Evaluation of Citation Graph Thematic Dataset Construction and Paper Filtering Methods for Research Literature Recommendation

Abdallah Farhat

A Thesis

in

The Department

of

Concordia Institute for Information Systems Engineering (CIISE)

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Quality Systems Engineering) at

Concordia University

Montréal, Québec, Canada

June 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Abdallah Farhat**

Entitled: **Evaluation of Citation Graph Thematic Dataset Construction and Paper Filtering Methods for Research Literature Recommendation**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Quality Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Yong Zeng

_____ External Examiner
Dr. Andrea Schiffauerova

_____ Examiner
Dr. Yong Zeng

_____ Supervisor
Dr. Chun Wang

Approved by _____
Martin D. Pugh, Chair
Department of Concordia Institute for Information Systems Engineering (CIISE)

_____ 2023

_____
Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Evaluation of Citation Graph Thematic Dataset Construction and Paper Filtering Methods for Research Literature Recommendation

Abdallah Farhat

One of the main challenges faced by new researchers is immersing themselves in the existing literature relevant to their field of interest. The vastness and continuous growth of knowledge in their field can be overwhelming, making it difficult to identify the most pertinent research papers within their research themes. To address this issue, research paper recommender systems have emerged as valuable tools. These systems allow researchers to find relevant papers based on their specific interests or research themes by analyzing various aspects such as titles, abstracts, and full texts. The quality of the dataset used is crucial for the development, testing, and refinement of these systems to ensure optimal results. Dataset quality directly impacts the accuracy and reliability of a recommender system. In this thesis, I propose a novel approach for constructing datasets using citation graph networks. These networks consist of nodes representing research papers and edges representing citations between them. By leveraging citation graph networks, we gain a more comprehensive understanding of the relationships and influences among different papers compared to traditional methods that rely solely on keyword searches. To evaluate the effectiveness of the citation graph network method, I compared it with the traditional keyword search approach for dataset construction. Additionally, I assessed the effectiveness of three recommender system algorithms: user-based collaborative filtering, combined with PageRank and personalized PageRank algorithms. The experimental findings provide clear evidence that utilizing citation graph network datasets significantly enhances the efficacy of research paper recommender systems. This improvement simplifies the process of finding relevant literature for researchers, potentially accelerating scientific discovery.

Acknowledgments

I am grateful to Professor Wang, my esteemed supervisor, for his invaluable guidance and unwavering support throughout my academic career. As a result of his mentorship, I have not only gained valuable knowledge in terms of critical thinking, efficient work, and effective writing, but I have also gained a deeper understanding of the subject matter. My thesis research has been shaped by Prof. Wang's expertise and dedication. He has provided me with unwavering commitment and guidance throughout the entire process.

My sincere gratitude is also extended to Concordia University and MITACS for their financial support, which helped me complete my thesis and pursue my research. Their contribution has helped facilitate my academic goals.

I am also thankful to my parents for their constant encouragement, support, and belief in my educational endeavors. Invaluable sources of inspiration and strength have been their unwavering faith in me and continuous motivation. Those who have shaped me as a person, through love, guidance, and sacrifice, are deeply appreciated. My future endeavors will be influenced by their influence.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Efficiently analyzing the literature requires researchers to identify seminal papers that have inspired investigations in their field. This ensures that their readings offer a comprehensive understanding of the topic at hand. The significance of providing new researchers with a high-quality reading list cannot be overstated Ekstrand et al. (2010). Such a foundation of knowledge enables them to grasp key concepts, debates, and approaches within their field. Without this foundation, meaningful engagement with the subject matter becomes challenging. A well-curated reading list offers researchers the most relevant and useful texts in their respective fields. This is particularly valuable for those who are just starting out and may be unfamiliar with key texts and studies. By exposing students to different perspectives and arguments in their field, a reading list encourages critical thinking and the development of analytical skills. It enables researchers to identify assumptions, strengths, and weaknesses, fostering greater engagement and confidence in their own research McGuinn, Stone, Sharman, and Davison (2017). By providing researchers with a thoughtfully selected reading list, they can avoid mistakes and overcome obstacles in their field. Directing them to appropriate resources saves time and prevents wasted effort on irrelevant materials. Numerous research paper recommender systems have been introduced to generate high-quality reading lists. These systems employ advanced algorithms and techniques to analyze literature and user preferences. By considering a researcher's interests and previous reading history, these systems can identify influential papers, detect patterns and trends, and offer personalized recommendations Beel et al. (2013). Consequently, researchers can save time, effort, and resources while gaining access

to a comprehensive, diverse, and up-to-date reading list that provides valuable insights into their research theme.

1.1 Research Paper Recommender Systems

Research paper recommender systems are tools or algorithms designed to provide users with relevant academic papers based on their interests, preferences, or research requirements Beel et al. (2013). These systems enable users to navigate the vast and ever-growing body of scientific literature more efficiently, ultimately allowing them to discover new insights and enhance their understanding of their research areas.

A research paper recommender system comprises two critical components: the research papers dataset and the recommender algorithm Beel, Gipp, Langer, and Breitinger (2016). The recommender algorithm generates personalized paper recommendations by analyzing user actions, paper content, citation connections, and user profiles, employing methods such as collaborative filtering, content-based filtering, graph-based techniques, or a combination of these approaches Aggarwal et al. (2016).

The dataset, on the other hand, refers to a comprehensive collection of research papers or metadata. In order to develop an effective research paper recommender system, it is crucial to have research papers of high quality and relevance to a particular theme Beel et al. (2016). Theme-specific or thematic datasets consist of research papers that focus on a specific aspect within a broader research field or topic, ensuring accurate and relevant recommendations. By concentrating on a particular theme, the dataset ensures that the system is robust, reliable, and adaptable to user preferences. It also addresses the unique nuances and perspectives within the chosen theme, providing valuable suggestions that enhance users' understanding and knowledge of the subject matter.

Constructing a research paper recommender system heavily relies on the dataset as it provides the necessary information to generate relevant recommendations. Currently, authors of research

paper recommender systems gather data by identifying appropriate sources and metadata, and manually curating topic-specific datasets primarily using keywords. Some popular sources include web-based search engines (e.g., Google Scholar[1], Microsoft Academic[2]), citation databases (e.g., Web of Science[3], Scopus[4]), academic databases (e.g., IEEE Xplore[5], PubMed[6]), digital repositories (e.g., arXiv[7], SSRN[8]), and domain-specific academic journals.

Web-based search engines enable users to locate academic and non-academic sources from various disciplines, utilizing keywords and advanced search options Valente, Holanda, Mariano, Furuta, and Da Silva (2022). Citation databases track and index citations in academic publications, providing citation counts, journal impact factors, and other bibliometric information. Researchers can use these databases to identify connections between publications, authors, and research topics across disciplines Valente et al. (2022). Academic databases offer scholarly literature in specific disciplines, providing advanced search options, tailored filters, and full-text access Valente et al. (2022). In digital repositories, researchers can share preprints, working papers, and published articles with their peers, facilitating the dissemination of research Wakeling et al. (2019). Academic journals regularly publish scholarly research within specific fields of study, including articles, reviews, and commentary authored by experts Wakeling et al. (2019).

Constructing theme-specific datasets for research paper recommenders can pose challenges due to issues related to data sources, search limitations, keyword ambiguity, time constraints, diversity, biases, citation analysis, and maintaining data quality and consistency across various sources. Overcoming these obstacles is crucial for building a robust recommender system that provides researchers with relevant and valuable paper recommendations Beel et al. (2016). By combining multiple data sources and leveraging citation networks, researchers can create more focused and comprehensive theme-specific datasets for research paper recommenders.

[1]https://scholar.google.com/
[2]https://academic.microsoft.com/
[3]https://clarivate.com/webofsciencegroup/solutions/web-of-science/
[4]https://www.scopus.com/
[5]https://ieeexplore.ieee.org/
[6]https://pubmed.ncbi.nlm.nih.gov/
[7]https://arxiv.org/
[8]https://www.ssrn.com/

1.2 Limitations of Existing Research Papers Dataset Construction Methods

Research paper datasets can be constructed from a variety of search engines and citation databases, including Google Scholar, PubMed, Web of Science, Scopus, and IEEE Xplore. To construct theme-specific research paper datasets, it's crucial for researchers to accurately predict or select pertinent keywords. Generally, these search engines and citation databases seek out documents bearing textual similarities to the input keywords. This approach makes it challenging for researchers to pinpoint papers directly related to their topics. A more exhaustive search is feasible by combining techniques such as keyword searches, citation analysis, and exploration of related papers. However, these methods, though comprehensive, are time-consuming and can become inefficient when handling vast amounts of information. Keyword-based searches can miss identical concepts named differently across fields or uncover unrelated ones. Additionally, keyword ambiguity may lead to the inclusion of irrelevant papers. Identifying foundational or highly-cited papers through these searches is also a challenge, as they do not prioritize papers based on their significance or influence. Moreover, keyword search results do not shed light on the relationships between papers, such as citations, thereby hindering the understanding of a research area's structure.

An alternate approach involves constructing research paper datasets using online academic databases and digital repositories. These repositories, curated by publishers, organizations, or institutions, contain preselected collections of research papers and scholarly publications. The content is typically peer-reviewed to assure quality and credibility and spans various subjects and disciplines. Users can search these repositories to find literature pertinent to their interests. While research paper datasets can be constructed using predefined repositories, they may offer limited coverage. Repositories focused on specific research fields or publishers might not encompass all relevant publications within a theme, resulting in the omission of essential papers from other sources, leading to a sparsity problem. In the thesis context, the sparsity problem is a concern that arises when there is insufficient data, particularly regarding interactions between research paper citations within a dataset. Furthermore, these repositories may offer limited search functionality in terms of keywords, metadata, or

search algorithms, making the discovery of relevant papers more difficult. The most recent publications and developments may not be featured in some predefined digital repositories if they are not regularly updated. It's also worth noting that online digital research paper repositories might not be applicable to all research areas, and their database construction process may lack transparency.

Ekstrand et al. (2010) used the ACM Computing Surveys (CSUR) journal reference list of papers to test the accuracy of their recommender systems algorithms. Journal papers are academic publications presenting original research findings or a review of existing research in a specific field. Typically, journal papers are peer-reviewed, which means they are evaluated by experts in the field before being published. However, the use of a single journal for building a research paper dataset can lead to several problems, such as a lack of diversity in ideas, theories, and findings. Furthermore, a journal may favor certain methodologies, theoretical frameworks, or research questions, which can introduce biases that affect the validity and generalization of the conclusions drawn from the dataset. Depending solely on one journal might also result in missing influential papers published elsewhere, resulting in a limited understanding of the research field's development. The dataset might also include lower quality papers if the chosen journal has a lower impact factor or less stringent peer-review process, which can affect the accuracy and reliability of any recommender system built upon it.

1.3 Contributions of This Thesis

This thesis introduces a methodology for constructing thematic research paper datasets via citation graph networks. It begins with a compact set of papers, which can comprise recent surveys on the theme or influential technical papers on the topic. Subsequently, datasets are constructed using citation graph networks, enabling authors to identify related papers based on citation patterns between papers. This includes references within the original publication and citations in the referenced paper. This approach allows authors to create a comprehensive and high-quality thematic dataset that accurately mirrors the research landscape of a specific theme. The method alleviates the sparsity issue while addressing challenges associated with data quality and availability, thereby enabling recommender systems to suggest top-notch research papers to new researchers in a specific

field of research. Moreover, I assess the performance of various research paper recommender algorithms on the datasets produced by the proposed citation graph-based dataset generation method and those created through keyword searches, a technique widely used in the literature.

1.4 Thesis Outline

In the following chapters, I will outline the design and research problems that form the basis of this study. I have carefully defined the scope of this research to ensure a focused and achievable approach, along with precise definitions of key terms to maintain consistency and clarity. The remaining thesis is organized as follows:

Chapter 2 provides an overview of the literature on recommender system algorithms employed in research paper recommendations. Moving forward, Chapter 3 discusses various approaches to constructing research paper datasets, highlighting their limitations. In Chapter 4, I introduce the proposed citation graph network algorithm for constructing thematic research paper datasets. Chapter 5 presents the methodology used to evaluate the construction of the research paper thematic dataset. Finally, Chapter 6 concludes the thesis and outlines potential avenues for future research.

Chapter 2

Recommender Systems for Research Paper Recommendation

The purpose of recommender systems is to provide users with suggestions and recommendations for products or content they are likely to find useful. These recommendations play a significant role in decision-making processes related to shopping, entertainment, and information consumption Patel, Desai, and Panchal (2017). Researchers can also benefit from recommender systems, especially when navigating through the vast amount of academic literature available. The initial research paper recommender system, developed by Giles, Bollacker, and Lawrence (1998), was part of the CiteSeer website project. This system utilized heuristics and web search engines to locate and download relevant documents. However, researchers often struggle to assess the relevance of information found through search engines like Google Scholar. In this context, recommender systems can serve as invaluable tools for researchers Bulut, Kaya, and Kaya (2019). In recent years, various approaches have been proposed and implemented to address the challenge of guiding researchers through the extensive academic literature in the field of research paper recommender systems. This literature review provides an overview of the key methodologies employed in recommender systems, discussing their strengths, limitations, and potential improvements. Content-based filtering (CBF) is a methodology that recommends papers based on their similarity to previously read or liked papers by the user. Collaborative filtering (CF) techniques, on the other hand, generate recommendations

based on user behavior patterns, including user-based and item-based CF approaches Aggarwal et al. (2016). Furthermore, the review explores graph-based methods that leverage citation networks to identify important papers and authors within a specific field. the review also delves into hybrid approaches that combine multiple techniques to improve recommendation performance and overcome limitations associated with individual methods. By examining these different methodologies and their applications in research paper recommender systems, I aim to provide a comprehensive understanding of the current state-of-the-art. This, in turn, will inspire further research and development efforts to enhance the support provided to researchers in their quest for relevant literature. In addition to providing a comprehensive review of recommender systems, this study delves further into research paper recommender systems, and focusing on three specific algorithms that serve as the foundation for our experiments.

2.1 Summary of Recommender Systems

2.1.1 Collaborative Filtering

Collaborative filtering is a technique used to make product recommendations by leveraging the feedback provided by other users. The underlying idea is that two users, A and B, who rate certain items similarly are likely to have similar interests. Collaborative filtering recommends products to User A based on User B's purchase history, if those items are not yet in User A's record Schafer, Frankowski, Herlocker, and Sen (2007). Memory-based collaborative filtering, which is the earliest form of collaborative filtering, can be classified into two subcategories: user-based and item-based. These methods use past user ratings to determine the similarity between two users or items. To do this, a similarity metric is defined, and the most comparable users or items are identified to suggest new, undiscovered items. Another approach is model-based collaborative filtering, which employs machine learning algorithms to predict a user's rating of an unrated item. This approach uses statistical models to represent user-item relationships and can generate accurate predictions even when data is sparse. The combination of both memory-based and model-based collaborative filtering methods can produce more robust and accurate recommendations Aggarwal et al. (2016).

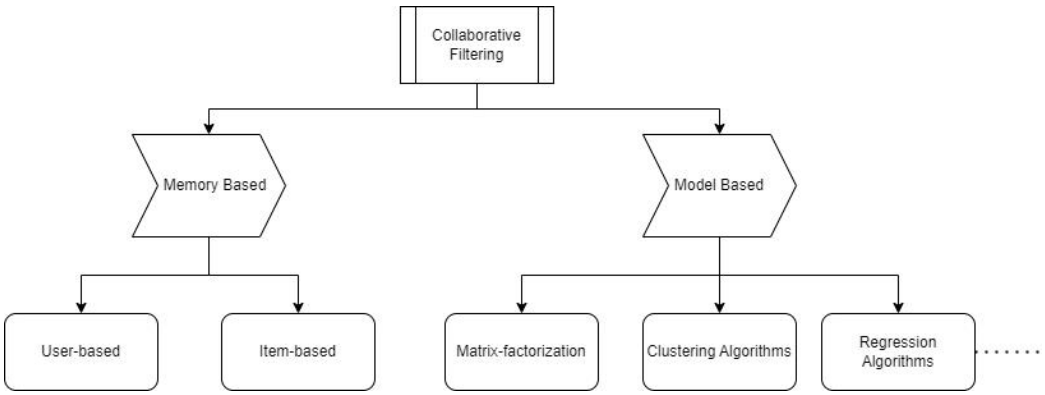


Figure 2.1: Collaborative filtering taxonomy

2.1.2 Content-Based Filtering

Content-based filtering is a recommender system approach that generates personalized recommendations for users based on the features and attributes of items they have previously interacted with or expressed interest in. The underlying principle of content-based filtering is that users are likely to be interested in items similar to those they have previously liked or engaged with. In this approach, user profiles are created by analyzing the content of the items they have interacted with, and these profiles are then used to recommend new items with similar content or characteristics Pazzani and Billsus (2007).

2.1.3 Graph-based Recommender Systems

In directed graphs, numerous algorithms have been developed to calculate node importance using the information contained within the graph structure. Although many of these algorithms were initially designed for web pages connected through hyperlinks, they can also be applied to other directed graph scenarios, such as citation networks Küçüktunç, Saule, Kaya, and Çatalyürek (2012). The primary focus of the graph-based approach is the construction of the graph, which can be built using citation networks, social networks, or other types of networks. PageRank is a widely used algorithm in this context. It operates on the premise that high-quality, authoritative pages will frequently receive links, particularly from other reputable pages, leading to better connectivity. PageRank utilizes a random walk to score the quality or authority of web pages Page, Brin, Motwani, and Winograd (1999). The HITS algorithm produces two graph rankings: the hub score, which

evaluates a node's ability to connect to trustworthy sources, and the authority score, which assesses a node's authority based on the extent to which it is linked by other nodes. These scores are computed as the fixed point of mutually reinforcing links Kleinberg (1999). A combination of text-based information retrieval methods and graph-based ranking algorithms is used by Kleinberg (1999), where the algorithm first gathers pages based on a query string search and constructs a website graph from the highest-ranked pages. The website graph includes pages that point to, and are pointed to by, the highest-ranked. HITS authority scores are then calculated, and the pages with the highest scores are selected.

2.1.4 Stereotyping Recommender Systems

Stereotyping represents one of the earliest approaches to user modeling and recommender systems. It was first introduced by Rich in the Grundy recommender system Rich (1979), which suggested novels to users. Rich drew inspiration from psychological stereotypes, which enabled psychologists to make quick judgments about individuals based on a few attributes. Rich referred to these stereotypes as "facets," which are sets of characteristics. For example, Grundy assumed that male users possess a relatively high tolerance for violence and suffering, along with a preference for thrills, suspense, fast-paced plots, and a disinterest in romance. As a result, Grundy recommended books that had been manually categorized to align with these facets.

2.1.5 Popularity-Based Recommender System

A popularity-based recommender system is an approach in which items are recommended to users based on their overall popularity or general appeal within a user base, rather than focusing on individual preferences Jannach, Zanker, Felfernig, and Friedrich (2010). These systems typically rely on aggregated information, such as the number of views, likes, or ratings, to determine the most popular items. While popularity-based recommender systems may not provide highly personalized suggestions, they can offer users valuable recommenders by capturing the preferences of a broader audience.

2.1.6 Hybrid Recommender System

A hybrid recommendation system is an approach that combines multiple recommendation techniques, such as content-based filtering, collaborative filtering, and knowledge-based methods, to provide more accurate and diverse suggestions to users Burke (2002). By integrating different algorithms, hybrid recommendation systems can leverage the strengths of each method while compensating for their weaknesses, potentially improving the overall quality of recommendations. Hybrid systems have been particularly effective in addressing common challenges like the cold-start problem and increasing the diversity of recommendations.

2.2 Research Paper Recommendation Approaches

Academic research stands out highly among the various domains in which recommender systems have demonstrated significant progress Beel et al. (2016). Research paper recommender systems have revolutionized the way researchers, students, and academics access the vast body of knowledge. Due to the exponential growth of academic publications, finding relevant and influential papers has become increasingly difficult Ekstrand et al. (2010). The issue of information overload has been addressed by research paper recommender systems. By analyzing various data points associated with each paper, they mitigate the challenge of finding relevant research resources. Researchers can easily navigate academic literature with the help of these systems by tailoring their recommendations, allowing them to discover relevant papers more efficiently. In this context, I highlight the approaches used to recommend research papers, some of which will be used to assess the thematic dataset construction approach presented in Chapter 4.

2.2.1 TFIDF CBF

CBF can represent scientific papers by extracting keywords from the title, abstract, and main text. The widely-used term frequency-inverse document frequenc (TFIDF) model Jomsri, Sanguansintukul, and Choochaiwattana (2010) converts paper content into weighted items. Ferrara, Pudota, and Tasso (2011) utilized CBF to compare two documents by extracting keyphrases, forming weight vectors of terms based on TFIDF, and determining the relevance of a newly extracted

paper to the user's preferred paper. CBF methods are commonly employed for recommending research papers. Dong, Tokarchuk, and Ma (2009) proposed three CBF techniques - pure, combined, and separated - for making research paper recommendations. Pure CBF identifies similar papers based on their text similarity to the active paper, utilizing TFIDF scores, and recommends the most closely related ones. Combined CBF consolidates the text of all citations into one large text block and recommends papers in the same manner as pure CBF. Separated CBF, on the other hand, examines the text of the target paper separately from the text of all its citations. The combined CBF method yielded the best results, as measured by the top-20 hit percentage evaluation metric.

2.2.2 CCIDF CBF

Inspired by TFIDF, Co-citation Inverse Document Frequency (CCIDF) was firstly introduced by Giles et al. (1998). CCIDF is a technique that combines co-citation and co-occurrence information to measure the importance of terms and calculate their weights in the recommendation process, particularly in the domain of research paper recommendation systems. Co-citation refers to the frequency with which two documents are cited together, while co-occurrence denotes the frequency of two terms appearing together in a document or a set of documents. By incorporating both co-citation and co-occurrence information, CCIDF aims to improve the effectiveness of content-based filtering methods and provide more accurate and relevant recommendations. Beel, Breitinger, and Langer (2017) conducted a comparison between CCIDF and CC (co-citation) alone. Interestingly, no significant differences were observed in the recommendation outcomes of the two approaches when evaluated using click-through rate in an online assessment. Tanner, Akbas, and Hasan (2019) expanded upon the existing CCIDF method by incorporating the number of citations as a weight for the corresponding edge in the unweighted citation network, which led to an improvement in overall recommendation performance.

2.2.3 Collaborative Filtering

Collaboration filtering is an established technique in recommender systems Schafer et al. (2007). According to CF, if two users have agreed or rated items similarly in the past, they are likely to have similar preferences in the future. CF systems use ratings matrices, also referred to as utility matrices,

to capture users' preferences. A dataset of recommendation matrices is shown in these matrices. The ratings matrix is divided into two dimensions: one for users, one for items (such as research papers). A user's rating or interaction with each item is represented in each cell of the matrix. There are however many empty cells and sparse matrices because not all users interact with every item. For recommending research papers, CF methods are widely used Beel et al. (2016), and rating matrices can be constructed in various ways. In collaborative filtering, memory-based techniques and model-based techniques are the two most common types. Using these methods, recommendations are generated by analyzing the ratings matrix and analyzing user-item similarities. CF has proven to be a powerful tool for recommending research papers, and the ratings matrix plays a crucial role in making accurate and relevant recommendations.

Construct Ratings Matrix

Memory-based and Model-based collaborative filtering requires a user-item matrix containing historical user ratings on items Schafer et al. (2007). Several algorithms have been proposed to recommend research papers for researchers, with memory-based CF being one of the most popular. However, to apply CF, a well-defined ratings matrix is needed. For research paper recommender systems, the ratings matrix can be created by collecting researchers' past opinions about papers as ratings Bogers and Van den Bosch (2008). Papers will be items, and researchers will be users. New users without any reading history may cause the cold start problem in this approach. User participation is crucial for CF, but users often lack motivation, leading to the cold-start problem Casadevall and Fang (2010). This problem arises when new users, items, or communities emerge and the algorithm struggles to generate recommendations due to insufficient information about user interests. The citation web can also be used to construct the ratings matrix. McNee et al. (2002) presented three methods using the citation web: treating users as papers and items as citations, considering users and items as citations and give a rating of 1 when they co-cite each other, or treating paper authors as users and items as their citations. However, this last approach may face generality problems since many authors publish in different fields. Ekstrand et al. (2010) used all citation web papers as users and items, with each user voting 1 for the item it cites. Similarly, Torres, McNee, Abel, Konstan, and Riedl (2004) built a ratings matrix based on paper citations and papers

in the CiteSeer repository. After creating the ratings matrix, CF, CBF, and hybrid combinations of CF and CBF recommendation systems have been applied. The top 20 hit percentage evaluation metric showed that pure user-based CF, which uses citations from a target paper as inputs, performed the best.

Memory-Based CF

Once the ratings matrix is clearly defined, user-based or item-based collaborative filtering can be easily implemented by selecting a target user. Torres et al. (2004) applied collaborative filtering, content-based filtering, and hybrid combinations of CF and CBF in their recommendation systems. Their evaluation using the top 20 hit percentage metric demonstrated that pure user-based CF, which takes citations from a target paper as inputs, performed optimally. Ekstrand et al. (2010) implemented standard item-based CF, as introduced by Karypis (2001), to generate reading lists for researchers, assuming they provided a set of important target papers. They compared this approach with a hybrid of PageRank and item-based CF, which improved the results when evaluated using the half-life utility metrics. Furthermore, McNee et al. (2002) compared item-based and user-based collaborative filtering. Item-based filtering took a set of paper citations as input and recommended similar citations, while user-based filtering used a single paper as the target user and provided recommended citations. Both user-based and item-based methods showed similarly favorable results using the rank metric.

Model-Based CF

Salakhutdinov and Mnih (2008) proposed Probabilistic Matrix Factorization (PMF) to address the scalability and sparsity issues in collaborative filtering. They demonstrated the efficacy of their approach on the SVD++ dataset, which included research paper ratings. C. Wang and Blei (2011) introduced the Collaborative Topic Regression (CTR) model, which combined Latent Dirichlet Allocation (LDA) and matrix factorization. They demonstrated the effectiveness of their method on CiteULike dataset which contained scientific articles. H. Wang, Wang, and Yeung (2015) proposed the Collaborative Deep Learning (CDL) model, which combined a Stacked Denoising Autoencoder (SDAE) with a matrix factorization technique. They demonstrated the effectiveness of their model

on the CiteULike dataset also, showing significant improvements in research paper recommendations.

2.2.4 Graph-Based Methods

In the field of research paper recommendations, relationships between researchers and papers or between the papers themselves can be regarded as edges between nodes. Graph-based (GB) ranking techniques are often employed to make research article recommendations. To leverage various graph-based methods, edges between nodes can be represented by the relationships between researchers and papers, or between the papers themselves. A citation graph must be constructed initially, with each node representing an article and the edges signifying citations between the papers White and Smyth (2003). As introduced by W. Lu, Janssen, Milios, Japkowicz, and Zhang (2007), a local citation graph is generated for each paper in the dataset using the papers' reference lists, which employs the authority vector metric approach. The HITS algorithm is used to compute the authority weights for each node in the citation graphs. Subsequently, a vector comprising nodes from the union of the two graphs is utilized to represent each paper, with the nodes authority weights in each paper's local citation graph serving as its elements. This vector is then used to calculate similarity scores between papers. The HITS algorithm can identify hub and authority papers in a document set Chang, Cohn, McCallum, et al. (2000), but its performance is mixed. It struggles to rank authorities in the presence of separate "communities" within a topic. An ideal system would learn to identify authorities by combining citation structure and user input.

2.2.5 Stereotype and Popularity Methods

In stereotype-based recommendation systems, individuals are judged based on a few characteristics. Beel, Langer, Kapitsaki, Breitinger, and Gipp (2015) employed a stereotype model as a fallback option when other recommendation approaches failed to provide satisfactory results. However, stereotype models possess two significant drawbacks: they can inaccurately group users, and they are labor-intensive, as each stereotype typically requires manual classification. On the other hand, global relevance recommendation systems suggest items with the highest overall relevance, often used as an additional ranking factor. Although no study has exclusively utilized popularity

relevance in the domain of research paper recommendations, it can be employed to re-rank the final list of suggested citations using PageRank scores, citation counts, and other factors Bethard and Jurafsky (2010).

2.2.6 Hybrid Methods

To enhance accuracy and performance, a scientific paper recommender system can use multiple recommendation techniques to provide personalized paper suggestions to researchers Tsolakidis, Triperina, Sgouropoulou, and Christidis (2016). One such technique, called combined CF and CBF defined by Dong et al. (2009), involves refining the reading list by identifying users with similar tastes through CF and subsequently applying content-based filtering to the list. Torres et al. (2004) propose several hybrid algorithms comprises two independent modules, a Collaborative Filtering and a Content-Based Filtering, which generate recommendations using different inputs - the text of the active paper for CBF and its citations for CF. Various hybrid algorithms, such as CF-CBF Separated, CF-CBF Combined, CBF Separated-CF, CBF Combined-CF, and Fusion, combine these recommendations in different ways. They utilize approaches like weighting and sorting based on the order of CF recommendations, aggregating the text of all recommendations from CF as input, augmenting the active paper's set of citations, and running both recommender modules in parallel to merge the results. By leveraging both CF and CBF modules, these hybrid algorithms offer a range of options for generating recommendations in research paper recommender systems. Ekstrand et al. (2010) introduced a hybrid approach that combines graph-based algorithms, collaborative filtering, and content-based filtering for research paper recommender systems. The authors developed a citation graph using web paper citations, and after constructing the graph, they determined node importance scores using PageRank, HITS, and SALSA algorithms. These scores were then employed to adjust the ratings matrix prior to applying the item-based collaborative filtering algorithm. The best performance, as measured by the half-life utility metric, was attained when PageRank scores influenced the item-based CF.

2.3 Commonly Used Research Paper Recommender Algorithms

From the mentioned research papers recommender systems, I have chose three algorithms to recommend a final reading list of research papers in order to evaluate the constructed thematic dataset described in 4. I compared the performances of the three algorithms, user-base collaborative filtering, PageRank combined with user-based CF, and personalized PageRank combined with user-based CF. Each algorithm takes an input of a target paper (user) and recommend a ranked list of items (citations) as output as shown in 2.2.

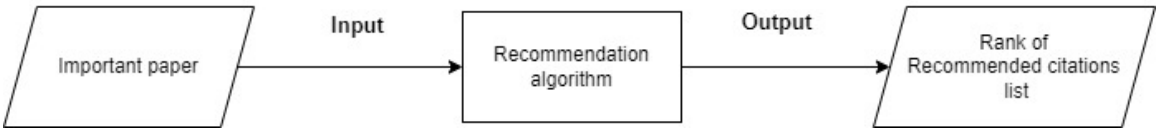


Figure 2.2: Recommendation Algorithms

In this section, I provide a comprehensive explanation of the implementation of each of the three methods.

2.3.1 User-Based Collaborative Filtering

I have applied user-based collaborative filtering as described by Aggarwal et al. (2016) with normalization to give users who cited fewer items a greater impact on the similarity of the items they have referred to. As shown in 2.3 This approach gives greater weight to users who have cited fewer items, enhancing their impact on the similarity of the referred items. The algorithm calculates user similarity and recommends items based on the preferences of similar users.

The process can be broken down into the following steps as stated by Aggarwal et al. (2016):

(1) Find the K most similar group of users to the target user u, by calculating the similarity scores between u and all the remaining users. To find the similarity scores between users, I can use various similarity or distance metrics, the most used are pearson correlation and cosine similarity.

- I used cosine similarity metric, to find the similarity scores since our data set is binary

(0,1).

$$Sim(u,v) = Cosine(u,v) = \frac{\sum_{k \in I_u \cap I_v} r_{uk} . r_{vk}}{\sqrt{\sum_{k \in I_u \cap I_v} r_{uk}^2} \cdot \sqrt{\sum_{k \in I_u \cap I_v} r_{vk}^2}} \tag{1}$$

I_u: The set of items indices for which ratings have been specified by user u.

I_v: The set of items indices for which ratings have been specified by user v.

r_{uk}: The rating given by user u for item k.

r_{vk}: The rating given by user v for item k.

$Sim(u,v)$: The the similarity between users u and v.

- Select the users with the K highest cosine similarity scores to our target user u.
- Store the similarity score for each peer user.

If our test set consists of multiple users, I do not consider them as similar to each other.

(2) Calculate the weighted average for each item $\in R_{m \times n}$.

$$\hat{r}_{ui} = \frac{\sum_{v \in P_u(i)} Sim(u,v) . r_{vi}}{\sum_{v \in P_u(i)} |Sim(u,v)|} \tag{2}$$

Where $P_u(i)$ represent the set of users who have rated item i and are similar to user u. And, $\sum_{v \in P_u(i)}$ is the sum over all v in the set $P_u(i)$.

I multiply each item vote (0 or 1) with the corresponding peer user cosine similarity and then divide the sum of the multiplication with the sum of cosine similarities of the peer users.

(3) Identify the final Top N recommended items.

I rank the items by their calculated weighted average to get The Top-N of recommended items to a target user u. The items already purchased by u are excluded from the final recommended rank list.

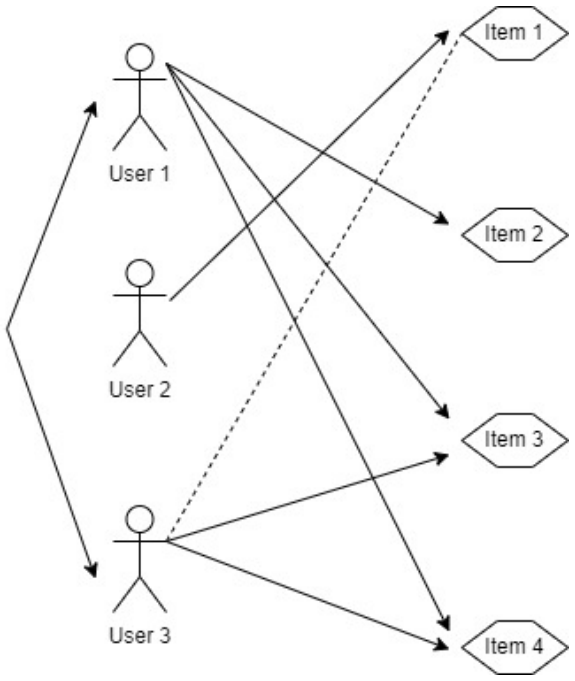


Figure 2.3: User Based CF

2.3.2 PageRank User-based Collaborative Filtering

Certain papers may hold greater significance than others due to their connections within the citation graph. I utilize the PageRank algorithm, as presented by Ekstrand et al. (2010), to compute the importance scores of the selected papers in the ratings matrix. Both chosen items and users accumulate nodes in the citation graph. When applied to a research papers citation graph, the PageRank algorithm ranks papers according to the importance of the citations they receive. This approach assigns a numerical value to each paper in the citation graph, indicating its prominence in the research papers network. The underlying concept of PageRank is that a paper cited by numerous influential papers is likely important itself. After determining the scores of the nodes in the citation graph network, I assign weights to the papers based on their graph importance score prior to creating the user-item similarity matrix. I replace the user-based vector normalization with a normalization procedure that multiplies each paper's citation vector by its importance score r(u), yielding r'_u = r(u)u.

$$PR(p) = \frac{1-d}{N} + d \sum_{q \in C_p} \frac{PR(q)}{L(q)} \tag{3}$$

Where:

- $PR(p)$ is the PageRank of paper p, which measures its importance in the citation network.
- d is the damping factor with value between 0 and 1, This factor represents the probability of a researcher continuing to follow citations in the graph.
- N is the total number of papers (nodes) in the citation graph.
- C_p is the set of papers that reference paper p (i.e., papers with outgoing citation links pointing to p).
- $L(q)$ is the number of outgoing references links from paper q.
- The summation $\sum_{q \in C_p}$ iterates over all papers q in the set C_p.

2.3.3 Personalized PageRank User-Based Collaborative Filtering

This algorithm is suggested by Ekstrand et al. (2010) without using it in any of their experiments for computational complexity with the motivation for experimenting is that they might be more effective in pinpointing the crucial papers that are highly pertinent to the user's inquiry. I tested the algorithm and compared the results with the other described algorithms. The way I used it as described by White and Smyth (2003) with giving priors to the pagerank basically gives even more importance to the target user in the citation graph network.

$$PR(p) = (1-d) \cdot P(p) + d \sum_{q \in C_p} \frac{PR(q)}{L(q)} \tag{4}$$

Where:

- $P(p)$ is the prior probability of paper p, which represents prior knowledge or preference for the paper in question.

The algorithm repeatedly calculates PageRank scores for all papers until a stable state is achieved. These PageRank scores are then used to rank papers by their significance in the citation network, taking into account both the citation structure and any prior preferences or knowledge. This approach helps identify impactful research and recommend papers that are more relevant to a specific context or field.

2.4 Summary and Discussion

In summary, the research paper recommender domain has witnessed the evolution and implementation of various methods aimed at facilitating the recommendation process. The selection of users and items in the ratings matrix can vary among different authors. In this study, I propose a novel approach for constructing the ratings matrix that takes into account the impact of the number of references and citing articles for each paper within the dataset. To evaluate the performance of hybrid methods, I assess two of the top-performing approaches suggested by Ekstrand et al. (2010), with minor modifications. For instance, I utilize user-based collaborative filtering instead of item-based CF to better leverage the similarity data among users. Additionally, I incorporate their proposed algorithm for combining personalized PageRank, a graph-based method, with CF. This combination was not attempted by the authors due to computational complexity. The objective is to evaluate the effectiveness of this approach in terms of overall recommendation performance. Table 2.1 provides an overview of the research paper recommender systems reviewed in this study, highlighting their key characteristics and features.

Reference	**Methodology**	**Features**
Jomsri et al. (2010)	TFIDF CBF	Introduce a framework for a tag-based research paper recommender system.
Ferrara et al. (2011)	TFIDF CBF	Compared the similarity of documents by extracting keyphrases and giving TFIDF weights.

Dong et al. (2009)	TFIDF CBF	Compared the similarity of documents by extracting the whole text of documents and giving TFIDF weights.
Giles et al. (1998)	CCIDF CBF	Introduced CCIDF approach.
Beel et al. (2017)	CCIDF CBF	Gave weights to the links between papers and the reference list.
Tanner et al. (2019)	CCIDF CBF	Compared between CCIDF and CC only.
Torres et al. (2004)	Memory-based CF	After determining the ratings matrix with users as papers and items as citations, user-based CF was compared with CBF and hybrid methods.
Ekstrand et al. (2010)	Memory-based CF	Upon establishing the ratings matrix with users and items represented by all the dataset papers, item-based CF was compared to a hybrid of GB and item-based CF.
McNee et al. (2002)	Memory-based CF	Demonstrated that both user-based and item-based CF produced comparable outcomes when recommending citations.
Salakhutdinov and Mnih (2008)	Model-based CF	Proposed Probabilistic Matrix Factorization (PMF) to address the scalability and sparsity issues in CF.
C. Wang and Blei (2011)	Model-based CF	Combined Latent Dirichlet Allocation (LDA) and matrix factorization.
H. Wang et al. (2015)	Model-based CF	Proposed the Collaborative Deep Learning (CDL) model.
White and Smyth (2003)	Graph-based CF	Defining and computing the importance of nodes in a graph relative to one or more root nodes.

W. Lu et al. (2007)	Graph-based CF	Local citation graph is generated for each paper in the dataset using the papers' reference lists and calculate similarity depending on the authority scores vectors.
Chang et al. (2000)	Graph-based CF	Calculate the authority scores using HITS algorithms for a set of documents.
Beel et al. (2015)	Stereotype	Employed a stereotype model as a fallback option when other recommendation approaches failed to provide satisfactory results.
Bethard and Jurafsky (2010)	Popularity	Include the popularity factors to re-rank the final recommended papers.
Dong et al. (2009)	Hybrid (CF + CBF)	Refining the reading list by identifying users with similar tastes through CF and subsequently applying content-based filtering to the list.
Torres et al. (2004)	Hybrid (CF + CBF)	Used the result of one model as the input of the other with 5 different ways.
Ekstrand et al. (2010)	Hybrid (GB + CF) and (GB + CBF)	Effectively Combined Graph based algorithms with CBF, and with CF.

Table 2.1: An Overview: Methodologies and Features of Various Research Papers recommender Systems

In the following chapters, I provide an overview of various approaches to construct research paper datasets, highlighting their strengths and limitations. Additionally, I introduce a novel approach of using citation graph network for assembling thematic datasets of research papers. By carefully selecting users and items for the ratings matrix, our aim is to enhance the accuracy and efficiency of recommender systems. Furthermore, I conduct a comparative evaluation between the proposed

citation network method and a keyword search approach that is limited to highly cited journals. I examine the performance of three recommender system techniques: user-based collaborative filtering, a combination of PageRank and user-based CF, and a fusion of personalized PageRank with user-based CF. Through this evaluation, I aim to assess the effectiveness and superiority of the citation network method in comparison to the keyword search approach. By investigating these approaches and their performance, I aim to contribute to the advancement of research paper recommender systems and provide valuable insights for researchers and developers in the field.

Chapter 3

Research Paper Dataset Construction

The importance of a well-constructed thematic dataset in the research paper recommendation domain cannot be overstated, as it forms the basis for the development, testing, and deployment of recommender algorithms. Researchers in the field of research paper recommender systems employ various methods to obtain thematic datasets, including keyword searches from search engines or citation databases, selecting digital repositories or academic databases, or focusing on specific journals to evaluate their proposed algorithms.

In this chapter, I offer an overview of current techniques utilized for dataset construction. I dissect the strengths and limitations of these methodologies, particularly when applied to the assembly of thematic datasets. Furthermore, I present a novel methodology for constructing the ratings matrix, which constitutes a crucial component of the user-based collaborative filtering algorithm. Through scrutinizing these techniques, I aim to pinpoint the challenges and constraints inherent in dataset construction, underscoring the necessity for more efficient methods. The ultimate goal is to devise strategies that will enable the construction of comprehensive and relevant thematic datasets.

3.1 Overview of Research Paper Dataset Construction Approaches

Datasets are crucial components of any research paper recommender systems Beel et al. (2016). Curation and construction of datasets specific to a given research field have been accomplished

through a variety of approaches. These methods are crucial for assembling a theme-oriented collection of research papers that serve as the basis for developing and testing recommender systems. Creating a relevant and comprehensive dataset can have a significant impact on a recommender system's performance. It is therefore crucial to understand the advantages and disadvantages of each dataset construction method. As a result, researchers are able to make informed decisions when creating datasets and avoid potential pitfalls, increasing the overall effectiveness of their recommender systems. With a particular focus on assembling thematic research paper datasets, this section highlights the most prominent strategies used for this purpose. It details their strengths and limitations.

Dataset Construction Using Journals

The dataset from Ekstrand et al. (2010) comprises around 201,145 papers from the ACM Digital Library as of April 2010. In order to evaluate the performance of various algorithms, they selected the journal "ACM Computing Surveys (CSUR)" and focused on its papers that had more than 15 references to other papers associated with the same journal. Moreover, they removed papers that lacked reference lists and cited mainly from outside the chosen journal. Afterward, they used all the papers as users and items to build the ratings matrix.

Dataset Construction Using Digital Repositories and Academic Databases

Tanner et al. (2019) utilized the 2014 ACL Anthology Network (AAN) dataset, which is a comprehensive collection of research papers in the field of computational linguistics. This dataset includes citation relationship information between papers, as well as paper metadata. The authors did not construct their own thematic dataset; they used all the papers existing ACL dataset to test various algorithms. Similarly, Bethard and Jurafsky (2010) The researchers evaluated their literature search models using the ACL Anthology Reference Corpus, a collection of 10,921 computational linguistics papers. They conducted retrieval experiments by creating queries from titles and abstracts and compared the results to the reference lists of the query articles. Ferrara et al. (2011) conduct an experimental evaluation of their content based filtering approach, focusing on tuning parameters and assessing performance. They are using a publicly available dataset containing 597

full papers from the ACL Anthology Reference Corpus (ACL ARC). Bulut et al. (2019) Used compiled dataset from multiple scientific datasets, The DBLP Computer Science Bibliography, IEEE Xplore Digital Library, ACM Digital Library (Association for Computing Machinery) and Pubmed databases, up to 4000000 papers was collected. After collecting data, data preprocessed by cleanup reduction, and parsing. Jomsri et al. (2010) In order to assess the suggested research paper recommender system, a crawler gathered data from CiteULike between March and May 2009. The dataset consists of 62,192 research papers, with 103 groups related to computer science. CiteULike website works as a social bookmarking and reference management tool that helps users discover, organize, and share academic papers while collaborating with others in their research community. McCallumzy, Nigamy, Renniey, and Seymorey (1999) introduce Cora a domain-specific search engine for computer science research papers that organizes over 50,000 papers into a topic hierarchy and maps citation links. It is built in three stages: collecting papers using a web-crawler, extracting relevant information from the papers, and presenting the information in a user-friendly web interface. The search engine supports standard search syntax and provides individual paper details, including citation maps for easy navigation.

Dataset Construction Using Keywords Search

In Torres et al. (2004), there is no explicit explanation of the dataset construction approach. The author collected over 500,000 papers and 2 million citations from the CiteSeer website an online repository of computer science research papers. CiteSeer operates by initially allowing users to search using specific keywords, authors, publication years, or research areas, and then expanding the search to include citations and cited papers of the resulting papers Giles et al. (1998). After gathering papers, the data was pre-processed by removing papers citing fewer than three other papers, and eliminating citations whose full text was unavailable. Subsequently, a ratings matrix was built, using papers as users and citations as items. McNee et al. (2002) utilized a dataset from the ResearchIndex website the old name of CiteSeer, consisting of over 500,000 papers. They refined the data by removing papers with fewer than two citations and excluding papers with insufficient connections to the rest of the dataset, specifically by discarding citations mentioned fewer than two times. In this context, a citation refers to a paper without an associated reading list. Finally, they treated papers

as users and citations as items to construct the final ratings matrix. Similarly, Dong et al. (2009) In order to assess the algorithms, a dataset was constructed using papers obtained from CiteSeer. The dataset was restricted in two ways: firstly, papers with less than two citations were excluded to minimize noise from weakly connected papers. Secondly, citations not present as papers within the dataset were removed, guaranteeing that every citation in the dataset was also a paper. This enabled both CF and CBF techniques to examine each paper in the dataset. Following these adjustments, the refined dataset consisted of 68,625 papers. J. Lu, Hoi, and Wang (2013) used a web robot to gather data from Citeseer, focusing on papers that were homogeneous and familiar to them. Despite the disadvantages, such as the need for expert judgment and the presence of clutter and incompleteness, they were able to construct a dataset specified for their study.

3.1.1 Discussion

Evaluating algorithm performance using a single journal can be advantageous as the citations are likely to be highly correlated. This allows for a more straightforward comparison between different algorithms. However, relying solely on citations from one journal for paper recommendations might overlook many relevant citations, as some papers related to a specific theme could be cited across various journals. Additionally, a single journal may not have a sufficient number of citations to effectively test recommender algorithms. The dataset would be larger and sparser, potentially leading to a decline in algorithm performance.

The ACL and Cora datasets compile papers related to a single topic and include a variety of research themes, making them suitable for effectively comparing different algorithms. However, the applicability of the ACL and Cora datasets is limited, as they cannot be generalized to other research topics. Furthermore, ACL lacks a clear explanation of the dataset construction mechanism. CiteUlike users' bookmarks can be used as ratings for papers, but it does not provide a comprehensive dataset of all papers related to a specified research theme since it mainly relies on users' collection of added paper metadata.

The Citeseer website can be utilized to expand the collected dataset using citation information. Citeseer primarily relies on keyword searches to compile paper datasets. However, using keyword searches may result in accumulating a large amount of irrelevant data, which can negatively impact

the performance of recommender algorithms. It focuses specifically on the fields of computer and information science.

Various approaches to constructing datasets for testing recommender algorithms have been explored. However, authors often overlook the importance of thematic dataset creation in the recommender system process. Relying solely on keyword searches for research paper collection can lead to an accumulation of irrelevant thematic research paper data, causing sparsity issues and diminishing algorithm performance. Predefined online repositories can be useful for testing different algorithms, but their results cannot be generalized across various research themes. Therefore, an approach to collect a thematic dataset across multiple research themes is still needed. Moreover, selecting papers from specific journals as a thematic dataset may exclude crucial papers published elsewhere. The proposed method eliminates the limitations of sourcing papers from specific journals, prevents the exclusion of significant works published in diverse outlets, and avoids the pitfalls associated with keyword-based search methods.

3.2 Ratings Matrix Construction

The first step in a typical e-commerce collaborative filtering recommender system is to build a customer-product matrix based on customer ratings or historical purchasing transactions of products as input data Lee, Kim, and Park (2007). The input data consists of preference scores from n customers for m products. This data is typically represented as an $R_{m \times n}$ customer-product matrix. If customer i purchased product j, then R_{ij} is 1; otherwise, R_{ij} is 0 Sarwar, Karypis, Konstan, and Riedl (2000).

Illustrative E-commerce Example:

In the unary ratings matrix example shown in Table 3.2, cells filled with 1 represent purchased items, and cells filled with 0 represent user-item pairs without recorded interactions. This matrix can be utilized by collaborative filtering algorithms to identify similar users or items and generate personalized recommendations for users.

UserID	Purchased itemID
U1	A
U1	D
U2	A
U2	C
U2	B
U3	D

Table 3.1: Small e-commerce dataset representing Users and their purchased items

	A	B	C	D
U1	1	0	0	1
U2	1	1	1	0
U3	0	0	0	1

Table 3.2: Unary ratings matrix of the e-commerce dataset

Research Papers Ratings Matrix

As mentioned in 2.2.3 In order to apply several research paper recommender systems we need to clearly define a ratings matrix. Once we have the thematic dataset of research papers, the next step is to create a ratings matrix that can be used to apply several research paper recommender algorithms. As mentioned in 2 The research paper recommendation system described in McNee et al. (2002) selected papers from the ResearchIndex website as users and their citations as items, and defined a citation as a reference to a research paper that do not have the full text for. Ekstrand et al. (2010) used all the papers in the citation web as users and items.

For constructing the ratings matrix, I used a different approach. The approach can add more value and produce more efficient recommendations of research papers. Starting by denoting the set of all papers in the dataset as $P = \{P1, P2, P3, ..., Pn\}$. Now, let's define the sets of "users" U and "items" I based on the citation behaviors of these papers within the dataset.

- Users U: This set consists of papers that reference the most other papers within the dataset. If we denote the number of references made by paper Pi as $ref(Pi)$, we can rank the papers based on $ref(Pi)$ in descending order, and select the top k papers to be in the users set: $U = \{Pu1, Pu2, ..., Puk\}$ such that $ref(Pui) \geq ref(Pui + 1)$ for all $1 \leq i < k$

- Items I: This set consists of papers that are most frequently cited by other papers in the

dataset. If we denote the number of times paper Pi is cited by other papers as $cite(Pi)$, we can rank the papers based on $cite(Pi)$ in descending order, and select the top l papers to be in the items set:

I = $\{Pi1, Pi2, ..., Pil\}$ such that $cite(Pij) \geq cite(Pij + 1)$ for all $1 \leq j < l$

This setup allows us to effectively analyze and recommend research papers based on their citation relationships and relevance within the thematic dataset.

Furthermore, I define a cut-off percentage to select a subset of the highest ranked papers for both identified "users" and "items" sets. This will determine the size of the ratings matrix and ensure the focus on the most valuable papers. By selecting a size for the chosen users and items, I aim to exclude less valuable papers while still ensuring that the items and users are large enough to provide useful recommendations.

- Users U: Let's denote the cut-off percentage for the "users" set as pu. We select the top pu percent of papers with the highest $ref(Pi)$ values to be in the users set: $U = \{Pu1, Pu2, ..., Puk\}$ such that $ref(Pui) \geq ref(Pui + 1)$ for all $1 \leq i < k$, where $k = n * pu/100$.

- Items I: Similarly, let's denote the cut-off percentage for the "items" set as pi. We select the top p_i percent of papers with the highest $cite(Pi)$ values to be in the items set: $I = \{Pi1, Pi2, ..., Pil\}$ such that $cite(Pij) \geq cite(Pij + 1)$ for all $1 \leq j < l$, where $l = n * pi/100$.

I define the percentage of the ratings matrix as the size of the selected pu and pi. For instance, when I state the percentage of the ratings matrix is 10%, that means $pu = pi = 10\%$.

Following the definition of the users and items set, we can construct a ratings matrix $R_{k\times l}$ to represent the citation relationships between these two sets. The entries in the matrix $R_{k\times l}$ are defined as follows: $R_{ij} = 1$, if "user" paper Pui references "item" paper Pij. Otherwise, $R_{ij} = 0$.

The goal is to focus on the most valuable papers while ensuring that the sizes of the "users" and "items" sets are large enough to provide useful reading lists recommendations. The values of pu and pi can be adjusted based on the dataset characteristics and requirements.

3.3 Summary

In this chapter, I have reviewed the diverse approaches used for constructing research paper datasets, discussing the advantages and shortcomings associated with each method. Existing techniques for developing datasets to test recommender algorithms, such as keyword searches, predefined online repositories, or a selection of papers from a particular journal, all carry inherent limitations. Moreover, I have introduced and described the approach of constructing the ratings matrix associated with the thematic dataset. In the subsequent chapter, I introduce a citation graph network approach that targets a specific theme of research and employs survey papers as its foundation. This method offers a comprehensive overview of related papers, ensuring the resulting dataset is exhaustive, relevant, and thematic. The citation graph connections influence the determination of a paper's significance, thus overcoming the aforementioned limitations inherent in the current approach to thematic research paper dataset construction.

Chapter 4

Citation Graph-based Research Papers Dataset Construction

The careful construction of our thematic dataset using the citation graph network represents a significant contribution to our research. It ensures the efficiency and accuracy of our recommendation system by providing a comprehensive and relevant list of high-quality papers related to a target theme. I leveraged the Scopus citation database as the data source, employing a meticulous selection process to incorporate the most pertinent publications and features for our selected paper recommendation algorithms. The performance of these recommendation algorithms hinges significantly on the quality and relevance of the thematic dataset used during development and evaluation. As such, our comprehensive dataset plays a pivotal role in enhancing the overall effectiveness of the recommendation systems.

In this chapter, I outline the data source used for constructing the thematic dataset for our research paper. Moreover, I delineate how I built the thematic research paper datasets derived from this data source using citation graph networks, supplementing our explanation with an illustrative Shared Mobility (MOD) dataset real case example.

4.1 Data Source

Scopus contains abstracts, citations, and abstracts of peer-reviewed literature in a variety of disciplines, including scientific journals, books, and conference proceedings. Elsevier, a world leader in scientific, technical, and medical information products and services, owns and operates it. Scopus covers a broad range of disciplines, indexes more than 70 million records from over 14,000 publishers, including physical sciences, life sciences, health sciences, and social sciences. Quality, accuracy, and comprehensiveness are all emphasized in the dataset, which contains a wide range of globally recognized publications. Elsevier provides an API that allows programmatic access to Scopus metadata. Using the Scopus API requires an API key from Elsevier, which I can obtain by registering with the Elsevier Developer Portal [1]. An API key will be generated after registration, which should be included in our API requests. Scopus metadata may be subject to usage restrictions, licensing agreements, or other restrictions imposed by Elsevier. Scopus data requirements and guidelines must be followed. It should be noted that full API access is only granted to researchers affiliated with certain organizations, such as Concordia University, who hold relevant subscriptions to Elsevier products. Each publication in our dataset is associated with a number of features, including the Scopus identification number, the date of publication, the specific journal it was published in, the author's affiliation, and the reference list. While all of these features provide valuable information, only some of them were useful for our thematic dataset construction method, and the used reading list research papers recommendation algorithms.

4.2 Citation Graph

Research papers graph networks are used to model and analyze the relationships between academic publications and their attributes, research paper graph networks consider multiple aspects of research papers, such as authors, keywords, topics, and institutions, in addition to citations Radicchi, Fortunato, and Vespignani (2011). Graphs of research papers can represent not only individual papers, but also authors, keywords, topics, and institutions. In order to indicate relationships between nodes, edges are used. Edges between an author node and a paper node may indicate that

[1] https://dev.elsevier.com/

the author wrote the paper, while edges between two paper nodes may indicate that the papers have been cited or have the same keywords. It is possible to study and analyze the connection and interaction between various elements of the academic research ecosystem with this more comprehensive approach. Analyzing the structure and properties of research paper graph networks enables identifying patterns and trends in research, identifying key influencers, and exploring the evolution of scientific fields.

Citation graphs network represent the relationships between academic papers through the citations they receive. The nodes in this network represent individual research papers, and the directed edges (arrows) represent citations between papers White and Smyth (2003). It indicates the flow of information from one paper to another by pointing from the citing paper to the cited paper. By analyzing the connections between papers, one can analyze the influence of specific papers or authors, the development of research fields over time, and the detection of emerging trends in research. Researchers can use citation graph networks for bibliometric analysis to gain insight into the scientific literature and to make informed decisions about the research landscape. Citation graphs are powerful representations of the intricate relationships among academic papers, based on the citations they exchange where citations are the most commonly utilized measure in assessing the worth of papers, investigators, research institutions, and educational institutions Tahamtan, Safipour Afshar, and Ahamdzadeh (2016). It is recommended to consider topics identified from citation contexts whenever possible Liu and Chen (2013).

4.2.1 Citation Graph Design

In a research papers citation graph network, a node represents a unique research paper, while directed edges or arrows represent citations connecting one paper to another White and Smyth (2003). Edges pointing from the citing paper to the cited paper illustrate the flow of information.

Using the citation graph network, I have constructed a thematic dataset that included only papers directly relevant to our target theme. To achieve this, I relied on survey papers As the foundation for our approach, survey papers prove valuable and effective when they tackle a significant subject that captures the attention of a substantial portion of the scientific community. Survey papers provide an overview of a particular research area, offering insight into the most relevant and influential papers.

it is typically written by an expert in the field who has a deep understanding of the subject matter. A well-curated and accurate representation of the research landscape is ensured by this expertise, lending credibility, and they contain a wealth references, making them an excellent starting point for further reading Batovski (2008). By using survey papers as a starting nodes for the citation graph network, I were able to identify a wide range of publications that provided valuable insights into our chosen target theme. To ensure the inclusion of relevant papers for the researcher, I identify the references and the citing papers of the initially selected survey papers. Then, extend this process by going a second level within the citation graph by identifying the references and the citing papers of the extracted papers. By doing so, I aim to capture a comprehensive list of papers that are potentially interesting to the researcher, while maintaining the main theme as the focal point. This enabled us to create a comprehensive thematic dataset that covers a broad spectrum of research on our target theme, while ensuring that only papers with the highest relevance and quality were included. Our approach not only allowed us to efficiently identify relevant papers but also provided a means of filtering out publications that were unrelated to our target theme, thereby ensuring the accuracy and validity of our thematic dataset. Which in turn can be used with recommendation systems algorithms to recommend highly relevant papers.

4.2.2 Illustrative Example

These steps can be followed to construct the thematic dataset of research papers using a two-level citation graph network and a survey paper as a seed as shown in 4.1:

(1) Select the survey paper ”An Overview of Shared Mobility” as seed paper and take its scopus id number, which is unique to each paper indexed in scopus database.

(2) Take first level citations: Extract all the 101 citing papers, and the 113 reference papers scopus ids of the selected survey paper. Citation graph network is formed by these papers. we end with 214 extracted research papers covering a theme of research, since the seed paper is a survey paper.

(3) Citations at the second level: Extract the scopus ids of the citing papers, and the reference papers at the second level for each paper of the 214 extracted papers in the first level citation.

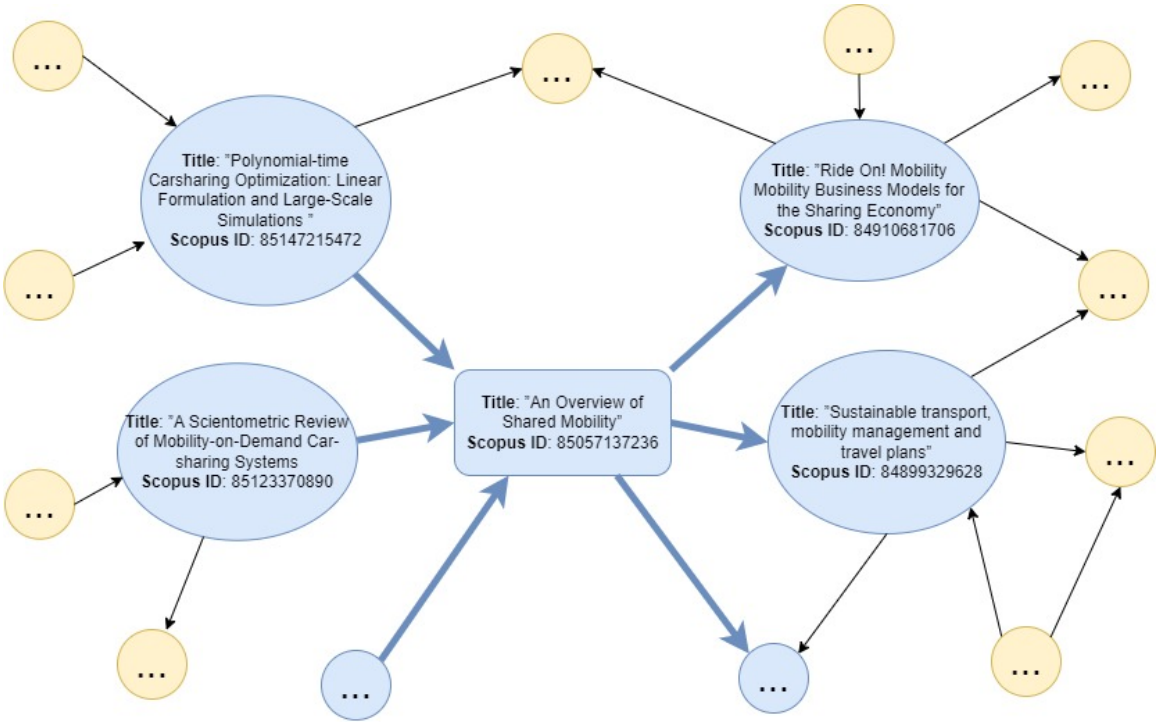


Figure 4.1: Illustrative example taking a survey paper as a starting node, with the citations between papers represented by the edges

A citation graph network is composed of these extracted papers at the second level. In this second level, we will cover papers that have been influential on the first level and may provide additional context or foundational knowledge.

(4) Collect all the papers as a thematic dataset that combines the seed paper, first-level citations, and second-level citations. A total of 37,493 papers are constructed. This dataset will represent the two-level citation graph network for the chosen survey paper.

4.2.3 Algorithm for Constructing The Citation Graph

I present our algorithm for constructing the thematic dataset through a citation graph network in Algorithm 1 and 4.1. The algorithm constructs a research papers dataset using the network of citations between research papers in a citation graph.

Algorithm 1 Two-Level Citation Graph Network Thematic Dataset Construction

Input: N - set of initial seed papers
Output: N^* - constructed set of thematic research papers dataset

Initialize $V \leftarrow \emptyset$
Initialize $E \leftarrow \emptyset$
Initialize $G(V, E) \leftarrow (V, E)$
Initialize $N^* \leftarrow \emptyset$
for $n \in N$ **do**
    Add n to V
    A_n = set of references for n
    B_n = set of citing papers for n
    for $a \in A_n$ **do**
        Add (n, a) to E
        Add a to N^*
    end for
    for each $b \in B_n$ **do**
        Add (n, b) to E
        Add b to N^*
    end for
end for
$N = N^*$
for each n in N **do**
    Add n to V
    A_n = set of references for n
    B_n = set of citing papers for n
    for $a \in A_n$ **do**
        Add (n, a) to E
        Add a to N^*
    end for
    for each $b \in B_n$ **do**
        Add (n, b) to E
        Add b to N^*
    end for
end for
Update $G(V, E) \leftarrow (V, E)$ **return** N^*

Notations	Descriptions
$G(V, E)$	Directed graph, with set of nodes V, and set of edges E
$n \in N$	Set of seed papers for the citation graph network
$a \subset A_n$	Set of references for n
$b \subset B_n$	Set of citing papers for n
N^*	Set of constructed research papers thematic dataset

Table 4.1: Citation Graph network Thematic Dataset Construction Notations

4.3 MOD Dataset Construction Using The Citation Graph Network

Researchers in our group focus mainly on Shared Movility (MOD) research theme. To construct MOD thematic datasets, I use the following steps:

(1) Selecting appropriate survey papers is important as we consider them to contain a wealth of references that provide a concise summary of the relevant theme. The choice of survey papers is primarily dependent on the theme I am investigating. My research group is focused on the subject of MOD. Therefore, I have identified four relevant survey papers, namely "Crowdsourced delivery: A review of platforms and academic literature", "Collaborative Urban Transportation: Recent Advances in Theory and Practice", "A survey of models and algorithms for optimizing shared mobility", and "An Overview of Shared Mobility", as the primary sources for the research papers thematic dataset.

(2) We identify the references and citing papers from the initial survey papers to ensure the inclusion of relevant papers for the researcher. By going a second level in the citation graph, we determine the references and the citing papers of the extracted papers. By doing so, we aim to capture a comprehensive list of papers that are potentially interesting to the researcher, while maintaining the main theme as the focal point. We were left with approximately 44,000 papers that were relevant to the shared mobility domain. We referred to this set of papers as the Search-set for the analysis. 4.2 show statistics of the search set papers. Where the average citations is defined as the mean number of citations each paper gets within for a dataset of research papers Moed (2006), and it can be calculated using the following equation:

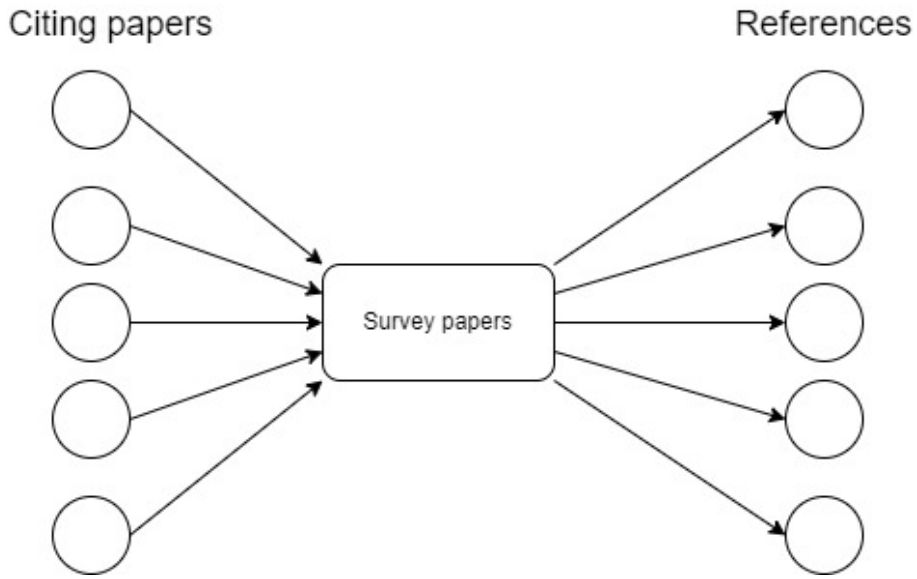


Figure 4.2: Survey papers citations illustration

$$\text{Average Citations} = \frac{\text{Total Citations}}{\text{Number of Papers}} \tag{5}$$

In addition to providing an overall picture of the citation impact of the papers in the dataset, this metric can be used to assess the impact of a particular paper in the academic community.

	Stats
Number of papers	43936.0
Average citations	2.62
papers with at least 15 references	2470.0
papers with at least 1 reference	10647.0
papers cited by at least 15 papers	1620.0
papers cited by at least 1 paper	20604.0

Table 4.2: Citation graph Search-set Statistics

(3) Once we have identified the search set, the next step is to create a ratings matrix as mentioned in 3.2 that can be used to apply several recommender system algorithms. This involves organizing the matrix in such a way that users are represented in the rows and items are represented in the columns.

(4) Identify the size of the ratings matrix $R_{k \times l}$, as stated in 3.2, the size of the ratings matrix

can vary. I have compared different algorithms based on various ratings matrix sizes. The user set comprises papers ranked by their number of references, while the item set consists of papers ranked by their citing papers count. To determine the size of the ratings matrix, we select a percentage of the highest-ranked papers from both the user and item sets. The chosen percentages include 10, 20, 40, 60, and 80 % of both sets 4.3. I then compare the density and degree of centrality of the resulting ratings matrices. A detailed comparison of the ratings matrix sizes based on the selected percentages can be found in Chapter 5. After selecting the users and items, each user in the users set votes 1 for the items it references and 0 otherwise. It should be noted that the dataset contains a high level of connections through citations, resulting in a dense ratings matrix. Therefore, it is expected that various recommendation system algorithms will perform well.

Percentage of the ratings matrix	Size of the ratings matrix
10	$R_{4400\times4400}$
20	$R_{8700\times8700}$
40	$R_{18000\times18000}$
60	$R_{26400\times26400}$
80	$R_{35200\times35200}$

Table 4.3: Corresponding sizes of the ratings matrix percentages of Citation graph search set

4.4 Summary

In this chapter, I have introduced and elucidated the technique for constructing a thematic dataset using the citation graph network. This method leverages the citation relations of survey papers targeting a specific research theme. In the forthcoming chapter on experiments, I will provide a thorough analysis of the experimental results from the recommendation systems derived from the proposed approach. These will be compared with results obtained from the keywords search approach for generating a research paper dataset, further illuminating the effectiveness and advantages of the proposed methodology.

Chapter 5

Experiments and Results

In this chapter, we examine the experimental framework applied to two distinct MOD datasets. The first dataset, detailed in Chapter 3, is constructed using the citation graph network, while the second dataset, outlined in this chapter, employs keyword searches, specifically constrained to certain academic journals. We commence by comparing the datasets in terms of their average citations, density, and degree of centrality. Following this, we describe the experimental setup and delve into the outcomes garnered from three chosen research paper recommender systems detailed in chapter 2. In the third section, we address potential validity concerns. To conclude, we collate the observations and insights derived from the experimental results.

5.1 Datasets Comparison

I have constructed MOD dataset using keywords search limited to academic journals to compare with the proposed citaion graph network MOD dataset in chapter 4. I compared the two datasets average citations, density, and average degree centrality, In addition, to evaluation of the three chosen recommender algorithms from 2.3 on both datasets.

5.1.1 Construct Thematic Dataset Using Keywords Search Limited to Specified Academic Journals

The most common way of constructing MOD datasets is through keywords search. It will be compared with the main citation graph network dataset construction. Based on the MOD theme keywords and journals in the data source, I created a search set. Here are the steps I followed to construct the data set:

(1) I conducted a keyword search to download every paper related to a particular theme of research from Scopus database. The choice of keywords is heavily influenced by the theme in question. For the MOD theme, I collected 562,407 publications using keywords such as ”shared mobility”, ”crowdsourcing”, ”collaborative transportation ”, ”ride sharing”, ”car sharing”, ”ride sharing”, and others. Collected from the keywords section in the mentioned surveys, These are the keywords commonly associated with the MOD topic.

(2) The initial MOD dataset, is both extensive and sparse. To navigate this, I narrowed the search to focus on papers from highly-cited journals within the dataset. I did this based on the principle that simple citation counts for journals are the most critical factor in their ranking, as suggested Davis (2008). I established a ranking of journals based on the number of citations they received from papers in the keywords search MOD dataset. This ranking was created by counting each journal's frequency of appearance in the data, and then arranging these journals in descending order. This process enabled us to assess the relative influence or significance of each journal within the context of the dataset. It also helped to mitigate the issue of data sparsity that I initially encountered with the keyword-derived data. Typically, higher-ranked journals are perceived as more authoritative or impactful. To gather a number of papers comparable to the citation graph approach, we can select papers from the top 15 highly-cited journals. If we need to expand the paper selection, we can extend the search to include a larger number of top-ranked journals.

Table 5.1 and graph 5.1 shows the effect of limiting the 562,407 collected from keywords search papers to the top ranked journals on the average citations between papers. Choosing to limit the initially keywords searched MOD dataset into a low number of journals can result in

Number of chosen top journals	Number of collected papers	Average citations
15	46,315	0.77
30	64,648	1.20
50	82,285	1.25
70	93,624	1.22
130	118,334	1.189

Table 5.1: Keywords and Journals Search-set Statistics.

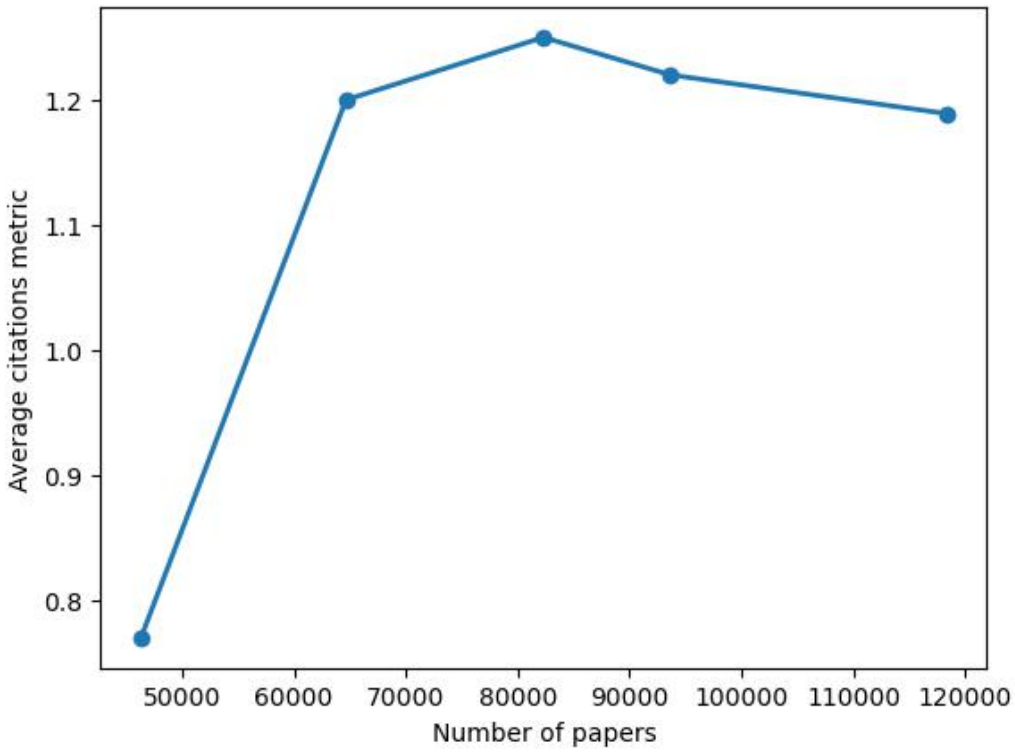


Figure 5.1: Average citations Metric

loosing some highly relevant papers, and choosing high number of journals can result in data sparsity as well and thus loose correlation between papers. Since the Average citations is the highest when we choose to limit the search to the top ranked 50 journals which comprises of 82,285 papers, I chose it as the search set for the comparison with the citation graph MOD dataset experiments.

Some of the Topic related journals: ”Transportation Research Part A: Policy and Practice”, ”Journal of Transport Geography”, ”Transportation Research Record”, ”Transport Policy”,

”Transportation Research Part C: Emerging Technologies”, ”Transportation Research Procedia”, ”Transportation Research Part D: Transport and Environment”.

(3) After identifying the search set from previous step, we obtain its papers citations. Afterwards, we identify users-set, items-set, and the ratings matrix as stated in 3.2.

Through a comparison of the search-set statistics presented in 4.2 and 5.1, it can be inferred that papers obtained through the citation graph method exhibit a higher level of correlation to one another via citations. This implies that the ratings matrix will contain more condensed ratings, which in turn means that various recommendation algorithms will perform better with this data.

Percentage of the ratings matrix	Size of the ratings matrix
10	$R_{8200\times8200}$
20	$R_{16500\times16500}$
40	$R_{32100\times32100}$
60	$R_{50000\times50000}$
80	$R_{65825\times65825}$

Table 5.2: Corresponding sizes of the ratings matrix percentages of choosing keywords limited to top50 journals search set

When examining tables 5.1 and 5.2, we observe that the sizes of the ratings matrix differ for various percentages across datasets. However, it is notable that the 10 percent size of the chosen keywords search set ratings matrix is equivalent to the 20 percent size of ratings matrix the search set derived from the citation graph network. To avoid any biases in the experiments, I have also included a comparison between the identical ratings matrix from both constructed datasets.

Comparison between the final chosen ratings matrix sizes between the two constructed datasets is presented in 5.3, these presented ratings matrices will be evaluated using recommender systems algorithms.

Datasets	Perecentage of ratings matrix	size of the ratings matrix
Citation Graph Network	10%	$R_{4400\times4400}$
	20%	$R_{8700\times8700}$
Keywords search limited to top 50 journals	10%	$R_{8200\times8200}$
	20%	$R_{16500\times16500}$

Table 5.3: Comparison of the final chosen matrix sizes

5.1.2 Compare The Density of The Ratings Matrix

In a research paper's citation recommendation system, the ratings matrix represents the relationship between papers and their citations. The matrix is typically sparse, meaning that most of the entries are zero or empty, indicating no citation relationship between two papers. Adomavicius and Zhang (2012) defined the density of the ratings matrix is a measure of how many of its entries are non-zero, and it can be computed using the following equation:

Let R be the $m \times n$ ratings matrix, where m represents the number users and n represents the number of items. The element r_{ij} represents the rating for the citation relationship between user i and item j. The density of the ratings matrix is calculated as:

$$\text{Density} = \frac{\text{Number of Non-Zero Entries}}{\text{Total Number of Entries}} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} I(r_{ij} \neq 0)}{m \times n} \tag{6}$$

Where:

- $I(r_{ij} \neq 0)$ is an indicator function that returns 1 if the rating r_{ij} is non-zero, and 0 otherwise.
- $m \times n$ is the total number of entries in the ratings matrix.

Figures 5.2 and 5.3 present a comparison of the density scores for two MOD datasets, the one obtained from a citation graph network and the other derived from keyword searches limited into highly-cited journals. Ratings matrix percentage is defined in 3.2. The figures reveal that the citation graph network approach generates a substantially denser ratings matrix, signifying a higher correlation and relevance among the papers for the given theme. In contrast, utilizing keyword searches for data collection leads to a considerable decrease in the density of the ratings matrix. This implies that the assembled dataset has limited inter-citation connections, which may adversely affect the performance of the recommendation algorithm, as elaborated in subsequent sections.

5.1.3 Compare The Degree Centrality Metric

Zhang and Luo (2017) Average degree centrality in a citation graph network reflects the overall connectedness of nodes, which represent academic papers. This metric is calculated by averaging the degree centralities, or the number of direct connections (citations) each node has. A higher

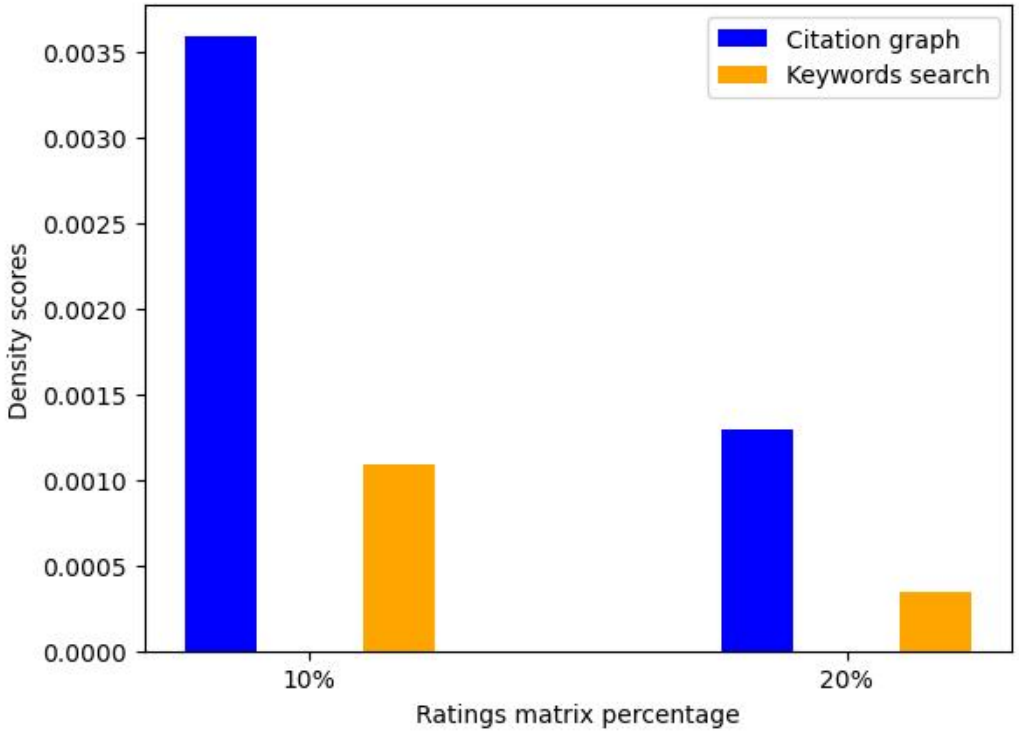


Figure 5.2: Density scores variation in terms of the ratings matrices percentage

average degree centrality indicates a dense, interconnected network where papers are more likely to cite and be cited by other papers within the network. Conversely, a lower value suggests fewer connections between papers. This measure helps us to understand the collaboration and influence within the academic papers network.

$$C_d(v) = \frac{\deg(v)}{n-1} \tag{7}$$

Where:

- $C_d(v)$ is the degree centrality of node v
- $deg(v)$ is the number of connections (edges) of node v
- n is the total number of nodes in the graph

Figures 5.4 and 5.5 display a comparison between various ratings matrices sizes for the citation

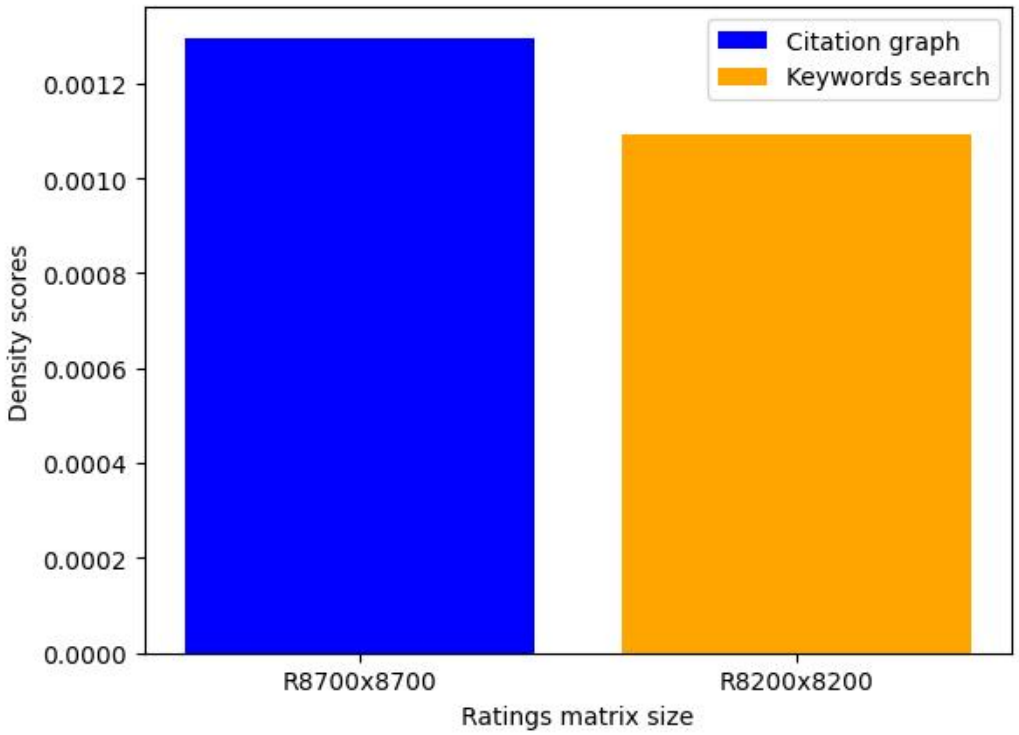


Figure 5.3: Density scores comparison with approximately the same number of ratings matrices papers

graph and keyword search-constructed MOD datasets in terms of average degree of centrality scores. The comparison reveals that the degree of centrality for the rating matrix derived from the citation graph network has stronger node connections, even when comparing the 20% size of the citation graph to the 10% size of the keyword search dataset. Consequently, the citation graph network dataset demonstrates a higher correlation between papers.

5.2 Experimental Procedure

In this section I compare the results of the three algorithms mentioned in chapter 2.3 used on the four chosen ratings matrices from table 5.3 of the two thematic datasets.

5.2.1 Recommendation Process Experiments

In the experiments I assumed a scenario, where a researcher possesses a significant paper within their research theme and seeks to find similar references. The input consists of a single paper, which

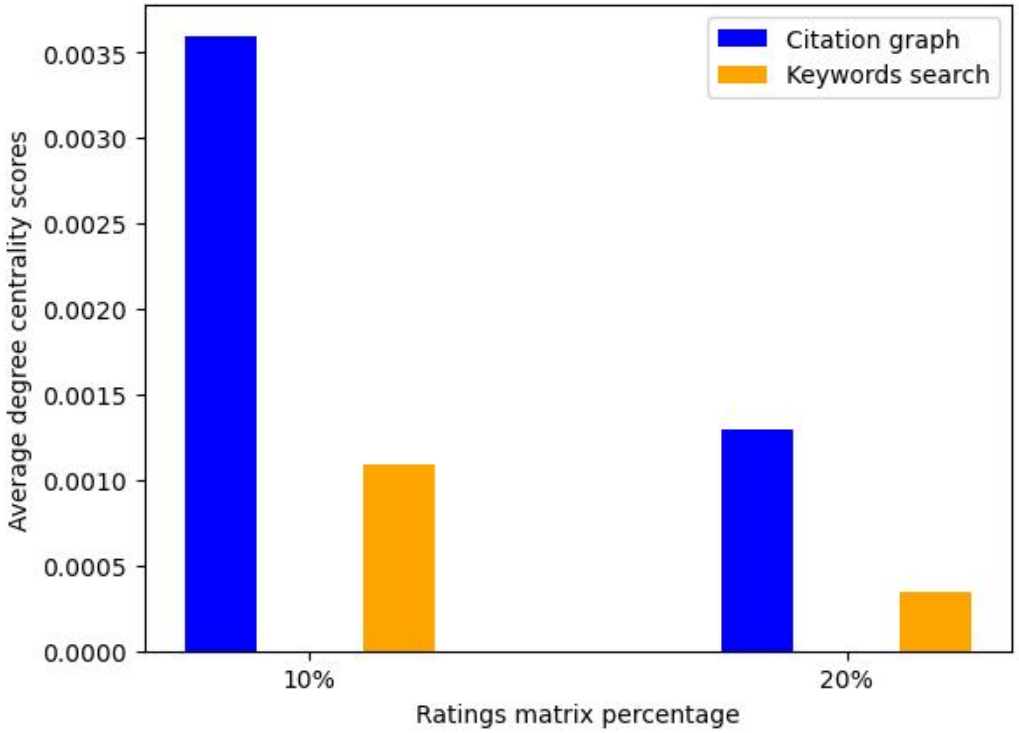


Figure 5.4: Average degree centrality scores variation in terms of the ratings matrices percentage

is treated as a user in the ratings matrix. After assigning the target user in the ratings matrix, one of the three algorithms is executed, generating a list of top N recommended items (references). The final size of the recommended list can be selected according to preferences.

To evaluate the recommendation algorithms performance, I randomly selected 20 users from the ratings matrix and hide 5 items for each user. I then evaluated if the algorithm effectively recommended the hidden items using Recall and Halflife utility metrics. In each run, I select one user, hide 5 items, and choose a size for the Top N recommendations. I tested with Top N sizes of 5, 10, 15, and 20 in the experiments. The subsequent chapter presents the experiment results and analysis.

5.2.2 Evaluation Metrices

To evaluate the results of the described recommendation algorithms, and compare the reslults on the two constructed datasets I chose two metrics, Recall and half life utility.

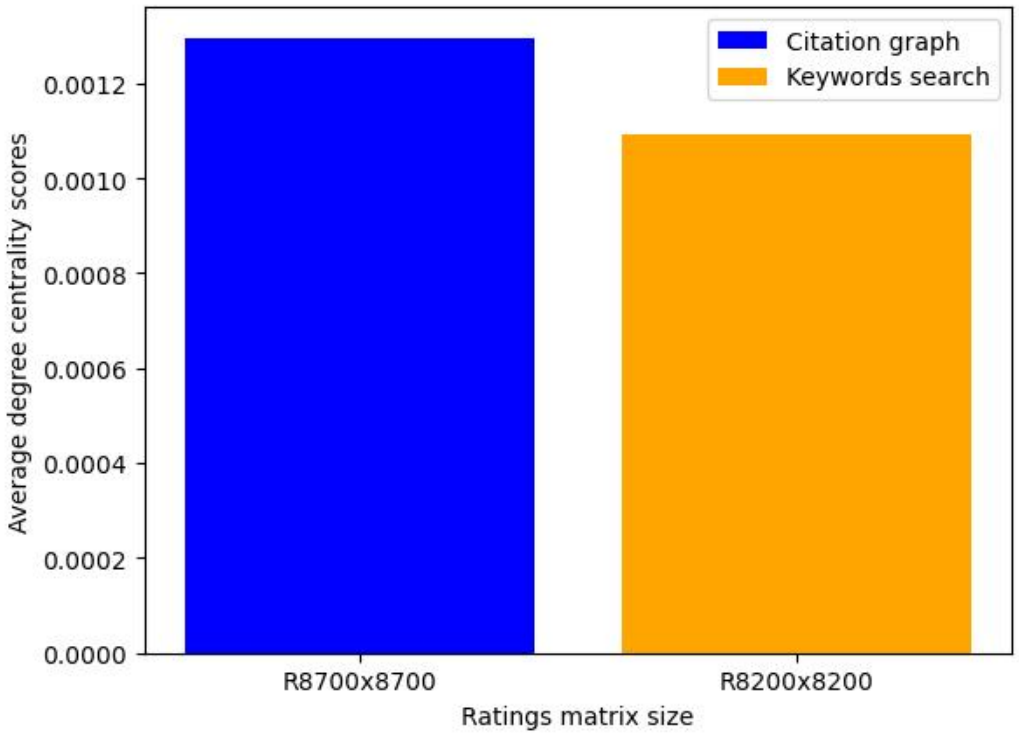


Figure 5.5: Average degree centrality scores comparison with approximately the same number of ratings matrices papers

Recall

In assessing the performance of recommendation systems, recall metrics are employed to calculate the percentage of pertinent items that have been successfully suggested Bellogin, Castells, and Cantador (2011). In the research papers recommendation context, recall is a performance metric used to evaluate the effectiveness of a recommender system. When some citations from a paper (user) are hidden, and the recommender system generates a list of recommended references (items), recall measures the proportion of relevant or true positive recommendations (hidden references) that were successfully retrieved by the system, out of the total number of relevant or true positive items (all hidden references).

Recall = (Number of relevant recommendations retrieved) / (Total number of relevant items)

Higher recall values indicate that the recommender system is more effective at identifying and suggesting the hidden citations.

Top 20 recommendations				
	10% Citation graph	10% Keywords search	20% Citation graph	20% Keywords search
User based CF	0.454	0.341	0.432	0.343
PageRank CF	0.442	0.342	0.447	0.346
Personalized Pagerank CF	0.394	0.285	0.4	0.33
Top 15 recommendations				
User based CF	0.412	0.298	0.384	0.28
PageRank CF	0.4	0.272	0.372	0.27
Personalized Pagerank CF	0.328	0.278	0.343	0.232
Top 10 recommendations				
User based CF	0.353	0.29	0.331	0.214
PageRank CF	0.344	0.284	0.341	0.218
Personalized Pagerank CF	0.296	0.236	0.23	0.218
Top 5 recommendations				
User based CF	0.254	0.154	0.202	0.188
PageRank CF	0.252	0.146	0.204	0.178
Personalized Pagerank CF	0.232	0.134	0.173	0.158

Table 5.4: Recall scores

$$\text{Recall} = \frac{\text{Number of relevant recommendations retrieved}}{\text{Total number of relevant items}} \tag{8}$$

Halflife utility metric

Top 20 recommendations				
Algorithms	10% Citation graph	10% Keywords search	20% Citation graph	20% Keywords search
User based CF	0.305367929	0.226897311	0.277420537	0.215988268
PageRank CF	0.290839872	0.215891753	0.285940153	0.211798133
Personalized Pagerank CF	0.248853559	0.178948839	0.243372723	0.214074648
Top 15 recommendations				
User based CF	0.314778487	0.221672607	0.27211507	0.206980918
PageRank CF	0.304584249	0.196947924	0.263200242	0.196112625
Personalized Pagerank CF	0.246833947	0.192460399	0.248089864	0.171511082
Top 10 recommendations				
User based CF	0.30644858	0.245705079	0.277420537	0.206218876
PageRank CF	0.297736766	0.237538431	0.285940153	0.21071434
Personalized Pagerank CF	0.254743695	0.212862181	0.252379609	0.200834842
Top 5 recommendations				
User based CF	0.272379958	0.170978341	0.219245889	0.212352287
PageRank CF	0.276737256	0.158241444	0.212005941	0.195646883
Personalized Pagerank CF	0.259247002	0.150384802	0.196606176	0.181492839

Table 5.5: HalfLife scores

In binary data sets, utility-based offline evaluation measures Aggarwal et al. (2016) can be used to assess the accuracy of ranking items. Top-ranked items are given greater importance. In utility-based measures, it is assumed that successive items on a recommended list are less likely to be viewed by a target user.

The evaluation is performed by randomly hold-out a number of cited items from a target user. Consequently, apply different recommendation algorithms and discover the location of the randomly hold-out items in the final recommended items list.

$$R_a = \sum_i \frac{u_{a,i}}{2^{(i-1)(\alpha-1)}} \tag{9}$$

R_a is the utility metric score for a target user a, with α number of hidden items, $u_{a,i}$ equal 1 if the item at position i of the final recommended list exists in the hidden items list, and equal zero otherwise. The maximum score of R_a depends of the number of hidden items α. for instance, R_a^{max} = 3.6426 for α = 5. While R_a^{max} = 1.5 if α = 2.

$$R = \frac{\sum_{a \in T} R_a}{|T| R^{max}} \tag{10}$$

R is used to normalize results and make it easier to compare between different α size. $|T|$ is used in case of multiple users.

5.3 Analysis of Results

This thesis present a method for creating thematic datasets of research papers and determining the corresponding reference ratings matrix. I evaluated and contrasted the performance of user-based collaborative algorithms and the impact of integrating them with PageRank and personalized PageRank, as recommended by Ekstrand et al. (2010). The findings indicate that the dataset construction using citation graphs yielded the best results in all instances, employing all the examined algorithms.

Figure 5.6 and table 5.4 present recall scores for generating top 5, 10, 15, and 20 recommendations for each of the three algorithms, based on the same chosen input scenarios. In all recall scores generation scenarios constructing MOD dataset using citation graph network has the highest recall scores using the three chosen algorithms, even when comparing the 20% ratings matrix of the citation graph MOD dataset to the 10% ratings matrix of the keyword search MOD dataset, which includes roughly the same number of papers. The recall scores increase with the increase in the number of output recommendations for all the recommender algorithms. As illustrated in Figure 5.6, the 20% citation graph ratings matrix has demonstrated superior performance over the 10% keywords search ratings matrix, as evidenced by their recall scores. For the user-based CF algorithm, the recall scores were 4.8%, 4.1%, 8.7%, and 9% respectively. When the Pagerank was integrated with the user-based CF algorithm, the recall scores were 5.8%, 5.7%, 10%, and 10.5% respectively. Lastly, when the Personalized-pagerank was combined with the user-based CF algorithm, the recall scores achieved were 3.9%, 0%, 6.5%, and 11.5% respectively.

The HalfLife metric scores for the three algorithms used on both constructed datasets are depicted in Figure 5.7 and 5.5. It's evident from the results that the citation graph approach consistently outperforms the keyword search approach in terms of HalfLife scores. This essentially implies that the true recommendations should appear closer to the top of the final list of recommended papers when using the citation graph dataset, compared to the keyword search approach. We also observe that HalfLife scores are higher for the 20% ratings matrix of the citation graph MOD dataset compared to the 10% ratings matrix of the keyword search MOD dataset. As illustrated in Figure 5.7, the 20% citation graph ratings matrix has demonstrated superior performance over the 10% keywords search ratings matrix, as evidenced by their HalfLife scores also. For the user-based CF algorithm, the HalfLife scores were 4.8%, 3.1%, 5%, and 5% respectively. When the Pagerank was integrated with the user-based CF algorithm, the HalfLife scores were 5.3%, 5.6%, 6.6%, and 4% respectively. Lastly, when the Personalized-pagerank was combined with the user-based CF algorithm, the HalfLife scores achieved were 4.6%, 3.9%, 5.6%, and 6.4% respectiely.

Moreover, when examining the performances of the employed algorithms, it is evident that user-based CF surpasses PageRank user-based collaborative filtering in the majority of cases, whereas

personalized PageRank user-based collaborative filtering yields the poorest performance. Consequently, incorporating PageRank or personalized PageRank may have a detrimental effect on collaborative filtering algorithms in this specific context.

5.4 Threat of Validity

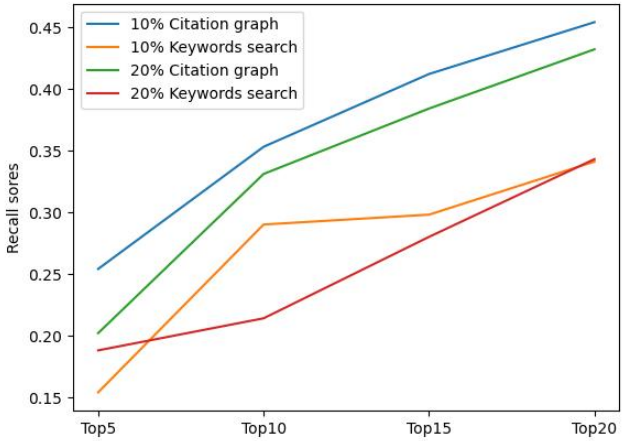
Results could be interpreted as partially invalid for a variety of reasons. The thesis must acknowledge them. Validity is threatened by the following factors:

- Author's interpretation of competing algorithms is the basis of experimental results. To determine which algorithm performs best in a particular situation, the author uses a variety of metrics and experiments. In order to determine which algorithm is most suitable for a particular task, these results are analyzed.
- During the experiment, I chose 20 papers randomly, hidden 5 of their citations randomly, and repeated this process for 5 iterations for each paper in order to reduce noise. The results may vary slightly depending on the experimental setup, but the general conclusion will remain the same.
- Four ratings matrices where used in the experiments two from the citation graph network constructed MOD dataset and two from the keywords search limited to journals MOD dataset. Testing on bigger ratings matrices sizes might lead to computational complexity. The goal was to select a ratings matrix that would not strain computing resources significantly, which would allow us to repeat and confirm the experiments with ease.
- In order to construct the thematic dataset, I used the same keywords that appeared in the survey papers. However, different keywords could be used by different researchers.

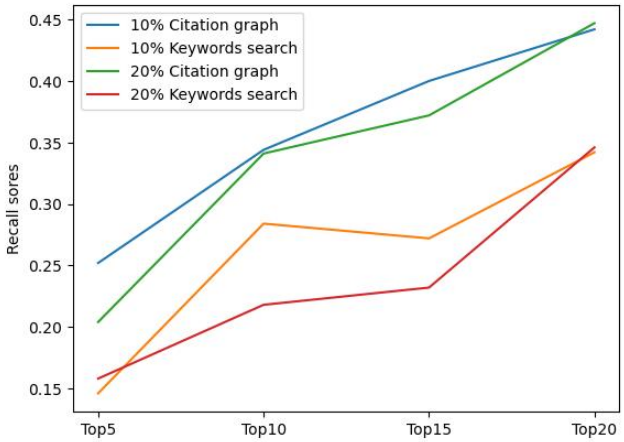
5.5 Summary

In this chapter, we have evaluated the thematic dataset construction technique utilizing the citation graph network as mentioned in 4. The primary enhanced approach capitalizes on citation

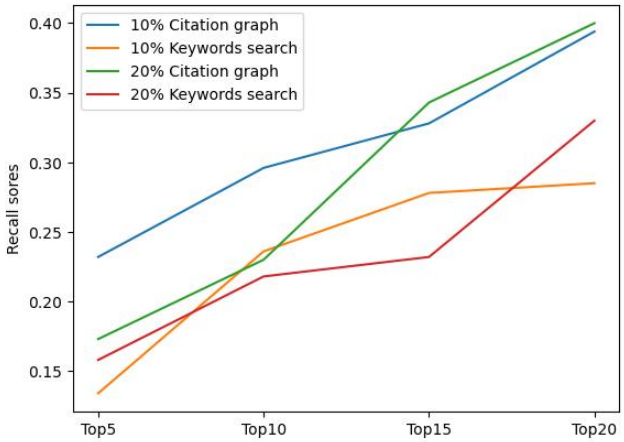
graph networks and survey papers, and it is compared with the traditional keyword search method, which is confined to highly relevant journals, as outlined in 5.1.1. I have demonstrated that the citation graph approach for constructing the research paper MOD dataset surpasses the keyword search method in terms of average citations, density, and degree of centrality of the nodes in the citation network. Furthermore, I compared the performance of user-based Collaborative Filtering, user-based CF combined with PageRank, and user-based CF combined with personalized PageRank on two constructed shared mobility datasets. The experiments provided evidence that the proposed dataset and ratings matrix construction approach yield more accurate, efficient, and robust recommendation results compared to the existing keyword search approach. Moreover, when comparing the three algorithms, we found that the user-based CF method outperforms the user-based CF combined with PageRank approach and the user-based CF combined with personalized PageRank method as suggested by Ekstrand et al. (2010). I will introduce a discussion on the conclusion and potential future research work in the following chapter.



(a) User-based CF Algorithm

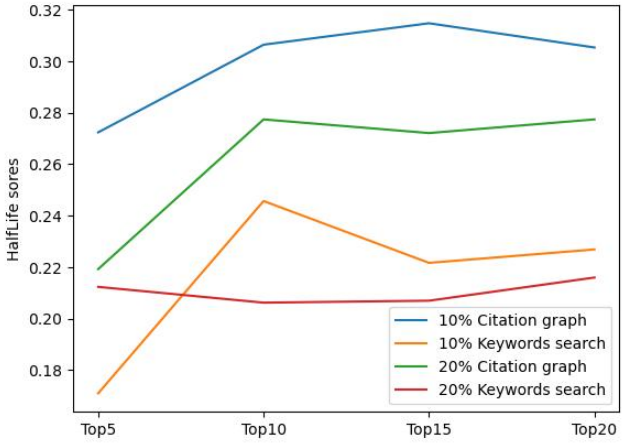


(b) Pagerank + User-based CF Algorithm

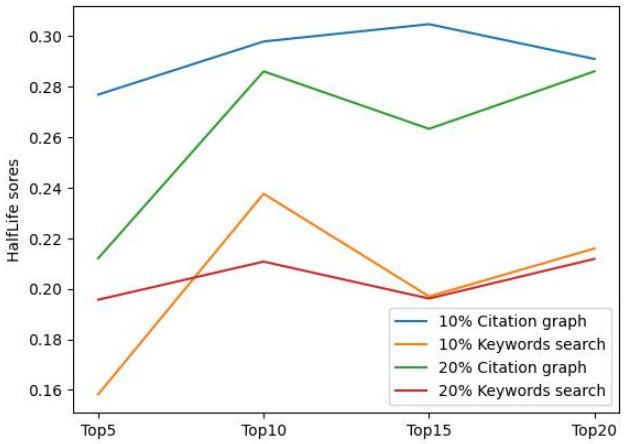


(c) Personalized-Pagerank + User-based CF Algorithm

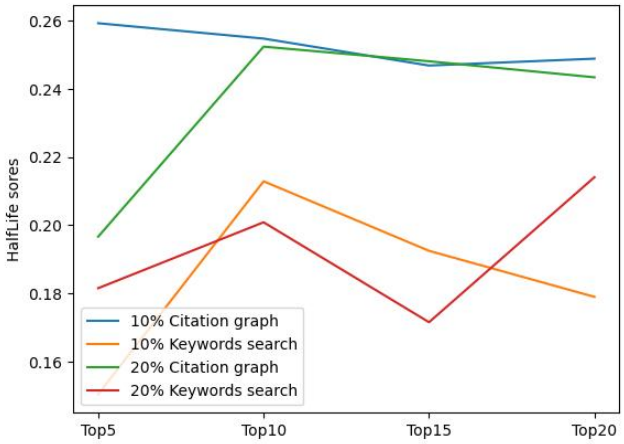
Figure 5.6: Comparison between Recall scores for different datasets



(a) User-based CF Algorithm



(b) Pagerank + User-based CF Algorithm



(c) Personalized-Pagerank + User-based CF Algorithm

Figure 5.7: Comparison between HalfLife scores for different datasets

Chapter 6

Conclusion

In this thesis, I present compelling evidence that showcases the substantial impact of constructing research paper thematic datasets on the outcomes of recommender systems. The investigation highlights the benefits of utilizing citation graph networks to address the issue of dataset sparsity by focusing on papers that are highly relevant to the subject matter. By leveraging the interconnections between research papers, the citation graph network approach facilitates a denser dataset, which in turn positively contributes to the performance of recommender systems. On the other hand, relying solely on keyword search to construct the dataset leads to a sparser representation, which limits the effectiveness of recommender systems.

The conclusion provides a comprehensive overview of this research, summarizing key conclusions and implications derived from the findings. However, it is important to acknowledge that this study has certain limitations. These limitations should be taken into consideration when interpreting the results and applying them in practice.

To address these limitations and further advance the field, I provide suggestions for future research. These suggestions aim to enhance my findings by exploring alternative approaches, considering additional factors, and conducting further evaluations. By addressing these areas of improvement, future research can build upon my work and contribute to the ongoing development of research paper recommender systems.

6.1 Overview

The primary objective of this study was to recommend research papers on specific themes for scholars to explore. To achieve this goal, curating a dataset rich in influential research papers was crucial. By leveraging survey papers as a foundation, I extracted valuable papers using the citation graph network, ensuring a dataset that was both rich and relevant. Selecting the most suitable ratings matrix for the recommender algorithms played a significant role in this process, improving the precision of the system's recommendations.

However, the findings indicated that merging the PageRank algorithm with user-based collaborative filtering did not significantly enhance the performance of the collaborative filtering algorithm. In fact, the overall recommendation performance declined when personalized PageRank was incorporated. Based on these results, we conclude that integrating PageRank with user-based collaborative filtering may not be the most effective approach in this specific case.

Therefore, for future research paper recommender systems, it is imperative to prioritize the exploration of alternative strategies that can enhance the performance of user-based collaborative filtering without compromising its effectiveness. Seeking more effective approaches that align with the specific requirements of the domain remains a priority in order to improve the overall performance and reliability of research paper recommender systems.

6.2 Limitations of The Thesis

Despite the fact that the solution proposed in this thesis meets the requirements and achieves the goals, there are still some limitations that should be acknowledged:

- Delving deeper into the citation graph network leads to a substantial increase in the number of papers. For instance, in our shared mobility topic, going three levels deep in the network resulted in approximately 4,000,000 papers. This can present challenges in terms of scalability and computational resources when working with large-scale datasets.

- The evaluation process involved applying algorithms to a subset of the constructed thematic dataset, specifically 10% and 20% of the dataset, which determined the size of the ratings

matrix. While choosing larger sizes may seem appealing, it is important to consider the potential increase in computational complexity that comes with it.

- In the proposed approach, I used paper IDs to calculate the importance of each paper node, relying heavily on the network of paper citations. It is important to note that this is just one possible approach among many. To assess importance, alternative nodes such as authors, academic journals, research groups, or abstracts can be selected to provide a different perspective. Incorporating these alternative nodes could enrich the understanding and potentially enhance the effectiveness of the recommender system.

- The Scopus API was utilized to access the metadata of various scholarly papers, including the reference list and citations. However, it is worth noting that there is a possibility that this information may not always be available in the Scopus metadata, potentially limiting the comprehensiveness of the dataset.

Acknowledging these limitations is crucial as they provide insights for future research and potential areas of improvement in the development of research paper recommender systems.

6.3 Future Work

In the future work, my objective is to develop a comprehensive recommendation framework that generates a ranked list of highly relevant papers within a specific domain, based on a target paper focusing on a specific theme. This framework will provide researchers with a valuable tool to explore and discover related research papers in their field of interest. As part of this effort, I plan to incorporate the trending dates of papers into the combined PageRank with user-based collaborative filtering approach. I believe that giving greater importance to recently trending papers will enhance the overall effectiveness of the recommendations.

Furthermore, I intend to evaluate the use of authors' names as nodes in the citation graph network, moving beyond the reliance solely on paper IDs. By considering the authors' influence within the citation network, we can assess the PageRank importance score of papers more accurately. Additionally, I will employ an algorithm that leverages the content of the citation graph to evaluate the

constructed thematic dataset. To further enhance this algorithm, I will integrate natural language processing techniques, enabling us to emphasize specific aspects and improve the analysis of the dataset. These enhancements will contribute to the accuracy and relevance of the recommendations.

References

Adomavicius, G., & Zhang, J. (2012). Impact of data characteristics on recommender systems performance. *ACM Transactions on Management Information Systems (TMIS)*, *3*(1), 1–17.

Aggarwal, C. C., et al. (2016). *Recommender systems* (Vol. 1). Springer.

Batovski, D. A. (2008). How to write a review article. *Assumption University Journal of Technology*, *11*(4), 199–203.

Beel, J., Breitinger, C., & Langer, S. (2017). Evaluating the cc-idf citation-weighting scheme: how effectively can 'inverse document frequency'(idf) be applied to references. *Proceedings of the 12th iConference*.

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Paper recommender systems: a literature survey. *International Journal on Digital Libraries*, *17*, 305–338.

Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., & Nürnberger, A. (2013). Research paper recommender system evaluation: a quantitative literature survey. In *Proceedings of the international workshop on reproducibility and replication in recommender systems evaluation* (pp. 15–22).

Beel, J., Langer, S., Kapitsaki, G., Breitinger, C., & Gipp, B. (2015). Exploring the potential of user modeling based on mind maps. In *User modeling, adaptation and personalization: 23rd international conference, umap 2015, dublin, ireland, june 29–july 3, 2015. proceedings 23* (pp. 3–17).

Bellogin, A., Castells, P., & Cantador, I. (2011). Precision-oriented evaluation of recommender systems: an algorithmic comparison. In *Proceedings of the fifth acm conference on recommender systems* (pp. 333–336).

Bethard, S., & Jurafsky, D. (2010). Who should i cite: learning literature search models from citation behavior. In *Proceedings of the 19th acm international conference on information and knowledge management* (pp. 609–618).

Bogers, T., & Van den Bosch, A. (2008). Recommending scientific articles using citeulike. In *Proceedings of the 2008 acm conference on recommender systems* (pp. 287–290).

Bulut, B., Kaya, B., & Kaya, M. (2019). A paper recommendation system based on user interest and citations. In *2019 1st international informatics and software engineering conference (ubmyk)* (pp. 1–5).

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, *12*, 331–370.

Casadevall, A., & Fang, F. C. (2010). *Reproducible science* (Vol. 78) (No. 12). Am Soc Microbiol.

Chang, H., Cohn, D., McCallum, A. K., et al. (2000). Learning to create customized authority lists. In *Icml* (pp. 127–134).

Davis, P. M. (2008). Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts? *Journal of the American Society for Information Science and Technology*, *59*(13), 2186–2188.

Dong, R., Tokarchuk, L., & Ma, A. (2009). Digging friendship: paper recommendation in social network. In *Proceedings of networking & electronic commerce research conference (naec 2009)* (pp. 21–28).

Ekstrand, M. D., Kannan, P., Stemper, J. A., Butler, J. T., Konstan, J. A., & Riedl, J. T. (2010). Automatically building research reading lists. In *Proceedings of the fourth acm conference on recommender systems* (pp. 159–166).

Ferrara, F., Pudota, N., & Tasso, C. (2011). A keyphrase-based paper recommender system. In *Italian research conference on digital libraries* (pp. 14–25).

Giles, C. L., Bollacker, K. D., & Lawrence, S. (1998). Citeseer: An automatic citation indexing system. In *Proceedings of the third acm conference on digital libraries* (pp. 89–98).

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: an introduction*. Cambridge University Press.

Jomsri, P., Sanguansintukul, S., & Choochaiwattana, W. (2010). A framework for tag-based research paper recommender system: an ir approach. In *2010 ieee 24th international conference on advanced information networking and applications workshops* (pp. 103–108).

Karypis, G. (2001). Evaluation of item-based top-n recommendation algorithms. In *Proceedings of the tenth international conference on information and knowledge management* (pp. 247–254).

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, *46*(5), 604–632.

Küçüktunç, O., Saule, E., Kaya, K., & Çatalyürek, Ü. V. (2012). Recommendation on academic networks using direction aware citation analysis. *arXiv preprint arXiv:1205.1143*.

Lee, H. J., Kim, J. W., & Park, S. J. (2007). Understanding collaborative filtering parameters for personalized recommendations in e-commerce. *Electronic Commerce Research*, *7*, 293–314.

Liu, S., & Chen, C. (2013). The differences between latent topics in abstracts and citation contexts of citing papers. *Journal of the American Society for Information Science and Technology*, *64*(3), 627–639.

Lu, J., Hoi, S., & Wang, J. (2013). Second order online collaborative filtering. In *Asian conference on machine learning* (pp. 325–340).

Lu, W., Janssen, J., Milios, E., Japkowicz, N., & Zhang, Y. (2007). Node similarity in the citation graph. *Knowledge and information systems*, *11*, 105–129.

McCallumzy, A., Nigamy, K., Renniey, J., & Seymorey, K. (1999). Building domain-specific search engines with machine learning techniques. In *Proceedings of the aaai spring symposium on intelligent agents in cyberspace* (pp. 28–39).

McGuinn, K., Stone, G., Sharman, A., & Davison, E. (2017). Student reading lists: evaluating the student experience at the university of huddersfield. *The Electronic Library*, *35*(2), 322–332.

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., ... Riedl, J. (2002). On the recommending of citations for research papers. In *Proceedings of the 2002 acm conference on computer supported cooperative work* (pp. 116–125).

Moed, H. F. (2006). *Citation analysis in research evaluation* (Vol. 9). Springer Science & Business Media.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web.* (Tech. Rep.). Stanford InfoLab.

Patel, B., Desai, P., & Panchal, U. (2017). Methods of recommender system: A review. In *2017 international conference on innovations in information, embedded and communication systems (iciiecs)* (pp. 1–4).

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web* (pp. 325–341). Springer.

Radicchi, F., Fortunato, S., & Vespignani, A. (2011). Citation networks. *Models of science dynamics: Encounters between complexity theory and information sciences*, 233–257.

Rich, E. (1979). User modeling via stereotypes. *Cognitive science*, *3*(4), 329–354.

Salakhutdinov, R., & Mnih, A. (2008). Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on machine learning* (pp. 880–887).

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd acm conference on electronic commerce* (pp. 158–167).

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In *The adaptive web* (pp. 291–324). Springer.

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, *107*, 1195–1225.

Tanner, W., Akbas, E., & Hasan, M. (2019). Paper recommendation based on citation relation. In *2019 ieee international conference on big data (big data)* (pp. 3053–3059).

Torres, R., McNee, S. M., Abel, M., Konstan, J. A., & Riedl, J. (2004). Enhancing digital libraries with techlens+. In *Proceedings of the 4th acm/ieee-cs joint conference on digital libraries* (pp. 228–236).

Tsolakidis, A., Triperina, E., Sgouropoulou, C., & Christidis, N. (2016). Research publication recommendation system based on a hybrid approach. In *Proceedings of the 20th pan-hellenic conference on informatics* (pp. 1–6).

Valente, A., Holanda, M., Mariano, A. M., Furuta, R., & Da Silva, D. (2022). Analysis of academic

databases for literature review in the computer science education field. In *2022 ieee frontiers in education conference (fie)* (pp. 1–7).

Wakeling, S., Spezi, V., Fry, J., Creaser, C., Pinfield, S., & Willett, P. (2019). Academic communities: The role of journals and open-access mega-journals in scholarly communication. *Journal of Documentation*, *75*(1), 120–139.

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining* (pp. 448–456).

Wang, H., Wang, N., & Yeung, D.-Y. (2015). Collaborative deep learning for recommender systems. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1235–1244).

White, S., & Smyth, P. (2003). Algorithms for estimating relative importance in networks. In *Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining* (pp. 266–275).

Zhang, J., & Luo, Y. (2017). Degree centrality, betweenness centrality, and closeness centrality in social network. In *2017 2nd international conference on modelling, simulation and applied mathematics (msam2017)* (pp. 300–303).