Academic Recommendation System Based On The Similarity Learning Of The Citation Network Using Citation Impact

Abdulrhman M. Alshareef

Thesis submitted
In partial fulfillment of the requirements
For the Ph.D. degree in
Computer Science

Ottawa-Carleton Institute for Computer Science
School of Electrical Engineering and Computer Science
University of Ottawa

Abstract

In today's significant and rapidly increasing amount of scientific publications, exploring recent studies in a given research area and building an effective scientific collaboration has become more challenging than any time before. Scientific production growth has been increasing the difficulties for identifying the most relevant papers to cite or to find an appropriate conference or journal to submit a paper to publish. As a result, authors and publishers rely on different analytical approaches in order to measure the relationship among the citation network. Different parameters have been used such as the impact factor, number of citations, co-citation to assess the impact of the produced research publication. However, using one assessing factor considers only one level of relationship exploration, since it does not reflect the effect of the other factors. In this thesis, we propose an approach to measure the Academic Citation Impact that will help to identify the impact of articles, authors, and venues at their extended nearby citation network. We combine the content similarity with the bibliometric indices to evaluate the citation impact of articles, authors, and venues in their surrounding citation network. Using the article metadata, we calculate the semantic similarity between any two articles in the extended network. Then we use the similarity score and bibliometric indices to evaluate the impact of the articles, authors, and venues among their extended nearby citation network.

Furthermore, we propose an academic recommendation model to identify the latent preferences among the citation network of the given article in order to expose the concealed connection between the academic objects (articles, authors, and venues) at the citation network of the given article. To reveal the degree of trust for collaboration between academic objects (articles, authors, and venues), we use the similarity learning to estimate the collaborative confidence score

that represents the anticipation of a prospect relationship between the academic objects among a scientific community. We conducted an offline experiment to measure the accuracy of delivering personalized recommendations, based on the user's selection preferences; real-world datasets were used. Our evaluation results show a potential improvement to the quality of the recommendation when compared to baseline recommendation algorithms that consider co-citation information.

Acknowledgments

First and foremost, I would like to thank Almighty *ALLAH*, the compassionate, the Merciful, for HIS tremendous blessings and guidance. I could have never completed this work without the faith I have in HIM, the Almighty.

I would like to acknowledge my sincere appreciation and respect is intended for my academic supervisor Professor Abdulmotaleb El Saddik, who has supported, guided, and encouraged me throughout my academic success. This thesis would not be a success without the assistance he has provided me. I am grateful for all his advice and guidance, not only in the field of research but also to build a future professional career.

Special thanks are due to my brother Dr. Mohammed Alhamid who support me with his strong background and in-depth knowledge. His valuable assistance, continuous support and careful revisions reflected in successful research achievement.

I am also thankful to all the members of the Multimedia Research Laboratory (MCRLab), for their feedback suggestions and cooperation during my research work.

Words lack to express my sincere appreciation, thanks, and love to my wife Hanan whom I owe her my sincerest gratitude for her continuous support and understanding. Her patience and sacrifice will remain my inspiration throughout my life. I am thankful also to my kids Shaden, Rusiyl, Mohammed, Ruwayna and Jawad whose consistent support, and endless love has given to me through all hardships incurred during my studies and research journey.

I feel a deep sense of gratitude for my grandmother, my father Mohammed and my mother Musbah who formed part of my vision and taught me good things that really matter in life. Their

infinite love and support have always been my strength. Their patience and sacrifice will remain my inspiration throughout my life. I am also very much grateful to all my brothers and sisters (Areej, Suha, Fahad, Rajeh and Ahmad) for their constant inspiration and encouragement. Also, I am grateful to my uncle Dr. Fasil for his support and guidance through my entire life.

Lastly, I want to express my gratitude to my friends and colleagues at the University of Ottawa and King Abdulaziz University. I have been fortunate to know and learn from M. Alowaidi, H. Hajar, B. Alfasi, M.Yamani, A Barnawi, S Alqarni, K. Alyoubi, A. Khadidos and all those who are honestly supporting me and had an impact on my life.

This thesis dedicated to

The memory of my beloved grandmother,

Mama Hamdah.

It is your shining example that I try to emulate in

all that I do.

Thank you for everything.

Table of Contents

List of Figures

List of Figures & Tables

List of Tables

List of Abbreviations

Academic Citation Impact	ACI
Local Academic Citation Impact	LACI
Global Academic Citation Impact	GACI
Cited Articles List	CAL
Venues Related Articles List	VARL
Article Weighted Similarity	AWS
Term Frequency	TF
Term Frequency-Inverse Document Frequency	TF-IDF
Latent Semantic Analysis	LSA
Natural Language Processing	NLP
Citation Count	CC
Author Impact Index	AII
Venue Impact Index	VII
Local Rank	LR
Global Rank	GR
Latent Preferences Identifier Model	LPIM
Collaborative Confidence Score	CCS
Cloud-Oriented Architecture	COA
Digital Object Identifier	DOI
Open Researcher and Contributor ID	ORCID
Java Server Pages	JSP
Application Programming Interface	API

List of Abbreviations

Unified Modeling Language	UML
Graphical User Interface	GUI

Chapter 1 . Introduction

The vision of digital twin as introduced by Prof. El Saddik [1] is a digital replication of a living or non-living physical entity. By bridging the physical and the virtual worlds, data are transmitted seamlessly, allowing the virtual entity to exist simultaneously with the physical entity. A digital twin facilitates the means to monitor, understand, and optimize the functions of the physical entity and provides continuous feedback to improve quality of life and wellbeing. A digital twin is hence the convergence of several technologies such as AI, AR/VR, and Haptics, IoT, Cybersecurity and Communication Networks. With the current rapid growth of Big Data, developments in the indexing of scientific publications and recommendations are receiving significant interest from researchers, who stand to benefit from advances in data analysis techniques. There is a marked increase in the volume of scientific publications and their corresponding metadata databases. Specifically, indexing articles that measure scientific impact and help build virtual communities of researchers with common interests are becoming more relevant than ever before [2], [3]. Scientific communities can be considered a particular kind of practice communities. They consist of various interdisciplinary groups that are often geographically distributed. A common area of research interests can bring them together. Scientific communities can be identified by analyzing the citation network for academic collaboration among

researchers. A citation network is created when two or more articles are in a citation relationship [4], and it represents scientific relations between the academic objects.

1.1. Background and Motivation

A component of the digital twin vision is to arranging a virtual collaboration between researchers within a research field. Hence, this work forming a virtual community to help young researchers to find proper collaborators, suitable venues, and related articles. The availability of a huge number of conferences and journals has increased the challenge of young researchers to reach and connect with the right community for either a literature survey or creating new connections. The lack of community connections in the new field affects the possibility of publishing a new idea outside the scope of expertise. Moreover, it is difficult for young researchers to determine the list of journals and conferences that are most relevant or the hot topics should they targeted for their research career.

The problem of determining academic impact is not new; there have been many attempts to create a list of metric indexes that is useful for evaluating a scientific article [5]. There are also different methodologies to analyze the citation impact of a specific article on a citation network [6], [7]. Mostly, the citation impact is a quantitative concept used to measure intellectual influence [8]. The Journal Impact Factor (JIF) and SCImago Journal Rank (SJR) are the best-known metrics intended to assess the journal's impact compared with other journals in any research field rest on bibliometric data [7]. The citation count is a widely-used bibliometric in the field of citations analysis to evaluate the citation impact [8], [9]. It represents a recognized metric at the article level that has been used to estimate an individual article's impact compared with other articles. The article considers the cornerstone of associations within the academic citation network. Measuring

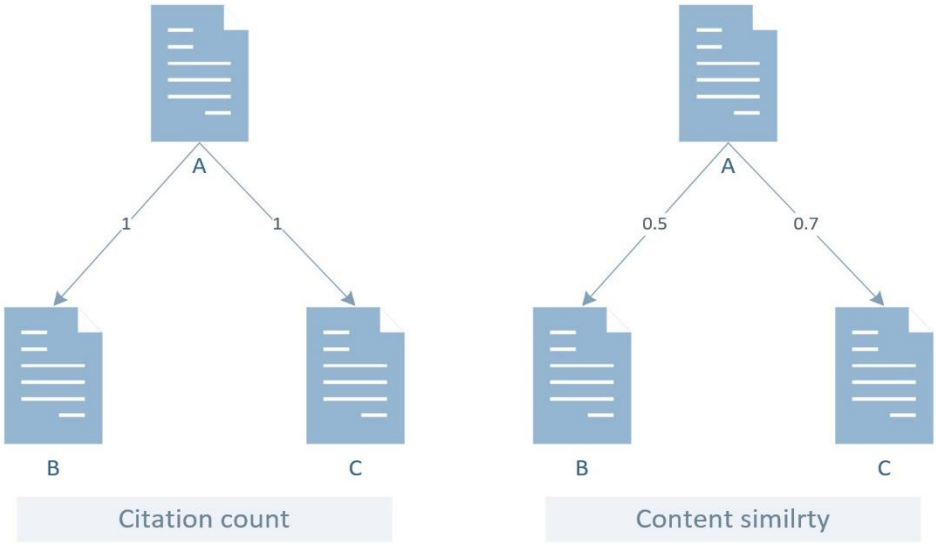

Figure 1-1: Article Citation relationship weighting practice.

the academic impact relies on the assessment of the impact of articles in the academic citation network.

Assessing the impact of an article using the bibliometric alone will not reveal the actual impact of the transferred knowledge. Content similarity represents the knowledge that has been perceived, discovered or learned from a cited article and transmitted to the citing article. Fundamentally, the relationship between a citing article and the list of cited articles is widely considered to be a granted scientific citation, according to scientific publication practices. However, each cited article has a different impact, depending on the amount of the knowledge transferred to the citing article. If we examine a citation relationship, we find that each cited article has a different impact on the content of the citing article. For example, as shown in Figure 1-1, if we have three articles that are in a citation relationship, Article A cites both article B and article C. Using the bibliometric data, the impact of both citied articles has the same weight regardless of their content semantic similarities with A; for instance, we could find that article C is more related to article A

than to article B, based on content semantic similarity. However, content semantic similarity alone does not reflect an article's impact among their citation network. Therefore, combining the bibliometric data and the content semantic similarity to analyze the citation impact of a specific article on an extended nearby citation network can facilitate the process of finding the most influential articles [10], [11]. Besides, it will benefit from exploring and analyzing current relationships and hidden preferences at the citation network to recommend future collaboration to build more effective scientific communities.

1.2. Motivating Scenarios

This section provides a few motivating scenarios that point out the benefits of the proposed system. These scenarios will be considered through the thesis and will subsequently be targeted as an example for the implementation of prototype applications, in order to assess the applicability of this system and its performance with the proposed algorithms.

In the first scenario, let us assume that Sarah is a fresh graduate student and has just started her Ph.D. program. During her initial reading, she came up with a brilliant idea, and want to complete her research towards this new idea. However, her professor on sabbatical leave and he is too busy to help her with her research. That is why she is looking for a solution that can help her to find relevant articles that she can read in regrades her topic and find suitable venues that she can suggest to her professor later on for paper submission. Once she starts the search for the available list of articles or venues, the search system would have already filtered out some articles or venues suggestions based on the received information, which, in this case, the user keywords related to the research idea. Hence, would such information be enough to make good interest predictions to help Sarah pick the right article or venue? The answer is no, since she may also be interested in

targeting top venues related to her topics. Also, she may be interested in reading and citing a most relevant article that has a higher impact on the targeting community.

Another scenario, let us assume that John is a researcher at a research center. He has the desire to return to the academic sector, which motivates him to work in publishing his work. However, he is very busy in his work and does not have the time to looking for people to collaborate with them or venues related to submitting his work at the earliest opportunity. That's why he is looking for a solution that can help him to find people to collaborate, related venues for submission. When the appropriate information such as the research title, the abstract, and the reference list, entered by the user, the system can capture user preferences based on his previous selection. The system does not only recommend the most related venues or the most appropriate authors but can also adjust the recommendation according to the user's preferences.

1.3. Definitions

This section provides a list of concepts and their definitions that will be exploited in this thesis. We use the term "scientific collaboration" to refer to the academic relationship between the authors in the scientific communities at the level of scientific citation, research participation, or joint presence at the same venue. The term "scientific communities" can be considered a particular kind of practice communities that consist of various interdisciplinary groups, bring them together with common areas of research interest.

We use the term "citation network" to refer to a social network that represents the relationships between the academic objects that represent the field of research and scientific publishing. The "academic objects" in the field of scientific research are the article, the author, and

the venue. We use the term “venue” to refer to the means by which an article was published, whether it be a scientific conference or a scientific journal.

We use the term “citation impact” to refer to a quantitative concept used to measure the intellectual influence of a specific academic object at a certain level in the citation network as a result of citation analysis or bibliometric data [8]. We use the term “bibliometric” to refer to a quantitative analysis of an academic object such as citation count at the article-level, journal impact factor at the journal- level and h-index at the author-level. We use the term “citation count” to refer to a quantitative index that represents the frequency number of an article was cited in other articles, books, or thesis. We use the term “journal impact factor” to refer to a quantitative index represents the average number of citations that articles published by a journal in the previous two years have received in the current year. We use the term “author impact index” to refer to a quantitative index based on the average of the author’s number of citations that they have received in other publications over the total publication of the same author among the extended nearby citation network. We use the term “semantic similarity” to refer to the strength of the relationship between two elements by assessing the relationship based on terminology relationships [12].

1.4. Research Statement

The study and analysis of scientific research and their relationships at the citation network will help to understand the direct and indirect impact of this relationship on the researchers, the publishers, and the institutions. In order to build an effective scientific community, Many users could benefit from citation network analysis and academic recommender systems to figure out an appropriate collaboration. Thus, it will help to optimize the use of existing resources to reach the expected objectives of the funders more efficiently.

Combining the content similarity in conjunction with bibliometric to analyze and evaluate the article citation impact is an important challenge in the citation network analysis [2], [3], [10], [13]. Our research goal is to identify different aspects of the citation network and select the most efficient features that are best suited to choose and deliver different collaboration modules. The focus of this thesis is to incorporate the semantic similarity, and bibliometric indexes into the recommendation process to enhance the citation impact analysis. Thus, it will be reflected in the selection quality of cited articles, author to team up, and publishing venues to build a more effective scientific community. This thesis tries to address the following requirements:

- Extract user preferences using similarity learning to analyze data more deeply.
- Customize the article, the author and the venue recommendations according to the user preferences.
- Support the impact assessment and the evaluation of different collaboration modules at the citation network.
- Analyze and compare the important aspects of the citation network with the surrounding scientific environment.
- Visualize the analyzed information to expose the hidden patterns.

This thesis also tries to address the following research questions:

- How can we predict scientific collaboration by analyzing the citation impact in the citation network of a given article?
- How to recommend a list of potential academic objects (articles, authors, venues) given an article title, abstract and references?

1.5. Contributions

The main objective of this thesis is to elaborate on the design, development, and evaluation of an academic recommendation system that incorporates semantic similarity and bibliometric data that can adapt to the user preferences regarding their citation impact and collaboration modules. The thesis contributions can be summarized by the following:

- The design and development of an academic recommendation system that incorporates the content semantic similarity, citation network analysis and the analysis of the latent preferences into the recommendation process to facilitate the capture, analysis, and demonstration of the academic collaboration in the extended nearby citation network.

- The modeling and development of a citation analysis algorithm that takes into account the content similarity and bibliometric indexes to measure the academic citation impact. We will use the proposed algorithm in the academic recommendation system to improve the level of user selections for an upcoming academic collaboration that will help to build an effective scientific community. The algorithm includes a technique to explore and learn the big data environment of the citation network considering the personalized user preferences.

- The design and implementation of a visual system interface to facilitate the awareness of the user's academic status. The visual interface will help the user to recognize the current academic collaborations in an interactive manner that helps to explore academic patterns efficiently.

1.6. Scholarly Achievements

- **A. Alshareef**, M. F. Alhamid, and A. El Saddik, “Article Impact Value for Nearby Citation Network Analysis," *2018 IEEE International Conference on Big Data and Smart Computing (BigComp '18)*, Shanghai, China, 2018, pp. 398-403. DOI: 10.1109/BigComp.2018.00065

- **A. Alshareef**, M. F. Alhamid, and A. El Saddik, “Recommending Scientific Collaboration Based on Topical, Authors and Venues Similarities," *2018 IEEE International Conference on Information Reuse and Integration (IRI '18)*, Salt Lake City, UT, USA, 2018, pp. 55-61. DOI: 10.1109/IRI.2018.00016

- **A. Alshareef**, M. F. Alhamid, and A. El Saddik, “Toward Citation Recommender Systems Considering the Article Impact at the Nearby Citation Network," *Peer-to-Peer Networking and Applications, Springer, 2018*. pp 1–10 DOI: 10.1007/s12083-018-0687-4

- **A. Alshareef**, M. F. Alhamid, and A. El Saddik, “Academic venue recommendation based on the similarity learning of the extended nearby citation network," *IEEE Access, IEEE,* vol. 7, pp. 38813-38825, 2019. DOI: 10.1109/ACCESS.2019.2906106

1.7. Thesis Organization

The thesis is organized as follows: In chapter 2, we review the current work and the research background in regards to the academic recommendation system and the citation network

analysis. In Chapter 3, we discuss in details the academic recommendation model. In Chapter 4, we discuss the proposed system framework and illustrate its major components. In Chapter 5, we presented the evaluation academic recommendation system; the achieved outcomes are analyzed and discussed in this chapter. In chapter 6 we summarize the thesis and presents ideas for future work.

Chapter 2 . Background and Related Work

In this chapter, we will discuss existing work in the field of the citation network analysis, academic recommendations, citation impact analysis, and scientific communities' recognition. In addition, we will explain the likenesses and variances of the proposed thesis briefly to the existing works.

2.1. Citation Network Analysis

In the current time of the rapid development of Big Data, the development of scientific publications indexing and recommendations is receiving significant interest from researchers to benefit from the advancement in data analysis techniques. There is an increase in the volume of scientific publications and their corresponding metadata databases. In specific, indexing articles, measuring the scientific impact of each, and building a virtual community of researchers with the same interest is becoming more relevant than it used to be [2], [3]. Figure 2-1 illustrates the citation network surroundings and relationships between articles where each line represent a relation between two articles, and each node represents an article. The increased volume of the publication pushed the research industry toward smarter Big Data environments.

The Citation Network is a social network that represents the relationships between the basic components that represent the field of research and scientific publishing. The basic components in

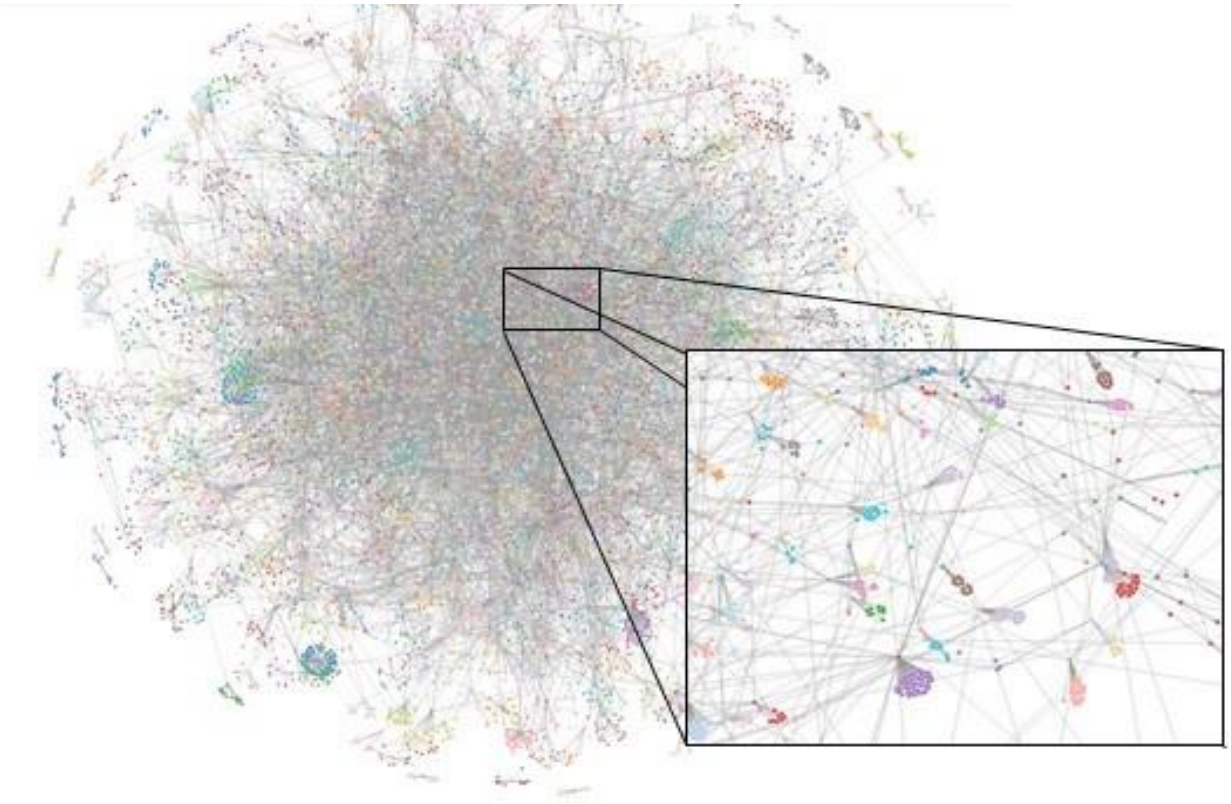

Figure 2-1: An overview of the Citation network surroundings and relationships.

the field of scientific research are the article, the author, and the publishing venue. Network analysis is a well-known practice in the field of scientific research, which is used to study the transferred knowledge and ideas, the influence of relations in the field of research, and evaluate the impact of the authors, publishers, and institutions [14].

The citation analysis helps researchers to identify the impact of the article among the citation network [15]. Citing another article is a way of giving credits to articles that influence research work. However, this influence is not explicitly and directly given in the citation process. Authors are giving credit to others by citing their works. Giving credit is a way of acknowledging other works impact on the research work. However, the amount of transferred knowledge is not explicitly and directly given in the bibliometric indexes. In other words, bibliometric indexes

considered an equal contribution for each cited article on the citing article. Thus, it leads to an inappropriate assessment of the article impact and the amount of transferred knowledge.

In a research study [16], Garfield discusses some issues that have been raised from the inconsistency in the citation standards of encompassing citation details that affect the citation evaluation process. For instance, John Smith and Josef Smith both can be cited as J. Smith that will lead to an incorrect citation count. According to the author [16], such inconsistency may lead to inaccuracy of the analysis process to distinguish whether the journal was cited or not. MacRoberts and MacRoberts in [17], [18] discuss the bibliography issues that can be potentially addressed in the objective of using citation analysis. The two papers discuss that the types of citation, the citation rate variations, the type of publications, the transferred knowledge, and the type of specialty should be addressed variable among others. Therefore, most of the researchers work in the field of citation analysis categorized under one of the following types: co-citation analysis, content analysis, and impact analysis. These three categories cover the areas that studying the citation network and the problems arising from the bibliography issues.

2.1.1. Co-Citation analysis

A good number of citation analysis approaches rely on the co-citation frequency for the scientific knowledge classification. For instance, Batagelj [19] proposes an algorithm to analyze the citation in a very large network using the Hummon and Doreian DNA theory [20]. The algorithm based on the co-citation analysis and bibliometric coupling to measure the importance of the nodes, and weights using three indices: Node Pair Projection Count (NPPC), Search Path Link Count (SPLC) and Search Path Node Pair (SPNP). Boyack and Klavans [21] discuss the citation-based mapping approach clustering accuracy. They use four standard approaches: co-

citation analysis, bibliographic coupling, direct citation, and one hybrid approach. The hybrid approach has boosted the overall accuracy by using the bibliographic coupling approach on both references and words that treated as a references list. Gipp and Beel [22] proposed an approach based on co-occurrences. The proposed approach uses the proximity of co-citations to calculate document relatedness. The closer the proximity of two references within an article, the more related the cited articles are assumed. Leydesdorff [23] proposes an information-theoretical approach which is used to calculate the co-citation similarity on the authors level. Leydesdorff argues that their approach gives better results due to the neglecting of the zero values in computing the similarity and considering the information about the redundancy provided in the matrix. Van Eck and Waltman [24] argues that using Pearson Correlation measure to find the similarity between two articles benefits the co-citation analysis. The paper shows that the personal correlation measure has some shortcomings as a measure of the similarity of the author co-citation.

Additionally, it shows that the similarity measures of the Jensen-Shannon divergence or the Bhattacharyya distance may be considered as better alternatives to personal correlation. More recent, Wang et al. [25] propose a cloud-based path analysis model using a component matrix. Moro et al. [26] propose an algorithm to explore citation sentences and the co-citation data to discover new relevant keywords to online articles.

2.1.2. Content analysis

Content analysis is one of the interesting topics discussed in parallel with the citation analysis [27], [28]. Liu and Chen [29] propose a co-citation proximity approach. The proposed approach considers for each cited article four proximity: the article level, the section level, the paragraph level, and the sentence level. If any two references cited at the same level we consider

that the citation based on that level. For example, if two articles cited in the same section; then these articles are considered section level co-cited articles. The sentence level considered the strongest bind between references while the weakest bond is at the article level. They discuss the content levels, but yet it focuses on the co-citation based on the article structure level while little attention is given to the content itself. Jeong et al. [27] propose an extended approach to analyze the similarity between authors. Citing sentence similarity measures the topical relatedness rather than the co-citation frequency. Furthermore, their method requires the full-text to be available to analyze the similarity. Jiang et al. [30] propose a recommending system based on the topical relatedness approach to analyze the similarity between articles. They aim to use the citation network to generate a set of possible relevant articles. Then, they calculate the similarities between candidate articles and the given article through a concept based topic model. Beel et al.[31] propose an algorithm based on a content-based approach using a term frequency-inverse document frequency (TF-IDF) technique. The proposed approach recommends a research paper considering the author preference is driven from its literature organization, comments and highlighted phrases.

2.1.3. Impact analysis

The citation impact has been examined at different levels, to evaluate the impact of the citation from different perspectives [7]. Commonly, the citation impact is measured by the total citation count in which the article has been cited in a separate article [8], [32]. However, many studies have utilized the citation count as a basic number to compute the impacts of the articles, authors or journals among their peers [33]. The article citation count can be considered the core element in the publication impact measurement to identify the publication's scientific influence. Moed [8] has examined the citation count accuracy in his book. The study targets the differences between the cited articles and the citing article. The used methodology is based on two rounds of

matching to match five elements (the author's family name, first name initial, the year of publication, volume number, and starting page number) and find the 'discrepant' article among the cited articles. The study outcomes show that the overall total number of differences between the cited articles and citing articles are very low. However, the used elements did not reflect the content's relatedness of the citing article among the cited articles appropriately. Fragkiadaki and Evangelidis [33] have reviewed many studies in a matter of measuring the citation impact. The reviewed studies covered the article-level metrics and author-level metrics. Most of the indicators in the article level, such as SCEAS Rank [34], [35] and PageRank [36], consider the citations for all of the cited articles to be equally valued concerning their impact on the citing article. Bornmann [37] discusses the new citation impact indicator for evaluating the article impact on the citation network. The new alternative metrics (altmetrics) approach aims to evaluate the article impact at the article level. This study focused on the correlation between altmetrics and the citation counts. It emphasizes the domination of the social media community participation on the altmetrics impact outcome. Thus, the altmetrics did not reflect the actual citation impact of the article under evaluation. Bornmann and Haunschildb [38] discusses another citation impact indicator to evaluating the article impact in the citation network. Their indicator based on the ratio between the given article citation count and the mean number of cited references in a field to normalize citations. The proposed indicator reflects the global relevancy for a given article in the research area of the citation network for that given article. Hutchins et al. [39] propose a Relative Citation Ratio (RCR) metric. This proposed metric uses citation rates to measure the citation's impact at the article level. The method that they follow to measure the citation rates is based on averaging the article citation count by the journal citation rate of the collection of articles in each co-citation network. Abramoa et al.[40] propose an Article Impact Index (AII). This proposed represents the average rate of the

total citations received by an article and the median of citations received for the publication in the same subject category. The study considers the subject category relation between in the citation analysis concerning the bibliometric indexes. However, they did not consider the content similarity analysis between the articles in the co-citation network. Colliander [41] proposes an approach to normalizing the citation count for a target article based on the similarity between the target article and the references list. The similarity estimated is based on shared cited references and shared technical terminology. The paper considers the content similarity and the commonly cited article. However, it neglects the relation to the co-cited articles and the research field related articles.

2.2. Academic Recommendation

The academic recommendation represents a good approach to support the researchers to take the most appropriate decisions on their future scientific research. Most of the current academic recommendation services cover the following areas: article citation, the collaboration between the authors, and the selection of suitable publishing venues. These services are designed to collect and analyze academic information, such as the similarities between authors, venues, and articles, in the citation network for academic recommendations. They rely on different recommendation techniques that considering the information about the “objects” (content-based), the information about the “ties” between the “objects” (collaborative, co-occurrence, graph-based), the information about the overall effect of “objects” and the “ties” in their environments (global relevance), or a mixture of all types (hybrid).

2.3. Academic Recommendation Techniques

Researchers are developing the academic recommendation system using different techniques. Those techniques targeting area such as content analysis, and relationships analysis.

2.3.1. Content-Based Filtering

The content-based filtering approach is the most frequently utilized technique by researchers in the area of academic recommendations[42]–[44]. It mainly utilized article, author, and venues (items) that had established ties with the researcher (user). Most studies rely on simple terms as a comparative element, and 70-83% of the research uses TF-IDF as the weighting scheme[43], [44]. In addition, Content-based filtering approach handles other information related to the objects such as n-grams linguistics model, layout information, social tag model, and topics relatedness. Vector Space Model has been used by most of the studies to deposit information that represents the items representations and user models [43], [44]. They use the cosine model to estimate the similarities between the user models and recommendation candidates [43], [44]. The main advantage of using this approach is that the recommendation system can benefit from the users' needs model which allows the system to personalize the recommendation outcomes for each user individually.

2.3.2. Collaborative Filtering

The collaborative filtering approach, in general, recommends items based on the analysis of users' behavior rather than items content. This approach highly considers the aspects of social network analysis, interests sharing and user's interaction behaviors. In the domain of academic recommendation, the dispersion and the diversity of the objects in this domain consider an effective factor in the recommendation outcomes using this approach. Vellino [45] discuss the items diffuse in the research industry (Mendeley) and the entertainment industry (Netflix). The study found that the items diffuse on the research industry was three orders of magnitude lower than on the entertainment industry. The different ratio of users and items represents the main factor for this difference. For instance, a typical recommender system has a higher number of users than

movies[46]. Usually, a movie has been watched by many users. Thus, users who share the same preferences can be given effective suggestions. However, it works differently in the research industry. Typically, the number of authors (users) is far lower than the number of articles (items). Besides, the article's rating has not been traded in the same way as the entertainment industry to capture the like-minded authors. Moreover, many articles are not cited by any authors and therefore cannot be recommended.

The “cold start” problem is a common problem of collaborative filtering that may arise in three conditions: new authors (users), new articles (items), and new venues (communities or disciplines) [47]. If a new author cites few or no articles, the system cannot find like-minded authors and hence cannot offer suggestions. If an article is new in the system and has not been cited yet by at least one author, it cannot be suggested. In a new venue, no articles have been published in that venue so that no suggestions can be offered, and as a result, the incentive for authors to cite articles is small.

2.3.3. Co-Occurrence

The co-occurrence approach depends on a frequent joint occurrence of two items together. In the research industry, the most common application reflected this methodology is co-citation analysis presented by Small [48]. The concept of the study is based on the more the same article cites the two articles, the higher the relatedness degree between those articles.

Co-occurrence approach focuses on the relatedness between the items instead of the similarity between them. This approach, unlike the content-based approach, supports the serendipitous recommendation that can benefit the recommendation systems [49]. The relatedness reflects the association degree between two items, which do not necessarily depend on their textual

characteristics. For example, if two articles are using the same words, then they are considered similar. On the contrary, the author and publishing venues are not similar but related, because the author cannot publish an article without a publishing venue and vice versa. Hence, the co-occurrence approach is more comparable to a collaborative filtering approach in terms of serendipitous recommendation. In addition, it minimizes the complexity of analysis by reducing access to the content and analyzing the lexical similarities between the objects. However, this approach reduces the reliance on user preferences in favor of the joint occurrence preferences to recommend items.

2.3.4. Graph-Based

The citation network, in general, represents the relationship (edges) between scientific articles (objects) that can be constructed based on the graph theory, which shows how the articles are linked based on the citations[50], [51]. The graph-based approach depends on the inherited links between the articles that exist in academia. Also, the graph can be built based on different objects such as authors, venues, and publication year [52]. The edges between the objects vary by object type; edges can be co-citations, co-authorship, co-publishing, attribute similarity, demographic information and purchases history [53], [54]. Once a graph was built, Random walks with restarts method, which is a well-known method in graph-based approach, utilized to discover the most popular objects in the graph to recommend the candidate objects [52], [55].

2.3.5. Global Relevance

The global relevance approach adopts a universal-fit methodology that recommends items with the highest relevance based on their surroundings. This approach does not depend on user preferences to calculate the relevance between the recommendation candidates. Instead, it depends

on global measurements such as overall popularity [43]. For instance, in the e-business industry, a retailer system, such as Amazon, could recommend those items that were most often bought at a specific time, for example, during the Christmas season. The basic assumption in this situation would be that users may prefer what most other users like in such an event.

Most of the studies in the academic recommendation have chosen to integrate the global relevance approach with another approach rather than using it exclusively. The vast majority of them use it as an additional ranking factor. Different type of popularity metrics has been used in rank with different approach [56], [57]. Recognizable metrics, such as PageRank, HITS, and Katz metric have been used to re-rank the candidates in the initial recommendation based on their global relevance [56], [58]. By considering the global relevance approach with different approaches can improve the outcome precision compared to content-based filtering approach [43].

2.3.6. Hybrid Recommendation

The hybrid recommendation approach is considered to take advantage of various algorithms from the different recommender systems approaches to make strong conclusions [59]. Aggarwal [59] suggests three main techniques for creating hybrid recommender systems: ensemble design, monolithic design, and mixed systems. Ensemble design got the results from the off-the-shelf algorithms and combined them in one output. The monolithic design aims to integrate existing recommendation algorithm from different approaches to build a comprehensive approach. This approach cannot be easily distinguished from off-the-shelf algorithms due to the integration of several data sources more closely. Mixed systems use multiple recommendation algorithms as a black-boxes like the ensemble design. However, the results of different systems are displayed side by side.

The hybrid recommender approach combines the strength of different types of recommendation approaches to boost the assumption outcomes. This combination can be achieved in several techniques, such as calculating the weighted average of several anticipates. Burke [60] classified the hybrid recommender systems based on the combination techniques into the following categories:

1. ***Weighted:*** The system combined the scores of various recommender approaches into a single score by calculating the weight of each individual component. The weighting method of components can be based on frequency weights, survey weights, analytical weights, or importance weights.
2. ***Switching:*** The system shifts between various approaches according to the current situation. For example, to avoid cold-start issues in the collaborative approach, the content-based approach can be used in the initial phases until more user rating are available. Therefore, the system can adapt itself to the approach that provides the most accurate recommendation.
3. ***Cascade:*** The system revises the candidate's list by an initial recommender approach using another. In general, the processing of data in the following approach is influenced by the outputs of the previous one, and the outcome of the last process is considered as a combined conclusion.
4. ***Feature augmentation:*** The system treats the output of an initial recommender approach as input features for the following approach. This technique is closer to the *Cascade* technique in sequencing two recommended approaches. However, it differs in the way the outputs of the previous approach are treated. This technique treats them as a feature in addition to the same inputs in the previous approach, while *Cascade* technique treats them as an independent input without looking at inputs of the previous approach.

5. ***Feature combination:*** The system treats different recommendation data as additional context features in one recommendation approach. For instance, in this technique, the system looks at collaborative and graph-based data without fully relying on it to reduce the system's sensitivity. Both data will be used as an additional context features to evaluate an item using content-based recommendation approach.
6. ***Meta-level:*** The system treats the recommendation model used in the initial recommender approach as an input for the following approach. This technique is closer to the *Feature Augmentation* technique in sequencing two recommended approaches. However, it differs in the way the outputs of the previous approach are treated. This technique treats the entire model as an input, while *Feature Augmentation* technique uses a learned model to generate features for input to a second approach.
7. ***Mixed:*** The system uses more than one technique to combine different recommendation approaches to make a large number of recommendations at the same time. Ordinarily, this technique is used when the recommendation is a composite entity in which multiple elements can be recommended as a relevant group. For instance, in the entertainment industry, textual descriptions of movies, the preferences of other users for a specific movie or similar category movies can be used to recommend a movie. This technique is most relevant to the areas of complex elements, and it is often used to avoid the 'cold-start' problem to recommend new items even if they have not been rated by anyone.

2.4. Academic Recommendation Applications

Some studies examine various aspects of academic recommendations in different applications. However, this section mainly reviews literature that uses recommendation systems,

particularly in academic, to enhance candidate recommendation, content-based search, and item visualization.

2.4.1. Article Recommendation

Article recommender system represents a good approach to support the researchers to identify the most appropriate articles to cite. Most of the existing systems rely on the topical relatedness which isolates the role of the citation network analysis to identify the most influential articles[61]–[63]. Huang et al.[62] propose a citation recommendation system using content-based approach. The proposed approach recommend an article for citation considering the topic relation between the citing and cited articles. However, it did not consider the relationship between articles in the citation network to determine the article impact using bibliometric and co-authorship information. Pera and Ng [63] discuss the author's research interests using content-based, collaborative and global relevance approaches. The proposed approach recommends a research paper considering the paper similarity, author rating score and the number of times an article stored in their personal libraries. Sugiyama and Kan [64] discuss the author's research interest and how it can be used to enhance the recommendation of articles using a collaborative approach. They recommend an article for citation considering the author’s research interest by building a user profile to enhance the results. Beel et al. [42] propose a research paper recommender system using content-based approach. The proposed approach recommends a research paper considering the author preference is driven from its literature organization, comments and highlighted phrases. Küçüktunç et al. [65] propose an academic recommender system using a graph-based approach. However, the approaches proposed by Pera and Ng, Sugiyama and Kan, and Beel et al. conceal the bibliometric information that can be used to examine the article impact among the given citation network. Thus, it did not reveal the article impact compared to peers in the citation network.

2.4.2. Author Recommendation

Author recommender system reflects a good approach to suggest expert researchers in order to cooperate with them or invite them as referees to review a research manuscript. Errami et al. [66] propose tools to recommend expert reviewers among other services. Their approach uses term-based similarity and the author's position to recommend experts in the given research field. However, they did not consider other relation that represents the citation network such as publication venue relation and the citation impact. Schuemie and Kors [67] propose tools for finding expert reviewers among other services. Their approach uses a content-based approach to find a similarity relationship between the search query and the candidate list. Then, they use a collaborative approach to determine the journal or author list. However, they neglect the co-authorship relation and publication venue relation that represents the citation network of a given research area. Cohen and Ebel [68] studies the author collaborators recommendation in a social network based on a given set of keywords. They are combining the co-occurrence approach, content-based approach, and global relevance approach to rank an author for collaboration. Their study concludes that combining the co-occurrence and content-based approaches have a positive effect on the analysis outcome. Paul et al. [69] propose an algorithm to rank an author based on temporal analysis of citation collaboration. Their approach uses co-occurrence, and global relevance approaches to analyzing author contributions over time. Although, they consider the co-authorship relation and citation relation in their analysis, yet they neglect the analysis of the content, which is an essential part of measuring the expected impact within the citation network of a given article.

2.4.3. Publishing Venue Recommendation

Papers submission to non-relevant publishing venue is one of the major reasons for rejection in the research industry. Publishing venue recommender system can assist a user to find the appropriate venue for his work to be submitted. Kang et al. [70] propose journal recommendation systems using content-based and global relevance approaches. The proposed systems use the traditional term frequency model and term weighting score to recommend related articles. However, they neglect the venue impact and relationship that are related to the given query. Yu et al. [71] propose a venue recommendation using co-occurrence and collaborative approaches. Although, they consider the co-authorship relation and venues relation in their analysis, yet they neglect the analysis of the content of the article, which is an essential part of measuring the expected impact within the citation network of a given article.

2.5. Comparison to the Existing Studies

Although certain recommendation and citation analysis parameters have already been incorporated in the academic recommendation process in a number of existing works, there is a lack of a general formulation on how to employ citation impact as dimensions in the traditional academic recommendation model. We do our research to cover the recommender systems in the academy. We came up with two main aspects that represent the system's features which are: society understanding and recommendation content as shown in table 2-1.

In the society understanding, most of the existing work has covered the subcategory under that aspect such as user preferences, time range topic based similarity. Another subcategory such as bibliometric citation analysis and popularity ranking has been partially covered by some of the existing work. In the other hand, up to our knowledge, there are some subcategory features that

have not been considered by the existing work such as semantic-based content similarity and citation analysis using citation impact which has been considering in this thesis in order to cover the gap in the existing work.

In the recommendation content, the main three academic objects that have been targeted by the researchers in the development of the academic recommendation systems are an article, author, and venue. Most of the existing work has covered each academic object individually. Some of the researchers have developed a system that covers two out of three academic objects. In the other hand and up to our knowledge we did not come up with any existing work that covers all three academic objects in the same time which has been considered in the preparation of this thesis in order to overcome the lack in the existing work.

2.6. Summary

With the current rapid growth of Big Data, developments in the indexing of scientific publications and recommendations are receiving significant interest from researchers. There is a marked increase in the volume of scientific publications and their corresponding metadata databases. Specifically, indexing articles that measure scientific impact and help build virtual communities of researchers with common interests are becoming more relevant than ever before. Authors and publishers wish to find their preferred content easily and efficiently, with minimal effort. Academic recommender systems try to increase the level of a user's satisfaction by analyzing the content of publications or the amount of information available about the citation network. However, the user's preferences and selections to cite different articles should not overcome the importance of citation impact. For instance, users' selections depend on various scenarios such as content similarity, bibliometric impact, author experience, or articles popularity.

Such information can increase the quality of the recommendation outcome and enhance the user experience [11]. Content-based filtering, collaborative filtering, co-occurrence, and global relevance are the fundamental recommendation approaches to handle the user's inquiries and find resources in the academic recommendation. Combining these four approaches supports the users to explore hidden links between the characteristics of content, trace the user's preferences and collaboration, and analyze the articles' impact compared to their peers in the citation network. We briefly bring attention to the different existing recommendation techniques in the research industry. We also review existing approaches that explore scientific communities' detection and citation impact analysis computing during the recommendation process. This chapter discussed the research works in academic recommendations and how they related to this thesis, and how this thesis differs from the current research in that filed.

Reference work	Society Understanding												Recommendation Content			Content Visualization
	Inferential Context				Content Similarity			Citation Analysis								
								Bibliometric								
	Approach	Hybrid Technique	User preferences	Temporal	Term based	Topic based	Semantic based	Article	Author	Venues	Citation Impact	Popularity Ranking	Article	Author	Venues	
Huang et al.[62]	CBF	–	●	x	x	●	x	x	x	x	x	x	●	x	x	x
Pera and Ng [63]	Hybrid (1,2,5)	Weighted	●	x	●	x	x	x	x	x	x	●	●	x	x	x
Sugiyama and Kan [64]	Hybrid (1,2)	Feature augmentation	●	x	●	●	x	x	x	x	x	x	●	x	x	x
Beel et al. [42]	CBF	–	●	x	x	●	x	x	x	x	x	x	●	x	x	x
Küçüktunç et al. [65]	GB	–	●	x	x	x	x	x	x	x	x	x	●	x	x	○
Errami et al. [66]	CBF	–	x	●	●	x	x	x	x	x	x	x	x	●	○	x
Schuemie and Kors [67]	Hybrid (1,2)	Cascade	x	x	x	●	x	●	x	x	x	x	x	●	○	x
Paul et al. [69]	Hybrid (3,5)	Weighted	x	●	x	●	x	x	x	x	x	x	x	●	x	x
Kang et al. [70]	Hybrid (1,5)	Feature combination	●	x	●	x	x	●	x	x	x	x	x	x	○	x
Yu et al. [71]	Hybrid (2,3,5)	Feature combination	●	x	x	●	x	x	x	x	x	x	x	x	●	x
Z. Chen et al. [82]	Hybrid (4,5)	Mixed	●	x	x	x	x	x	x	x	x	●	x	x	●	x
F. Xia et al. [85]	Hybrid (2,3,5)	Feature augmentation	●	●	x	x	x	x	x	x	x	x	x	●	x	x
This thesis	**Hybrid (1,2,3,5)**	**Mixed**	●	●	●	●	●	●	●	●	●	●	●	●	●	●

Table 2-1: A summary of related literature for context-based recommender systems

(CBF (1): Content-Based Filtering, CF (2): Collaborative Filtering, CO (3): Co-Occurrence, GB (4): Graph Based, GR (5): Global Relevance, ●: information is considered, ○: information is partially considered, x: information is not considered,−: information is not applicable)

Chapter 3 . Academic Recommendation Model

In this chapter, we propose a new algorithm to measure the citation impact named Academic Citation Impact (ACI). Our ACI algorithm enables us to identify the impact of articles at their extended nearby citation network for a given article. We combine the article contents with its bibliometric to evaluate the citation impact of the articles in their surrounding network. We utilize the article metadata to calculate the semantic similarity between the two articles in the extended network. The article similarity scores along with the bibliometric scores are used to assess the impact of the article among their extended nearby citation network. After that, the algorithm builds three similarity matrices using the citation network matrices for the given article to find the similarity between the articles, authors, and venues. After that, the algorithm builds three latent models using the articles, author, and venue similarity matrices. The latent models are i) the latent author preference model to determine the author's latent preferences toward their association to articles and venues, ii) the latent article preference model to determine the article's latent preferences toward their association to authors and venue, and iii) the latent venue preference model to determine the venue's latent preferences toward their association to articles and authors.

Then a ranking algorithm using similarity matrices with citation network matrices to recommend the academic objects according to the user's request by computing the confidence score in order

to find the most suitable venues, potential authors to team up with, and most relevant articles to cite according to the user needs.

Table 3-1: Summary of notations and their meaning

Notations	Meaning
$\boldsymbol{H}$	Set of Authors.
$\boldsymbol{A}$	Set of Articles.
$\boldsymbol{V}$	Set of Venues.
$\boldsymbol{T}$	Set of Terms.
$\boldsymbol{E}$	Set of Article Elements.
$\mathbf{C}(\tilde{\mathbf{C}})$	Author-Article matrix (Normalized matrix of $\mathbf{C}$).
$\mathbf{P}(\tilde{\mathbf{P}})$	Venue-Article matrix (Normalized matrix of $\mathbf{P}$).
$\mathbf{L}(\tilde{\mathbf{L}})$	Author-Venue matrix (Normalized matrix of $\mathbf{L}$).
$\mathbf{R}(\mathbf{R}^{\boldsymbol{k}})$	Author-Author similarity matrix (Contains k most similar authors).
$\mathbf{G}(\mathbf{G}^{\boldsymbol{k}})$	Article-Article similarity matrix (Contains k most similar Articles).
$\mathbf{Z}(\mathbf{Z}^{\boldsymbol{k}})$	Venue-Venue similarity matrix (Contains k most similar Venues).
$\overrightarrow{\mathbf{HA}}$	Latent preferences of authors toward articles matrix.
$\overrightarrow{\mathbf{HV}}$	Latent preferences of authors toward venues matrix.
$\overrightarrow{\mathbf{AH}}$	Latent preferences of articles toward authors matrix.
$\overrightarrow{\mathbf{AV}}$	Latent preferences of articles toward venues matrix.
$\overrightarrow{\mathbf{VA}}$	Latent preferences of venues toward articles matrix.
$\overrightarrow{\mathbf{VH}}$	Latent preferences of venues toward authors matrix.

3.1. Definitions and Notations

In this chapter, we introduce a common set of concepts that will be presented in this thesis. Upper-case bold letters, such as **J,** represent matrices; while the corresponding lower-case italicized letters, such as *a,* represent entries in the matrices and capital italic letters, such as ***A*,** represent sets of entries. The lower case letter subscripts, for example A_y, represents an entry y from the set A. Table 3-1 summarizes notations used in the rest of this chapter.

3.2. Building the Academic Recommendation Model

The citation network represents a group of authors, articles, and venues that have relationships with one another. These relationships represented in the citation network by one or more of the following academic association: articles citation, co-citation, co-authorship, and co-publishing. For a set of authors, articles, and venues corresponding to our academic recommendation model, the set of authors $H = \{h_1, h_2, \ldots, h_{|H|}\}$, the set of articles $A = \{a_1, a_2, \ldots, a_{|A|}\}$, and a set of venues $V = \{v_1, v_2, \ldots, v_{|V|}\}$, the citation network formalized as a tuple $G := (H, A, V, Q)$ where $Q \subseteq H \times A \times V$ a trilateral relationship. Therefore, a citation network can be viewed as a three-dimensional space of authors, articles, and venues; consequently, this three-dimensional space can be projected onto three two-dimensional matrices, as shown in Figure 3-1.

We first define the three matrices obtained by aggregating over authors, articles, and venues, as follows:

- Author-Article matrix $\mathbf{C} = [c_{h,a}]_{|H|\times|A|}$, where $c_{h,a}$ represents the academic citation impact of article a that cited articles written by author h.

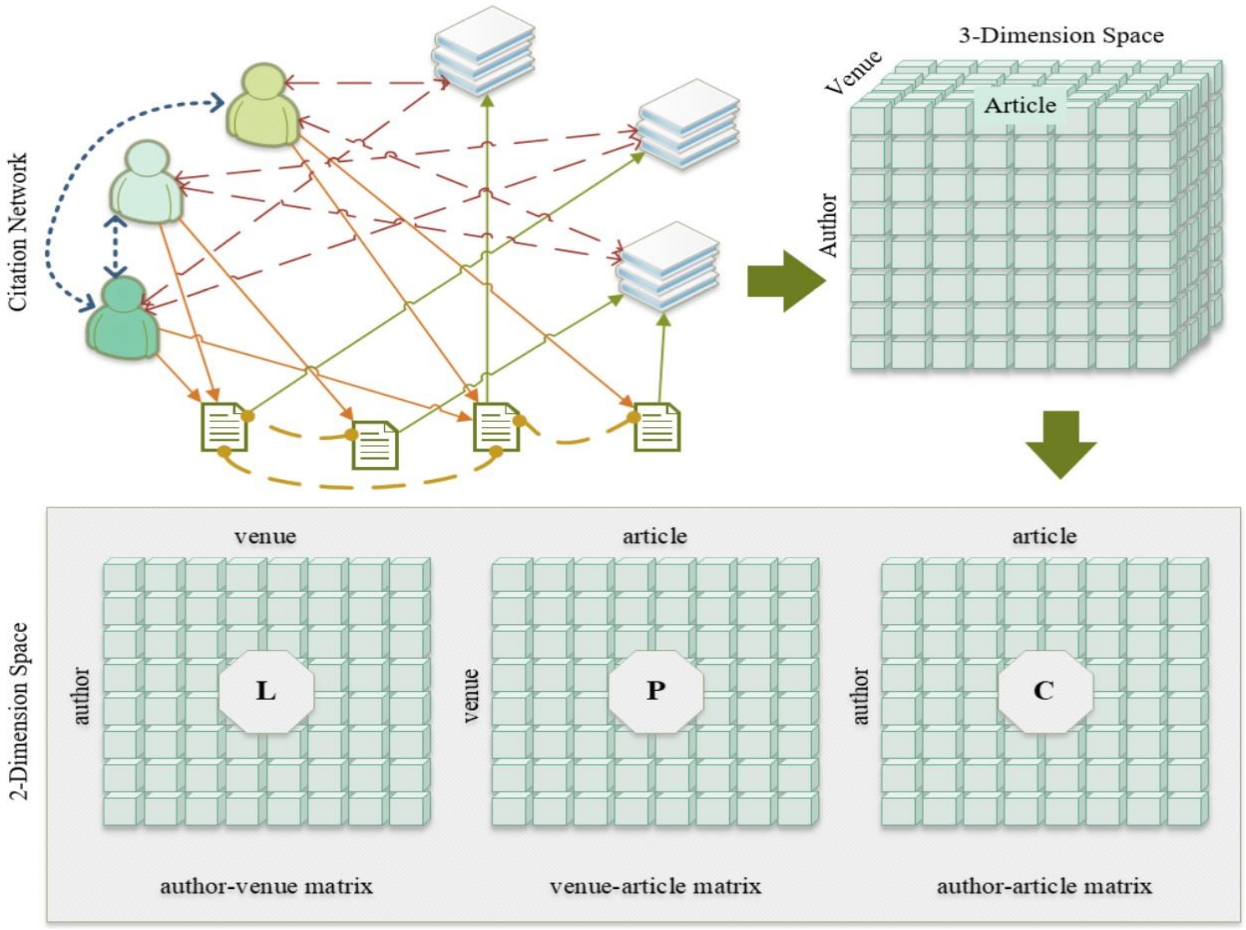

Figure 3-1: Illustrating the Citation Network in three-dimensional space and in two-dimensional space

- Venue-Article matrix $\mathbf{P} = [p_{v,a}]_{|V|\times|A|}$, where $p_{v,a}$ represents the academic citation impact of article a that cited articles published in venue v.
- Author-Venue matrix $\mathbf{L} = [l_{h,v}]_{|H|\times|V|}$, where $l_{h,v}$ represents the academic citation impact of articles that author h have published in venue v.

3.3. Academic Citation Impact analysis

The citation impact is a quantitative concept that is utilized to measure the intellectual influence [8]. Measuring the academic impact by using the bibliometric alone will conceal the actual impact of the transferred knowledge. The content similarity represents the knowledge that

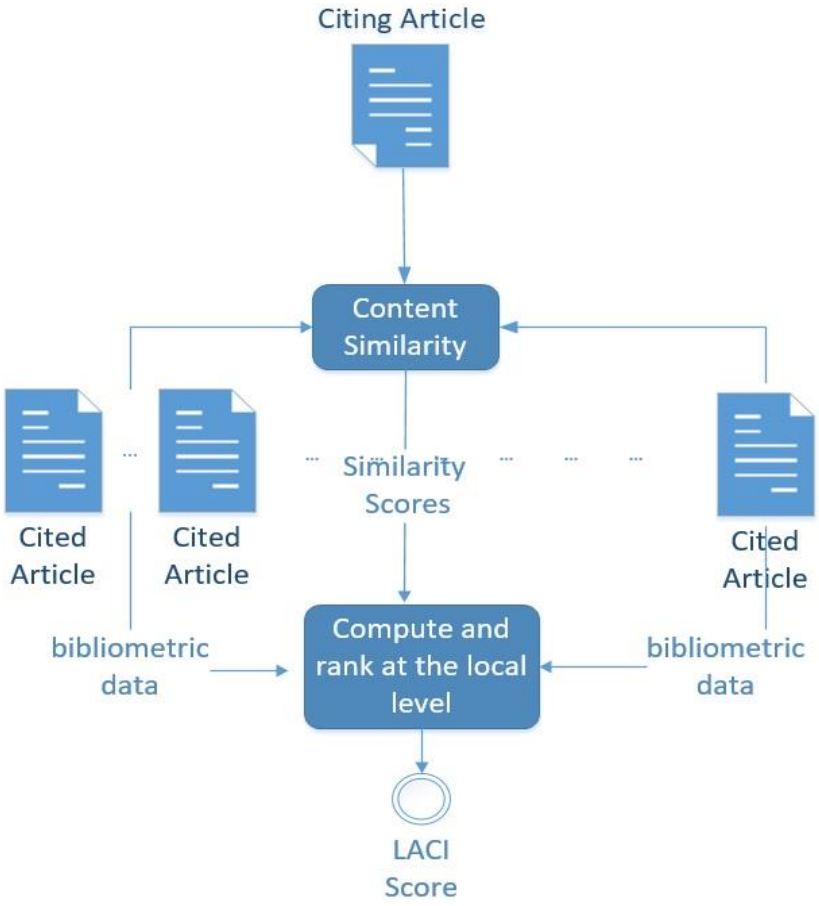

Figure 3-2: Extracting and calculating the assessment parameters at the local dimension.

has been perceived, discovered, or learned from a cited article and transmitted to the citing article. Fundamentally, the relation between the citing article and the list of cited articles is widely considered a granted citation relationship according to scientific publications practices. However, each cited article has a different impact depending on the amount of the transferred knowledge to the citing article. We argue that combining the bibliometric data, and the content semantic similarity together to evaluate a cited article's impact will improve the evaluation process of citation network analysis. Thus, it will facilitate the process of finding the most influential articles. In order to find the academic citation impact, we combine the bibliometric data and the content

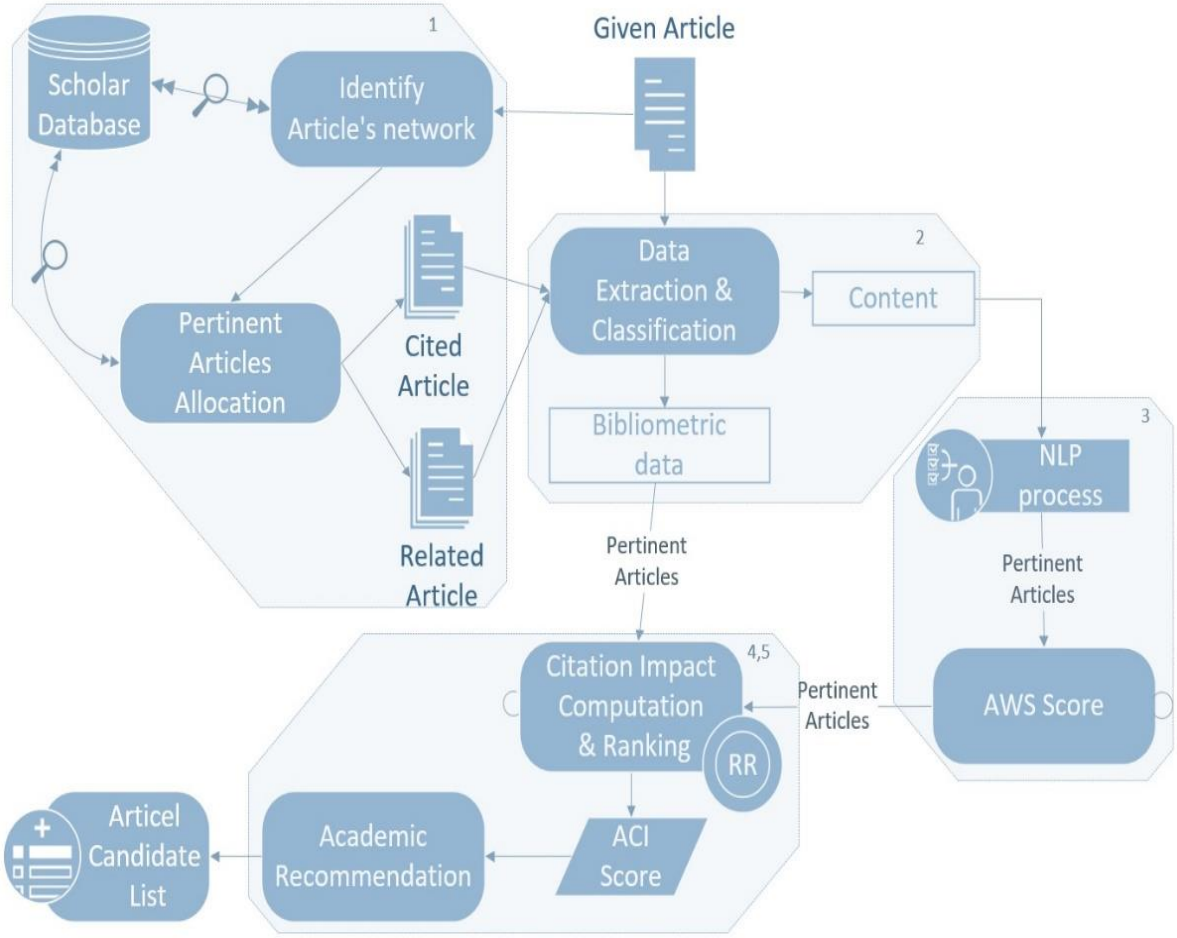

Figure 3-3: An overview of the system architecture of the proposed algorithm to measure the citation impact of a specific article in the collection of pertinent articles.

semantic similarity to evaluate the academic's impact on the given article (citing article) in two dimensions: local and global. At the local dimension, we examine the academic's impact on the given article among the co-cited articles to measure the Local Academic Citation Impact (LACI), as shown in Figure 3-2. At the global dimension, we examine the academic's impact on the given article among the related articles in its research scope field, to measure the Global Academic Citation Impact (*GACI*). We also use the semantic similarity score and bibliometric parameters to analyze an academic's impact on the given article among the extended nearby citation network,

based on the proposed dimensions. These are used to compute the final Academic Citation Impact (*ACI*) of each article in the collection of pertinent articles, as described in Section 3.3.3.

The system architecture of the proposed algorithm to measure the citation impact of a specific article in a collection of pertinent articles is illustrated in Figure 3-3.

The system applies the following five steps:

1. Searching for and retrieving the collection of pertinent articles containing the cited articles in the given article, and the related articles from the publishing venue (conference or journal).
2. Extracting and identifying the article's main metadata content (i.e., title, abstract and keywords), and the bibliometric information in the metadata for each article.
3. Determine the semantic similarity between each article in the collection of pertinent articles and the given article.
4. Ranking the collected articles based on their impact on the given article, using the computed semantic similarity and the bibliometric data.
5. Computing both the *LACI* and *GACI* for each article in the pertinent articles collection, in order to calculate the *ACI* score. The *ACI* score is used to recommend a list of articles with higher impacts on the given article. Users can then explore the recommended list to find more information related to their work.

3.3.1. Pertinent Articles Allocation

The collection of pertinent articles has two distinct types of articles: the cited articles and related articles. The cited articles include all the articles that have been cited together in a given article. The related articles include all the articles that have been published in the given venue of

the citing article and its corresponding cited articles. We use the following parameters in our citation analysis: the semantic similarity score of the main metadata content; and the bibliometric data consisting of the citation count and the publisher impact index.

We construct the cited articles list (CAL) and the venues related articles list (VARL) to distinguish between the two types of articles in the collection of pertinent articles. To retrieve the CAL, which is used to compute LACI, we search and retrieve the cited articles for each given article using the article id. We collect the VARL, which is used to compute GACI, as follows: first, from the metadata of the given article and each of its corresponding cited articles, we extract their venue names. Then, we search and retrieve an initial list that includes all the articles from each of the publishing venues. Next, we extract and prepare the articles main metadata content from the given article and the cited article to build a topic corpus. Later, we use the topic corpus to search all the articles that are topically related to the cited articled and, hence build an extended nearby citation network. We categorize the venues related articles based on a relevancy score using TF-IDF. We measure the similarity for each article in the initial list collected from given venues based on the topic corpus. The returned articles are sorted based on their relevancy scores. The relevancy score is computed using equation (*3-1*)

$$tfidf(t,d,D) = \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}} . \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (3\text{-}1)$$

where $f_{t',d}$ denotes the total number of all terms t' used in the article d, and $f_{t,d}$ denotes the total number of a particular term t in that article d. $|D|$ denotes the total number of articles in the collection, whereas $|\{d \in D : t \in d\}|$ denotes the total number of articles d in the venues related articles collection D that contain a particular term t from the subject corpus. To

formulate the topic corpus for the citing articles, we extract and classify the article terms. The set of terms is denoted as $T = \{t_1, t_2, \cdots, t_n\}$, which represents the citing article's topic terms. The term frequency (TF) for each term t in the given article d is computed using $tf(t,d) = \frac{f_{t,d}}{\max\{f_{t',d}:t' \in d\}}$. The terms t is clustered and normalized based on the terms t' frequency in the given article to identify the top ranked terms. We use the top ranked terms to formulate the topic corpus to conduct an extensive search to retrieve the articles that have the most common terms compared to the citing article and the cited articles, as introduced previously.

3.3.2. Article Semantic Similarity

We collect the title, abstract and keywords for each article in the CAL and the VARL lists, and use this information as comparable elements set E to compute the semantic similarity between the articles. We calculate the semantic similarity score between the source article X (given article) and the target article Y (cited/related article) using the proposed comparable set E, then we compute the Article Weighted Similarity (AWS) score, which represents the semantic similarity degree between two articles.

To prepare the elements for comparison, the stemmed and tokenized techniques from the NLP approach are applied to maximize the benefits of the used elements. To compute the AWS, we compute the similarity distance between the elements in both the source and target articles as element-element vectors. Several similarity techniques can be used [72], including Euclidean Distance [72], Cosine Similarity [27], [72] and Pearson Correlation Coefficient [24], [72]. We chose to use the Cosine Similarity technique due to its high performance, which has been proven in previous studies [24], [72]. Cosine Similarity represents the cosine angle between two sets of vectors.

The similarity between a source element e_x and a target element e_y is calculated using the list of terms $\{t_1, t_2, \cdots, t_n\}$. For example, the title "Signature Verification Algorithm Using Transition Model" is mapped to the element set of e_x, and the title "Verification Protocol for handwritten images," is mapped to the element set e_y. Table 3-2 illustrates the semantics corpus buildup for the elements e_x and e_y.

Table 3-2: The semantics corpus of two articles on one element (article title).

Terms	verification	signature	algorithm	protocol	transition	model	handwritten	images
$\boldsymbol{e_x}$	1	1	1	0	1	1	0	0
$\boldsymbol{e_y}$	1	0	0	1	0	0	1	1

Using the Latent Semantic Analysis technique (LSA), the elements' semantic space is constructed using the semantics corpus of each element in both articles [42]. LSA is a well-known algorithm to capture the semantic relations among elements from the semantic corpus under optimal conditions [14, 43]. The use of the comparable elements increases the LSA technique capability to capture the semantic relations from the terms semantic space. The similarity between each pair of elements is calculated using Equation (3-2).

$$Sim_{(e_x,e_y)} = \frac{\sum_{i=1}^{|\mathrm{E}|} e_{x,i} . e_{y,i}}{\sqrt{\sum_{i=1}^{|\mathrm{E}|} e_{x,i}^2 \cdot \sum_{i=1}^{|\mathrm{E}|} e_{y,i}^2}} \tag{3-2}$$

where i represents a particular comparable element e in the source and target article, and $|\mathrm{E}|$ represents the total number of comparable elements in the source and target article. The variable $e_{x,i}$ represents the ith comparable element in the source article x, and $e_{y,i}$ represents the ith corresponding element in the target article y.

The *AWS* score represents the article weighted similarity score between article x and article y, where the score ranges between 0 and 1. For example, if the *AWS* score is equal to zero, then there is a lack of relationship between the articles. In contrast, if the score is equal to 1, then the two articles are identical. A score between 0 and 1 represents partial semantic similarity between the two compared articles. The score is calculated by giving a specific weight to each element, based on a precomputed optimal weight combination W for each, as shown in Equation (3-3).

$$AWS(x,y) = \sum_{i=1}^{|E|} ow_i . Sim_{(e_x,e_y)_i} \tag{3-3}$$

where $Sim_{(e_x,e_y)_i}$ represents a non-negative cosine similarity degree range value from 0 to 1 between the compared elements $e_1, e_2, \cdots, e_{|E|}$ for a given article x and a target article y. The variable ow_i represents the weight of each element-element similarity score $Sim_{(e_x,e_y)_i}$, and the set of weights is denoted as $W = \{ow_1, ow_2, \cdots, ow_{|W|}\}$, where the total sum of the weights $\sum_{i=1}^{|W|} w_i = 1$. The *AWS* score is calculated for each article in the *CAL* and *VARL* lists, based on its relationship to the citing article. Thus, the score is used to compute the relation relevancy of the given article among the *CAL* and *VARL* lists to evaluate its impact.

3.3.3. Pertinent Articles Ranking

We evaluated the impact of the collection of pertinent articles based on their relevancy ratio to the given article, and we used reversal rank to compute the relation relevancy for the collection of pertinent articles based on their standing among the counterpart articles in the collection using different parameters. The parameters include the citation count, author impact index, publisher

impact index, and *AWS* score computed and extracted previously. We use the parameters to calculate the citation impact for each article in the collection. We aim to use the calculated *ACI* to recommend the article for the user based on the given article.

An academic's impact is assessed in two dimensions: the local dimension and the global dimension. We measure an academic's local impact by considering the articles in the *CAL* list to reflect the academic impact on the given article at the local dimension. Moreover, we measure the academic global impact by considering the articles in the *VARL* list to reflect the academic impact on the given article at the global dimension. Accordingly, the local dimension reflects each academic's impact on the given article with respect to the co-cited articles, and the global dimension reflects each academic's impact on the given article with respect to the related articles from the given publishing venues. *ACI* is calculated using Equation (3-4).

$$ACI\ (y|x) = \frac{\left(LACI(y|x) + GACI\ (y|x)\right)}{2} \tag{3-4}$$

where the *ACI* score denotes the article y academic citation impact on the given article x, *LACI* denotes the local academic citation impact and *GACI* denotes the global academic citation impact. Both scores are calculated for each article in the the collection of pertinent articles, using Equations (3-5) and (3-6) to evaluate an academic's impact on the given article among their counterpart articles.

$$LACI\ (y|x) = \frac{\left(\frac{1}{|P|}\sum_{i=1}^{|P|}\frac{1}{LR(P(y)_i)}\right) + \left(\frac{1}{LR\left(AWS_{(x,y)}\right)}.AWS_{(x,y)}\right)}{2} \tag{3-5}$$

The *LACI* score denotes the local academic impact value of article y on a given article x, and i denotes a particular parameter P for article y. The set of parameters is represented as $P(y) = \{CC_y, AII_y, VII_y\}$, where CC_y denotes the total number of times article y has been cited, AII_y denotes the total of author's impact for authors who write article y, VII_y denotes the venue's impact where article y was published and P denotes the total number of parameters that have been used in the ranking process. AWS denotes the weighted similarity score between article y and the given article x. $LR\left(P(y)_i\right)$ denotes the local ranking position (k value) of article y for each parameter p among its counterpart articles in the *CAL* list and $LR(AWS_y)$ denotes the local ranking position (k value) of the article y similarity score among its counterpart articles in the *CAL* list.

$$GACI\ (y|x) = \frac{\left(\frac{1}{|P|}\sum_{i=1}^{|P|}\frac{1}{GR(P(y)_i)}\right) + \left(\frac{1}{GR\left(AWS_{(x,y)}\right)}.AWS_{(x,y)}\right)}{2} \quad (3\text{-}6)$$

The *GACI* score denotes the global academic impact value of article y on the given article x, and i denotes a particular parameter P for the article y. The set of parameters is represented as $P(y) = \{CC_y, AII_y, VII_y\}$, where CC_y denotes the total number of times article y has been cited, AII_y denotes the total of author's impact for authors who write article y, VII_y denotes the venue's impact where article y was published and P denotes the total number of parameters that have been used in the ranking process. AWS denotes the weighted similarity score between article y and the given article x. $GR\left(P(y)_i\right)$ denotes the global ranking position (k value) of article y for each parameter p among its counterpart articles in the *VARL* articles list and $LR(AWS_y)$ denotes the global ranking position (k value) of the article y similarity score among its counterpart articles in

the *VARL* articles list.We use the *ACI* score to propose a citation recommendation list that represents a user candidate recommended list.

3.4. Citation Network Similarities

In this thesis, we determine three types of similarities in order to calculate the collaborative confidence score. The similarities types are the author-author similarity, article-article similarity, and venue-venue similarity. The similarity measure can be computed using different matrices. For instance, to compute the similarities between authors $\mathbf{D}_{|H|\times|H|}$, we use two matrices the author-article matrix $\mathbf{C}_{|H|\times|A|}$ and the author-venue matrix $\mathbf{L}_{|H|\times|V|}$. As mentioned earlier, If two authors frequently appear together in multiple venues, those authors have greater potential for collaboration based on the closely related semantics [71]. In this case, we could successfully measure the authors' similarities in terms of venues. The authors' similarities in terms of articles will be measured based on the co-occurrence of the event of co-authorship. In this thesis, we examine the use of both approaches to determine the similarities.

3.4.1. Author-Author Similarity

To compute the author-author similarity matrix $\mathbf{R}_{|H|\times|H|}$, we start by utilizing the author-article matrix $\mathbf{C}_{|H|\times|A|}$ and the author-venue matrix $\mathbf{L}_{|H|\times|V|}$. The idea behind relying on the detection of similar authors who attend the same venues is to use the list of venues attended by given authors to discover other interesting venues attended by similar authors. On the other hand, noticing the authors who are cited the same articles help to discover the degree of interest in cooperation between the authors, which helps to predict the possibility of future cooperation between them. We adopted the cosine-based similarity approach to determine the similarity

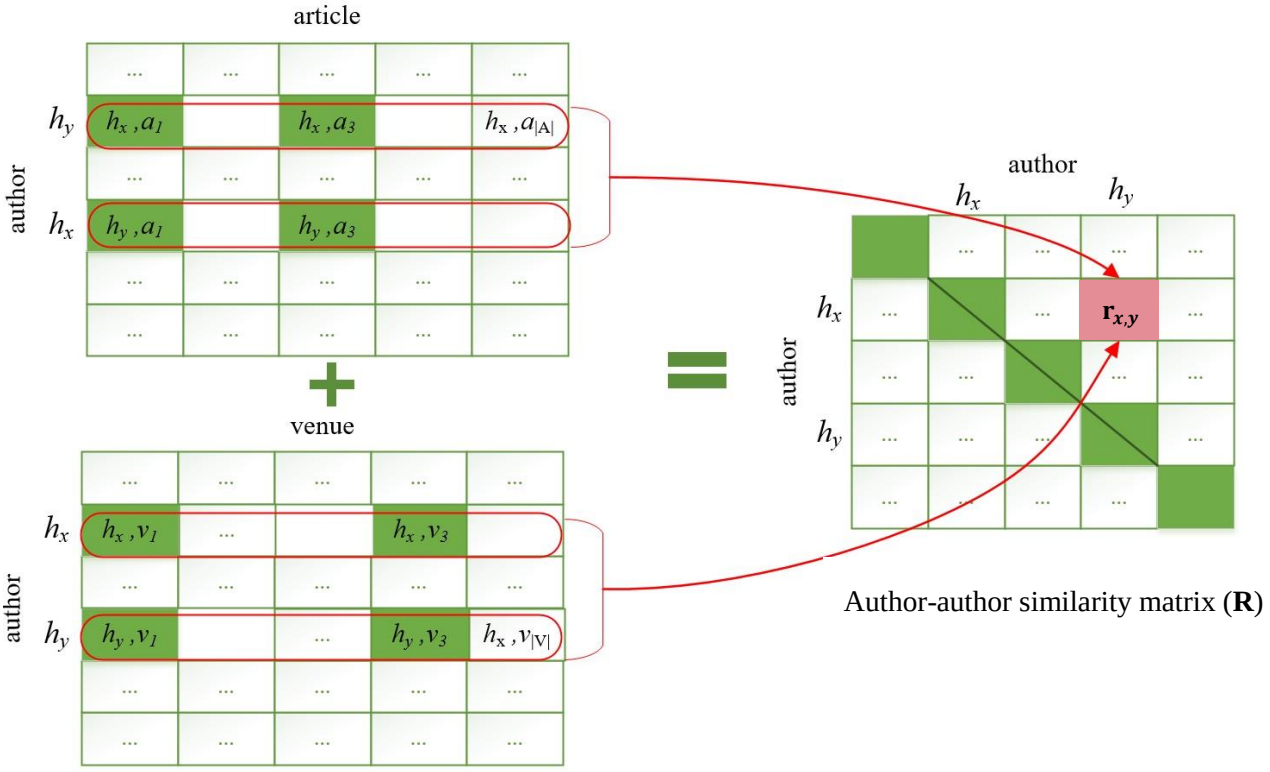

Figure 3-4: An illustration of the author-author similarity matrix **R**.

between two authors. The cosine-based similarity takes two vectors of shared venues and articles of authors h_x, and h_y, and computes their similarity according to their angle, as in Equation (3-7).

$$r_{(h_x,h_y)} = \cos(h_{x,v}, h_{y,v}) + \cos(h_{x,a}, h_{y,a}) = \frac{h_{x,v} \cdot h_{y,v}}{\|h_{x,v}\| \cdot \|h_{y,v}\|} + \frac{h_{x,a} \cdot h_{y,a}}{\|h_{x,a}\| \cdot \|h_{y,a}\|} \quad (3\text{-}7)$$

The author-author similarity matrix $\mathbf{R}_{|H|\times|H|}$ represents the similarity between authors where both rows and columns represent authors. We consider the top k nearest authors for each author in order to lower the computational time. Accordingly, if the corresponding similarity value between a pair of authors x and y is greater than the kth highest similarity value, we set $r_{x,y}$ to the similarity value otherwise, we set $r_{x,y}$ value to zero. We use the non-zero entries to build the called the $\mathbf{R}^{k}$ matrix that contains the k most similar authors. Figure 3-4 shows an example of building the similarity matrix **R**.

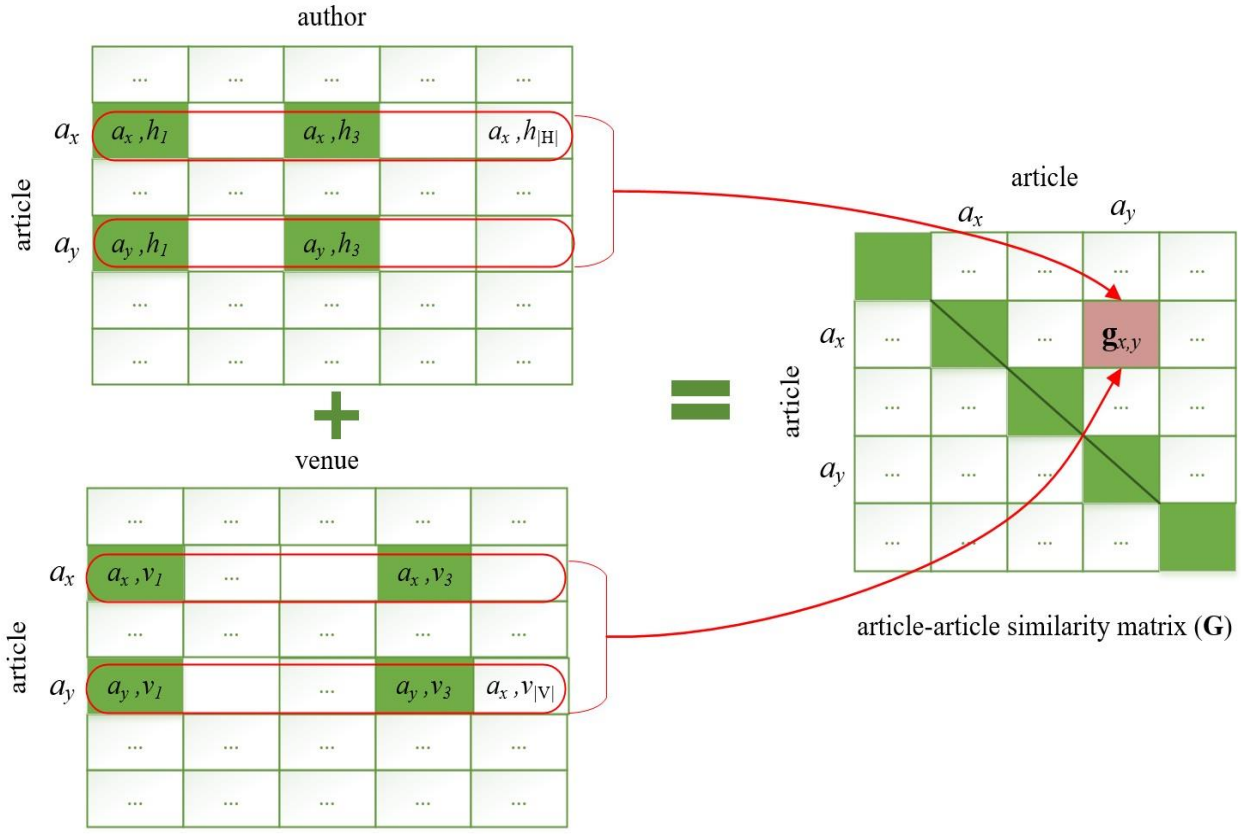

Figure 3-5: An illustration of the article-article similarity matrix **G**.

3.4.2. Article-Article Similarity

To compute the article-article similarity matrix $\mathbf{G}_{|A|\times|A|}$, we consider the author-article matrix $\mathbf{C}_{|H|\times|A|}$ and the article-venue matrix $\mathbf{P}_{|A|\times|V|}$. To discover venues that most suitable to publish a given article, we detect the similar cited articles that have been published in the same venues to identify the list of venues that has the most similar article to the given one. Moreover, detecting the similar articles that have been cited by same authors help to forecast the possibility of future cooperation between authors that have a "common collaborator" (shared author) [73], [74]. The cosine-based similarity approach has been selected to determine the similarity between

two articles a_x, and a_y based on the shared venues and authors. The cosine-based computes their similarity according to the angle between two vectors of articles as in Equation (3-8).

$$g_{(a_x,a_y)} = \cos(a_{x,v}, a_{y,v}) + \cos(a_{x,h}, a_{y,h}) = \frac{a_{x,v} \cdot a_{y,v}}{\|a_{x,v}\| \cdot \|a_{y,v}\|} + \frac{a_{x,h} \cdot a_{y,h}}{\|a_{x,h}\| \cdot \|a_{y,h}\|} \quad (3\text{-}8)$$

Figure 3-5 shows an example of creating the article-article similarity matrix **G**. The $\mathbf{G}_{|A|\times|A|}$ matrix represents the similarity between articles where both rows and columns represent articles. We consider the top k nearest articles for each article to build the $\mathbf{G}^{\boldsymbol{k}}$ matrix. Accordingly, if the corresponding similarity value between a pair of authors *x* and *y* is greater than the *k* highest similarity value, we set $g_{x,y}$ to the similarity value otherwise, we set $g_{x,y}$ value to zero. The non-zero entries will be consider to build the matrix that contains the k most similar articles.

3.4.3. Venue-Venue Similarity

To compute the venue-venue similarity matrix $\mathbf{Z}_{|V|\times|V|}$, we utilize the venue-article matrix $\mathbf{P}_{|V|\times|A|}$ and the author-venue matrix $\mathbf{L}_{|H|\times|V|}$. The observation of similar authors who attended different venues highlights the relationship between different venues because the author usually has an interest in venues that are in his field of specialization. Besides, if two different venues has been cited by the same article, it can help to identify the relationship between these two venues. The idea is that a given article generally cited articles that is related to its research scope, thus the venues that published the cited articles are within the same research scope. In order to determine the similarity between the two venues, we consider the cosine-based similarity approach. The cosine-based similarity takes two vectors of shared authors and articles of venues and computes their similarity according to their angle, as in Equation *(3-9*).

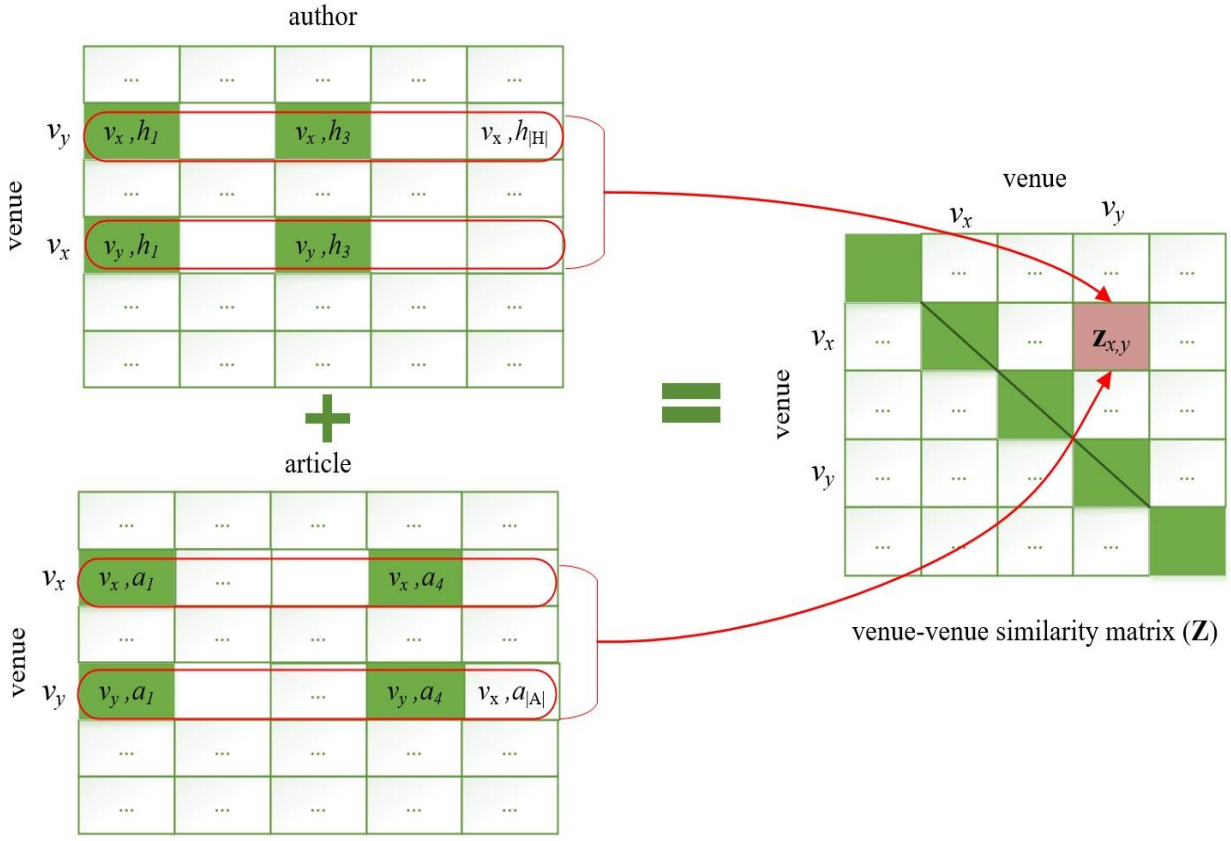

Figure 3-6: An illustration of the article-article similarity matrix **Z**.

$$z_{(v_x,v_y)} = \cos(v_{x,h}, v_{y,h}) + \cos(v_{x,a}, v_{y,a}) = \frac{v_{x,h} \cdot v_{y,h}}{\|v_{x,h}\| \cdot \|v_{y,h}\|} + \frac{v_{x,a} \cdot v_{y,a}}{\|v_{x,a}\| \cdot \|v_{y,a}\|} \quad (3\text{-}9)$$

Figure 3-6 shows an example of constructing the similarity matrix **Z**. The $\mathbf{Z}_{|V|\times|V|}$ matrix represents the similarity between venues where both rows and columns represent venues. As mentioned previously, we consider the top k nearest venues for each venue to build the matrix of the most similar venues. Hence, if the corresponding similarity value between a pair of venues x and y is greater than the k highest similarity value, we set $z_{x,y}$ to the similarity value otherwise, we set $z_{x,y}$ value to zero. The non-zero entries will be consider to build $\mathbf{G}^{k}$ matrix that contains the k most similar venues.

3.5. Latent Preferences Identifier Model

The idea of the proposed Latent Preferences Identifier Model (LPIM) is to reveal the concealed connection between the academic objects of the research industry (articles, authors, and venues). For instance, authors who attended certain venues are more likely to collaborate in the production of joint scientific articles. This possibility increases if there is a common collaborator among them (joint author). In order to identify the hidden preferences for each object, we analyze the bibliography information associated with the citation network of h, a, v in the dataset. In order to do that, we construct three latent models that represent the hidden preferences of the authors, the articles, and the venues.

3.5.1. Identify Latent Preference for Author

Determine the authors' latent preferences toward their selection of articles and venues facilitate the process of recommending an academic object (author, article, and venue). In order to do that, we analyze the hidden preferences of the authors' selection toward articles and venues. We utilize the matrix $\mathbf{C}$ and the matrix $\mathbf{G}$ to capture the authors' latent preferences toward articles. We construct a new author-article latent preference matrix $\overrightarrow{\mathbf{HA}}$, which represents the product of matrix $\mathbf{C}$ and $\mathbf{G}$, as in Equation (3-10):

$$\overrightarrow{\mathbf{HA}} = \tilde{\mathbf{C}}\,\mathbf{G}^{\boldsymbol{k}} \tag{3-10}$$

Where the matrix $\tilde{\mathbf{C}}$ denotes a normalized version of the matrix $\mathbf{C}$, and the matrix $\mathbf{G}^{\boldsymbol{k}}$denotes the top k nearest articles. We consider the top k nearest neighbors for each article in order to minimize the computational cost. The normalized author-article matrix $\tilde{\mathbf{C}}$ can be defined as $\tilde{\mathbf{C}} = \left[\tilde{c}_{h,a}\right]_{|H|\times|A|}$, where $\tilde{c}_{h,a}$ is obtained as follows:

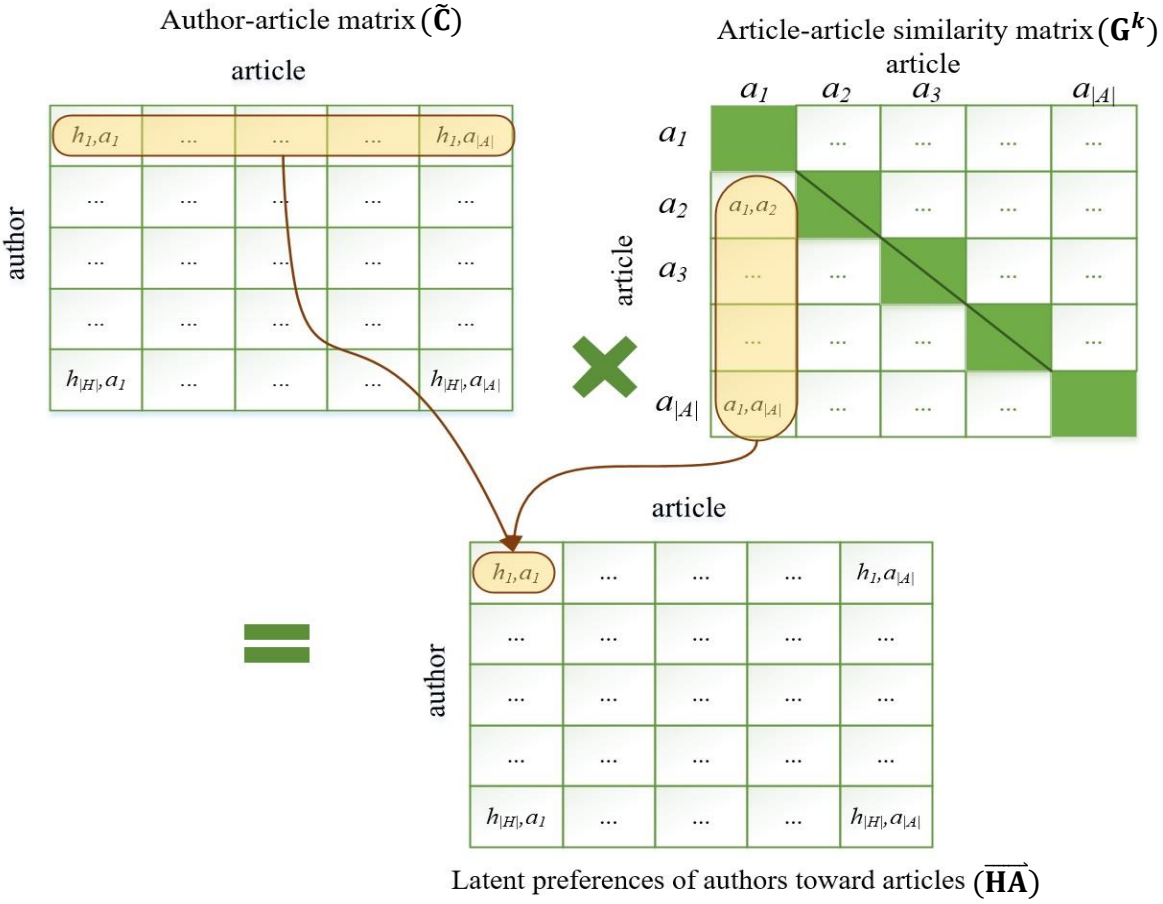

Figure 3-7: An illustration of the process of computing the author-article matrix $\overrightarrow{\boldsymbol{HA}}$

$$\tilde{c}_{h,a} = c_{h,a} / \sqrt{\sum_{i}^{|H|} \left(c_{i,a}\right)^2} \tag{3-11}$$

In the matrix $\overrightarrow{\mathbf{HA}}$, each entry is represented by the multiplication of the h-th row by the a-th column entails to identify the latent preferences of authors h, on article a with respect to the articles k nearest neighbor. Figure 3-7 shows the details of the construction of the new matrix $\overrightarrow{\mathbf{HA}}$.

3.5.1.1. Example: building a latent preference model for an author

To illustrate a simple example of building the latent tag preference model, let us consider there are six different authors citing seven different articles in the same citation network. We use

article A_1 as reference examples thru this section. When we aggregate the author's association over the article, we can obtain the author-article matrix shown in Table 3-3.

From the example matrix C, we can easily derive the normalized matrix $\tilde{C}$. For example, the normalized value of A_1 for Abed is calculated as $\tilde{c}_{Abed,A_1} = 3/\sqrt{3^2+2^2+3^2+1^2+1^2+1^2} = 0.12$, whereas the value of article A_5 is $\tilde{c}_{Abed,A_5} = 3/\sqrt{3^2+1^2+3^2+1^2} = 0.15$. Even though the number of citation by Abed for article A_1 equals the citation number for article A_5, the article A_5, retains more influence than the former article does with regard to his preferences.

Table 3-3: An example of an author-article matrix, C

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
Abed	3	0	2	0	3	1	0
Moh	2	1	0	0	1	2	2
Basem	3	1	2	3	0	1	3
Jone	1	2	3	2	3	1	2
Nam	1	0	2	3	1	2	3
Sam	1	2	3	2	0	1	2

To calculate Abed's latent preference toward the article A_2, when we determine the article-article similarity between the article A_2 and every other article in the citation network. We consider the top five most similar articles to the article A_2. We determine the articles and similarities shown in Table 3-4. From the normalized and similarity values, Abed's latent preference value toward the

article A_2, $\overrightarrow{\mathbf{HA}}_{Abed,A_2}$, is calculated as follows: $\overrightarrow{\mathbf{HA}}_{Abed,A_2} = (0 * 1) + (0.07 * 0.88) + (0.15 * 0.67) + (0.08 * 0.65) + (0.12 * 0.45) = 0.27$.

Table 3-4: Article similarities between the article A_2 and its five most similar articles

	$\boldsymbol{A_2}$	$\boldsymbol{A_3}$	$\boldsymbol{A_5}$	$\boldsymbol{A_6}$	$\boldsymbol{A_1}$
A_2	1	0.88	0.67	0.65	0.45

Although Abed has not previously cited the article A_2, yet the latent value for the association between Abed and the article A_2 can be expected as 0.27. For the reason that, Abed cite articles that are similar to article A_2 and there are in the same citation network. Using the same calculation concept, we will calculate the latent preference in the following sections.

We utilize the matrix **L** and the matrix **Z** to capture the authors' latent preferences toward venues. We constructed the new author-venues latent preference matrix $\overrightarrow{\mathbf{HV}}$. The matrix represents the product of matrix **L** and **Z**, as in Equation (3-12):

$$\overrightarrow{\mathbf{HV}} = \tilde{\mathbf{L}}\,\mathbf{Z}^{\boldsymbol{k}} \tag{3-12}$$

Where the matrix $\tilde{\mathbf{L}}$ denotes a normalized version of the matrix **L**, and the matrix $\mathbf{Z}^{\boldsymbol{k}}$ denotes the top k nearest venues as explained formerly. The multiplication of the h-th row by the v-th column entails to identify the latent preferences of authors h, on venues v with respect to the venues k nearest neighbor. Figure 3-8 shows the details of the construction of the new matrix $\overrightarrow{\mathbf{HV}}$.

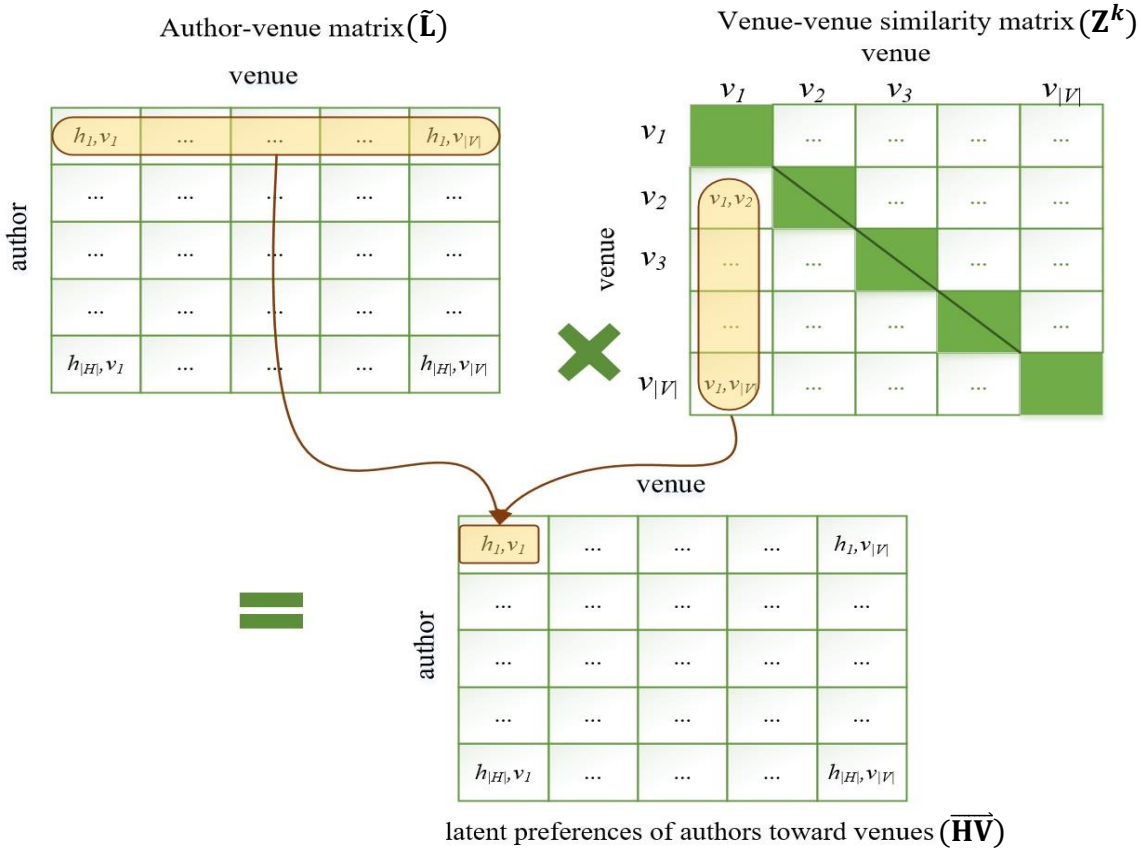

Figure 3-8:An illustration of the process of computing the author-venue matrix $\overrightarrow{\boldsymbol{HV}}$

3.5.2. Identify Latent Preference for Article

In order to facilitate the process of recommending an academic object, we recognize the articles' latent preferences toward their selection of authors and venues. We analyze the hidden preferences of the articles' choice toward authors and venues. We utilize the matrix $\mathbf{C}$ and the matrix $\mathbf{R}$ to capture the articles latent preferences toward authors. We constructed the new author-article latent preference matrix $\overrightarrow{\mathbf{AH}}$. The matrix represents the product of matrix $\mathbf{C}$ and $\mathbf{R}$, as in Equation *(3-13)*:

$$\overrightarrow{\mathbf{AH}} = \tilde{\mathbf{C}}\,\mathbf{R}^{k} \tag{3-13}$$

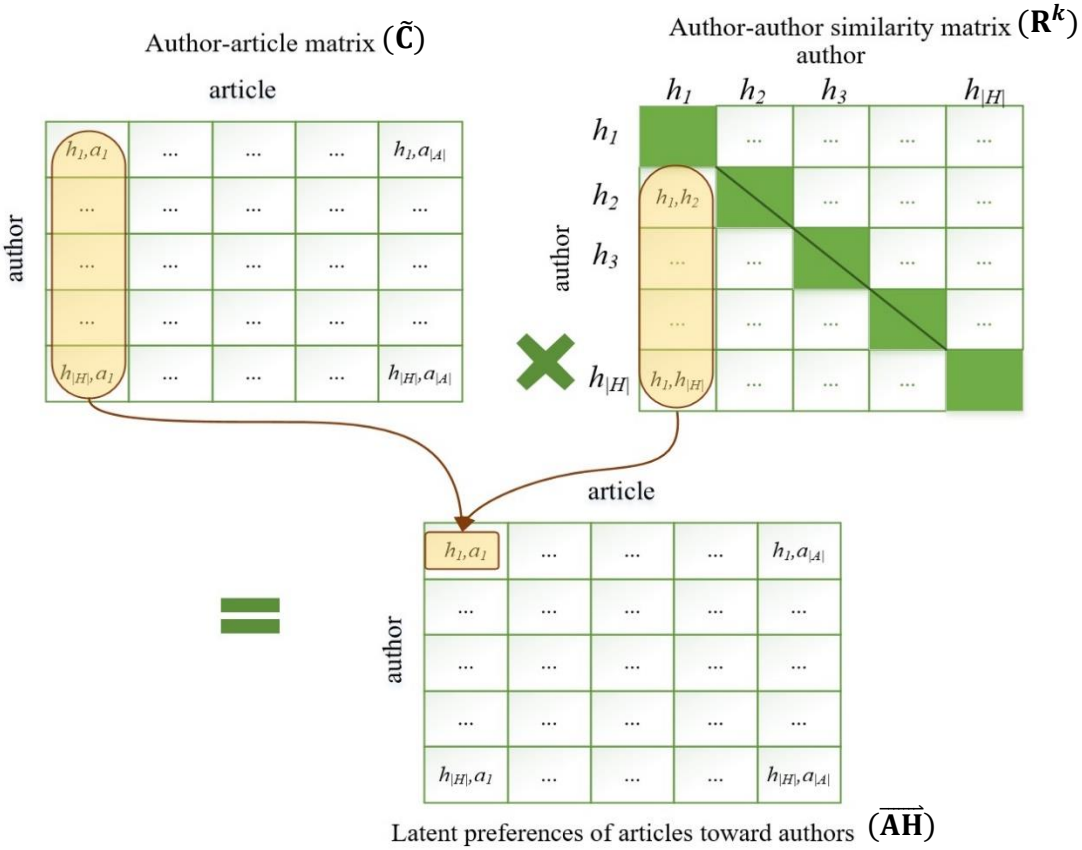

Figure 3-9: An illustration of the process of computing the article-author matrix $\overrightarrow{\boldsymbol{AH}}$

Where the matrix $\tilde{\mathbf{C}}$ denotes a normalized version of the matrix $\mathbf{C}$, and the matrix $\mathbf{R}^{k}$denotes the top k nearest authors as explained formerly. The multiplication of the a-th row by the h-th column entails to identify the latent preferences of articles a, on authors h with respect to the authors k nearest neighbor. *Figure 3-9* shows the details of the construction of the new matrix $\overrightarrow{\mathbf{AH}}$.

We utilize the matrix **P** and the matrix **Z** to capture the articles latent preferences toward venues. We constructed the new article-venue latent preference matrix $\overrightarrow{\mathbf{AV}}$. The matrix represents the product of matrix **P** and **Z**, as in Equation*(3-14)*:

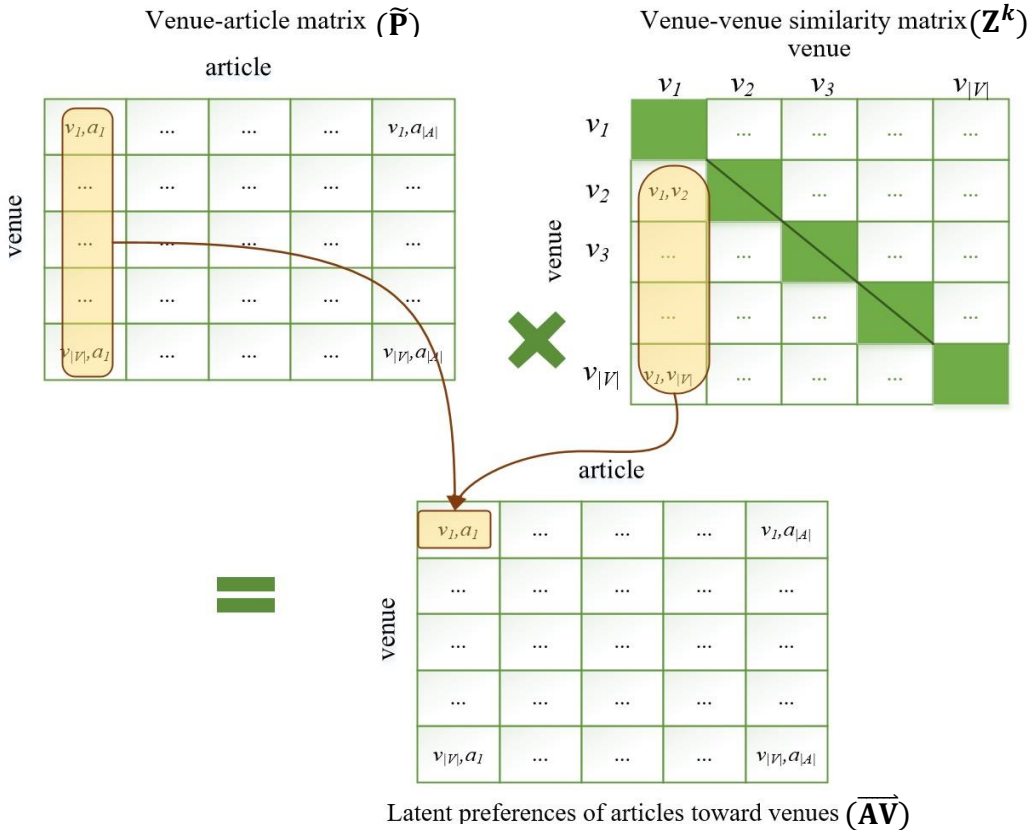

Figure 3-10: An illustration of the process of computing the article-venue matrix $\overrightarrow{\boldsymbol{AV}}$

$$\overrightarrow{\mathbf{AV}} = \widetilde{\mathbf{P}}\,\mathbf{Z}^{k} \tag{3-14}$$

Where the matrix $\widetilde{\mathbf{P}}$ denotes a normalized version of the matrix $\mathbf{P}$, and the matrix $\mathbf{Z}^{k}$ denotes the top k nearest venues as explained formerly. The multiplication of the a-th row by the v-th column entails to identify the latent preferences of articles a, on venues v with respect to the venues k nearest neighbor. Figure 3-10 shows the details of the construction of the new matrix $\overrightarrow{\mathbf{AV}}$.

3.5.3. Identify Latent Preference for Venue

We analyze the hidden associations of venues' to identify the venues' latent preferences toward their articles and authors. In order to do that; we utilize the matrix $\mathbf{P}$ and the matrix $\mathbf{G}$ to

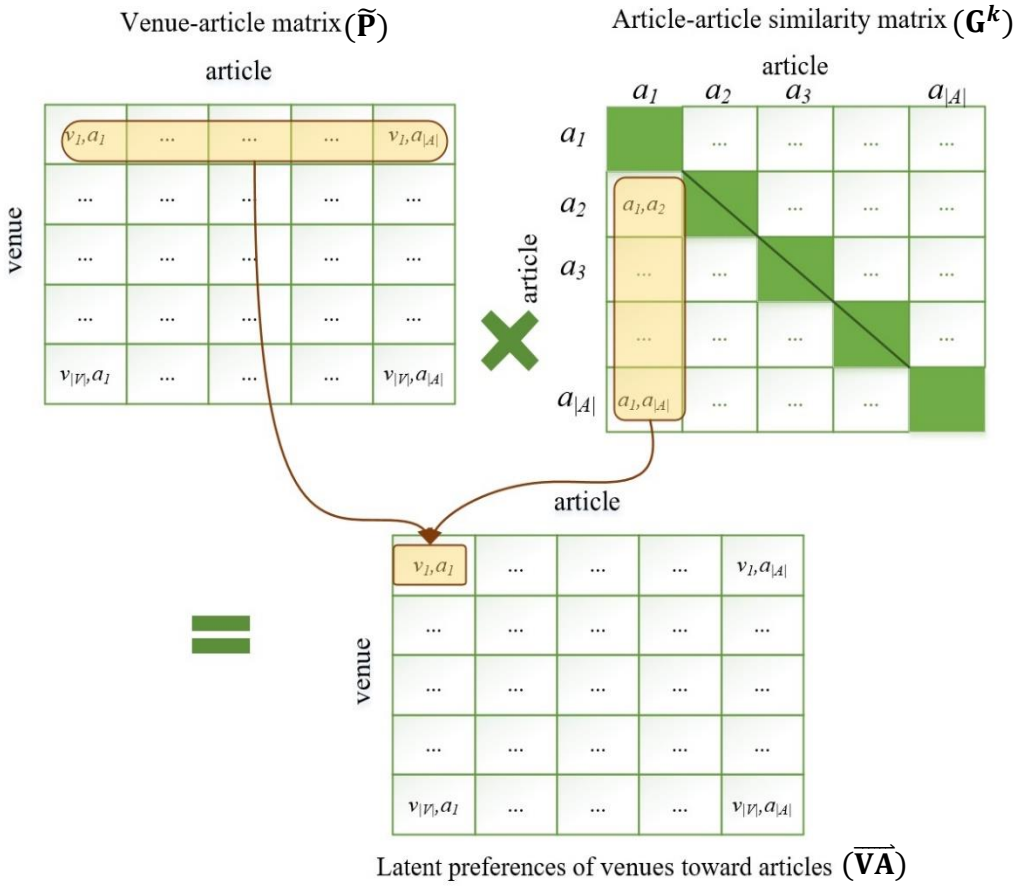

Figure 3-11: An illustration of the process of computing the venue-article matrix $\overrightarrow{\boldsymbol{VA}}$.

capture the venues' latent preferences toward articles. We constructed the new venue-article latent preference matrix $\overrightarrow{\mathbf{VA}}$. The matrix represents the product of matrix $\mathbf{P}$ and $\mathbf{G}$, as in Equation *(3-15)*:

$$\overrightarrow{\mathbf{VA}} = \widetilde{\mathbf{P}}\,\mathbf{G}^{k} \tag{3-15}$$

Where the matrix $\widetilde{\mathbf{P}}$ denotes a normalized version of the matrix $\mathbf{P}$, and the matrix $\mathbf{G}^{k}$denotes the top k nearest articles as explained formerly. The multiplication of the v-th row by the a-th column entails to identify the latent preferences of venues v, on articles a with respect to the articles k nearest neighbor. Figure 3-11 shows the details of the construction of the new matrix $\overrightarrow{\mathbf{VA}}$.

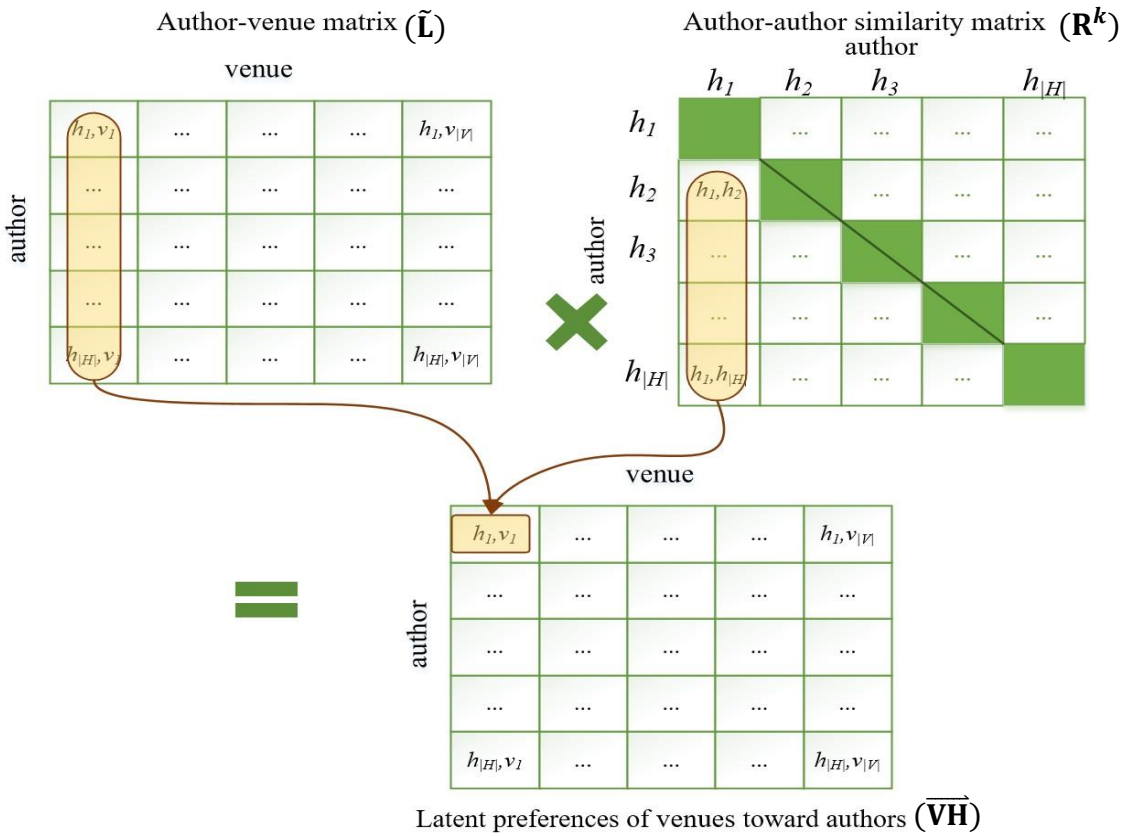

Figure 3-12: An illustration of the process of computing the venue-author matrix $\overrightarrow{\boldsymbol{VH}}$

We utilize the matrix **L** and the matrix **R** to capture the venues' latent preferences toward authors. We constructed the new venue-author latent preference matrix $\overrightarrow{\mathbf{VH}}$. The matrix represents the product of matrix **L** and **R**, as in Equation *(3-16)*:

$$\overrightarrow{\mathbf{VH}} = \tilde{\mathbf{L}}\,\mathbf{R}^k \tag{3-16}$$

Where the matrix $\tilde{\mathbf{L}}$ denotes a normalized version of the matrix **L**, and the matrix $\mathbf{R}^k$ denotes the top k nearest authors as explained formerly. To minimize the computational cost, we consider the top k nearest neighbors for each author. The multiplication of the v-th row by the h-th column entails to identify the latent preferences of venues v, on authors h with respect to the authors k nearest neighbor. Figure 3-12 shows the details of the construction of the new matrix $\overrightarrow{\mathbf{VH}}$.

3.6. Academic Collaborative Ranking

The Academic Collaborative Ranking uses the previously formed latent preferences matrices to reveal the degree of trust for collaboration between academic objects in the citation network. This proposed method produces objects (author, article, venues) recommendations relevant to a given query. For example, if an author wants a list of related articles that can be used for reference, the algorithm will compute the confidence score based on the user's requested list and produce a candidate recommendation list containing articles.

3.6.1. Author Collaborative Confidence

According to our proposed model, we use the four latent models $\overrightarrow{\mathbf{HV}}$, $\overrightarrow{\mathbf{VA}}$, $\overrightarrow{\mathbf{HA}}$and $\overrightarrow{\mathbf{AV}}$ to build our personalized algorithm for rating authors according to a given query. We compute the Collaborative Confidence Score (CCS) to propose authors candidate list. The model will compute the CCS based on the user requested list for each author based on its influence on the articles and venues using Equation *(3-17*).

$$CCS\,(h|q) = \frac{\sum_{v=1}^{|V|} hv_{h,v} \cdot \mu_v}{|V|} + \frac{\sum_{a=1}^{|A|} ha_{h,a} \cdot \mu_a}{|A|} \tag{3-17}$$

where $hv_{h,v}$ denotes the latent preferences value of author h toward a specific venue v in the matrix $\overrightarrow{\mathbf{HV}}$ and $ha_{h,a}$ denotes the latent preferences value of author h toward a specific article a in the matrix $\overrightarrow{\mathbf{HA}}$. μ_v denotes the mean of latent preferences value of a specific venue v toward the articles in the citation network where $\mu_v = \frac{\sum_{a=1}^{|A|} va_a}{|A|}$. μ_a denotes the mean of latent preferences value of a specific article a toward the venues in the citation network where $\mu_a = \frac{\sum_{v=1}^{|V|} av_v}{|V|}$. The main idea here is to consider the value of the other dimension in the citation network

and its effect to close the relation loop. $|V|$ represents the total number of venues, $|A|$ represents the total number of articles, and $|H|$ represents the total number of authors. As we mentioned earlier, a latent preference value reflects the authors' choices towards venues and articles. Thus, it can give the influence of a potential expansion of options according to particular user's interests and the relation among the citation network between the academic objects. Accordingly, by utilizing the latent models $\overrightarrow{\mathbf{HV}}$, $\overrightarrow{\mathbf{VA}}$, $\overrightarrow{\mathbf{HA}}$ and $\overrightarrow{\mathbf{AV}}$,that represents the relation between the academic object at the citation network, authors who are likely to collaborate in the future with the user are ranked higher in the potential search list, considering the relation among the academic objects, regardless of whether there is prior collaboration.

3.6.2. Article Collaborative Confidence

To build a personalized algorithm for ranking articles according to a given query; we use the four latent models $\overrightarrow{\mathbf{AH}}$, $\overrightarrow{\mathbf{HV}}$, $\overrightarrow{\mathbf{AV}}$ and $\overrightarrow{\mathbf{VH}}$. We compute the CCS for each article in the articles candidate list based on its influence on the authors and venues using Equation (*3-18*).

$$CCS\,(a|q) = \frac{\sum_{h=1}^{|H|} ah_{a,h} \cdot \mu_h}{|H|} + \frac{\sum_{v=1}^{|V|} av_{a,v} \cdot \mu_v}{|V|} \tag{3-18}$$

where $av_{a,v}$ denotes the latent preferences value of article a toward a specific venue v in the matrix $\overrightarrow{\mathbf{AV}}$ and $ah_{a,h}$ denotes the latent preferences value of article a toward a specific author h in the matrix $\overrightarrow{\mathbf{AH}}$. μ_v denotes the mean of latent preferences value of a specific venue v toward the authors in the citation network where $\mu_v = \frac{\sum_{h=1}^{|H|} vh_h}{|H|}$. μ_h denotes the mean of latent preferences value of a specific author h toward the venues in the citation network where $\mu_h = \frac{\sum_{v=1}^{|V|} hv_v}{|V|}$. As we mentioned earlier, a latent preference value reflects the articles selection towards venues and

authors. Thus, it reflect the impact of articles selection to of a potential expansion according to article's production. Accordingly, by utilizing latent models, $\overrightarrow{\mathbf{AH}}$, $\overrightarrow{\mathbf{HV}}$, $\overrightarrow{\mathbf{AV}}$ and $\overrightarrow{\mathbf{VH}}$, articles that are most fitting to be cited in the given article are ranked higher in the potential search list considering the other dimension impact on the relation loop.

3.6.3. Venue Collaborative Confidence

We use the four latent models $\overrightarrow{\mathbf{VH}}$, $\overrightarrow{\mathbf{HA}}$, $\overrightarrow{\mathbf{VA}}$ and $\overrightarrow{\mathbf{AH}}$ to build a tailored algorithm for rating venues according to a given query. The CCS for each venue will be calculated to propose venues candidate list. The model will compute the CCS each venues based on its relation to the authors and venues using Equation (*3-19*).

$$CCS\,(v|q) = \frac{\sum_{h=1}^{|H|} vh_{v,h} \cdot \mu_h}{|H|} + \frac{\sum_{a=1}^{|A|} va_{v,a} \cdot \mu_a}{|A|} \tag{3-19}$$

where $vh_{v,h}$ denotes the latent preferences value of venue v toward a specific author h in the matrix $\overrightarrow{\mathbf{VH}}$ and $va_{v,a}$ denotes the latent preferences value of venue v toward a specific article a in the matrix $\overrightarrow{\mathbf{VA}}$. μ_h denotes the mean of latent preferences value of a specific author h toward the articles in the citation network where $\mu_h = \frac{\sum_{a=1}^{|A|} ha_a}{|A|}$. μ_a denotes the mean of latent preferences value of a specific article a toward the authors in the citation network where $\mu_a = \frac{\sum_{h=1}^{|H|} ah_h}{|H|}$. As we mentioned earlier, a latent preference value reflects the venues' effects toward authors and articles. Thus, it can increased the potential expansion of possibilities according to the venue's research scope. By utilizing latent models, $\overrightarrow{\mathbf{VH}}$, $\overrightarrow{\mathbf{HA}}$, $\overrightarrow{\mathbf{VA}}$ and $\overrightarrow{\mathbf{AH}}$, venues that are suitable according to the user queries are ranked higher in the potential search list. We consider the other

dimension impact on the relation loop that reflects the relation between the academic objects at the citation network of the given query upon user request.

3.7. Summary

In this chapter, we have presented a new academic recommendation model in an extended nearby citation network that aims to search and rank academic objects (author, article, venue) relevant to users' inquiry to build an effective scientific community. The recommendation model combines the bibliometric data and the content semantic similarity to evaluate the academic impact among the extended network using the content-based filtering, and global relevance approach. Then, it computes the cosine similarity between the authors, articles, and venues using the collaborative filtering approach to estimate the connection between the objects. After that, it builds the latent preferences models for each academic object to be used in the computation of the ranking score. The academic collaborative ranking algorithm uses the latent preferences models to estimate the collaborative confidence score for each object, depending on a user's query, to determine the most suitable venues, potential authors to team up with, and most relevant articles to cite that are personalized to the user's requests.

Chapter 4 . Design and Implementation of an Academic Recommender System

This chapter presents the design and architecture of the academic recommendation system and the design and implementation of the prototype applications. We present the proof-of-concept of our proposed algorithms using a web-based application developed using Cloud-Oriented Architecture to enhance the system usability and accessibility.

4.1. The System Overview

Exploring the scientific communities, research collaborations, and determining the direction of research for recommendations still needs further studies [75], [76]. Mainly, it might be due to the semantic gap and the data inconsistencies in the citation systems [16], [18]. While the value of research-oriented systems is well recognized, finding ways to help authors to discover new research trends or collaborate with others is still a challenge.

The proposed system aims to discover and analyze the scientific relationships among virtual communities to help the users to establish future collaborations with new authors or to find venues to submit their research work. In addition, it helps them to find their strength research area

to promote their works for a research grant. Furthermore, it helps new researchers to reach the key scholars who are closer to their research area quickly and efficiently.

The primary objective of our approach is to enable authors to know their current network of research more efficiently and the prospective venues (whether they are journals or conferences) that best suit their work. We build a visualization interface that provides an in-depth analysis of the article metadata. This interactive interface is designed to create an informative environment that helps the users to explore the hidden relationships among virtual scientific communities.

The system provides the abilities to search for articles using an author's network, or a publishing venue network. It supports different search strategies and scenarios:

- ***Direct Search:*** this service supports the search using keywords or metadata such as title, author name, and bibliographic metadata fields.

- ***Scientific Community Recommendation:*** this service uses the semantic similarity and the common relationship between authors, articles, and venues building a citation network that surrounds articles. The recommendation includes most related articles to cite, most potential authors to collaborate with, and most suitable venues to publish a given research work.

- ***Scientific Community Visualization:*** this service uses learning techniques with visual charts to generate reliable visual information and reveal hidden knowledge through learning from existing relationships in a virtual scientific community.

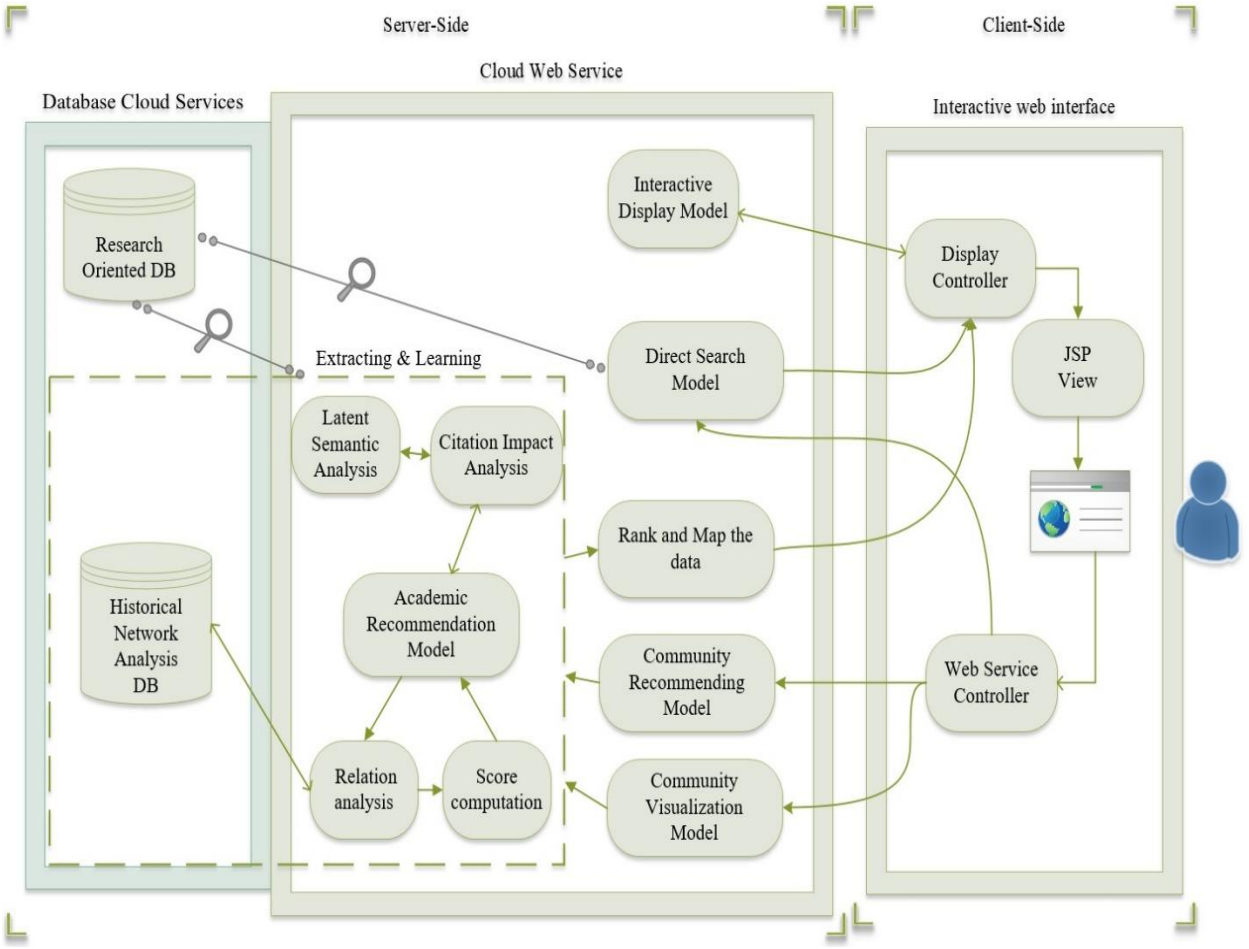

Figure 4-1: The system framework for the high-level architecture.

4.2. System Architecture

The system relies on semantic analysis and the similarity analysis between academic objects in the citation network to achieve the proposed services. The semantic analysis model deals with topical relatedness between articles, authors, and venues. We recognize the collaborative relationships among the communities by combining topical semantic similarities, author similarities, and venue similarities. The system is developed using Cloud-Oriented Architecture (COA) to enhance system usability and accessibility [77]. Figure 4-1 shows the high-level

architecture of the system framework and its components. It also explains the internal interactions between the components to handle user requests. The architecture is comprised of two components: the client side, and the server side. The client side is composed of an interactive web interface components that provide the means to request inquiries and display the retrieved information to the user, in a meaningful way according to its requirements. The server performs a deep analysis of the collected data and provides storage and distribution capabilities to provide the information to the recommending process.

The framework integrates algorithms and processing tools to obtain useful interactive inferences from the collected bibliography data. Furthermore, this interactive interface is designed to create an informative environment that can benefit the users to reveal the concealed connections among virtual scientific communities.

4.2.1. System Module

4.2.1.1. Direct Search

This function provides a search service based on a given keyword(s). If a researcher is looking for more information about a specific academic object (article, author, venue), the system can help her/him to provide more information for that specific segment with the given keywords. The given keywords can be an article title, an author name or, a venue name. For example, if the researcher plan to write a paper about building a smart city application, they will need to search using a few keywords related to that topic. The system would be able to return relevant articles using the provided keywords. The resulting page includes the article title(s), the author name(s), the article identification number(s), and their bibliometric data. The system enables the user to

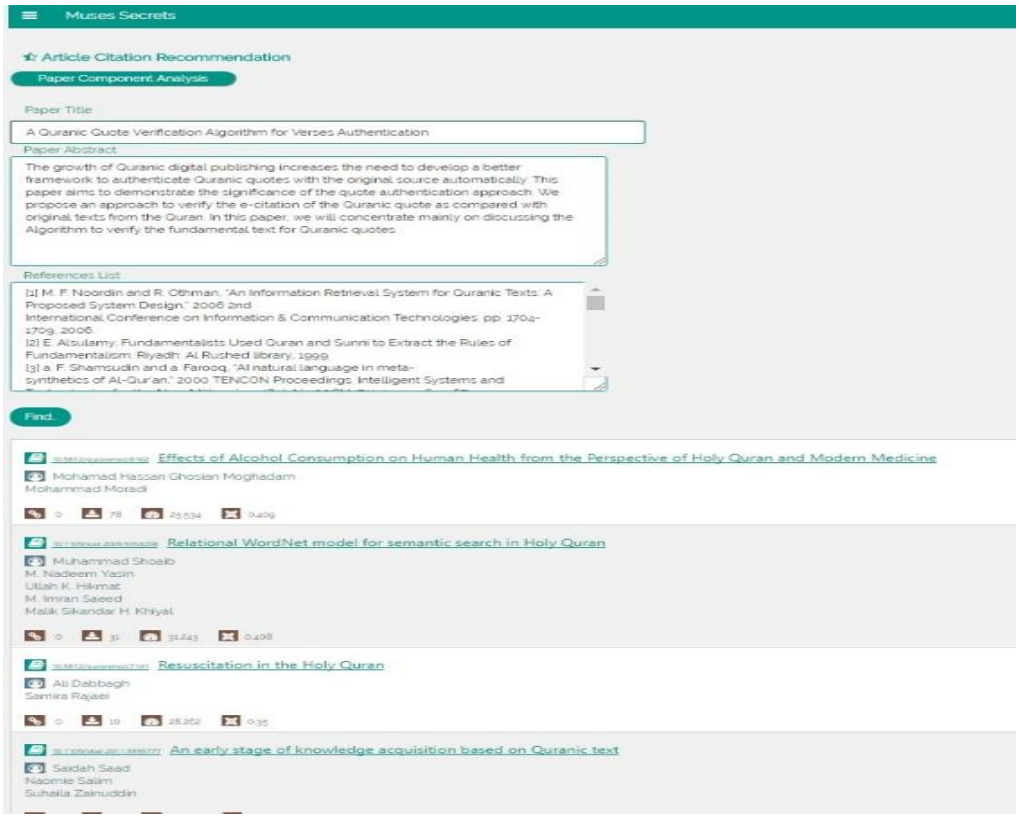

Figure 4-2: Community recommendation function GUI (Article Citation)

explore further information related to the research categories, authors commonly publishing in that area and venues related to the research scope.

4.2.1.2. Scientific Community Recommendation

This function provides a recommendation service based on given article information to help a user to find the most relevant articles to cite, most appropriate researchers to collaborate, or most suitable venues for article submission. For example, if a researcher starts working on a new article, the system will be able to assist her to look for further articles to cite. It gives her the option to enter the title, the abstract and the references list of the proposed article as depicted in Figure 4-2. The system will process the article information to identify the article topic, and the citation network based on the author's research behavior. The identified information helps the system to

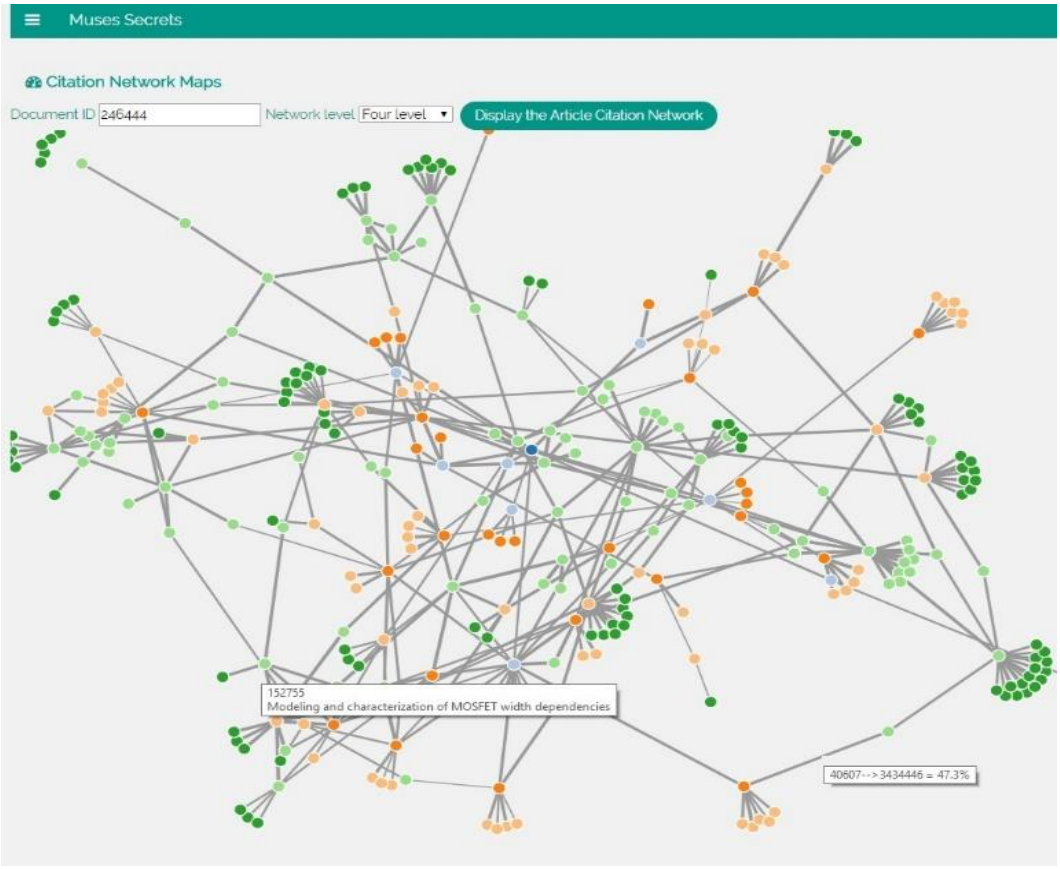

Figure 4-3: Visualization function to map the articles in citation network.

suggest a list of the most related articles tailored to the proposed article and the author's research preferences. The presented information includes the article title, the author name(s), the article identification number, the bibliometric data, and the relatedness score.

4.2.1.3. Scientific Community Visualization

This function provides a visualization service to visualize the current status based on the given information. The primary objective is to visualize the current situation of different parts of the scientific community such as authors, venues, and articles. This function enables the exploration of different patterns related to the targeted query. For instance, Figure 4-3 illustrates an example of using a visualization function to visualize the current citation relation in a network

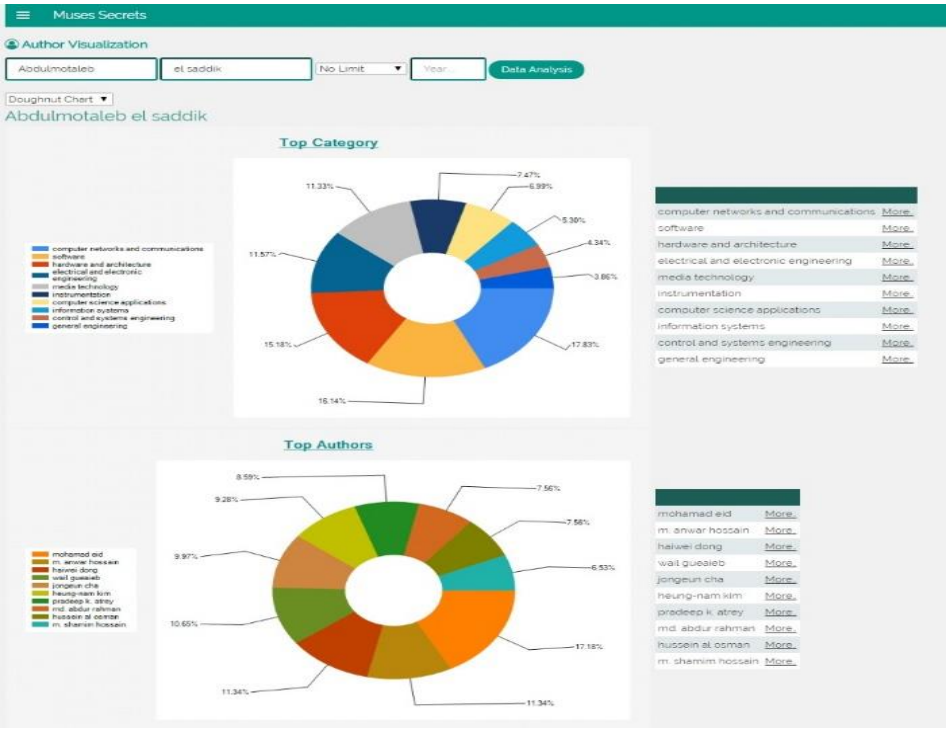

Figure 4-4: Community visualization function GUI (Author)

map. Each circle represents an article node, and each line represents a citation relationship between two articles in the citation network. It shows different group color for the nodes. Each group color represents a relation level in the citation network. For instance, Figure 4-3 visualizes a five levels citation network map for a specific article ***A***. The node with the dark blue color represents that article. The nodes with the light blue color represent the first level of the citation network (the cited articles of the specific article ***A***). The nodes with the orange color represent the second level of the citation network (the references list of the cited articles). The node continues to change as the number of levels in the network map increases. When the pointer moved over the node, some information about the article will appear. The information represents the article title and article id as shown in Figure 4-3.

Figure 4-4 illustrates an example of using a visualization function. The primary objective is to visualize the current situation of a different area of the scientific community. This function enables the exploration of different patterns related to the targeted query. For instance, Figure 4-4 visualizes information about a specific author. The visual content includes a list of the categories in which the work of the author is classified, the list of authors most collaborating with him, and a list of the publishing venues where his works were published. The user can take advantage of the displayed information to explore the scientific classification and the collaboration patterns quickly and easily.

4.2.1.4. Databases

The system uses two databases to retrieve and store relevant data about academic objects. These databases are detailed below:

- **Research Oriented DB**: We use open access data via API interfaces to access the research-oriented data source. We build data crawler tool to retrieve the article's main metadata content (i.e., title, abstract, keywords, and DOI), and the bibliometric information in the metadata for each article upon user request. We have identified one main information source and several alternative data sources to compensate for the lack of information. We use Crossref API [78] to get the Initial data. Crossref is a well-known research-oriented data source that offers richer metadata that enables the proposed engine to extract and run the learning algorithm. The alternative data sources that provide API freely access includes among others IEEE Xplore [79], ACM Digital Library[80], DBLP[81]. The size of the information that can be accessed through the crawling tool exceeds 90 million records between articles, authors and venues

information. Using the data crawling tool in the proposed system enables it to maintain access to the up-to-date information so that the user can get the results based on the updated information.

- **Historical Network Analysis DB**: This database stores the information for each academic object and the proposed system analysis results for each object. The stored information includes among others the proposed index, the recommendation results, the author states results, the venues analysis results. In addition, it stores the information for each user request to be used in the future to improve the system outputs based on the previous research and to compare the previous and current results to reach the optimal result. The information stored in the historical database is updated using two basic rules: as per user request, every four months. As per user request update: after the user submits the request to the system, the system will ensure that the information stored does not exceed more than one month, which is the minimum routine period from the date of the beginning of the conference to the publication of the paper electronically. If more than one month has passed, the user request will be processed using the current information. Every four months update, which is the time needed to analyze data for all records retrieved from the main data source due to the size of the analyzed information.

4.2.2. System Dynamic Behavior

The dynamic behavior is presented in two UML behavior diagrams: state chart diagrams and interaction diagrams. The statechart diagrams figures help to demonstrate how the different entities are interacting by describing their activity behavior. In specific, Figure 4-6 describes the user direct search behavior to query a variety of information available as per user request. The

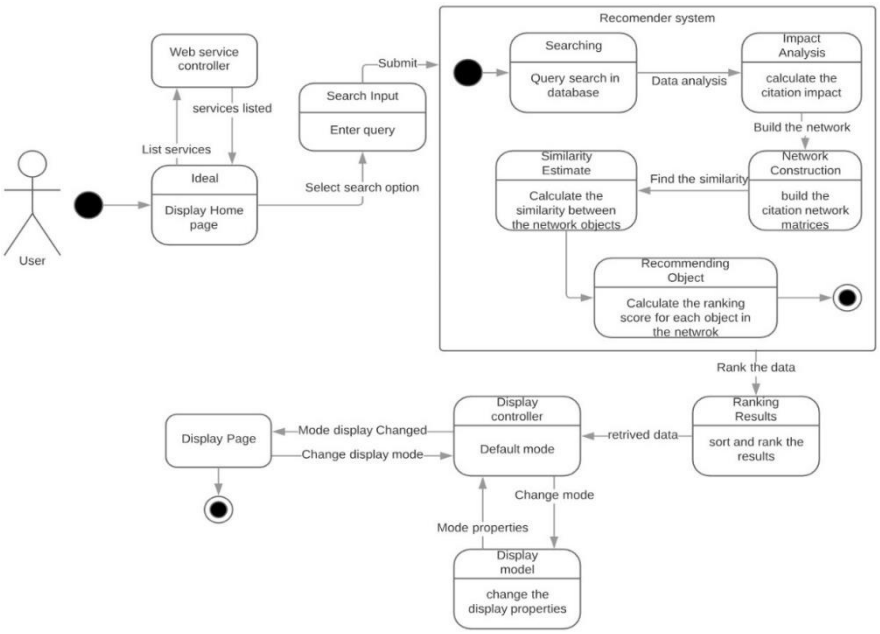

Figure 4-5: Statechart diagram of a user searching a candidate academic objects using recommendation model.

successful user scenario starts by loading the home page with all available services are listed. Then, the user can review the retrieved information using various display modes.

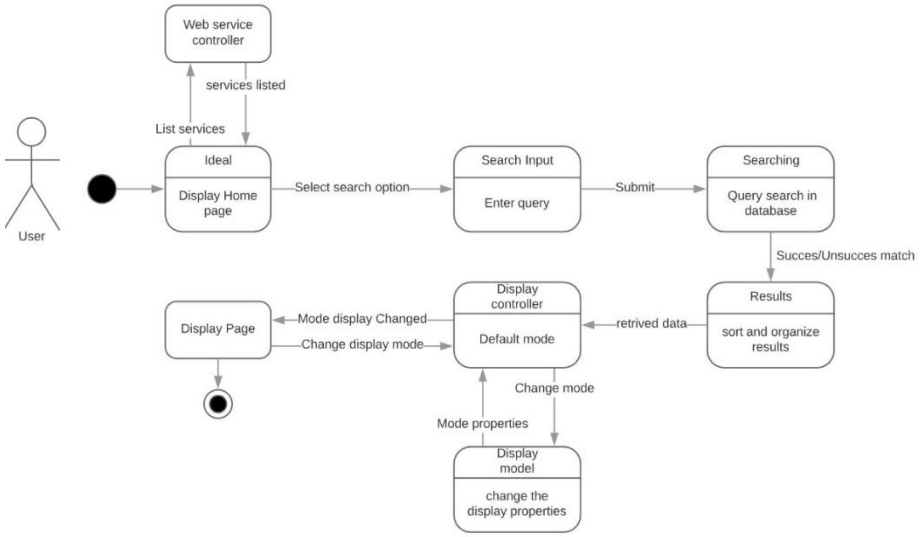

Figure 4-6: Statechart diagram of a user searching and displaying an information.

On the other hand, Figure 4-5 describes different activity behaviors that the system can initiate when the user submits a query request to recommend academic candidate objects that related to the user query. The recommendation system container presents the status that the user requests pass through to recommend a candidate list. The three states - impact analysis, network construction, and similarity estimate - describe the behavior of the system to analyze the citation network of the given article based on the data and relationships retrieved from the database according to the user request. These three states can enable the system to calculate the confidence score to rank the academic objects in the candidate list by sending only valuable information to the ranking results states.

We use the interaction diagrams to highlight the mechanism streams among the system component. Interaction diagrams are a division of the behavior diagrams that illustrate the system's workflow in terms of functions, controls, and data among the structure of the system. The sequence diagram is an interaction diagram that will be used to describe the system's workflow. The main tasks of the server side of the system will be described using two sequence diagrams. Figure 4-7: demonstrates the sequence of procedures which the system attempted when a user search for an article, author, or venues based on an entered search string. In addition, it illustrates the sequence of events when a user wants to change the display mode for the information demonstration. Figure 4-8 demonstrates the sequence of procedures which the system attempted when the user aims to determine the most suitable venues, potential authors to team up with, and most relevant articles to cite according to the user request.

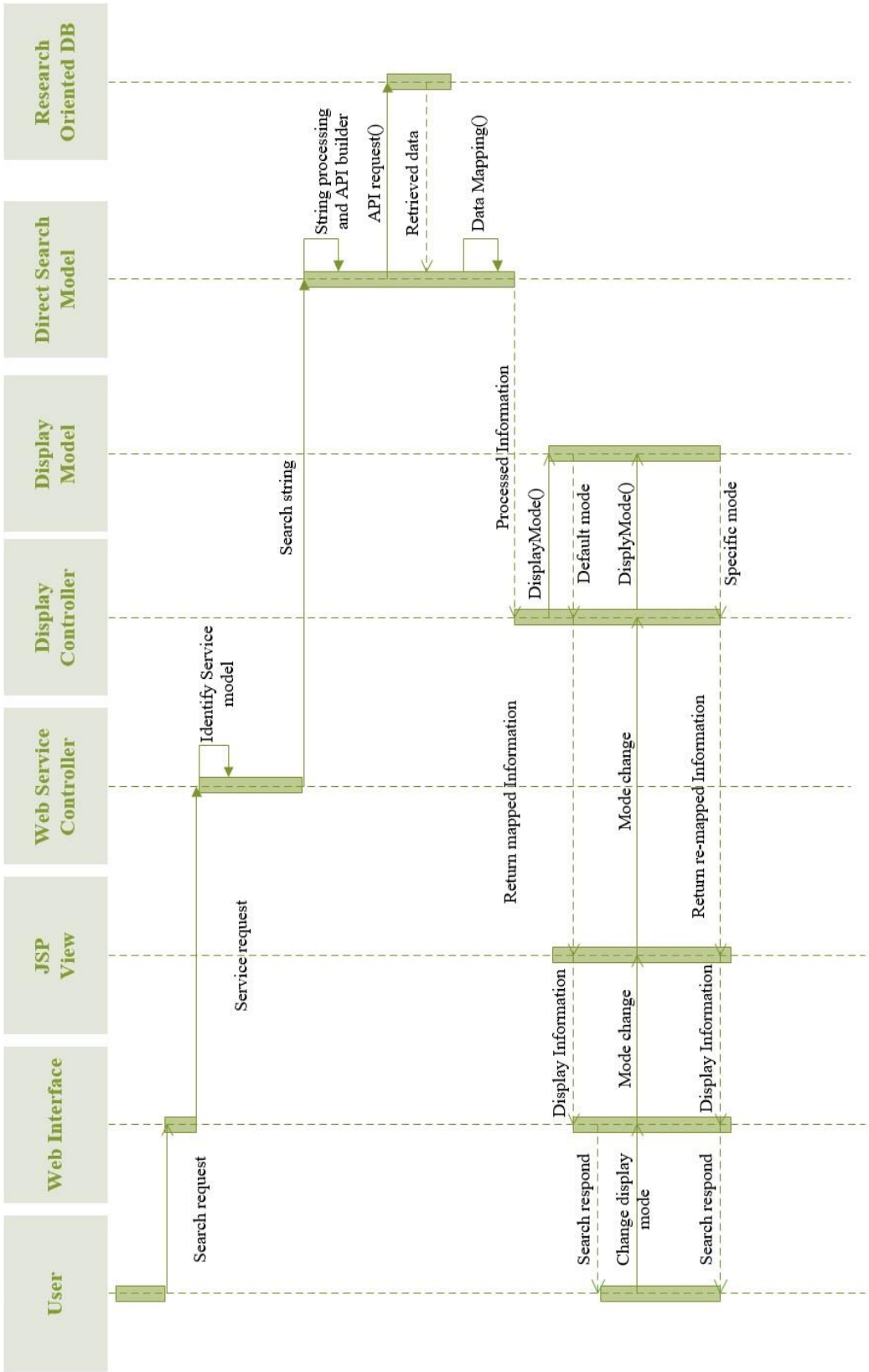

Figure 4-7: Interaction diagram of a user searching and displaying an academic object (article, author, and venue) information.

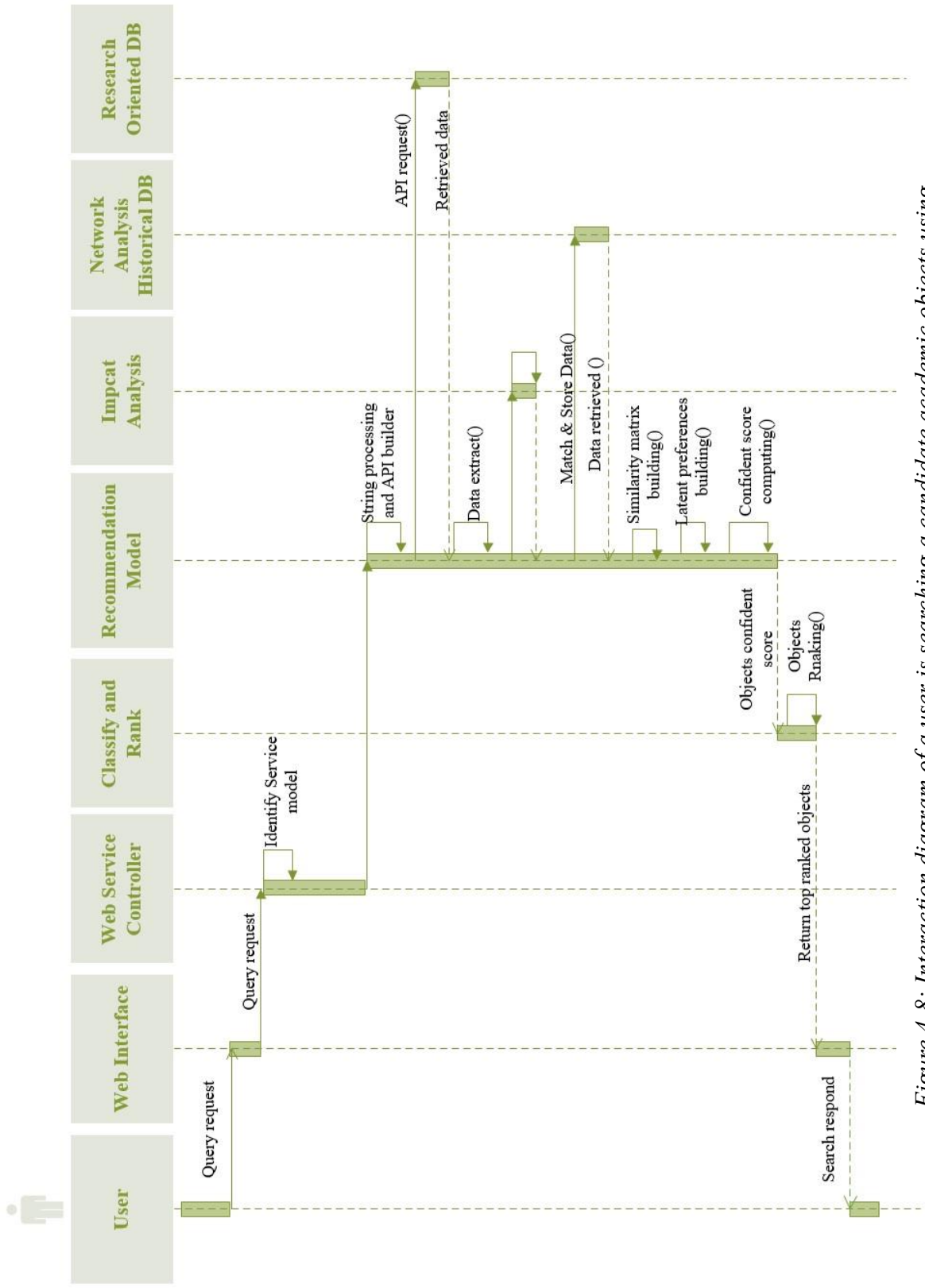

Figure 4-8: Interaction diagram of a user is searching a candidate academic objects using recommendation model.

4.3. System Implementation

We employ the semantic analysis model along with social network analysis to analyze and visualize the extracted metadata from articles, authors, and venues. The aim is to identify and categorize the relationships among a community for recommendation and visualization. We use an open access data via API interfaces to access the research-oriented data source. We consider analyzing the metadata due to the availability of such information on a large scale. The stored resulted relationships include but not limited to author-author, venue-venue, article-article relations as explained earlier.

4.3.1. Client Side

The client-side handles most of the user request and the data visualization associated with the system. It is considered as a mean of communication between the users and the system. The system was developed based on Model View Controller approach in order to handle the user request and to process the retrieved information in a more meaningful manner.

The client side includes controllers to handle user requests and display the results. The web service controller will handle the user request by submitting the query to the right service model according to the user request. The display controller will handle the way the information stored and displayed on the user side. The JSP view will display the information based on the received data and the displayed properties from the display controller.

4.3.2. Server Side

The server-side handles logical entities, service workflow, communications functions and data mapping associated with the application as previously described in Chapter 3 . It represents the core layer of the system as it contains the main components to extract, process, learn, and

analysis the information according to the user's requests. Primary, it extracts the bibliography information and the citation network data from Crossref and builds the citation network matrices considering the academic citation impact among the extended nearby citation network as described in section 3.2. Afterward, it computes the cosine similarity and the latent preferences matrices for the authors, articles, and venues. Later on, the CCS ranking scores will be calculated in order to identify the candidate list for each academic object (authors, articles, and venues) according to the user request and objects preferences.

4.4. Summary

In this chapter, we have presented the design of our proposed framework for the system. The system uses an interactive interface that could simplify the academic exploration toward effective scientific communities collaboration. We propose developing a cloud-oriented architecture that provides access to different types of bibliography information, and citation network analysis that allows users to access them effortlessly. This application applied our ACI algorithm to compute the academic impact among the extended nearby citation network; also the application applied the academic collaborative ranking algorithm to compute the ranking score based on the collaborative confidence score.

Chapter 5 . Evaluation and Outcomes

In order to evaluate the proposed recommendation model, we aim in this chapter to presents the evaluation methodology to measure the accuracy of our proposed algorithms described previously in Chapter 3 . The research question that we aim to consider to evaluate the system are, first, can the use of the latent preferences helps to enhance the recommendation outcomes; second, how the model performs compared to baseline methods. We also specify the technique used to analyze the proposed recommendation algorithms, and the procedure used to gather the related datasets. We also indicate the evaluation measurements used to analyze the proposed algorithms.

5.1. Evaluation Methodology

Most studies in the area of recommender system rely on different evaluation methods to determine the effectiveness of the recommendation approaches. Indicating the appropriate evaluation method is one of the basic requirements for a more effective assessment of the used approach. Among other requirements, the assessment of the recommendation systems can benefit from a decent number of participants in a user study and compare the proposed approach against one or more state-of-the-art approaches.

5.1.1. Accuracy verification

To measure the accuracy of a recommender system; typically the offline evaluations will be used based on ground truth. Offline experiment considers the most common evaluation methods to assess the recommender system. We use offline experiments on different datasets crawled from online academic databases to measure the accuracy of the academic recommendation model. Different metrics can be used to express how many items of the ground truth are recommended within the top n recommendations list. Common evaluation metrics include precision, recall, and F-measure.

In order to evaluate the recommendation accuracy, we will consider the evaluation procedure described in [58], [82]. We randomly divided the experiment datasets into two groups: a training set that represents 80% of the original dataset, and the remaining 20% used as a test set. In order to build the test set, we randomly withheld objects to use them later as test-queries for each request. For instance, in the article recommendation, the original reference list, which was not present while training, will be used as the ground truth.

In the field of scientific research, the reference list of an article is carefully chosen by the authors based on the close relationship between the produced article and the referenced articles. We aim to compare the system prediction to the author's actual selections for the articles in the reference list for each tested object. We assume that any paper other than the actual author's selections is not relevant. To ensure that our results were not sensitive to a particular test query, we conducted five different runs using different training-test portioning datasets to avoid the possibility that the test set is not biased.

5.1.2. Dataset

One of a critical challenge, in order to build an academic recommender system, is to find a publicly available scholar data source that is less sparse. Such lack of availability challenges the design of any academic recommendation algorithm. In addition, due to the Intellectual property and websites blocking, we face a challenge to get access to some of the big research-oriented data source such as Google Scholar. As a solution, our dataset is crawled from open access data via API interfaces to access different research-oriented data sources. We build data crawler tool to retrieve the article's main metadata content and the bibliometric information in the metadata for each article upon user request. The metadata includes among others title, abstract, keywords, DOI, authors' names and affiliations, and venues information. We have identified one main information source and several alternative data sources to compensate for the lack of information. For instance, one author can be identified in the same data source with four different ways: "Abdulmotaleb El Saddik," "A. El Saddik," "Abdulmotaleb Elsaddik," and "A. Elsaddik." In order to overcome this issue, we aim to use the alternative sources and the metadata information like author's affiliations or ORCID [83] and Publons [84] to fill the gap.

Initially, the crawled dataset was too sparse, and we had to clean it before starting the experiments. The cleaning process involved the removal of articles that had a very small number of citation.Similarly, we removed articles that do not have a reference list, author information (affiliations or ORCID), and venue information (title, publisher). We used the publicly available data source from the digital libraries of IEEE, ACM, DBLP, and Crossref to build the dataset. The size of the information that can be accessed on the data sources tool exceeds 90 million records between articles, authors and venues information. We build different dataset based on the target

outcome of the evaluation experiment. The dataset will be explained for each experiment later on the evaluation strategy.

5.2. Evaluation Measurements

Precision, recall and $F_1 - \text{measure}$ considered the most widely adopted evaluation measurements to evaluate recommender systems. These three parameters are used to calculate the effectiveness of the resulting list of recommendations. Precision computes the ratio of the retrieved recommended objects to the actual relevant objects according to the actual selection made on the tested article as in Equation (5-1). Recall computes the number of relevant objects among the recommended objects(articles) as in Equation (5-2). For instance, for each query in the test set, we use the original set of references list as the ground truth RL. Assume that the set of the candidate recommended articles for citation are RC , so that the correctly recommended citations are $RC \cap RL$.

$$precision = \frac{|RC \cap RL|}{|RC|} \tag{5-1}$$

$$recall = \frac{|RC \cap RL|}{|RL|} \tag{5-2}$$

In addition to precision and recall, we calculate the $F_1 - \text{measure}$ using Equation (5-3) in order to compare our method to the other benchmark algorithms. $F_1 - \text{measure}$ integrate the precision and recall performance into one comparable value. In order to measure the ranking positions of each recommended object, we varied the number of objects retrieved (top k values).

$$F_1 = \frac{2(precision\,.recall)}{precision + recall} \quad (5\text{-}3)$$

5.3. Evaluating the Academic Citation Impact Algorithm

In this section, we evaluate our proposed approach using the Academic Citation Impact index and compare its performance against baseline textual, co-occurrence and global relevance approach and their related indices. The main purpose of this experiment is to measure the effectiveness of the proposed algorithm in the discovery of influential articles compared to the baseline approaches used in scientific search engines. The experimental dataset used in this experiment was collected randomly from the IEEE digital library. The original dataset was too sparse to be used for experiments. Therefore, we cleaned the dataset to carry out experiments that were more meaningful. The cleaning process involved the removal of articles that had a very small number of citation. In addition, we cleaned articles that had been cited less than five times. Similarly, we removed articles that don't have a reference list, author information (affiliations or ORCID), and venue information (title, publisher). The cleaned dataset used in this study contained 480,432 articles, where 67% of the dataset represents conference articles, and 33% represents journals.

5.3.1. Evaluation strategy

To evaluate the proposed algorithm, we followed the evaluation procedure described in [58]. We randomly divided the experiment datasets into two groups: a training set that represents 80% of the original dataset, and the remaining 20% used as a test set. In order to build the test set, we randomly withheld object for testing later on. We use the original reference list, which was not present while training, as the ground truth. In the field of scientific research, the reference list of

an article is carefully chosen by the authors based on the close relationship between the produced article and the referenced articles. We aim to compare the system prediction to the author's actual selections for the articles in the reference list for each tested object. We assume that any paper other than the actual author's selections is not relevant. To ensure that our results were not sensitive to a particular test query, we conducted five different runs using different training-test portioning datasets to avoid the possibility that the test set is not biased. Three types of evaluation metrics were adopted: precision at top- k using Equation (5-1), recall at top- k using Equation (5-2) and F1 measure (at top 10) using Equation (5-3) to assess the accuracy of *ACI* as an index to recommend articles.

5.3.2. Baseline algorithms

We compare our proposed approach with three well-known baseline approaches: I) content-based relevance approach described in [28], [31]; II) co-occurrence based relevance approach using the references citation relation approach presented in [48], [49]; and III) the global relevance approach presented in [38]. Because those algorithms were mostly tailored to our search scenario with the academic recommendation, we implemented them to the best of our knowledge, based on the published papers.

The content-based approach using different NLP techniques to apply it considering the textual content of the target manuscript. We chose the most utilized baseline technique which is term frequency-inverse document frequency (denoted TF-IDF) technique in the comparison [43]. The final ranked list was created based on the TF-IDF score between the user's query (given article) and the venues related articles at the nearby citation network for a given user's query. The co-occurrence approach (denoted CO-OC) depends on a frequent joint occurrence of two objects

together. In the research industry, the most common application reflected this methodology is co-citation analysis. The concept is based on the more appearance of the same two or more objects together in the same place, the higher the correlation between these objects. The resulted candidate list was created based on the total number of joint citation references between the user's query (given article and the venues related articles at the nearby citation network of that given user's query. The global relevance approach (denoted GR) adopts a universal-fit methodology that recommends items with the highest relevance based on their surroundings. This approach does not depend on user preferences to calculate the relevance between the recommendation candidates. Instead, it depends on global measurements such as overall popularity. Therefore, it considers the size of the search area surrounding the citation network. The outcome recommended list was created based on the citation count of a focal article (venue related) is divided by the mean number of cited references (total citation count divided by the total number of articles in a specific venue) at the citation network for a given user's query. We have changed the number of returned k items to examine the accuracy of each algorithm's based on the positions of the relevant item ranking at the recommendation candidates list for a given user's query.

5.3.3. Comparison with other ranking algorithms

As mentioned earlier, we evaluated the performance of the proposed algorithm by comparing it to the other three alternative methods based on the resulted indicators. The purpose of the experiment is to examine the accuracy of the proposed approach for retrieving the most influential articles to use as references. In the field of scientific research, the reference list of an article is carefully chosen by the authors based on the close relationship between the produced article and the referenced articles. The reference list selection was used for each article in the experimental dataset, and we consider this list to be the basis for evaluating the recommendation

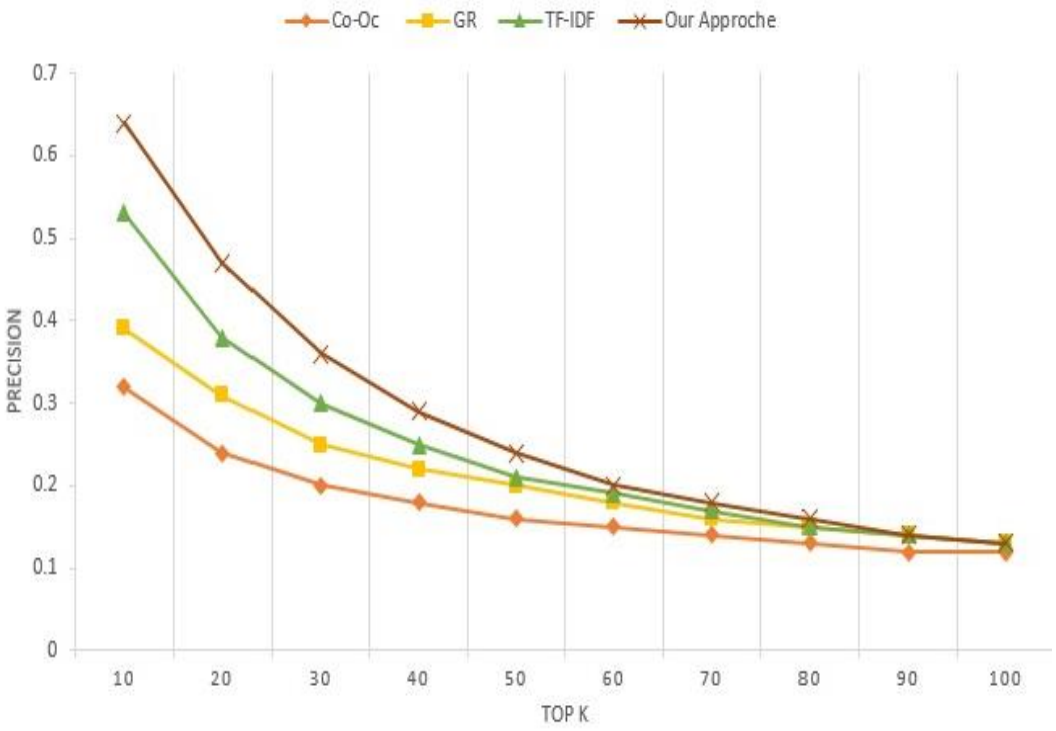

Figure 5-1: Precision at top k for each recommendation approach.

quality. Figure 5-1 shows the results of the precision performance with respect to different values of k (top item recommended), according to Equation (5-1). The results show that the CO-OC and GR approaches have the worst performance and that our approach outperforms all other methods.

We continue to inspect the recall and F1 measures of each algorithm, as shown in Figure 5-2 and Figure 5-3, using Equation (5-2) and Equation (5-3). Our proposed approach obtained approximately 29.61%, 17.34%, and 11.8% improvement on recall and 47.62%, 33.93%, and 17.35%, improvement on the F1 measure (at top 10, 20, and 30), compared to CO-OC, GR, and TF-IDF respectively. When comparing the reported results, we observe that finding latent semantics of content and latent relations of the articles in the extended nearby citation network revealed more relevant article than term-based approaches. We also notice that a higher the impact of the pertinent article list on the given article using different parameters in a different dimension,

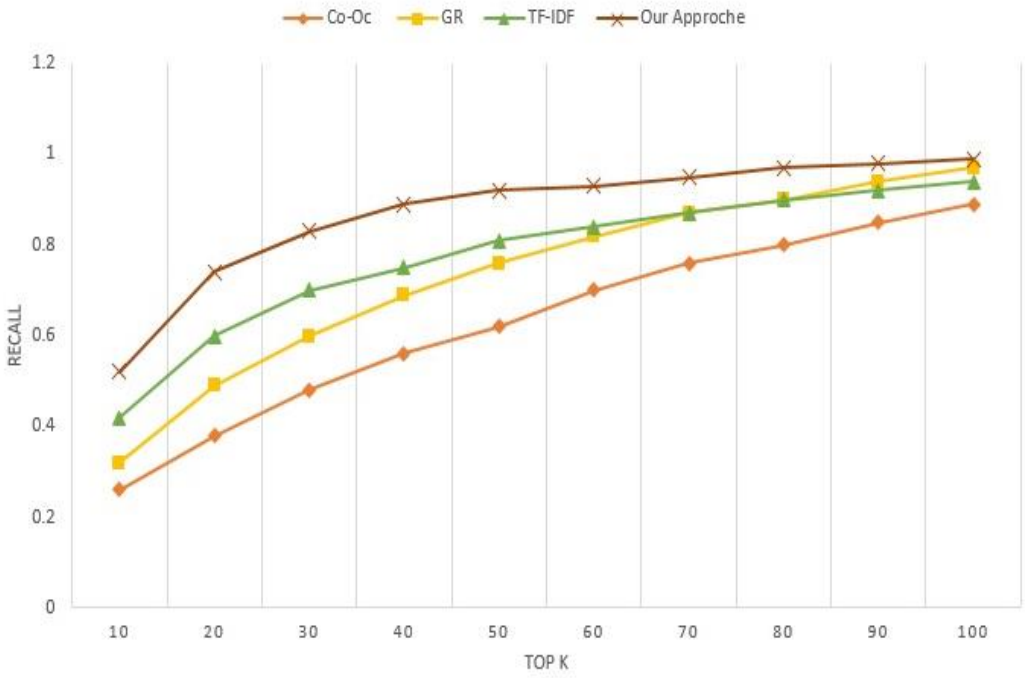

Figure 5-2: Recall at top k for each recommendation approach.

the better the recommendation would be. It indicates a positive influence on the identification of relation accuracy between the articles in the pertinent article list.

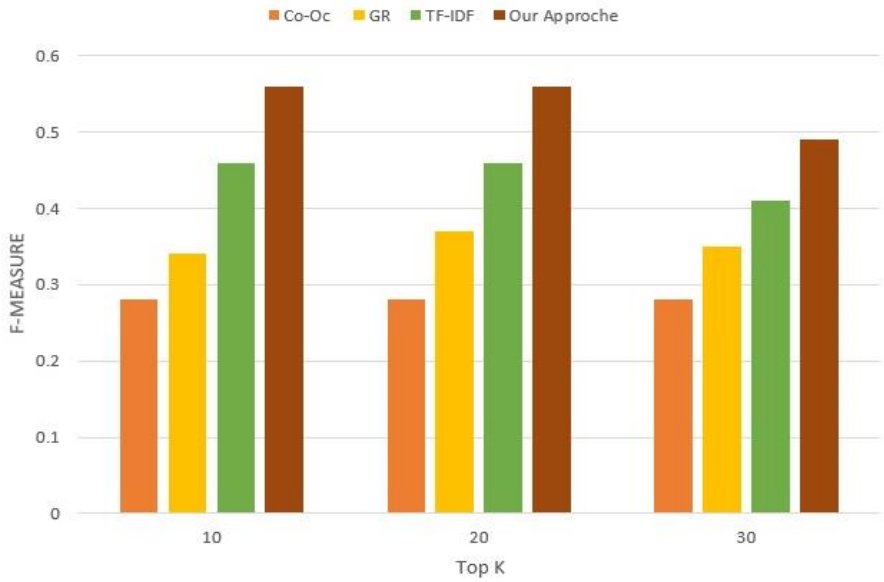

Figure 5-3: The F-measure at top k=10, 20 and 30 for each recommendation approach.

Dataset	Number of articles	Number of authors	Number of venues
IEEE	4,673,183	2,606,867	28,967
ACM	2,835,548	2,339,783	26,935
DBLP	4,340,893	2,169,653	11,529

Table 5-1: Statistics of the original dataset from different digital libraries

5.4. Evaluating the Academic Collaborative Recommendation

In this section, we evaluate our proposed Academic Collaborative Recommendation based on the LPIM model and compare its performance against well-known baseline recommender systems. The main purpose of this experiment is to measure the effectiveness of the proposed system in the discovery of influential objects compared to the baseline approaches used in academic recommender systems. The experimental dataset used in this experiment was collected from different IEEE, ACM, and DBLP digital libraries. Table 5-1 defines the size of the crawled datasets from different libraries. The original datasets were too sparse to be used for experiments.

Therefore, we cleaned the datasets to carry out experiments that were more meaningful. The cleaning process involved the removal of articles that had a very small number of citation. In addition, we cleaned articles that had been cited with less than five. Similarly, we removed articles that don't have a reference list, author information (affiliations or ORCID), and venues' information (title, publisher). In addition, we cleaned venues information by grouping the venues that have the same name with a different publishing year. also, we cleaned authors information by

grouping the authors with a different written style based on the affiliations or ORCID information if applicable.

5.4.1. Evaluation strategy

We divided the evaluation process into three different procedures to cover each academic object (articles, authors and venues) that represents the academic publishing industry. To evaluate the proposed algorithm, we followed the evaluation procedures described in [58], [71]. We randomly divided the experiment datasets into two groups: a training set and a test set. In order to build the test set, we randomly withheld objects as a test-queries for each request. We use the original object information, which was not present while training, as the ground truth. We aim to compare the system prediction to the actual object selections in the original object list. To ensure that our results were not sensitive to a particular test query, we conducted five different runs using different training-test portioning datasets to avoid the possibility of biased. We will explain the evaluation process for each different experiment procedures for each academic object in the following sections.

5.4.2. Article Recommender Evaluation

To evaluate the proposed system aiming article recommendation, we followed the evaluation procedure described in [58]. Due to the processing time-consuming, we collect randomly two sub-datasets from the IEEE and ACM digital libraries. The IEEE dataset contains 1,500 source (citing) articles, and 516,907 target (cited and related) articles and the ACM dataset contains 1,500 source (citing) articles and 480,432 target (cited and related) articles. We randomly divided the experiment datasets into two groups: a training set that represents 80% of the original dataset, and the remaining 20% used as a test set. In order to build the test set, we randomly

withheld one object as a test-query. To ensure that our results were not sensitive to a particular test query, we conducted ten different runs using different training-test portioning datasets to avoid the possibility that the test set is not biased. Three types of evaluation metrics were adopted: precision at top- k using Equation (5-1), recall at top- k using Equation (5-2) and F1 measure (at top 10) using Equation (5-3) to assess the accuracy of the proposed approach.

5.4.2.1. Baseline algorithms

We compare our proposed approach with three well-known baseline approaches: I) topic based relevance approach (denoted RefSeer) described in [62]; II) content and global relevance approach (denoted PubRec) presented in [63]; and III) the Social Recommender approach (denoted SR) presented in [85]. IV) Because those algorithms were mostly tailored to our search scenario with the academic recommendation, we implemented them to the best of our knowledge, based on the published papers.

The purpose of the experiment is to examine the accuracy of the proposed approach for retrieving the most influential articles to use as references. In the field of scientific research, the reference list of an article is carefully chosen by the authors based on the close relationship between the produced article and the referenced articles. The reference list selection was used for each article in the experimental dataset, and we consider this list to be the basis for evaluating the recommendation quality.

5.4.2.2. Comparison with other baseline algorithms

We aim to compare the system prediction to the author's actual selections for the articles in the reference list for each tested article. We merge the articles from the reference list and the related articles in a testing pool to identify the candidate list of articles to be cited for the tested

Figure 5-4: Precision comparison for different approaches with the IEEE dataset

article. For each given article, we examine how well the systems predicted the actual articles in the reference list based on the given article.

Figure 5-4 and Figure 5-5 illustrate the experimental results compared our proposed approach to the other four baseline approaches with the IEEE dataset. We retrieve the top k ranked articles that should be suggested to the user. The precision and recall curves show how our proposed approach outperformed the other baseline approaches. The precision and recall values of a specific top k retrieved article are shown as data points of the graph curves, with the utmost left points on the curves denoting top $k = 10$, and the lower points on the right denoting the top $k = 120$. Our proposed approach achieved 46.2%, 41.9%, 22.2% and 10% on precision improvement and 40.4%, 30.8%, 27.9% and 11.2% on recall improvement compared to SR, TVS, PubRec, and RefSeer, respectively. We continue to examine the F-measures of each algorithm, as shown in

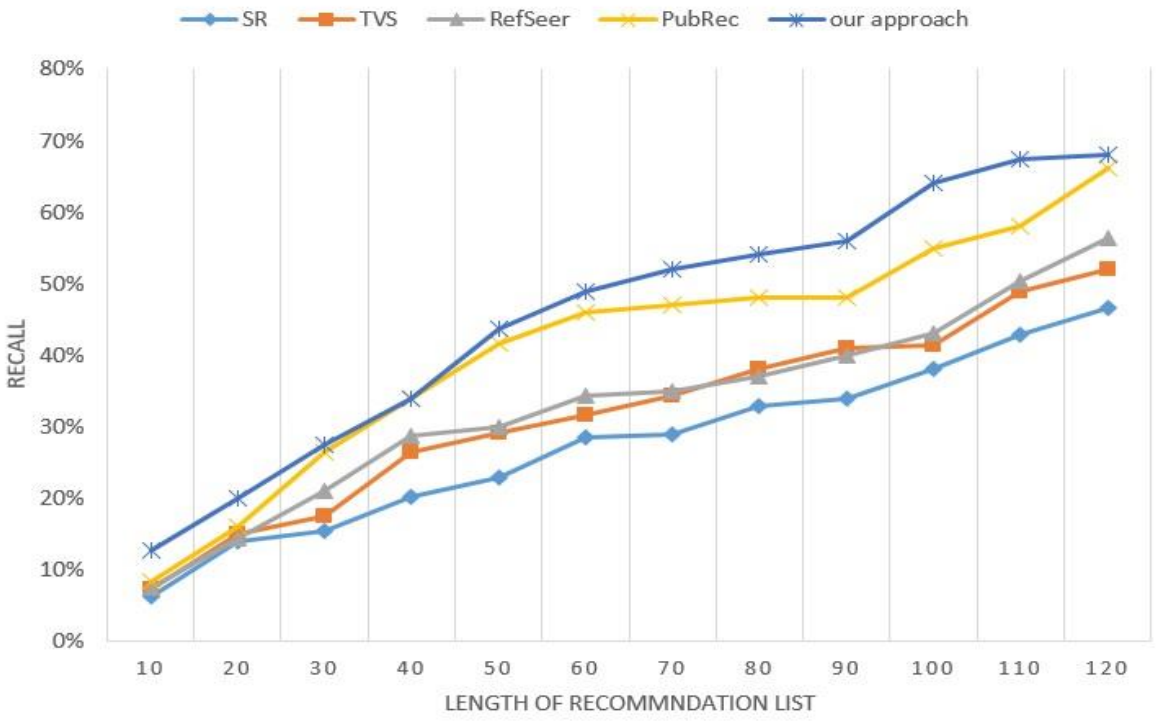

Figure 5-5: Recall comparison for different approaches with the IEEE dataset

Figure 5-6, the proposed system obtained approximately 43.3%, 36.9%, 28.7% and 17.6% improvement on the F-measures average at top 10, 20 and 30.

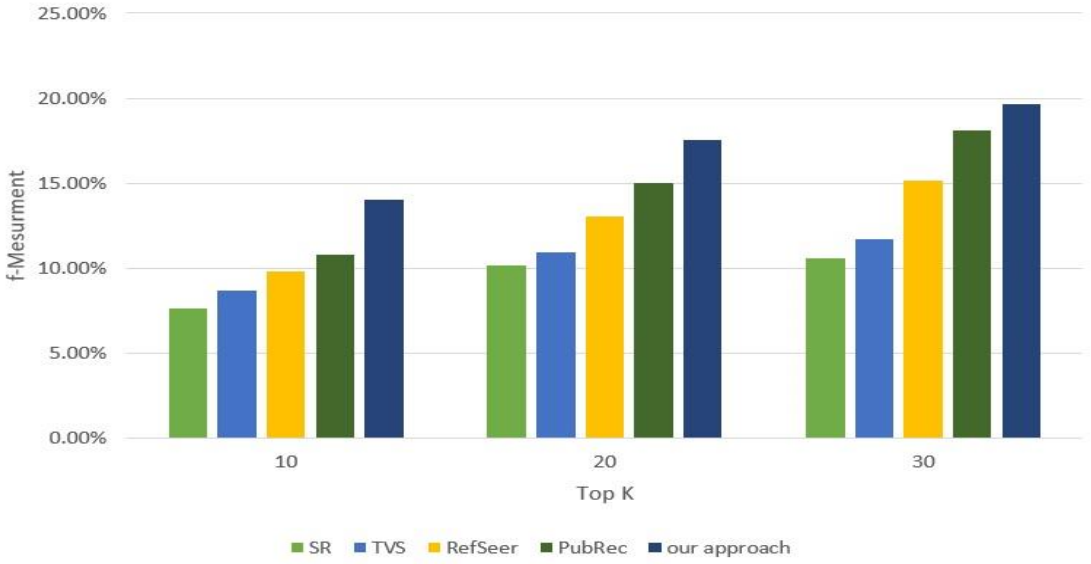

Figure 5-6: F-measurement comparison for different approaches with the IEEE dataset

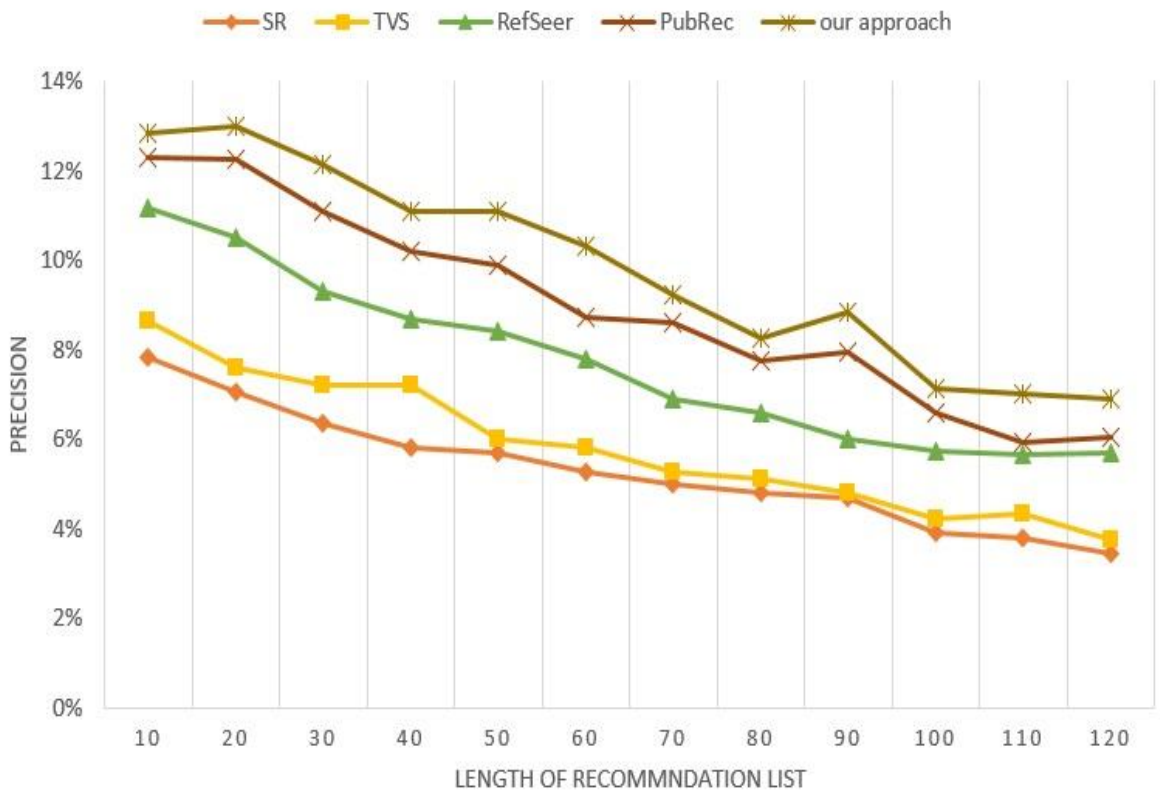

Figure 5-7: Precision comparison for different approaches with the ACM dataset

Figure 5-7 and Figure 5-8 illustrate the experimental results compared our proposed approach to the other four baseline approaches with the ACM dataset. As we mention early, we retrieve the top k ranked articles that should be suggested to the user. The precision and recall curves show how our proposed approach outperformed the other baseline approaches. The precision and recall values of a specific top k retrieved article are shown as data points of the graph curves, with the utmost left points on the curves denoting top $k = 10$, and the lower points on the right denoting the top $k = 120$. Our proposed approach achieved 46.1%, 40.9%, 21.6% and 9.3% on precision improvement and 40.2%, 31.2%, 27.6% and 10.1% on recall improvement compared to SR, TVS, PubRec, and RefSeer, respectively. We continue to examine the F-measures of each

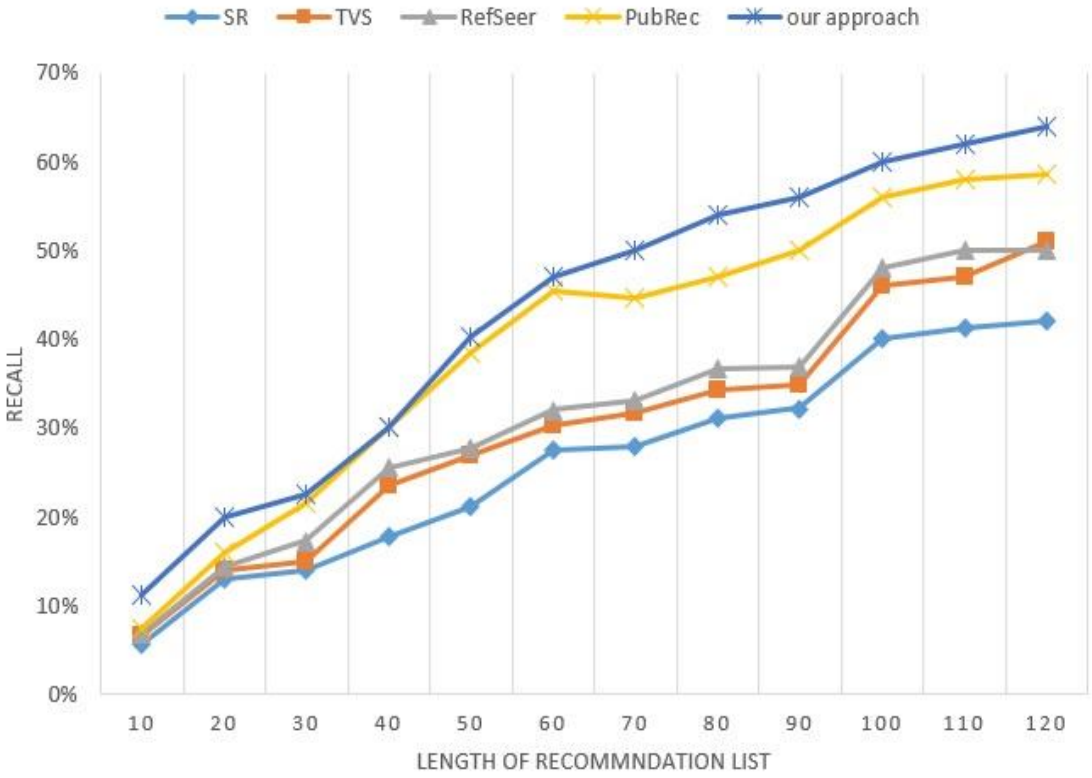

Figure 5-8: Recall comparison for different approaches with the ACM dataset

approach, as shown in Figure 5-8, the proposed system obtained approximately 42.9%, 36.7%, 28.1% and 16.8% improvement on the F-measures average at top 10,20 and 30.

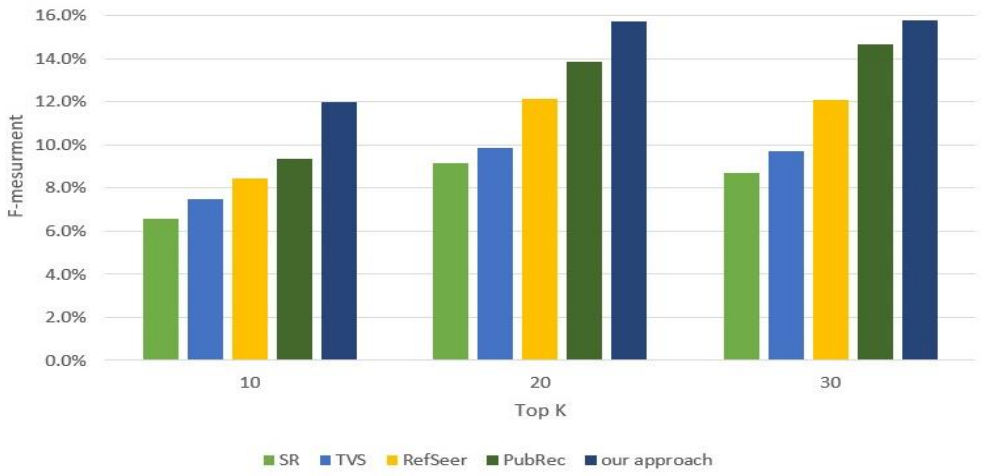

Figure 5-9: F-measurement comparison for different approaches with the ACM dataset

Dataset containing	Articles $\|A\|$	Authors $\|H\|$	Venues $\|V\|$
"database"	143,524	3,563	130
"multimedia"	95,156	11,086	204

Table 5-2: Statistics of the IEEE datasets used to evaluate the recommendation system

When comparing the reported results, we observe that finding latent semantics of content and latent relations of the articles in the extended nearby citation network revealed more relevant article than term-based approaches. We also notice that a higher the impact of the pertinent article list on the given article using different parameters in a different dimension, the better the recommendation would be. It indicates a positive influence on the identification of relation accuracy between the articles in the pertinent article list. The results suggest that the ILPM model recommends more relevant items in any given nearby citation network.

5.4.3. Author recommender evaluation

To evaluate the proposed system aiming author recommendation, we followed the evaluation procedure described in [82]. Due to the high processing time, we used two subsets of the IEEE digital library for our experiments, containing the terminologies "database" and "multimedia" citation network. Table 5-2 defines the size of the crawled datasets from the data source. The original datasets were too sparse to be used for experiments. Therefore, we cleaned the datasets to carry out experiments that were more meaningful as we mentioned early. We divided each dataset into two groups: a randomly withheld 20% of the dataset as the test set and the remaining 80% as the training set; then we conducted ten runs using different training-test apportioning datasets to avoid the possibility that the test set was biased. Three types of evaluation

metrics were adopted: precision at top- k using Equation (5-1), recall at top- k using Equation (5-2) and F1 measure using Equation (5-3) to assess the accuracy of the proposed system.

5.4.3.1. Baseline algorithms

We compared our proposed approach with four recommendation models: most valuable collaborators walker model (denoted MVCW) described in [86], Random walks with restarts model (denoted RWR) described in [87], topic-based model, and co-occurrence model. MVCW model proposed based on the random walk model where they weighted the edges between the nodes in the network based on some academic factors. In this implementation, we consider a total number of collaboration and the coauthor order to purpose to estimate the edge's weight between the nodes. Random walks with restarts method, which is a well-known method in the graph-based approach, utilized to discover the most relevant item in a weighted graph to recommend the candidate objects. Different academic factors used to estimate the edge's weight between two nodes in the graph. In this implementation, we regard the co-authorship to purpose to estimate the edge's weight between the nodes. The topic-based model represents the content-based approach. It aims to compute the similarity between the authors and venues using cosine similarity based on the topic of the articles as feature vectors. The co-occurrence model represents the collaborative filtering approach. It depends on a frequent joint occurrence of two items together. For instance, if two authors are citing the same article, those authors can be recommended to collaborate. The purpose of the experiment is to examine the accuracy of the proposed approach to predict the closest authors who can be recommended to collaborate with them.

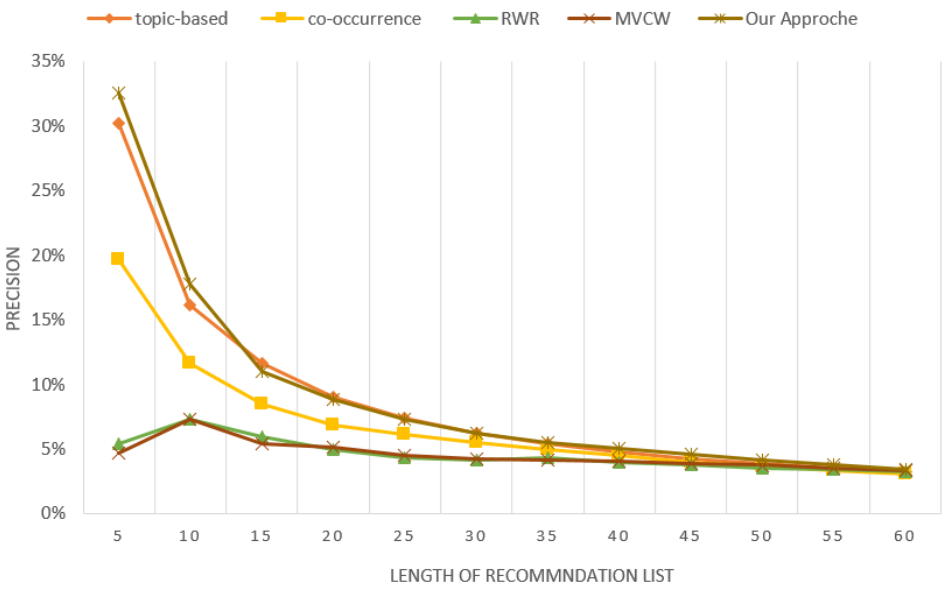

Figure 5-10: Precision comparison for different approaches to recommend authors.

5.4.3.2. Comparison with other baseline algorithms

The reference list was used for each article in the experimental dataset to extract the bibliographic information about the authors. We considered the author's names for each tested article to be the basis for evaluating the recommendation quality. We aim to compare the system prediction to the actual collaborations. We combine the author's names from different articles in a testing pool to identify the candidate list of authors for a given article. For each given article, we tested how well the algorithm predicted the authors of the article in which the article was published based on the given query.

Figure 5-10 and Figure 5-11 illustrate our experimental results compared to the other three state-of-the-art approaches based on author recommendation. We retrieve the top k authors ranked that should be recommended to the user. The precision and recall curves show how our proposed

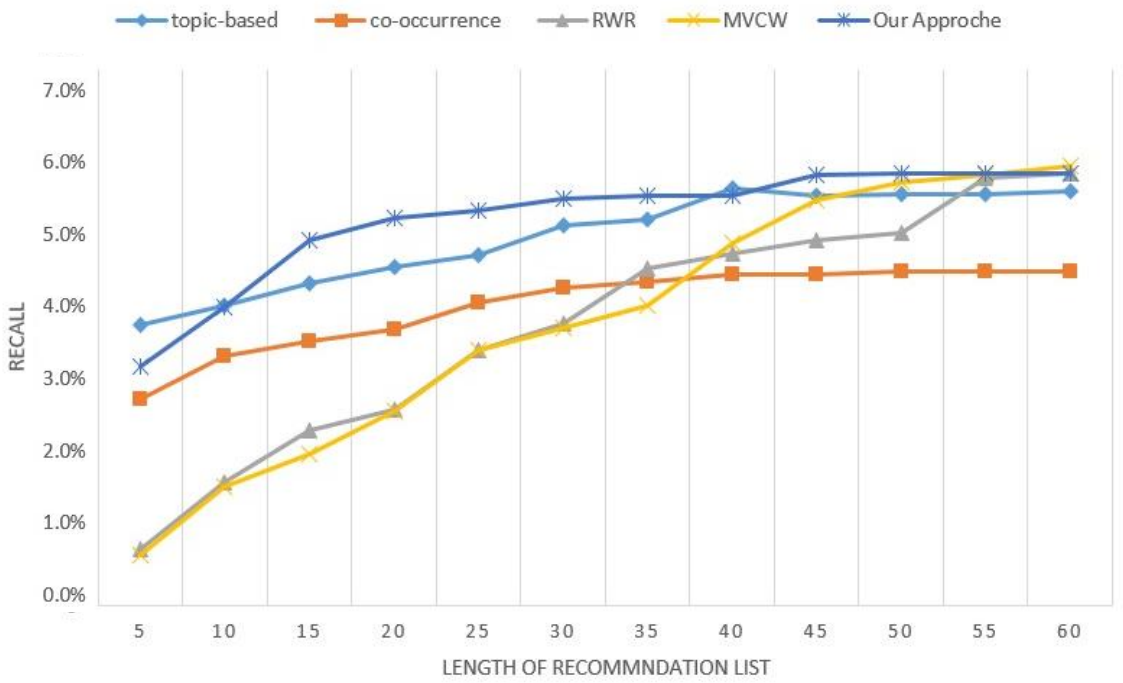

Figure 5-11: Recall comparison for different approaches to recommend authors.

approach outperformed the other approaches with both datasets. The precision and recall values of

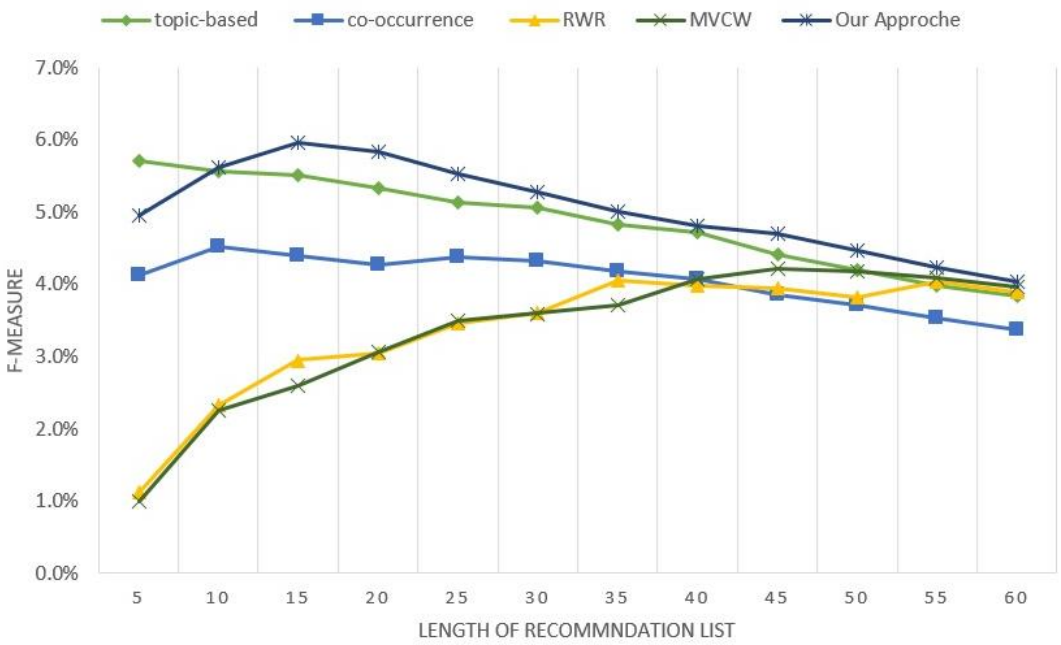

Figure 5-12: F-measurement comparison for different approaches to recommend authors.

a specific top k retrieved authors are shown as data points of the graph curves, with the left points on the curves denoting top k = 5, and the points on the right denoting the top k = 60. Our approach showed 3.4%, 18.0%, 32.7%, and 32.0% precision improvement and 3.9%, 21.9%, 30.3, and 30.0% recall improvement on the dataset average compared to topic-based, co-occurrence, RWR and MCVW models respectively. We continue to examine the F-measures of each algorithm, as shown in Figure 5-12, the proposed system obtained approximately 3.5%, 19.0%, 31.6%, and 31.2% improvement on the F-measures on the average of the dataset compared to topic-based, co-occurrence, RWR, and MCVW models respectively.

We found that the accuracy of our approach has higher average precision and recall values than other approaches when retrieving a different number of authors. When comparing the reported results, we observed that topic-based model could accomplish better result than MCVW, RWR, and co-occurrence models in the long run. Furthermore, the use of hybrid latent approaches combining latent semantics of content and latent relations of the authors using multiple dimension revealed more relevant authors than different filtering approaches. It indicates a positive influence on makes more efficient recommendation.

5.4.4. Venues recommender evaluation

To evaluate the proposed system aiming article recommendation, we followed the evaluation procedure described in [82]. We used previously mentioned two sub-datasets of the IEEE digital library for our experiments that defined in Table 5-2.We divided each dataset into two groups: a randomly withheld 20% of the dataset as the test set and the remaining 80% as the training set; then we conducted ten runs using different training-test apportioning datasets to avoid the possibility that the test set was biased. Three types of evaluation metrics were adopted:

precision at top- k using Equation (5-1), recall at top- k using Equation (5-2) and F1 measure using Equation (5-3) to assess the accuracy of the proposed system.

5.4.4.1. Baseline algorithms

We compared our proposed approach with four recommendation models: Academic Venue Recommendation model (denoted AVER), Random Walks with Restarts model (RWR), topic-based model, and co-occurrence model all described in [82]. AVER is an extended version of random walks with restarts model where the edge's weight between two nodes based on the co-publication frequency and researcher academic level. Random walks with restarts method, which is a well-known method in the graph-based approach, utilized to discover the most relevant item in a weighted graph to recommend the candidate objects. Different academic factors used to estimate the edge's weight between two nodes in the graph. In this implementation, we regard the co-publishing to purpose to estimate the edge's weight between the nodes. The topic-based model represents the content-based approach. It aims to compute the similarity between the venues using cosine similarity based on the topic of the articles as feature vectors. The co-occurrence model represents the collaborative filtering approach. It depends on a frequent joint occurrence of two items together. For instance, if author *S* and author *F* are attending the same venue *X*, and author *S* attended venue *Y*. Thus, venue *Y* can be recommended to author *F* to attend. The purpose of the experiment is to examine the accuracy of the proposed approach to predict the most related publication venue in which the article can be submitted.

5.4.4.2. Comparison with other baseline algorithms

The reference list was used for each article in the experimental dataset to extract the bibliographic information about the authors and venues. We considered the venues names for each

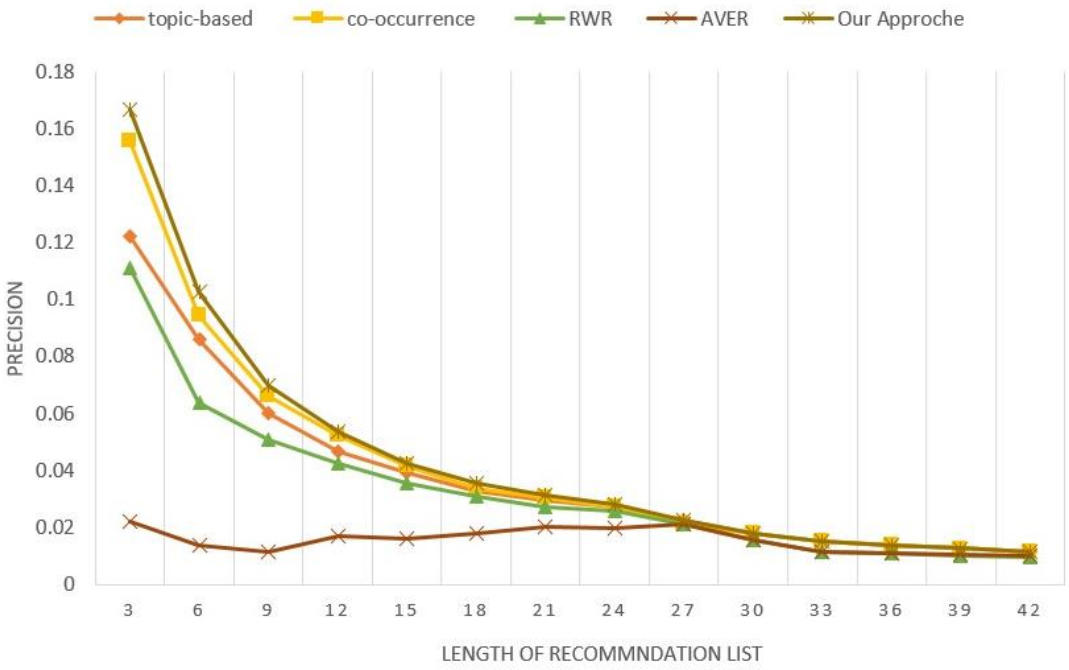

Figure 5-13: Precision comparison for different approaches to recommend venues @ top k venue.

tested article to be the basis for evaluating the recommendation quality. We aim to compare the system prediction to the actual venue selections. We combine the venue's names from different articles in a testing pool to identify the candidate list of venues for the given article. For each given article, we tested how well the algorithm predicted the venues in which the given article was actually published in based on the given query.

Figure 5-13 and Figure 5-15 illustrate our experimental results compared to the other three state-of-the-art approaches based on venue recommendation. We retrieve the top k venues ranked that should be recommended to the user. The precision and recall curves show how our proposed approach outperformed the other approaches with both datasets. The precision and recall values of a specific top k retrieved venues are shown as data points of the graph curves, with the left points on the curves denoting top k = 3, and the points on the right denoting the top k = 42. Our approach

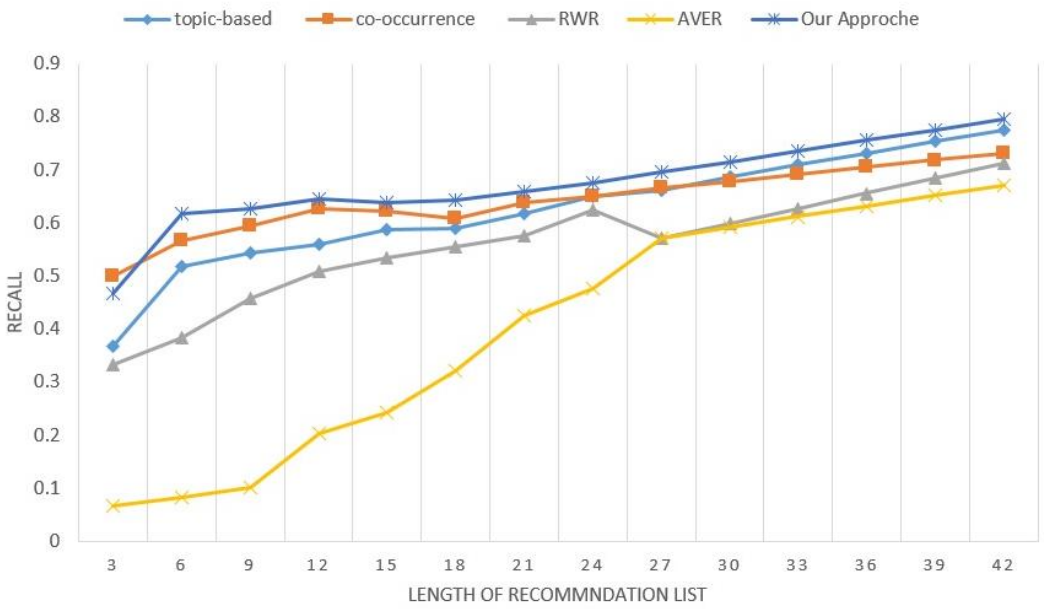

Figure 5-15: Recall comparison for different approaches to recommend venues.

showed 6.9%, 3.1%, 19.2%, and 42.6% precision improvement and 8.0%, 4.4%, 17.8%, and 43.0% recall improvement on the dataset average compared to topic-based, co-occurrence, RWR, and AVER models respectively. We continue to examine the F-measures of each algorithm, as shown

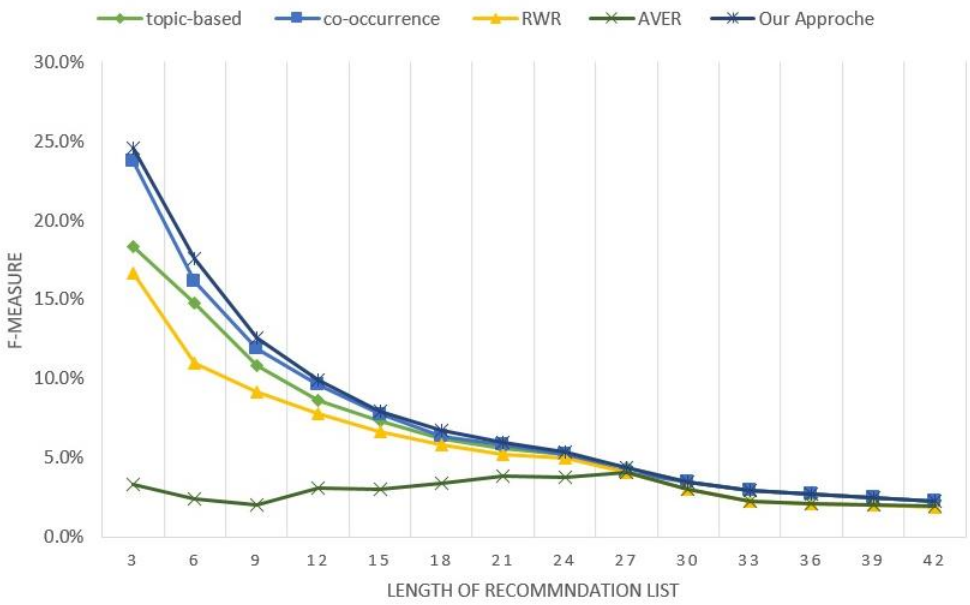

Figure 5-14: F-measurement comparison for different approaches to recommend venues

in Figure 5-14, the proposed system obtained approximately 6.8%, 2.9%, 19.1%, and 42.6% improvement on the F-measures.

We found that the accuracy of our approach has higher average precision and recall values than other approaches when retrieving a different number of venues. When comparing the reported results, we observed that the co-occurrence model could accomplish better result than AVER, RWR, and topic-based models in the long run. Furthermore, the use of an LPIM model based on multiple dimension revealed more relevant venues than different filtering approaches. It shows a positive impact to recommend more efficient candidate list.

5.5. Computational Analysis

5.5.1. Computing Similarity

We analyze the computational complexity of the objects similarity matrices according to the number of articles$|A|$, the number of authors$|H|$, and the number of venues$|V|$. In order to decrease the time and cost of computing the offline similarities, we only consider the most similar neighbors for each similarity computation. Accordingly, we exclude the computed similarities of the academic objects that share few connections with others, and assign a zero similarity value if the similar objects is not among the top k nearest neighbors. In addition, we ignore computing the similarities of the same vector entries, For instance, we employ the matrix $\mathbf{G}$, where $\mathbf{G} = \mathbf{G^k}$ to form the article-article similarity matrix. For any two column vectors $sim(g_x, g_y) = 1$, if $x = y$. Therefore, the outcome similarity matrix will be represented as a diagonal similarity matrix. Then, only the non-zero entries are used in the process of building the recommendation model. The worst-case computational cost of building the similarity matrices ($\mathbf{R}_{|H|\times|H|}$, $\mathbf{G}_{|A|\times|A|}$ and $\mathbf{Z}_{|V|\times|V|}$) is $O(|H|^2|A||V|)$, $O(|A|^2|H||V|)$, and $O(|V|^2|A||H|)$ respectively.

5.5.2. Identifying the latent preferences

We analyze the computational complexity of our Latent Preferences Identifier Model algorithm according to$|A|, |H|, |V|$, the number of similar articles k_a, the number of similar authors k_h, and the number of similar venues k_v. Identifying the hidden preferences is completed offline in the recommender system. In the worst-case scenario, computing the six hidden preferences models $\overrightarrow{\mathbf{HA}}_{|H|\times|A|}$, $\overrightarrow{\mathbf{HV}}_{|H|\times|v|}$, $\overrightarrow{\mathbf{AH}}_{|A|\times|H|}$, $\overrightarrow{\mathbf{AV}}_{|A|\times|V|}$, $\overrightarrow{\mathbf{VH}}_{|V|\times|H|}$ and $\overrightarrow{\mathbf{VA}}_{|V|\times|A|}$ requires $O(k_a|H||A|)$, $O(k_v|H||V|)$, $O(k_h|A||H|)$, $O(k_v|A||V|)$, $O(k_h|V||H|)$ and $O(k_a|V||A|)$ respectively. However, this step depends on the creating the similarity matrix. Thus, the total cost of building each model contains the cost creating the similarity matrix in Section 5.5.1 and the cost of this step. Therefore, calculating each model costs $O(|H|^2|A||V| + k_a|H||A|)$, $O(|H|^2|A||V| + k_v|H||V|)$, $O(|A|^2|H||V| + k_h|A||H|)$, $O(|A|^2|H||V| + k_v|A||V|)$, $O(|V|^2|A||H| + k_h|V||H|)$ and $O(|V|^2|A||H| + k_a|V||A|)$ for models $\overrightarrow{\mathbf{HA}}$, $\overrightarrow{\mathbf{HV}}$, $\overrightarrow{\mathbf{AH}}$, $\overrightarrow{\mathbf{AV}}$, $\overrightarrow{\mathbf{VH}}$ and $\overrightarrow{\mathbf{VA}}$ respectively. Thus, we have considered the worst-case scenario for calculating the computational cost.

5.5.3. Computing the Academic Collaborative Ranking

All the previous steps to build the citation network and find the latent preferences using the similarity matrices computation can be done offline whenever new information available. When building the latent model, objects scores for a given user query will be calculated based on the extended nearby citation network. To find the authors, articles and venues candidate list for a given query, the required computational cost are $(|h|^2|a||v|)$, $O(|a|^2|h||v|)$ and $O(|v|^2|a||h|)$, respectively, where $|h|, |a|$ and $|v|$ represent a collection of authors, articles and venues in the extended nearby citation network for given article.

5.6. Summary

In this chapter, we provided an evaluation procedure that has been considered to assess the outcomes of the proposed recommendation system to adopt the proposed system. We divided the evaluation procedure is divided into two parts. The first part is evaluating the effectiveness of the proposed ACI algorithm, and the second part is evaluating the LPIM model for the recommendation system. We inspected whether or not the proposed recommendation model improves the accuracy of the item prediction regardless of performance. The goal of the experiments is to evaluate the accuracy of the proposed recommendation system: the utilization of bibliometric data and content semantic similarity of the articles in the extended nearby citation network to recommend a different number of items. The LPIM model demonstrates with positive advantages the accuracy of prediction and the provision of more suitable objects to the user based on the given query. The reported outcomes show that our ACI algorithm and academic recommendation system outperforms the other baseline systems in terms of accuracy to recommend candidate list according to the authors' actual selections. According to the experiment, our proposed system can be considered to retrieve more relevant and influential articles, considering the analysis of the extended nearby citation network of given article.

Chapter 6 . Conclusion and Future Work

In today's large and rapidly increasing amount of scientific publications, exploring recent studies in a given research area has become more challenging than any time before. Scientific production growth has been increasing the difficulties for finding the most relevant paper to cite or to find an appropriate conference or journal to submit a paper for publishing of to find a researcher to team up with them. Exploring and arranging a virtual collaboration between researchers within a research field is one of the components that formulated by the digital twin vision. Hence, this work developed a system to help young researchers to find proper collaborators, suitable venues, and related articles. The enormous amount of conferences and journals consider a challenge to young researchers to connect the right community. Besides, they face difficulties to determine the list of articles that are most relevant for their research work.

6.1. Thesis Summary

In this thesis, we proposed the design and implementation of an academic recommendation system that incorporates the content semantic analysis and citation network analysis into the recommendation process to facilitate the capture, analysis, and demonstration of the academic collaboration in the extended nearby citation network. The proposed system facilitates the awareness of the impact of academic objects (articles, authors, and venues) in scientific communities by adopting a multi-model citation analysis algorithm to measure the academic

citation impact. We purpose to design and implement a multi-dimensional academic recommendation model. This model incorporates the awareness of the citation network and the academic latent preferences to predict what the user wants to use afterward. The proposed model identifies the similarity between the objects in the citation network in order to find the hidden preferences for each object from similar ones.

We propose a new algorithm to measure the citation impact. Our ACI algorithm allows us to detect the impact of articles at their extended nearby citation network for a given article. We combine the article contents with its bibliometric to calculate the citation impact of the articles at the surrounding citation network. We employ the article metadata includes among others title, abstract, and keywords to calculate the semantic similarity between the articles in the citation network. We use the article similarity scores along with the bibliometric indices are to evaluate the impact of the article among their extended nearby citation network. After that, we proposed an algorithm to build the similarity matrices using the citation network matrices for the given article to find the similarity between the articles, authors, and venues. The main goal of the similarity is to determine the article-article, author-author and venue-venue similarity relationship among citation network. Then, we use similarity matrices to builds the latent models for each academic object in order to identify the hidden preferences among the citation network. The author's latent model aims to determine the author's preferences toward their association with articles and venues. The article latent model aims to determine the article’s preferences toward their association to authors and venue. The venue latent model aims to determine the venue's preferences toward their association with articles and authors. Then we use the latent preferences matrices to estimate the amount of reliance for collaboration between academic objects in the citation network. We proposed the academic collaborative ranking algorithm to predict recommendations candidate list

for academic objects (author, article, venues) that are relevant to a given query. The algorithm will compute the confidence score based on the users requested and produce a candidate recommendation list containing academic object ranked based on the confidence score for each object.

The advantage of our proposed recommendation model is that it considers the similarity learning information reflected by the content semantic similarity and the bibliographic information available online to explore the hidden relations among the citation network of a given article as well as by applying collaborative filtering to find hidden preferences from similar objects. To evaluate the citation impact of articles, authors, and venues in their surrounding citation network, we combine the content similarity with the bibliometric indexes in the multi-model analysis algorithm. The experimental results demonstrate that the proposed hybrid recommendation technique offers promising gains in enhancing the accuracy of the prediction and defining the most suitable venue, and potential authors to team up with and retrieve more relevant and influential articles upon the user request. The lack of available consistence bibliographic information has become an essential challenge in the design of any academic recommendation system. Therefore, we built a prototype algorithm that collects bibliographic information from different research-oriented data sources. By building such an algorithm, we collected the required metadata and bibliographic information.

6.2. Future Work

The scope of the academic recommendation is rich in opportunities to increase the effectiveness of the proposed system in a real-world environment. Potential enhancements will be addressed in regrades the influences of object impact index among a particular research scope, a

particular institution, and self-citations. We intend to enrich the analysis further by enhancing the recommender system so it can simultaneously suggest different types of collaboration in the scientific community considering the current and former affiliation information for the authors. In addition, we aim to enhance the analysis further by considering different factors affecting the venue impact such as the journal editorial board, the conference committee, and the publisher reputation. Most of the conferences have a limitation in regrades the number of the accepted paper. This factor can be considered in the future work for recommending venues in order to help users to identify the most appropriate conferences that can accept his article with the current number of pages.

We intend to identify the future research topic based on learning from the current trend. Which enables the system to recommend appropriate articles that mimic future topics, thus helping the researcher to adjust the direction of his current research, if he wishes, to comply with the future direction. We intend to identify the relation between the published given article and the extended version of it the citation impact analysis in order to enhance the recommendation outcomes. Although our recommender system offers a more accurate recommendation, it is computationally expensive to calculate the confidence score for each academic object in the citation network in a large-scale network. Consequently, we need to study more efficient and scalable techniques that can reduce the time needed to reach better results which helps improve the system performance for future work. In addition, we aim to enhance the algorithm that collects the metadata and more bibliographic information. The time we spent in collecting and organizing the data was very high. Thus, considering a standard bibliographic registering producer is one of the main targets in future development in the research industry. Accordingly, we need to work with the leaders in the research industry to build more reliable databases to facilitate future data analysis.

References

[1] A. El Saddik, “Digital Twins: The Convergence of Multimedia Technologies,” *IEEE Multimed.*, vol. 25, no. 2, pp. 87–92, Apr. 2018.

[2] R. Vogel, “The Visible Colleges of Management and Organization Studies: A Bibliometric Analysis of Academic Journals,” *Organ. Stud.*, vol. 33, no. 8, pp. 1015–1043, 2012.

[3] L. Bornmann and W. Marx, “Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts?,” *J. Informetr.*, vol. 9, no. 2, pp. 408–418, 2015.

[4] F. Musau, G. Wang, and M. B. Abdullahi, “Group formation with neighbor similarity trust in P2P E-commerce,” *Peer-to-Peer Netw. Appl.*, vol. 7, no. 4, pp. 295–310, 2014.

[5] R. Todeschini and A. Baccini, *Handbook of Bibliometric Indicators: Quantitative Tools for Studying and Evaluating Research*, vol. 128, no. 40. 2016.

[6] M. van Wesel, “Evaluation by Citation: Trends in Publication Behavior, Evaluation Criteria, and the Strive for High Impact Publications,” *Sci. Eng. Ethics*, vol. 22, no. 1, pp. 199–225, 2016.

[7] L. Waltman, “A review of the literature on citation impact indicators,” *Journal of Informetrics*, vol. 10, no. 2. pp. 365–391, 2016.

[8] H. F. H. Moed, *Citation analysis in research evaluation*, vol. 9. 2006.

[9] M. Thelwall and R. Fairclough, “The influence of time and discipline on the magnitude of correlations between citation counts and quality scores,” *J. Informetr.*, vol. 9, no. 3, pp. 529–

541, 2015.

[10] C. A. Sula and M. Miller, “Citations, contexts, and humanistic discourse: Toward automatic extraction and classification,” *Lit. Linguist. Comput.*, vol. 29, no. 3, pp. 452–464, 2014.

[11] A. M. Alshareef, M. F. Alhamid, and A. El Saddik, “Article Impact Value for Nearby Citation Network Analysis,” in *IEEE International Conference on Big Data and Smart Computing*, 2018, pp. 398–403.

[12] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, “Semantic Similarity from Natural Language and Ontology Analysis,” *Synth. Lect. Hum. Lang. Technol.*, vol. 8, no. 1, pp. 14–15, May 2015.

[13] F. Radicchi, A. Weissman, and J. Bollen, “Quantifying perceived impact of scientific publications,” *J. Informetr.*, vol. 11, no. 3, pp. 704–712, 2017.

[14] C. Hurter, “Analysis and Visualization of Citation Networks,” *Synth. Lect. Vis.*, vol. 3, no. 2, 2015.

[15] A. Calma and M. Davies, “Studies in Higher Education 1976 – 2013 : a retrospective using citation network analysis,” *Stud. High. Educ.*, vol. 40, no. 1, pp. 4–21, 2013.

[16] E. Garfield, “Citation analysis as a tool in journal evaluation.,” *Science*, vol. 178, no. 60, pp. 471–479, 1972.

[17] M. H. MacRoberts and B. R. MacRoberts, “Problems of citation analysis: A critical review,” *J. Am. Soc. Inf. Sci.*, vol. 40, no. 5, pp. 342–349, Sep. 1989.

[18] M. H. MacRoberts and B. R. MacRoberts, “Problems of citation analysis,” *Scientometrics*, vol. 36, no. 3, pp. 435–444, 1996.

[19] V. Batagelj, “Efficient Algorithms for Citation Network Analysis,” *Networks*, pp. 1–27, 2003.

[20] N. P. Hummon and P. Dereian, “Connectivity in a citation network: The development of DNA theory,” *Soc. Networks*, vol. 11, no. 1, pp. 39–63, 1989.

[21] K. Boyack and R. Klavans, “Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately?,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2389–2404, 2010.

[22] B. Gipp and J. Beel, “Citation Proximity Analysis (CPA) – A new approach for identifying related work based on Co-Citation Analysis,” *ISSI ’09 Proc. 12th Int. Conf. Sci. Inf.*, vol. 2, no. July, pp. 571–575, 2009.

[23] L. Leydesdorff, “Similarity measures, author cocitation analysis, and information theory,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 56, no. 7, pp. 769–772, 2005.

[24] N. J. Van Eck and L. Waltman, “Appropriate similarity measures for author co-citation analysis,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, no. 10, pp. 1653–1661, 2008.

[25] N. Wang, H. Liang, Y. Jia, S. Ge, Y. Xue, and Z. Wang, “Cloud computing research in the IS discipline: A citation/co-citation analysis,” *Decis. Support Syst.*, no. April, 2015.

[26] R. Moro, M. Vangel, and M. Bielikova, “Identification of Navigation Lead Candidates Using Citation and Co-Citation Analysis,” in *SOFSEM 2016: Theory and Practice of Computer Science: 42nd International Conference on Current Trends in Theory and Practice of Computer Science, Harrachov, Czech Republic, January 23-28, 2016, Proceedings*, vol. 9587, M. R. Freivalds, G. Engels, and B. Catania, Eds. Berlin, Heidelberg:

Springer Berlin Heidelberg, 2016, pp. 556–568.

[27] Y. K. Jeong, M. Song, and Y. Ding, “Content-based author co-citation analysis,” *J. Informetr.*, vol. 8, no. 1, pp. 197–211, 2014.

[28] Y. Ding, G. Zhang, T. Chambers, M. Song, X. Wang, and C. Zhai, “Content-based citation analysis: The next generation of citation analysis,” *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 9, pp. 1820–1833, 2014.

[29] S. Liu and C. Chen, “The proximity of co-citation,” *Scientometrics*, vol. 91, no. 2, pp. 495–511, 2012.

[30] Y. Jiang, A. Jia, Y. Feng, and D. Zhao, “Recommending academic papers via users’ reading purposes,” *Proc. sixth ACM Conf. Recomm. Syst. - RecSys ’12*, p. 241, 2012.

[31] J. Beel, S. Langer, M. Genzmehr, and A. Nürnberger, “Introducing Docear’s research paper recommender system,” *Proc. 13th ACM/IEEE-CS Jt. Conf. Digit. Libr. - JCDL ’13*, p. 459, 2013.

[32] D. C. Cavalcanti, R. B. C. Prudêncio, S. S. Pradhan, J. Y. Shah, and R. S. Pietrobon, “Good to be bad? Distinguishing between positive and negative citations in scientific impact,” *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, pp. 156–162, 2011.

[33] E. Fragkiadaki and G. Evangelidis, “Review of the indirect citations paradigm: theory and practice of the assessment of papers, authors and journals,” *Scientometrics*, pp. 1–28, 2013.

[34] A. Sidiropoulos and Y. Manolopoulos, “A citation-based system to assist prize awarding,” *ACM SIGMOD Rec.*, vol. 34, no. 4, pp. 54–60, 2005.

[35] A. Sidiropoulos and Y. Manolopoulos, “Generalized comparison of graph-based ranking

algorithms for publications and authors," *J. Syst. Softw.*, vol. 79, no. 12, pp. 1679–1700, 2006.

[36] N. Ma, J. Guan, and Y. Zhao, "Bringing PageRank to the citation analysis," *Inf. Process. Manag.*, vol. 44, no. 2, pp. 800–810, 2008.

[37] L. Bornmann, "Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics," *Scientometrics*, vol. 103, no. 3, pp. 1123–1144, 2015.

[38] L. Bornmann and R. Haunschild, "Citation score normalized by cited references (CSNCR): The introduction of a new citation impact indicator," *J. Informetr.*, vol. 10, no. 3, pp. 875–887, 2016.

[39] B. I. Hutchins, X. Yuan, J. M. Anderson, and G. M. Santangelo, "Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level," *PLoS Biol.*, vol. 14, no. 9, pp. 1–25, 2016.

[40] G. Abramo, T. Cicero, and C. A. D'Angelo, "A sensitivity analysis of researchers' productivity rankings to the time of citation observation," *J. Informetr.*, vol. 6, no. 2, pp. 192–201, 2012.

[41] C. Colliander, "A novel approach to citation normalization: A similarity-based method for creating reference sets," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 3, pp. 489–500, 2015.

[42] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breitinger, and A. Nürnberger, "Research paper recommender system evaluation: a quantitative literature survey," *RepSys*, vol. 20, no. April, pp. 1–35, 2013.

[43] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a

literature survey," *Int. J. Digit. Libr.*, no. June, 2015.

[44] J. Beel, "Towards Effective Research-Paper Recommender Systems and User Modeling based on Mind Maps," 2015.

[45] A. Vellino, "Usage-based vs. Citation-based Methods for Recommending Scholarly Research Articles," *arXiv:1303.7149*, 2013.

[46] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, 2004.

[47] J. Ben Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative Filtering Recommender Systems," in *The Adaptive Web*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 291–324.

[48] H. Small, "Co-Citation in Scientific Literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science*, vol. 24. pp. 265–269, 1973.

[49] K. Sugiyama and M.-Y. Kan, "Serendipitous Recommendation for Scholarly Papers," *Proc. 11th Annu. Int. ACM/IEEE Jt. Conf. Digit. Libr.*, pp. 307–310, 2011.

[50] O. Küçüktunç, E. Saule, K. Kaya, and Ü. V. Çatalyürek, "Recommendation on Academic Networks using Direction Aware Citation Analysis," *Vldb*, p. 10, 2012.

[51] Y. Liang, Q. Li, and T. Qian, "Finding Relevant Papers Based on Citation Relations," vol. 7923, no. November, 2011, pp. 403–414.

[52] N. Lao and W. W. Cohen, "Relational retrieval using a combination of path-constrained random walks," *Mach. Learn.*, vol. 81, no. 1, pp. 53–67, 2010.

[53] Z. Huang, W. Chung, T.-H. Ong, and H. Chen, “A graph-based recommender system for digital library,” *Proc. Second ACM/IEEE-CS Jt. Conf. Digit. Libr. - JCDL ’02*, p. 65, 2002.

[54] M. Baez, D. Mirylenka, and C. Parra, “Understanding and supporting search for scholarly knowledge,” *7th Eur. Comput. Sci. Summit*, 2011.

[55] M. Gori and A. Pucci, “Research paper recommender systems: A random-walk based approach,” *Proc. - 2006 IEEE/WIC/ACM Int. Conf. Web Intell. (WI 2006 Main Conf. Proceedings), WI’06*, pp. 778–781, 2007.

[56] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, “Context-aware citation recommendation,” *Proc. 19th Int. Conf. World wide web - WWW ’10*, p. 421, 2010.

[57] A. Woodruff, R. Gossweiler, J. Pitkow, E. H. Chi, and S. K. Card, “Enhancing a digital book with a reading recommender,” *Proc. SIGCHI Conf. Hum. factors Comput. Syst. - CHI ’00*, pp. 153–160, 2000.

[58] L. Rokach, P. Mitra, and S. Kataria, “A Supervised Learning Method for Context-Aware Citation Recommendation in a Large Corpus,” *LSDS-IR Large-Scale Distrib. Syst. Inf. Retr.*, pp. 17–22, 2013.

[59] C. C. Aggarwal, *Recommender Systems*, vol. 40, no. 3. Cham: Springer International Publishing, 2016.

[60] R. Burke, “Hybrid Recommender Systems : Survey and Experiments,” *User Model. UserAdapted Interact.*, vol. 12, no. 4, pp. 331–370, 2002.

[61] Q. He, D. Kifer, J. Pei, P. Mitra, and C. L. Giles, “Citation recommendation without author supervision,” *Proc. fourth ACM Int. Conf. Web search data Min. - WSDM ’11*, p. 755, 2011.

[62] W. Huang, Zhaohui Wu, P. Mitra, and C. L. Giles, “RefSeer: A citation recommendation system,” in *IEEE/ACM Joint Conference on Digital Libraries*, 2014, pp. 371–374.

[63] M. S. Pera and Y.-K. Ng, “A personalized recommendation system on scholarly publications,” *Proc. 20th ACM Int. Conf. Inf. Knowl. Manag. - CIKM ’11*, p. 2133, 2011.

[64] K. Sugiyama and M.-Y. Kan, “Exploiting Potential Citation Papers in Scholarly Paper Recommendation,” *Proc. 13th ACM/IEEE-CS Jt. Conf. Digit. Libr. - JCDL ’13*, pp. 153–162, 2013.

[65] O. Küçüktunç, E. Saule, K. Kaya, and Ü. V. Çatalyürek, “Towards a personalized, scalable, and exploratory academic recommendation service,” *Proc. 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. - ASONAM ’13*, no. August, pp. 636–641, 2013.

[66] M. Errami, J. D. Wren, J. M. Hicks, and H. R. Garner, “ETBLAST: A web server to identify expert reviewers, appropriate journals and similar publications,” *Nucleic Acids Res.*, vol. 35, no. SUPPL.2, pp. 12–15, 2007.

[67] M. J. Schuemie and J. A. Kors, “Jane: Suggesting journals, finding experts,” *Bioinformatics*, vol. 24, no. 5, pp. 727–728, 2008.

[68] S. Cohen and L. Ebel, “Recommending collaborators using keywords,” *Proc. 22nd Int. Conf. World Wide Web - WWW ’13 Companion*, pp. 959–962, 2013.

[69] P. S. Paul, V. Kumar, P. Choudhury, and S. Nandi, “Temporal analysis of author ranking using citation-collaboration network,” *2015 7th Int. Conf. Commun. Syst. Networks, COMSNETS 2015 - Proc.*, 2015.

[70] N. Kang, M. A. Doornenbal, and R. J. A. Schijvenaars, “Elsevier Journal Finder,” *Proc. 9th*

ACM Conf. Recomm. Syst. - RecSys '15, pp. 261–264, 2015.

[71] S. Yu, J. Liu, Z. Yang, Z. Chen, H. Jiang, A. Tolba, and F. Xia, "PAVE: Personalized Academic Venue recommendation Exploiting co-publication networks," *J. Netw. Comput. Appl.*, vol. 104, no. December 2017, pp. 38–47, 2018.

[72] A. Huang, "Similarity measures for text document clustering," *Proc. Sixth New Zeal.*, no. April, pp. 49–56, 2008.

[73] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 2, pp. 404–409, 2001.

[74] M. E. J. Newman, "Scientific collaboration networks: I. Network construction and fundamental results," *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 64, no. 1, pp. 1–8, 2001.

[75] A. Fiallos, K. Jimenes, C. Vaca, and X. Ochoa, "Scientific communities detection and analysis in the bibliographic database: SCOPUS," in *2017 4th International Conference on eDemocracy and eGovernment, ICEDEG 2017*, 2017, pp. 118–124.

[76] P. Tuire and E. Lehtinen, "Exploring Invisible Scientific Communities: Studying Networking Relations within an Educational Research Community. A Finnish Case," *High. Educ.*, vol. 42, no. 4, pp. 493–513, 2001.

[77] M. F. Alhamid, M. Eid, A. Alshareef, and A. El Saddik, "MMBIP: Biofeedback system design on Cloud-Oriented Architecture," *2012 IEEE Int. Symp. Robot. Sensors Environ. ROSE 2012 - Proc.*, no. November 2014, pp. 79–84, 2012.

[78] Crossref, "Crossref Digital Library." [Online]. Available:

https://www.crossref.org/.[Accessed: 11-Feb-2018]

[79] IEEE, “API Xplore Digital Library.” [Online]. Available: https://developer.ieee.org/. [Accessed: 11-Feb-2018].

[80] ACM, “ACM Digital Library.” [Online]. Available: https://dl.acm.org/. [Accessed: 11-May-2017].

[81] Schloss Dagstuhl LZI, “DBLP.” Available: https:// http://dblp.uni-trier.de/ [Accessed: 11-Feb-2018] .

[82] Z. Chen, F. Xia, H. Jiang, H. Liu, and J. Zhang, “Aver:Random Walk Based Academic Venue Recommendation,” in *Proceedings of the 24th International Conference on World Wide Web - WWW ’15 Companion*, 2015, pp. 579–584.

[83] “ORCID.” [Online]. Available: https://orcid.org/

[84] “Publons.” [Online]. Available: https://publons.com/about/home/.

[85] C. Basu, H. Hirsh, W. W. Cohen, and C. Nevill-Manning, “Technical paper recommendation: A study in combining multiple information sources,” *J. Artif. Intell. Res.*, vol. 14, pp. 241–262, 2001.

[86] F. Xia, Z. Chen, W. Wang, J. Li, and L. T. Yang, “MVCWalker: Random walk-based most valuable collaborators recommendation exploiting academic factors,” *IEEE Trans. Emerg. Top. Comput.*, vol. 2, no. 3, pp. 364–375, 2014.

[87] H. Tong, C. Faloutsos, and J. Y. Pan, “Random walk with restart: Fast solutions and applications,” *Knowl. Inf. Syst.*, vol. 14, no. 3, pp. 327–346, 2008.