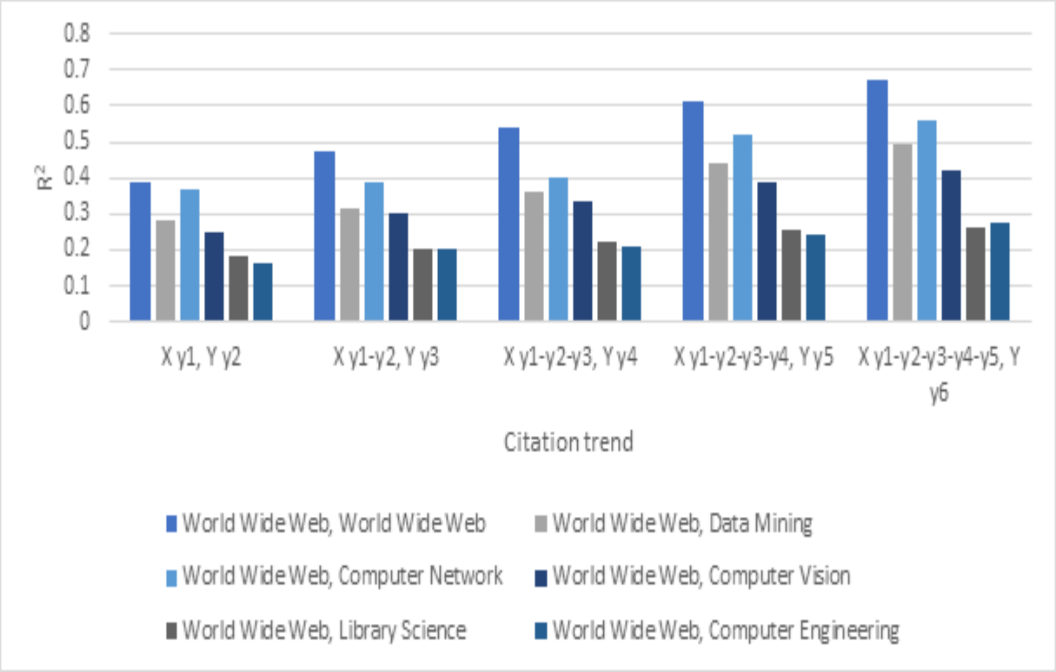


FIGURE 4.39: ANN models trained on World Wide Web FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figure 4.39 trained on World Wide Web FoS data set and tested with different FoS data sets from 1970-1975 using ANN.

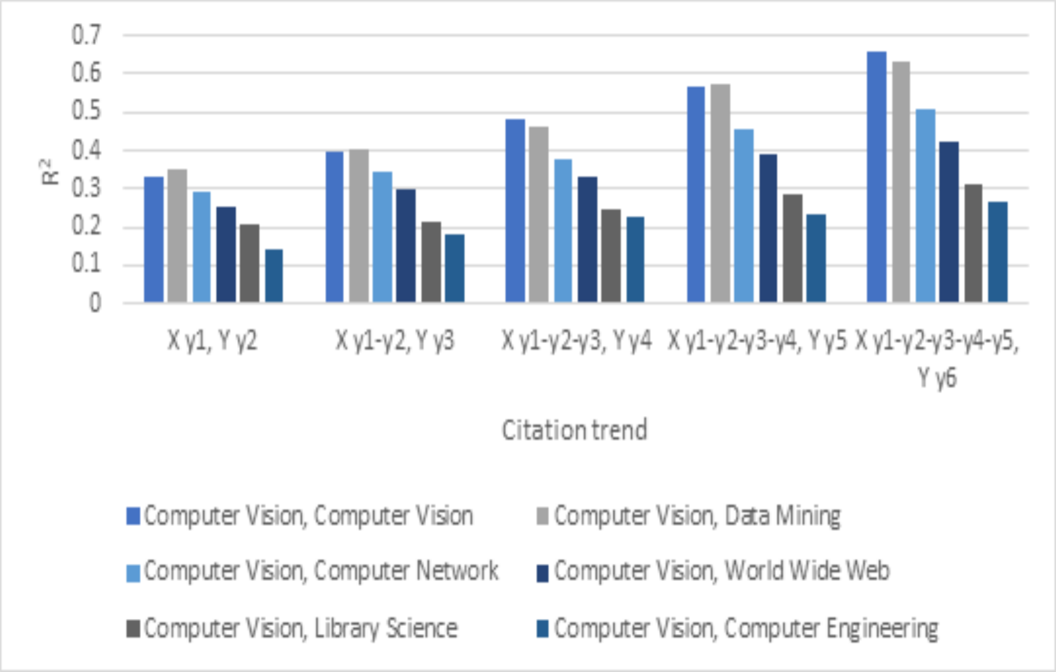


FIGURE 4.40: ANN models trained on Computer Vision FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figure 4.40 trained on Computer Vision FoS data set and tested with different FoS data sets from 1970-1975 using ANN.

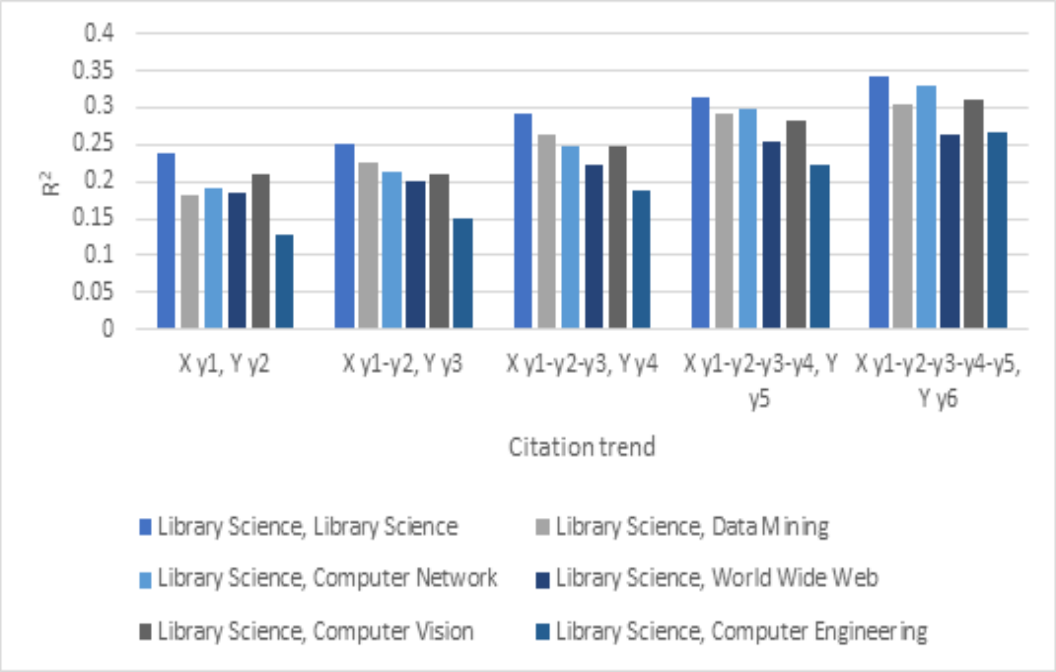


FIGURE 4.41: ANN models trained on Library Science FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figure 4.41 trained on Library Science FoS data set and tested with different FoS data sets from 1970-1975 using ANN.

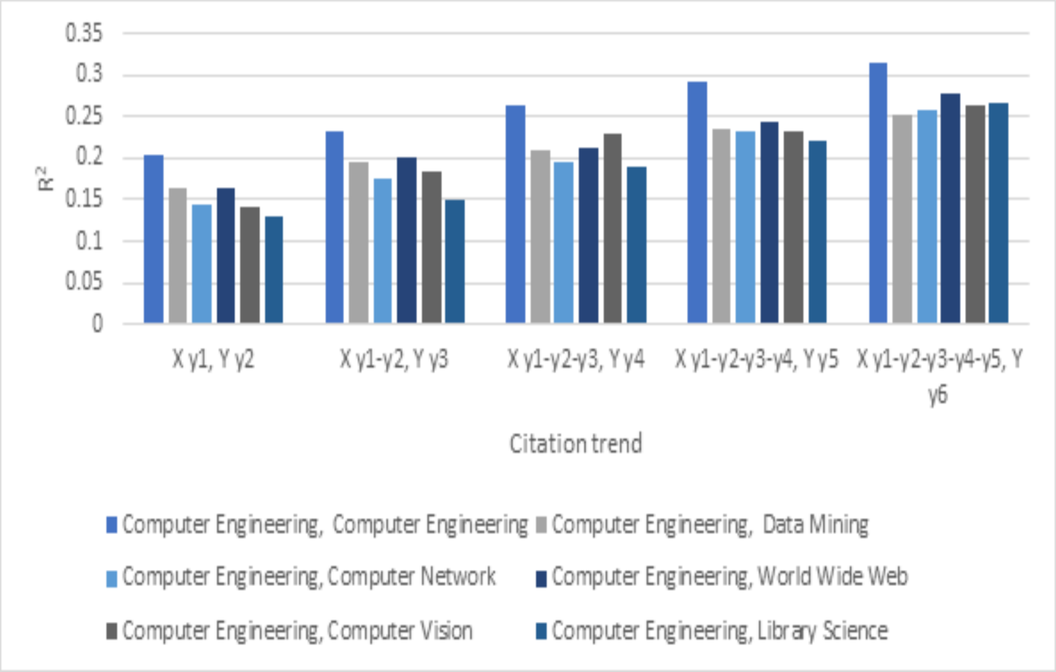


FIGURE 4.42: ANN models trained on Computer Engineering FoS data set and tested with different FoS data sets from 1970-1975.

The results shown in Figure 4.42 trained on Computer Engineering FoS data set and tested with different FoS data sets from 1970-1975 using ANN.

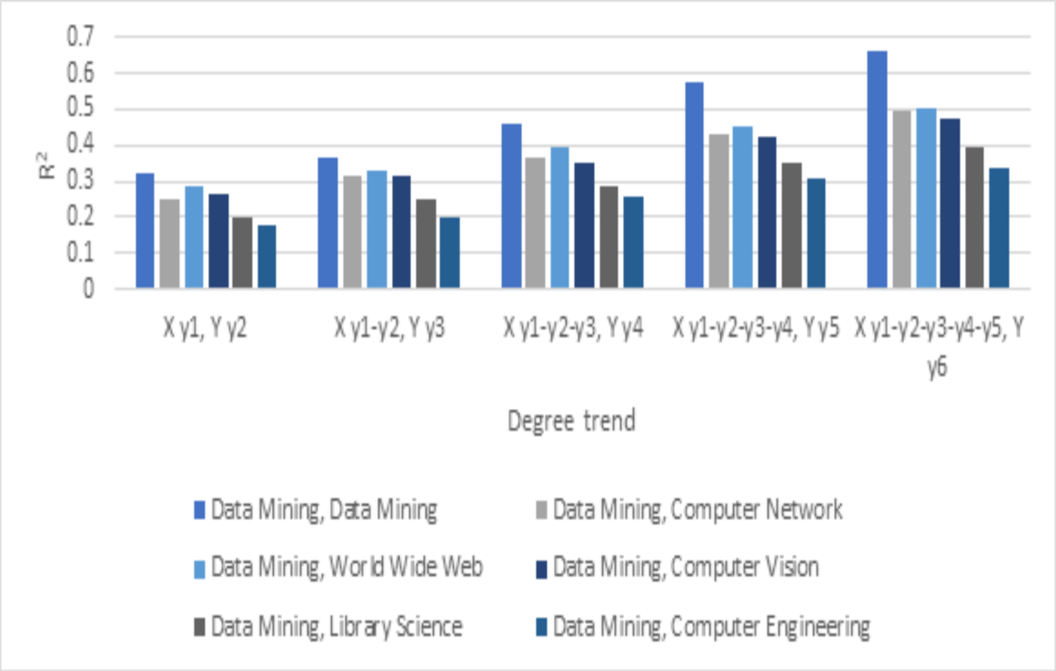


FIGURE 4.43: ANN models trained on Data Mining FoS data set and tested with different FoS data sets from 1990-1995.

The results shown in Figure 4.43 trained on Data Mining FoS data set and tested with different FoS data sets from 1990-1995 using ANN.

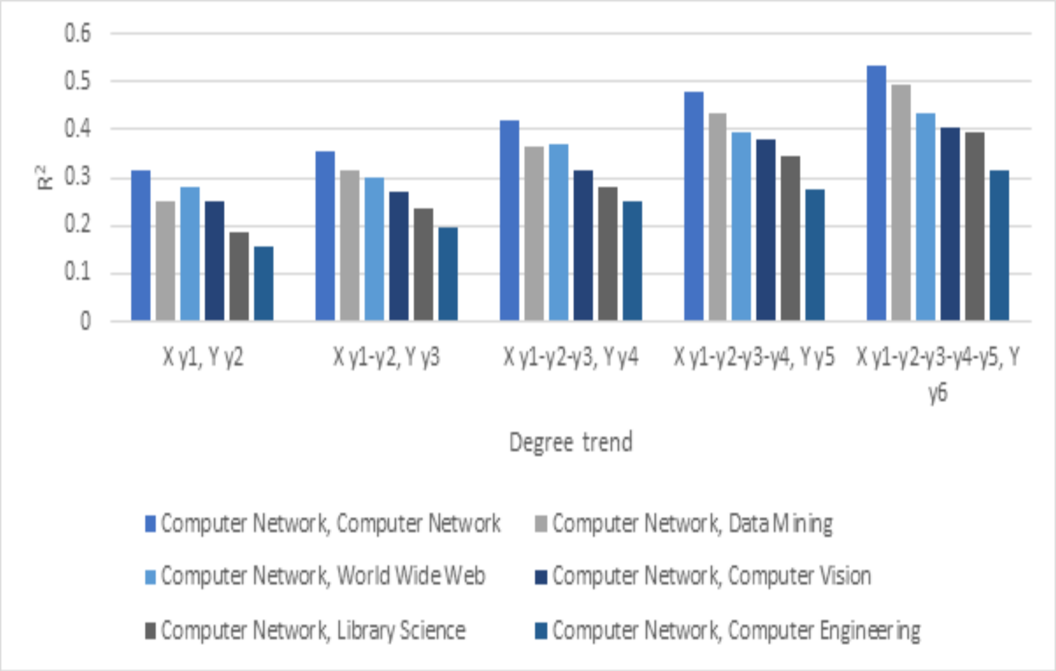


FIGURE 4.44: ANN models trained on Computer Network FoS data set and tested with different FoS data sets from 1990-1995.

The results shown in Figure 4.44 trained on Computer Network FoS data set and tested with different FoS data sets from 1990-1995 using ANN.

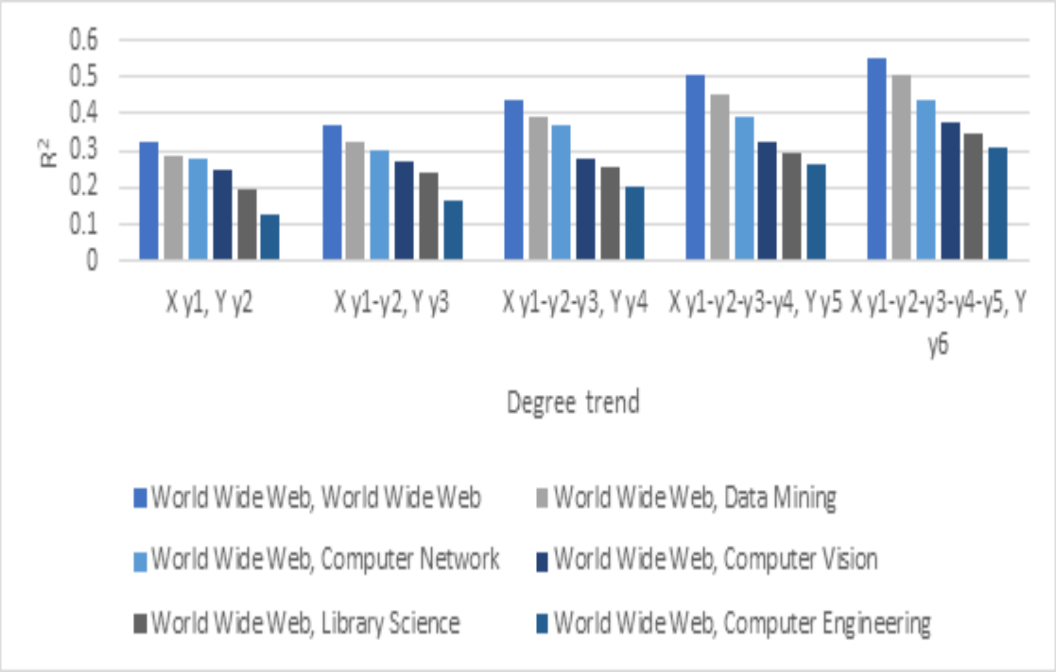


FIGURE 4.45: ANN models trained on World Wide Web FoS data set and tested with different FoS data sets from 1990-1995.

The results shown in Figure 4.45 trained on World Wide Web FoS data set and tested with different FoS data sets from 1990-1995 using ANN.

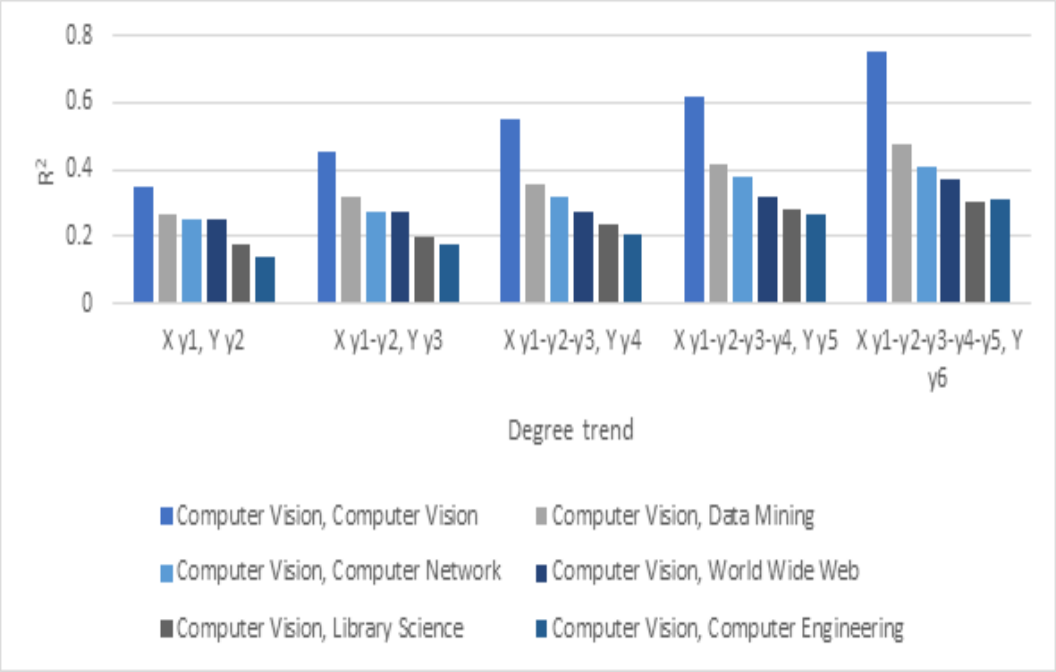


FIGURE 4.46: ANN models trained on Computer Vision FoS data set and tested with different FoS data sets from 1990-1995.

The results shown in Figure 4.46 trained on Computer Vision FoS data set and tested with different FoS data sets from 1990-1995 using ANN.

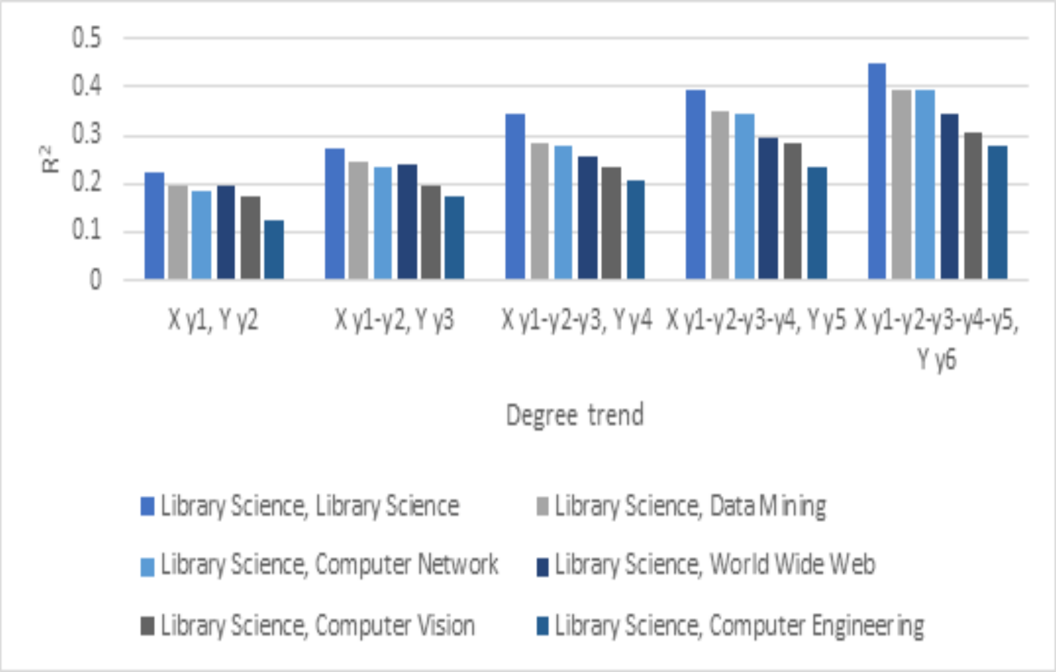


FIGURE 4.47: ANN models trained on Library Science FoS data set and tested with different FoS data sets from 1990-1995.

The results shown in Figure 4.47 trained on Library Science FoS data set and tested with different FoS data sets from 1990-1995 using ANN.

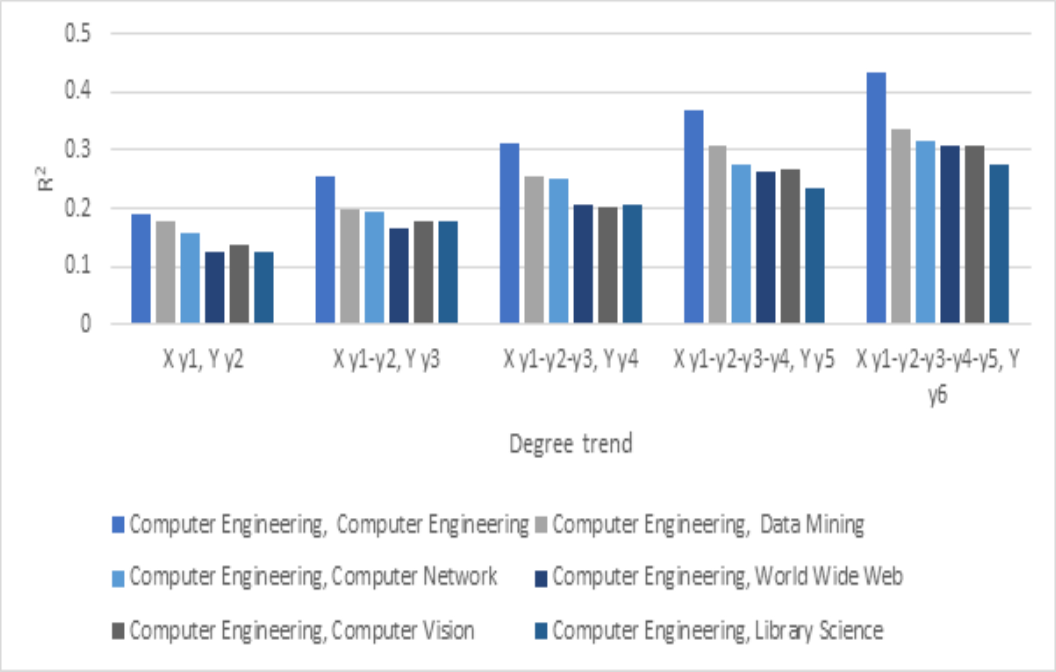


FIGURE 4.48: ANN models trained on Computer Engineering FoS data set and tested with different FoS data sets from 1990-1995.

The results shown in Figure 4.48 trained on Computer Engineering FoS data set and tested with different FoS data sets from 1990-1995 using ANN.

TABLE 4.9: Training and Test FoS datasets R^2 for predicting 6th year as output using MLR and ANN.

Training	Test FoS Dataset											
FoS	Data		Computer		World		Computer		Library		Computer	
Dataset	Mining		Networks		Wide Web		Vision		Science		Engineering	
	MLR	ANN	MLR	ANN	MLR	ANN	MLR	ANN	MLR	ANN	MLR	ANN
Data Mining	**0.5934**	**0.7932**	0.4451	0.4287	0.4587	0.5023	0.5285	0.6567	0.3865	0.4587	0.2989	0.3976
Computer Networks	0.4212	0.3911	**0.5971**	**0.7539**	0.5617	0.5487	0.4012	0.5745	0.3426	0.4598	0.2456	0.3675
World Wide Web	0.4438	0.4961	0.5451	0.5234	**0.5756**	**0.5946**	0.4034	0.3623	0.2876	0.3822	0.2561	0.3546
Computer Vision	0.4933	0.6012	0.3876	0.5465	0.3954	0.3425	**0.5693**	**0.58652**	0.2865	0.3999	0.2719	0.3743
Library Science	0.3412	0.3956	0.3013	0.3954	0.2865	0.3654	0.2876	0.3876	**0.4567**	**0.4945**	0.2986	0.3534
Computer Engineering	0.2568	0.3561	0.2176	0.3467	0.2472	0.3087	0.2754	0.3654	0.2844	0.3551	**0.3965**	**0.4364**

Figure 4.9 shows a comparative analysis of MLR and ANN models and clearly presents improvement in results. Figures 4.25-4.48 results show that there is a similarity between citation trend of authors that belong to the same FoS as compared to different FoS and achieved consistent R^2 values. FoS degree trend results are much better than citation trend results. FoS degree trend results also expose the impact and worth of the scientific research field. This shows the FoS trend following has a certain impact on the citation count of authors. Further, the result proves that if authors follows the same FoS trend, they have similar citation trend.

In this chapter we discussed experiments and results of our three research questions. For RQ-1, the RI results show a similar level of similarity between clustering based on FoS and four different measures, i.e., frequency, degree, betweenness, closeness. Frequency and degree centrality have relatively higher values of RI as compared to the other two and out of these two- degree centralities have the highest RI values across multiple years. As the results indicate, the degree has achieved the highest RI value 0.69.

The results indicate that if the papers belong to the same FoS, then there are 69% of chances, that they have the same citation trend. This proves that a field of study has a certain impact on citation count of a paper and researchers should also contemplate on the trend of a field of study while selecting a particular research area. Also, the degree centrality is a more suitable metric to measure the trend of an FoS than a simple citation count.

Further, in RQ-2 we identified researchers in the early stage of an FoS trend, that is, trend setters for the FoS of Semantic Search. Finally, for RQ-3 the results proved that if a person publishes in a particular FoS, then the citation trend of this author's work resembles more to the overall citation trend of that particular FoS than that of some other FoS. The R^2 value is though not very high, but it is higher in all the cases in the same FoS than the other ones. This gives enough evidence to believe the similarity of an author's citation trend with that of the particular FoS.

The results also show that FoS degree trend results are much better than citation trend results. FoS degree trend results also expose the impact and worth of the scientific research field. This shows the FoS trend following has a certain impact on the citation count of authors. Further, the result proves that if authors follows the same FoS trend, they have similar citation trend.

The next section presents the comprehensive conclusion of this study.

Chapter 5

Conclusion and Future Work

This chapter concludes the thesis by presenting the main contributions related to impact of following FoS trend on research papers and authors citation count. Further, future work scopes and issues in this area are also discussed.

5.1 Summary and Contributions of the Work

FoS trend following is significant for Computer Science researchers as they will be able to smartly guess at what could be coming ahead of the research fields in Computer Science and will help to make positive and insight decisions for the future of their research fields in this area. Similarly, researchers will aptly be able to anticipate the new FoS trends coming in Computer Science field, because they will have already have a good idea of what is already coming. FoS trend following also provides researchers a good advantage in the fast-paced world and to be able to connect well professionally and also it creates a high impact on their paper citations and their careers. The significance of identifying FoS trends, a researcher could determine fields of interest with respect to its success or impact. The ability to recognize FoS trends is noteworthy for anyone involved in the research environment, including researchers, academic publishers, journal editors, institutional funding bodies and other relevant stakeholders.

In this thesis, we have analyzed the impact of following an FoS trend. We developed an approach to identify the impact of FoS trend on research paper citations. This research question will establish the similarity between the citation trends of authors belonging to same FoS. To answer this question, we collected papers from MAG dataset belonging to different FoS, and listed the citation patterns of all papers for five years. We then performed clustering on FoS and citations trend of papers separately. We compared the similarity between these two clusters. Rand Index (RI) has been used to compare the similarity between the two clusters. We performed another experiment for the same RQ with a novelty that we developed an FoS Multigraph (FoM) from where we computed different centrality measures. Then, we used these centrality measures in the same experiment with the objective to find a better metric to establish similarity in the trend of papers belonging to same FoS. Once again we used RI for the purpose and correlation coefficient have been employed to find the relationship between FoS citations pattern.

The experimental results show that there is a similarity between clusters formed on the basis of FoS and citations pattern and there exists a relationship between FoS citations pattern that belong to the same FoS. The results indicate that FoS hold a certain impact on the citation count. Further, if the papers belong to the same FoS, then there are 66% of chances that they hold a same citation trend as they achieved high correlation value. This proves that an FoS has a certain impact on the citation count of a research paper and researchers need to consider the trend of an FoS while selecting a particular research area.

We have developed another approach to detect researchers who are involved at the early stage of an FoS trend. First, we calculated the debut year of an FoS. Then, we have computed the FoS publication count, its author count and FoS trend by using FoM with degree centrality measure. Afterwards, we applied Rogers for the detection of trend setters and followers. Lastly, we have compared our list of researchers (trend setters) with two existing lists that contain highly recognized Computer Science scientists. The lists are as follows; (i) top 10 influential authors and (ii) an existing list of Computer Science scientists with H-index of 40 or higher. The result shows that our approach identifies many of the influential researchers as appeared on their lists. There are cases where there is an exact match of recognition that have been given with respect to the FoS where they are detected as trend setters.

Finally, to detect the impact of FoS on authors citation trend, we have developed an approach that detect the FoS trend of an individual author in his/her career years by characterizing the relations between his/her publications and citations. We detected the FoS of authors, then we selected those authors who follow the maximum trend of an FoS. Further, we calculated the citation count of authors and we computed the citation trend of authors. We used the citation trend of authors as input and predict the next year citation trend as output by using Multiple Linear Regression (MLR) and Artificial Neural Network (ANN).

The experimental results show that there is a similarity between citation trend of authors that belong to the same FoS as compared to different FoS and achieved consistent R2 value. FoS trend following has a certain impact on the citation count of authors. The result also shows that if an author publishes in a particular FoS, then the citation trend of this author's work resembles more to the overall citation trend of that particular FoS than that of some other FoS.

The main contributions of this thesis are as follows;

- We developed a novel method Field of Study Multigraph (FoM), by using centrality measures, degree, betweenness and closeness to analyze the field of study trend, citation trend, the relation between research areas.
- We developed an approach to identify trend setters and followers at the early stage of an FoS.
- We developed another approach to identify the FoS trend that an individual author is involved and the impact of FoS trend on authors careers.

5.1.1 Future Work

There are various future directions which could be explored. These future directions are as follows:

- Extra/additional features such as authors collaboration and venues can be further explored in future studies. This may allow the opportunity to explore other types of dynamics that could be linked with new FoS trends, such as the pace of collaboration between prominent researchers, or new FoS popularity in scientific venues.

- To further our research in this area, we therefore intend to take into account more recent scientific information.

- The study shows that the established approaches are for Computer Science field and might be practical to achieve knowledge of different research fields such as Engineering, Mathematics, Physics etc. The approach developed in this thesis can be straightforwardly applied to other fields.

- Our goal is to build a generic approach that will be useful for researchers, academic editors, publishers and policy makers to achieve the better understanding of the dynamics of new FoS trends. The approach developed in this thesis can be straightforwardly applied to other fields.

Bibliography

[1] A. E. Jinha, “Article 50 million: an estimate of the number of scholarly articles in existence,” *Learned publishing*, vol. 23, no. 3, pp. 258–263, 2010.

[2] S. Effendy and R. H. Yap, “Analysing trends in computer science research: A preliminary study using the microsoft academic graph,” in *Proceedings of the 26th international conference on world wide web companion*, pp. 1245–1250, 2017.

[3] A. Hoonlor, B. K. Szymanski, and M. J. Zaki, “Trends in computer science research,” *Communications of the ACM*, vol. 56, no. 10, pp. 74–83, 2013.

[4] A. A. Salatino, F. Osborne, and E. Motta, “How are topics born? understanding the research dynamics preceding the emergence of new areas,” *PeerJ Computer Science*, vol. 3, p. e119, 2017.

[5] D. Hicks, P. Wouters, L. Waltman, S. De Rijcke, and I. Rafols, “Bibliometrics: the leiden manifesto for research metrics,” *Nature*, vol. 520, no. 7548, pp. 429–431, 2015.

[6] C.-T. Li, Y.-J. Lin, R. Yan, and M.-Y. Yeh, “Trend-based citation count prediction for research articles,” in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 659–671, Springer, 2015.

[7] A. C. Sparavigna and R. Marazzato, “Using google ngram viewer for scientific referencing and history of science,” *arXiv preprint arXiv:1512.01364*, 2015.

[8] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th international conference on world wide web*, pp. 243–246, 2015.

[9] M. Callon, J.-P. Courtial, W. A. Turner, and S. Bauin, "From translations to problematic networks: An introduction to co-word analysis," *Social science information*, vol. 22, no. 2, pp. 191–235, 1983.

[10] J. Courtial, "A coword analysis of scientometrics," *Scientometrics*, vol. 31, no. 3, pp. 251–260, 1994.

[11] M. Callon, J.-P. Courtial, and F. Laville, "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry," *Scientometrics*, vol. 22, no. 1, pp. 155–205, 1991.

[12] Y. Ding, "Community detection: Topological vs. topical," *Journal of Informetrics*, vol. 5, no. 4, pp. 498–514, 2011.

[13] N. Coulter, I. Monarch, and S. Konda, "Software engineering as seen through its research literature: A study in co-word analysis," *Journal of the American Society for Information Science*, vol. 49, no. 13, pp. 1206–1223, 1998.

[14] Y. Tian, R. A. Hankins, and J. M. Patel, "Efficient aggregation for graph summarization," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 567–580, 2008.

[15] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report," 1998.

[16] D. Eichmann, M. Ruiz, P. Srinivasan, N. Street, C. Culy, and F. Menczer, "A cluster-based approach to tracking, detection and segmentation of broadcast news," in *Proceedings of the DARPA Broadcast News Workshop*, pp. 69–76, 1999.

[17] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[18] J. M. Schultz and M. Liberman, "Topic detection and tracking using idf-weighted cosine coefficient," in *Proceedings of the DARPA broadcast news workshop*, vol. 1892192, Citeseer, 1999.

[19] I. Roche, D. Besagni, C. François, M. Hörlesberger, and E. Schiebel, "Identification and characterisation of technological topics in the field of molecular biology," *Scientometrics*, vol. 82, no. 3, pp. 663–676, 2010.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[21] S. Deerwester, S. Dumais, T. Landauer, *et al.*, "Indexing by latent semantic analysis journal of the american society for information science," 1990.

[22] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, 1999.

[23] D. Blei and J. Lafferty, "Correlated topic models," *Advances in neural information processing systems*, vol. 18, p. 147, 2006.

[24] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 424–433, 2006.

[25] Z.-Y. Shen, J. Sun, and Y.-D. Shen, "Collective latent dirichlet allocation," in *2008 Eighth IEEE International Conference on Data Mining*, pp. 1019–1024, IEEE, 2008.

[26] T. Griffiths, M. Jordan, J. Tenenbaum, and D. Blei, "Hierarchical topic models and the nested chinese restaurant process," *Advances in neural information processing systems*, vol. 16, 2003.

[27] J. Chang and D. M. Blei, “Hierarchical relational models for document networks,” *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 124–150, 2010.

[28] S. Morinaga and K. Yamanishi, “Tracking dynamics of topic trends using a finite mixture model,” in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 811–816, 2004.

[29] J. Kleinberg, “Bursty and hierarchical structure in streams,” *Data mining and knowledge discovery*, vol. 7, no. 4, pp. 373–397, 2003.

[30] D. Chavalarias and J.-P. Cointet, “Phylomemetic patterns in science evolution—the rise and fall of scientific fields,” *PloS one*, vol. 8, no. 2, p. e54847, 2013.

[31] H. Sayyadi and L. Raschid, “A graph analytical approach for topic detection,” *ACM Transactions on Internet Technology (TOIT)*, vol. 13, no. 2, pp. 1–23, 2013.

[32] 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.

[33] H. G. Small, “A co-citation model of a scientific specialty: A longitudinal study of collagen research,” *Social studies of science*, vol. 7, no. 2, pp. 139–166, 1977.

[34] K. W. Boyack and R. Klavans, “Creation of a highly detailed, dynamic, global model and map of science,” *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 670–685, 2014.

[35] H. Small, “Visualizing science by citation mapping,” *Journal of the American society for Information Science*, vol. 50, no. 9, pp. 799–813, 1999.

[36] S. Upham and H. Small, “Emerging research fronts in science and technology: patterns of new knowledge development,” *Scientometrics*, vol. 83, no. 1, pp. 15–38, 2010.

[37] H. Small, K. W. Boyack, and R. Klavans, “Identifying emerging topics in science and technology,” *Research policy*, vol. 43, no. 8, pp. 1450–1467, 2014.

[38] C. Chen, “Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature,” *Journal of the American Society for information Science and Technology*, vol. 57, no. 3, pp. 359–377, 2006.

[39] C. Chen, Z. Hu, S. Liu, and H. Tseng, “Emerging trends in regenerative medicine: a scientometric analysis in citespace,” *Expert opinion on biological therapy*, vol. 12, no. 5, pp. 593–608, 2012.

[40] C. Chen, Y. Chen, M. Horowitz, H. Hou, Z. Liu, and D. Pellegrino, “Towards an explanatory and computational theory of scientific discovery,” *Journal of Informetrics*, vol. 3, no. 3, pp. 191–209, 2009.

[41] Y. Jo, C. Lagoze, and C. L. Giles, “Detecting research topics via the correlation between graphs and texts,” in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 370–379, 2007.

[42] E. Abrahamson, “Management fashion,” *Academy of management review*, vol. 21, no. 1, pp. 254–285, 1996.

[43] D. N. McCloskey, “The rhetoric of economics,” *Journal of economic literature*, vol. 21, no. 2, pp. 481–517, 1983.

[44] A. Duvvuru, S. Kamarthi, and S. Sultornsanee, “Undercovering research trends: Network analysis of keywords in scholarly articles,” in *2012 Ninth International Conference on Computer Science and Software Engineering (JCSSE)*, pp. 265–270, IEEE, 2012.

[45] A. Duvvuru, S. Radhakrishnan, D. More, S. Kamarthi, and S. Sultornsanee, “Analyzing structural & temporal characteristics of keyword system in academic research articles,” *Procedia Computer Science*, vol. 20, pp. 439–445, 2013.

[46] F. Osborne and E. Motta, "Klink-2: integrating multiple web sources to generate semantic topic networks," in *International Semantic Web Conference*, pp. 408–424, Springer, 2015.

[47] S. Yi and J. Choi, "The organization of scientific knowledge: the structural characteristics of keyword networks," *Scientometrics*, vol. 90, no. 3, pp. 1015–1026, 2012.

[48] C. Erten, P. J. Harding, S. G. Kobourov, K. Wampler, and G. Yee, "Exploring the computing literature using temporal graph visualization," in *Visualization and Data Analysis 2004*, vol. 5295, pp. 45–56, SPIE, 2004.

[49] M. Herrera, D. C. Roberts, and N. Gulbahce, "Mapping the evolution of scientific fields," *PloS one*, vol. 5, no. 5, p. e10355, 2010.

[50] G. Palla, A.-L. Barabási, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, no. 7136, pp. 664–667, 2007.

[51] R. Ohniwa, A. Hibino, and K. Takeyasu, "Trends in research foci in life science fields over the last 30 years monitored by emerging topics," *Scientometrics*, vol. 85, no. 1, pp. 111–127, 2010.

[52] S. L. Decker, B. Aleman-Meza, D. Cameron, and I. B. Arpinar, *Detection of bursty and emerging trends towards identification of researchers at the early stage of trends*. PhD thesis, University of Georgia Athens, 2007.

[53] F. Osborne, E. Motta, and P. Mulholland, "Exploring scholarly data with rexplore," in *International semantic web conference*, pp. 460–477, Springer, 2013.

[54] F. Osborne, A. Salatino, A. Birukou, and E. Motta, "Automatic classification of springer nature proceedings with smart topic miner," in *International Semantic Web Conference*, pp. 383–399, Springer, 2016.

[55] F. Osborne, T. Thanapalasingam, A. Salatino, A. Birukou, and E. Motta, "Smart book recommender: A semantic recommendation engine for editorial products," 2017.

[56] F. Osborne, G. Scavo, and E. Motta, "Identifying diachronic topic-based research communities by clustering shared research trajectories," in *European Semantic Web Conference*, pp. 114–129, Springer, 2014.

[57] D. Zhou, X. Ji, H. Zha, and C. L. Giles, "Topic evolution and social interactions: how authors effect research," in *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 248–257, 2006.

[58] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," *arXiv preprint arXiv:1207.4169*, 2012.

[59] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 306–315, 2004.

[60] L. Bolelli, Ş. Ertekin, and C. L. Giles, "Topic and trend detection in text collections using latent dirichlet allocation," in *European conference on information retrieval*, pp. 776–780, Springer, 2009.

[61] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 990–998, 2008.

[62] S. Yi and J. Choi, "The organization of scientific knowledge: the structural characteristics of keyword networks," *Scientometrics*, vol. 90, no. 3, pp. 1015–1026, 2012.

[63] J. Wang, X. Hu, X. Tu, and T. He, "Author-conference topic-connection model for academic network search," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 2179–2183, 2012.

[64] X. Sun, K. Ding, and Y. Lin, “Mapping the evolution of scientific fields based on cross-field authors,” *Journal of Informetrics*, vol. 10, no. 3, pp. 750–761, 2016.

[65] K. W. Boyack, R. Klavans, and K. Börner, “Mapping the backbone of science,” *Scientometrics*, vol. 64, no. 3, pp. 351–374, 2005.

[66] M. C. Pham, R. Klamma, and M. Jarke, “Development of computer science disciplines: a social network analysis approach,” *Social Network Analysis and Mining*, vol. 1, no. 4, pp. 321–340, 2011.

[67] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[68] L. Leydesdorff and I. Rafols, “Interactive overlays: A new method for generating global journal maps from web-of-science data,” *Journal of Informetrics*, vol. 6, no. 2, pp. 318–332, 2012.

[69] L. Leydesdorff, I. Rafols, and C. Chen, “Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal–journal citations,” *Journal of the American society for Information science and Technology*, vol. 64, no. 12, pp. 2573–2586, 2013.

[70] E. Mohammadi and A. Karami, “Exploring research trends in big data across disciplines: A text mining analysis,” *Journal of Information Science*, vol. 48, no. 1, pp. 44–56, 2022.

[71] A. A. Salatino, A. Mannocci, and F. Osborne, “Detection, analysis, and prediction of research topics with scientific knowledge graphs,” in *Predicting the Dynamics of Research Impact*, pp. 225–252, Springer, 2021.

[72] S. M. J. Jalali, H. W. Park, I. R. Vanani, and K.-H. Pho, “Research trends on big data domain using text mining algorithms,” *Digital Scholarship in the Humanities*, vol. 36, no. 2, pp. 361–370, 2021.

[73] N. Van Eck and L. Waltman, “Software survey: Vosviewer, a computer program for bibliometric mapping,” *scientometrics*, vol. 84, no. 2, pp. 523–538, 2010.

[74] K. Boyack, K. Börner, and R. Klavans, “Mapping the structure and evolution of chemistry research,” *Scientometrics*, vol. 79, no. 1, pp. 45–60, 2009.

[75] D. Fried and S. G. Kobourov, “Maps of computer science,” in *2014 IEEE Pacific Visualization Symposium*, pp. 113–120, IEEE, 2014.

[76] L. Di Caro, M. Guerzoni, M. Nuccio, and G. Siragusa, “A bimodal network approach to model topic dynamics,” *arXiv preprint arXiv:1709.09373*, 2017.

[77] L. M. Bettencourt, D. I. Kaiser, and J. Kaur, “Scientific discovery and topological transitions in collaboration networks,” *Journal of Informetrics*, vol. 3, no. 3, pp. 210–221, 2009.

[78] Y. Jo, C. Lagoze, and C. L. Giles, “Detecting research topics via the correlation between graphs and texts,” in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 370–379, 2007.

[79] N. Shibata, Y. Kajikawa, Y. Takeda, and K. Matsushima, “Detecting emerging research fronts based on topological measures in citation networks of scientific publications,” *Technovation*, vol. 28, no. 11, pp. 758–775, 2008.

[80] J. Kleinberg, “Bursty and hierarchical structure in streams,” *Data mining and knowledge discovery*, vol. 7, no. 4, pp. 373–397, 2003.

[81] J. C. Ho, E.-C. Saw, L. Y. Lu, and J. S. Liu, “Technological barriers and research trends in fuel cell technologies: A citation network analysis,” *Technological Forecasting and Social Change*, vol. 82, pp. 66–79, 2014.

[82] H. Guo, S. Weingart, and K. Börner, “Mixed-indicators model for identifying emerging research areas,” *Scientometrics*, vol. 89, no. 1, pp. 421–435, 2011.

[83] T. Furukawa, K. Mori, K. Arino, K. Hayashi, and N. Shirakawa, “Identifying the evolutionary process of emerging technologies: A chronological network

analysis of world wide web conference sessions," *Technological Forecasting and Social Change*, vol. 91, pp. 280–294, 2015.

[84] S. Morinaga and K. Yamanishi, "Tracking dynamics of topic trends using a finite mixture model," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 811–816, 2004.

[85] K. K. Mane and K. Börner, "Mapping topics and topic bursts in pnas," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5287–5290, 2004.

[86] W. Ke, K. Borner, and L. Viswanath, "Major information visualization authors, papers and topics in the acm library," in *IEEE symposium on information visualization*, pp. r1–r1, IEEE, 2004.

[87] K. W. Boyack, K. Mane, and K. Borner, "Mapping medline papers, genes, and proteins related to melanoma research," in *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.*, pp. 965–971, IEEE, 2004.

[88] D. He, X. Zhu, and D. S. Parker, "How does research evolve? pattern mining for research meme cycles," in *2011 IEEE 11th International Conference on Data Mining*, pp. 1068–1073, IEEE, 2011.

[89] C. Chen, "Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature," *Journal of the American Society for information Science and Technology*, vol. 57, no. 3, pp. 359–377, 2006.

[90] K. Börner, W. Huang, M. Linnemeier, R. Duhon, P. Phillips, N. Ma, A. Zoss, H. Guo, and M. Price, "Rete-netzwerk-red: analyzing and visualizing scholarly networks using the network workbench tool," *Scientometrics*, vol. 83, no. 3, pp. 863–876, 2010.

[91] A. A. Salatino, F. Osborne, and E. Motta, "Augur: forecasting the emergence of new research topics," in *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, pp. 303–312, 2018.

[92] S. A. Morris, G. Yen, Z. Wu, and B. Asnake, "Time line visualization of research fronts," *Journal of the American society for information science and technology*, vol. 54, no. 5, pp. 413–422, 2003.

[93] H. Small and P. Upham, "Citation structure of an emerging research area on the verge of application," *Scientometrics*, vol. 79, no. 2, pp. 365–375, 2009.

[94] S. A. Morris, "Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution," *Journal of the American Society for Information Science and Technology*, vol. 56, no. 12, pp. 1250–1273, 2005.

[95] Y. Takeda and Y. Kajikawa, "Optics: A bibliometric approach to detect emerging research domains and intellectual bases," *Scientometrics*, vol. 78, no. 3, pp. 543–558, 2009.

[96] F. Åström, "Changes in the lis research front: Time-sliced cocitation analyses of lis journal articles, 1990–2004," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 947–957, 2007.

[97] H. Guo, S. Weingart, and K. Börner, "Mixed-indicators model for identifying emerging research areas," *Scientometrics*, vol. 89, no. 1, pp. 421–435, 2011.

[98] C. R. Rao, "Diversity: Its measurement, decomposition, apportionment and analysis," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 1–22, 1982.

[99] A. Stirling, "A general framework for analysing diversity in science, technology and society," *Journal of the Royal Society Interface*, vol. 4, no. 15, pp. 707–719, 2007.

[100] L. Di Caro, M. Guerzoni, M. Nuccio, and G. Siragusa, "A bimodal network approach to model topic dynamics," *arXiv preprint arXiv:1709.09373*, 2017.

[101] S. Morinaga and K. Yamanishi, "Tracking dynamics of topic trends using a finite mixture model," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 811–816, 2004.

[102] I. Rafols, A. L. Porter, and L. Leydesdorff, "Science overlay maps: A new tool for research policy and library management," *Journal of the American Society for information Science and Technology*, vol. 61, no. 9, pp. 1871–1887, 2010.

[103] A. Zeng, Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang, and H. E. Stanley, "The science of science: From the perspective of complex systems," *Physics Reports*, vol. 714, pp. 1–73, 2017.

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

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

[106] A. Rzhetsky, J. G. Foster, I. T. Foster, and J. A. Evans, "Choosing experiments to accelerate collective discovery," *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. 14569–14574, 2015.

[107] M. De Domenico, E. Omodei, and A. Arenas, "Quantifying the diaspora of knowledge in the last century," *Applied Network Science*, vol. 1, no. 1, pp. 1–13, 2016.

[108] A. Clauset, D. B. Larremore, and R. Sinatra, "Data-driven predictions in the science of science," *Science*, vol. 355, no. 6324, pp. 477–480, 2017.

[109] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, *et al.*, "Science of science," *Science*, vol. 359, no. 6379, p. eaao0185, 2018.

[110] T. Kuhn, M. Perc, and D. Helbing, "Inheritance patterns in citation networks reveal scientific memes," *Physical Review X*, vol. 4, no. 4, p. 041036, 2014.

[111] A. M. Petersen, M. Riccaboni, H. E. Stanley, and F. Pammolli, "Persistence and uncertainty in the academic career," *Proceedings of the National Academy of Sciences*, vol. 109, no. 14, pp. 5213–5218, 2012.

[112] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, "Quantifying the evolution of individual scientific impact," *Science*, vol. 354, no. 6312, p. aaf5239, 2016.

[113] L. Liu, Y. Wang, R. Sinatra, C. L. Giles, C. Song, and D. Wang, "Hot streaks in artistic, cultural, and scientific careers," *Nature*, vol. 559, no. 7714, pp. 396–399, 2018.

[114] B. F. Jones and B. A. Weinberg, "Age dynamics in scientific creativity," *Proceedings of the national academy of sciences*, vol. 108, no. 47, pp. 18910–18914, 2011.

[115] A. M. Petersen, S. Fortunato, R. K. Pan, K. Kaski, O. Penner, A. Rungi, M. Riccaboni, H. E. Stanley, and F. Pammolli, "Reputation and impact in academic careers," *Proceedings of the National Academy of Sciences*, vol. 111, no. 43, pp. 15316–15321, 2014.

[116] A. M. Petersen, "Quantifying the impact of weak, strong, and super ties in scientific careers," *Proceedings of the National Academy of Sciences*, vol. 112, no. 34, pp. E4671–E4680, 2015.

[117] P. Deville, D. Wang, R. Sinatra, C. Song, V. D. Blondel, and A.-L. Barabási, "Career on the move: Geography, stratification and scientific impact," *Scientific reports*, vol. 4, no. 1, pp. 1–7, 2014.

[118] A. M. Petersen, "Multiscale impact of researcher mobility," *Journal of The Royal Society Interface*, vol. 15, no. 146, p. 20180580, 2018.

[119] R. K. Merton, "The matthew effect in science: The reward and communication systems of science are considered.," *Science*, vol. 159, no. 3810, pp. 56–63, 1968.

[120] M. De Domenico, E. Omodei, and A. Arenas, "Quantifying the diaspora of knowledge in the last century," *Applied Network Science*, vol. 1, no. 1, pp. 1–13, 2016.

[121] P. Bourdieu, "The specificity of the scientific field and the social conditions of the progress of reason," *Social science information*, vol. 14, no. 6, pp. 19–47, 1975.

[122] J. G. Foster, A. Rzhetsky, and J. A. Evans, "Tradition and innovation in scientists' research strategies," *American Sociological Review*, vol. 80, no. 5, pp. 875–908, 2015.

[123] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," *arXiv preprint arXiv:1207.4169*, 2012.

[124] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," *arXiv preprint arXiv:1206.3298*, 2012.

[125] A. Hoonlor, B. K. Szymanski, and M. J. Zaki, "Trends in computer science research," *Communications of the ACM*, vol. 56, no. 10, pp. 74–83, 2013.

[126] T. Jia, D. Wang, and B. K. Szymanski, "Quantifying patterns of research-interest evolution," *Nature Human Behaviour*, vol. 1, no. 4, pp. 1–7, 2017.

[127] A. Zeng, Z. Shen, J. Zhou, Y. Fan, Z. Di, Y. Wang, H. E. Stanley, and S. Havlin, "Increasing trend of scientists to switch between topics," *Nature communications*, vol. 10, no. 1, pp. 1–11, 2019.

[128] S. Shinde and B. Tidke, "Improved k-means algorithm for searching research papers," *Int. J. Comput. Sci. Commun. Networks*, vol. 4, no. 6, pp. 197–202, 2014.

[129] R. L. Miller, "Rogers' innovation diffusion theory (1962, 1995)," in *Information seeking behavior and technology adoption: Theories and trends*, pp. 261–274, IGI Global, 2015.

[130] F. Boudin, “A comparison of centrality measures for graph-based keyphrase extraction,” in *International joint conference on natural language processing (IJCNLP)*, pp. 834–838, 2013.

[131] E. F. Codd, “Extending the database relational model to capture more meaning,” *ACM Transactions on Database Systems (TODS)*, vol. 4, no. 4, pp. 397–434, 1979.

[132] T. Berners-Lee, R. Cailliau, A. Luotonen, H. F. Nielsen, and A. Secret, “The world-wide web,” *Communications of the ACM*, vol. 37, no. 8, pp. 76–82, 1994.

[133] A. A. Salatino, F. Osborne, and E. Motta, “How are topics born? understanding the research dynamics preceding the emergence of new areas,” *PeerJ Computer Science*, vol. 3, p. e119, 2017.

[134] K. Y. Yeung and W. L. Ruzzo, “Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data,” *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[135] I. H. Wenno, “The correlation study of interest at physics and knowledge of mathematics basic concepts towards the ability to solve physics problems of 7th grade students at junior high school in ambon maluku province, indonesia,” *Education Research International*, vol. 2015, 2015.

[136] C. Holden, “Random samples: Data point-impact factor,” *Science*, vol. 309, no. 8, p. 1181c, 2005.

A MLR models trained on same and different FoS from 1990-1995

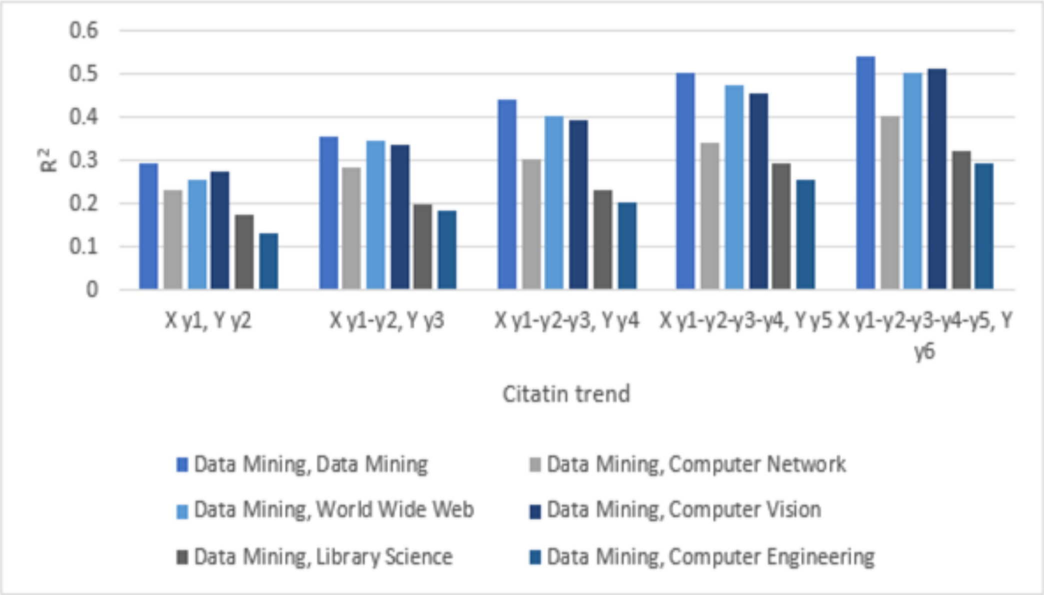


FIGURE A.1: MLR models trained on Data Mining data set and tested with different FoS data sets from 1990-1995.

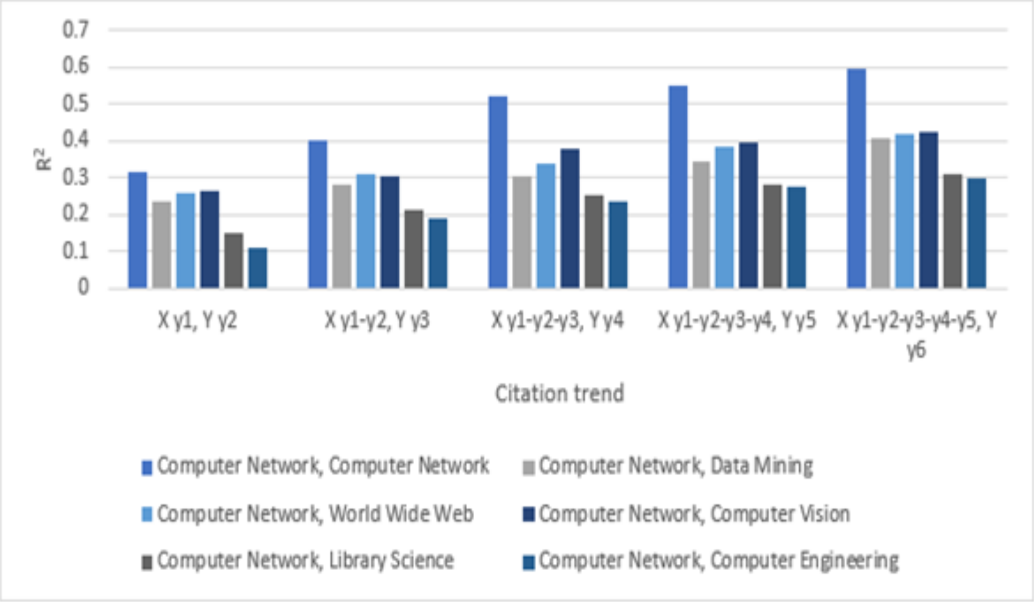


Figure A.2: MLR models trained on Computer Network data set and tested with different FoS data sets from 1990-1995.

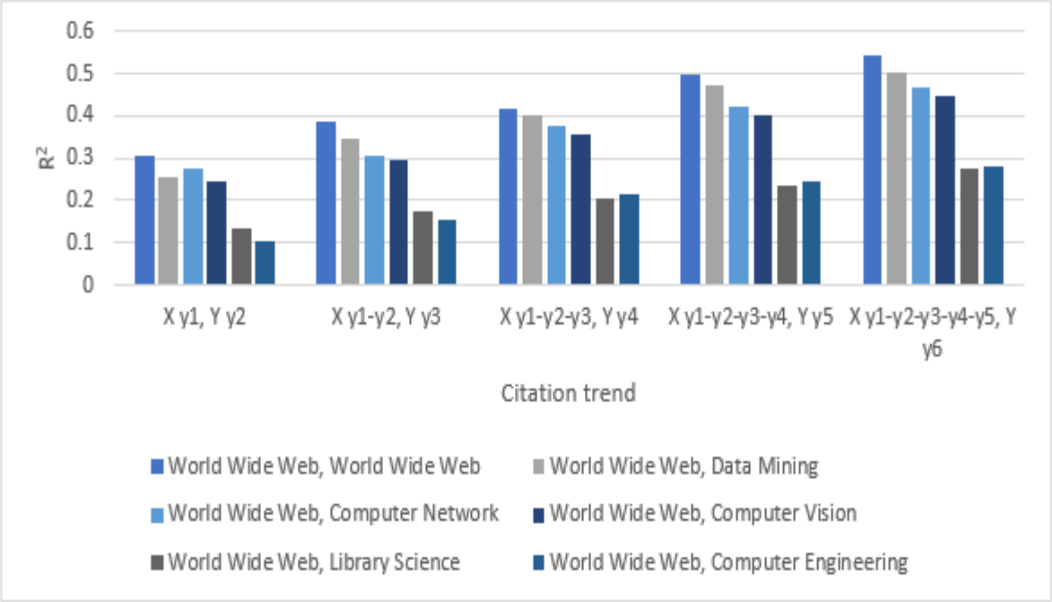


Figure A.3: MLR models trained on World Wide Web data set and tested with different FoS data sets from 1990-1995.

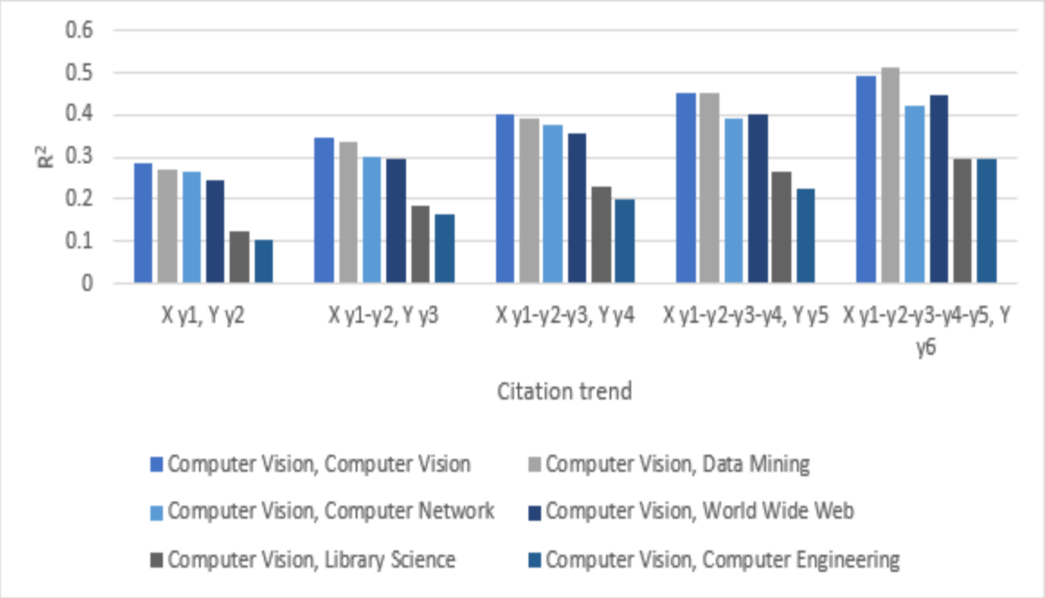


FIGURE A.4: MLR models trained on Computer Vision data set and tested with different FoS data sets from 1990-1995.

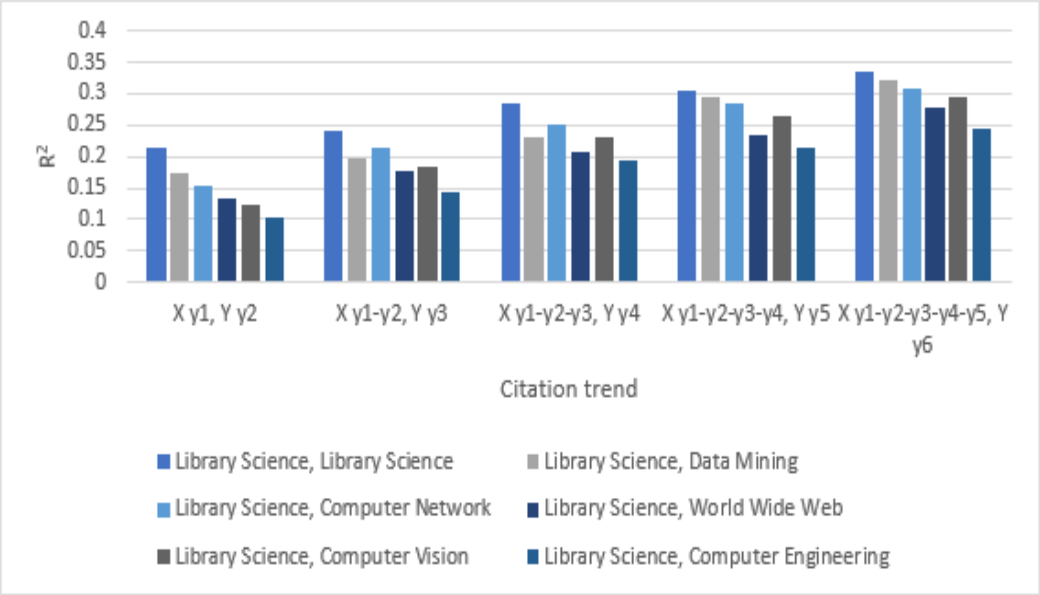


FIGURE A.5: MLR models trained on Library Science data set and tested with different FoS data sets from 1990-1995.

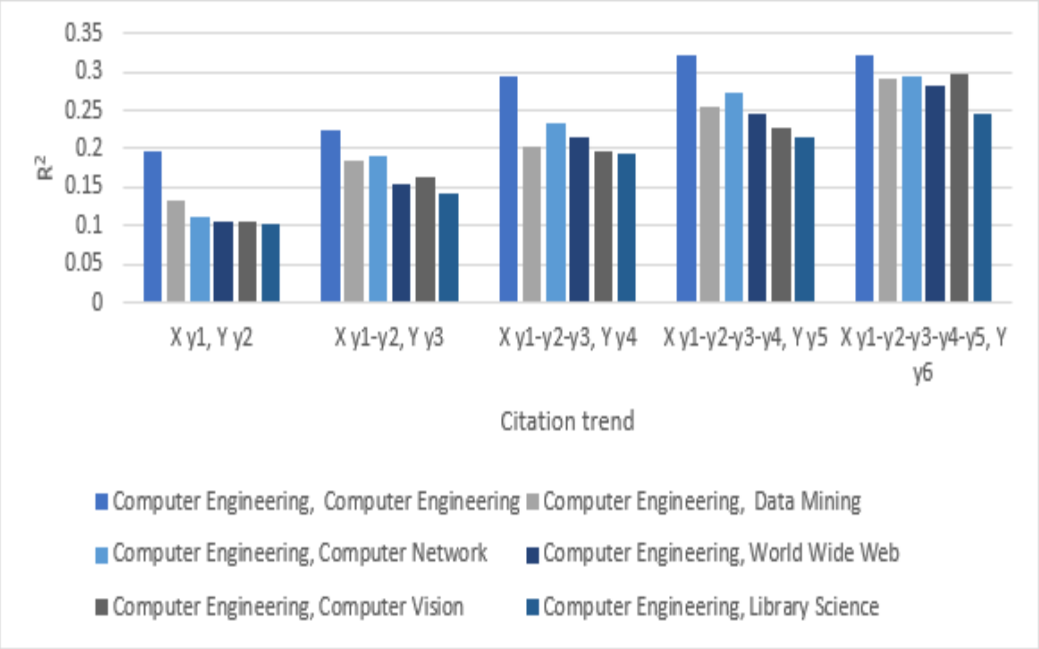


Figure A.6: MLR models trained on Computer Engineering data set and tested with different FoS data sets from 1990-1995.

B ANN models trained on same and different FoS from 1990-1995

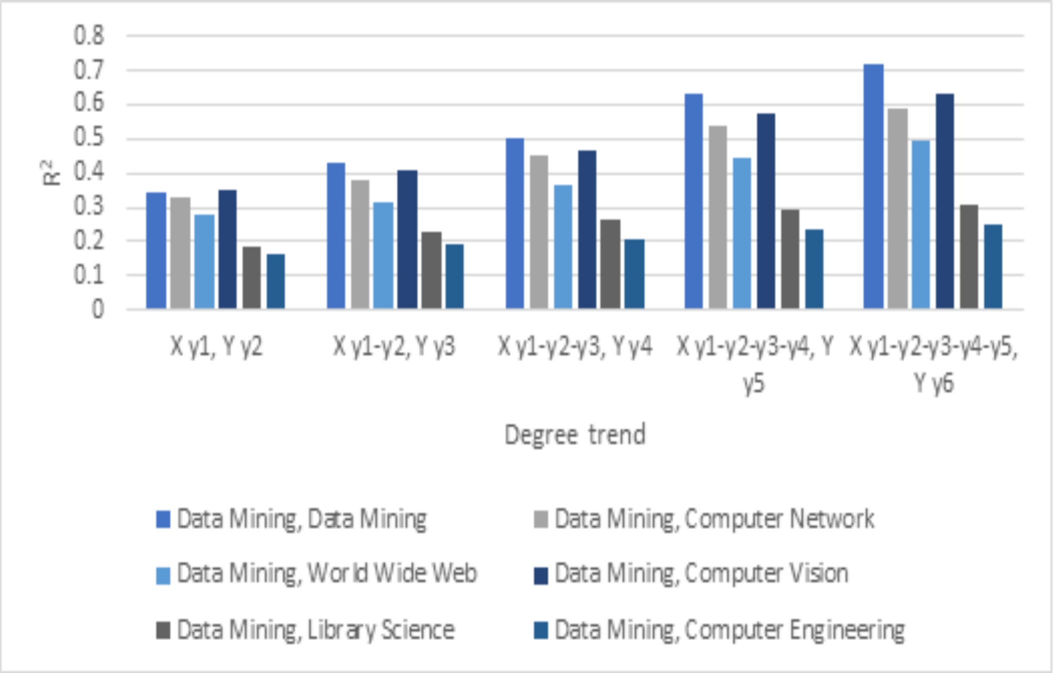


FIGURE B.1: ANN models trained on Data Mining data set and tested with different FoS data sets from 1990-1995.

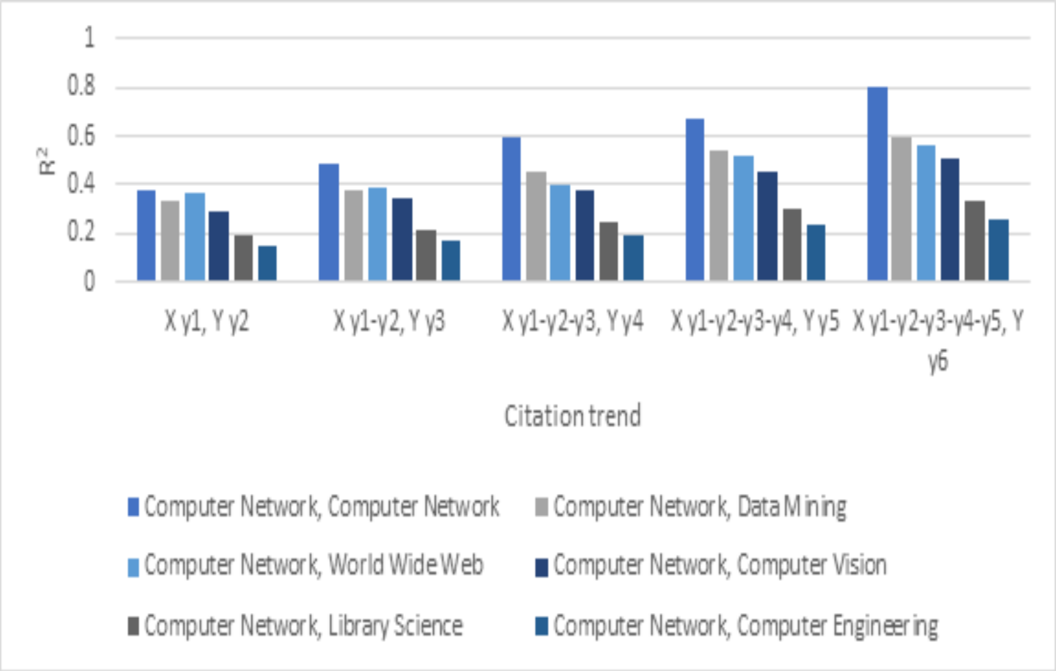


Figure B.2: ANN models trained on Computer Network data set and tested with different FoS data sets from 1990-1995.

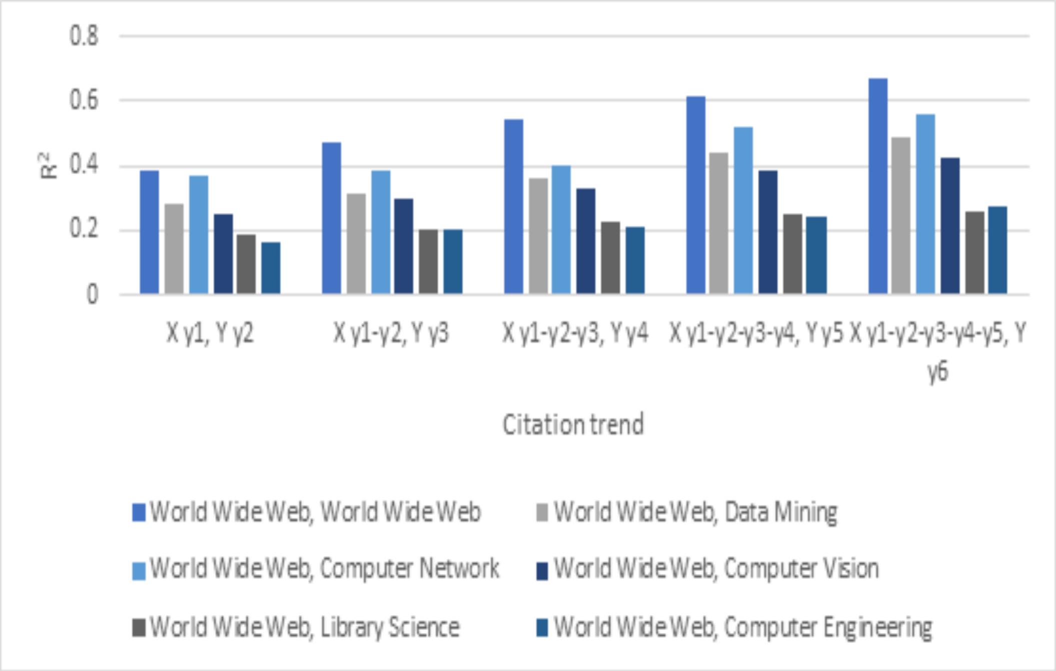


Figure B.3: ANN models trained on World Wide Web data set and tested with different FoS data sets from 1990-1995.

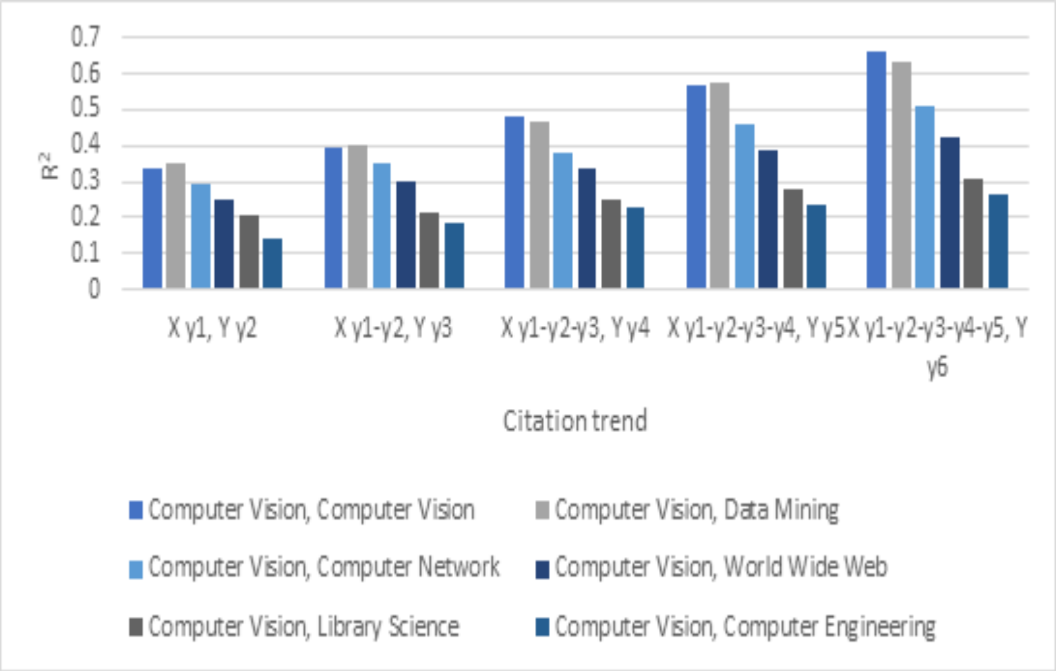


Figure B.4: ANN models trained on Computer Vision data set and tested with different FoS data sets from 1990-1995.

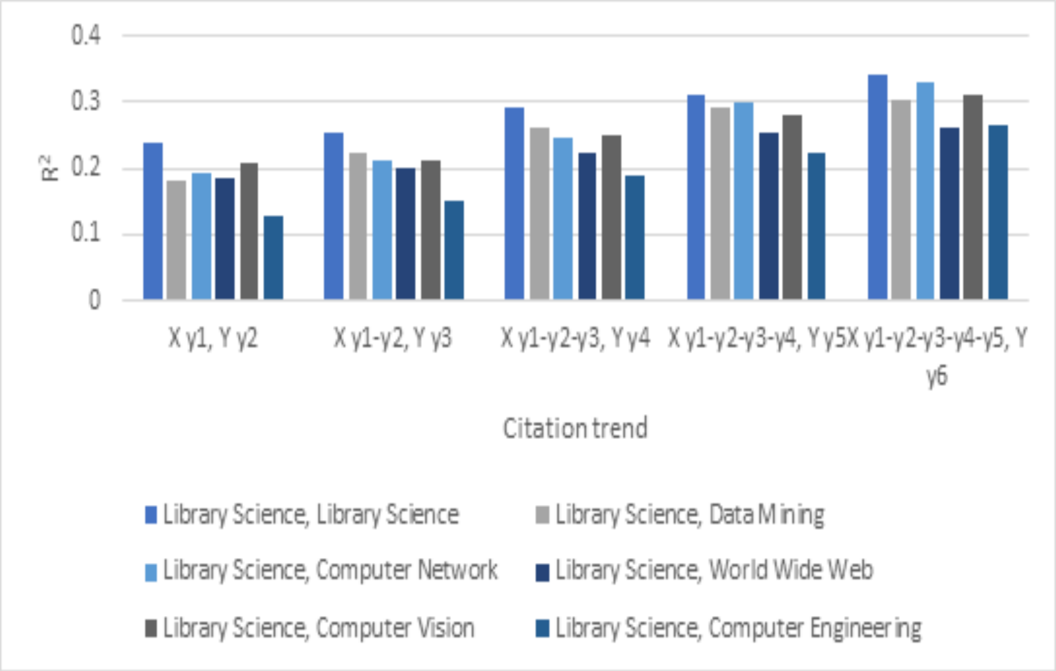


Figure B.5: ANN models trained on Library Science data set and tested with different FoS data sets from 1990-1995.

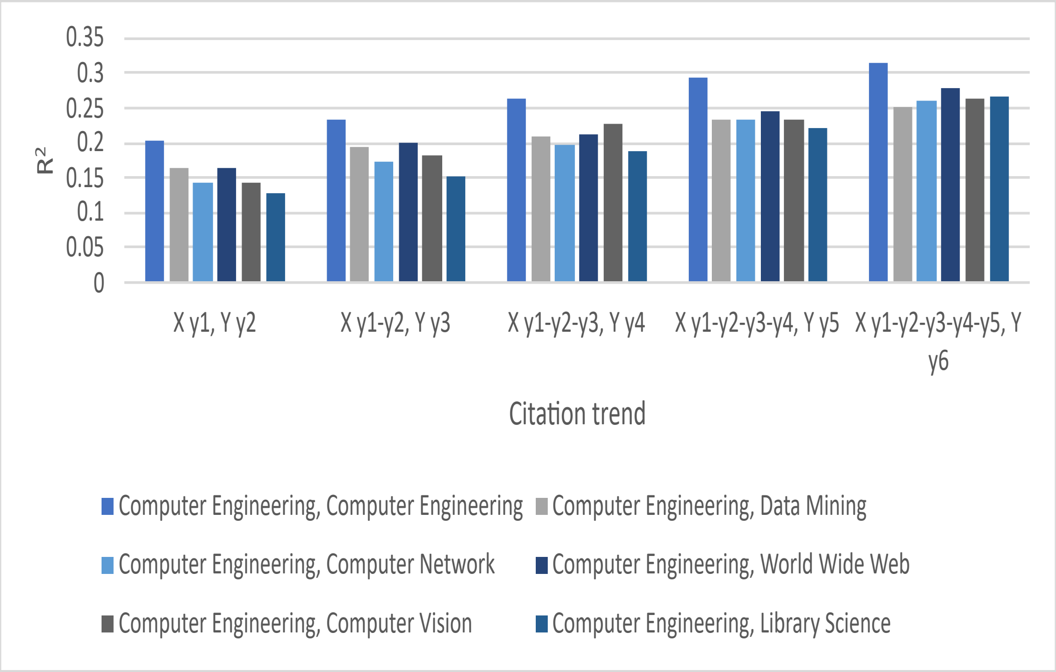


Figure B.6: ANN models trained on Data Mining data set and tested with different FoS data sets from 1990-1995.