Towards partial fulfillment for Undergraduate Degree Level Programme
Bachelor of Technology in Computer Engineering

A First Stage Project Evaluation Report on:

Research Paper Recommendation System

Prepared by :

Admission No.	Student Name
U17CO099	Pankhi Khandelwal
U17CO101	Himanshu Choudhary
U17CO109	Shubhi Agarwal
U17CO112	Aman Mishra

Class : B.TECH. IV (Computer Engineering) 7th Semester

Year : 2020-2021

Guided by : Dr. Rupa G. Mehta

DEPARTMENT OF COMPUTER ENGINEERING
SARDAR VALLABHBHAI NATIONAL INSTITUTE OF TECHNOLOGY,
SURAT - 395 007 (GUJARAT, INDIA)

Student Declaration

This is to certify that the work described in this project report has been actually carried out and implemented by our project team consisting of

Sr.	Admission No.	Student Name
1	U17CO099	Pankhi Khandelwal
2	U17CO101	Himanshu Choudhary
3	U17CO109	Shubhi Agarwal
4	U17CO112	Aman Mishra

Neither the source code there in, nor the content of the project report have been copied or downloaded from any other source. We understand that our result grades would be revoked if later it is found to be so.

Signature of the Students:

Sr.	Student Name	Signature of the Student
1	Pankhi Khandelwal	
2	Himanshu Choudhary	
3	Shubhi Agarwal	
4	Aman Mishra	

Certificate

This is to certify that the project report entitled Research Paper Recommendation System *is prepared and presented by*

Sr.	Admission No.	Student Name
1	U17CO099	Pankhi Khandelwal
2	U17CO101	Himanshu Choudhary
3	U17CO109	Shubhi Agarwal
4	U17CO112	Aman Mishra

Final Year of Computer Engineering and their work is satisfactory.

SIGNATURE :

GUIDE JURY HEAD OF DEPARTMENT

Abstract

With the ever-increasing amount of data being accessible over the web, the need to make recommendation systems more and more accurate is pressing. The approaches involving the semantic comparison of documents tend to become infeasible when querying a very large amount of data. This problem is not just restricted to a few domains but is slowly and gradually becoming a part of almost all the domains involving look-up for bulk data. The same goes for the research community as well. Research paper recommendation aims to recommend new articles that match researchers' interests. It has become an attractive area of study since the number of scholarly papers increases exponentially. There are already approaches that make use of personalized suggestions based on user information but these approaches deal with the issues of lack of data for a newly registered user. This problem is referred to as cold-start. Cold-start problem occurs when trying to suggest a newly registered user regarding whom we don't have much data and hence the recommendations systems are not able to figure out what to suggest.

The aim of this project is to devise a graph-based network involving the use of Natural Language Processing based techniques for filtering to reduce the search space for documents. This will be followed by a personalized look-up in the search space and ultimately a semantic comparison in the resultant search space. The goal is to reduce the search space for semantic comparisons from as large as 1,00,000 documents to a few thousand documents. This will lead to really fast searches. Moreover, the addition of personalized search approaches will also contribute to increased accuracy in search results.

Keywords: Semantic Comparisons - Cold-Start - Graph - Search Space - Personalized Search - Keyword Networks - Natural Language Processing

Contents

List of Figures

List of Acronyms

CF	Collaborative Filtering
KCK	Keyword-Citation-Keyword
NLTK	Natural Language Toolkit
RAKE	Rapid Automatic Keyword Extraction
TF-IDF	Term Frequency Inverse Document Frequency

List of Symbols

δ	Greek Symbol Delta
$\cap$	Intersection of Sets
$\cdot$	Dot Product of Vectors

Chapter 1

Introduction

The researchers around the globe find it an extensively time-consuming task to search for the related work and articles throughout the course of their research and dissertation from the digital repositories. Due to an upsurge in the number of publications (growth is exponential at a rate of 3.7% per annum) during the last few decades, the corpus for this search procedure has become manifold. The problem also gets intensified when the researchers possess very skimmed knowledge of operating these repositories. Performing these search operations manually is also a very time-consuming and tedious task making it almost infeasible to conduct.

In this project, we aim to build a graphical-network-model based on keywords, which co-relates the input publication to its n-neighbors on the basis of their semantic relationships and associations. The scope of this project provides a lead to the related research works in which only the quantitative correlation between the co-existing keywords is analyzed. Basically the proposed recommender is a software application that helps a user to search for the related and relevant documents automatically based on certain preferences described in the query thereby reducing the manual load.

1.1 Applications

A researcher in the current scenario has to manually search for the similar publications corresponding to his research to use them for preparing and studying the content related to his research or to perform the corresponding literature survey or to quote the citations. Therefore, a system that can automatically query an existing database to structure the co-related documents in a network and then look-up for the required one in the search space by reducing the search cost as compared to that of traditional means is of great significance and importance for practitioners.

Along with researchers, any educational and research institution can also customize this module in order to provide an efficient e-library system for the members of their organization.

1.2 Motivation

The structure of the research papers has made it significant to devise a system that checks for both the weighted and semantic similarity between documents before recommendation.

The bulk of publications present and its exponential growth rate have forced the existing search engines or recommendation systems to use enormous storage and search-space thereby increasing the search cost.

Moreover, inexperienced researchers find it quite challenging to deal with digital repositories. Sometimes, due to unfamiliarity of the search criteria incorporated in these search engines, land them in a situation where they get irrelevant data resulting in unnecessary delay in their progress.

Hence, a handy system that is easy to understand and exploit will receive an appreciation among the practitioners and researchers.

1.3 Objectives

The primary objective of this project is to explore innovative and efficient graphical network-based approaches to reduce the search space for semantic comparisons of documents while at the same time maintaining a high level of accuracy. A research paper almost all of the time follows a fixed format - a title then an abstract followed by the content of the paper and in the end a list of all the related citations. Networks based on co-occurrence of citations or keywords among papers are being explored. At the same time, efficient preprocessing techniques to extract the desired data for instance keywords from papers are also being looked into.

1.4 Contribution

To fulfill the project objectives hitherto discussed, in this project report, we have conducted an in-depth survey of pre-existing literature regarding the topics of document structuring and similarity analysis of legal documents. Furthermore, we have performed initial levels of pre-processing involving NLP based techniques like lemmatization, stemming, and created two different networks. At the same time, we have also tested one of the many approaches for keyword extraction from research papers that is to be used as a part of the preprocessing pipeline while creating the network.

1.5 Organization of project report

Chapter 1 includes a brief introduction of the project, its application in the real world, the motivation behind choosing this topic, its objectives, and our contribution towards this topic.

Chapter 2 includes the literature survey which gives an overview of the research work implemented in this domain and gives an overview of some of the concepts relevant to our project. A detailed emphasis has been laid on the existing methodologies and modules and the scope and technologies involved in developed systems have been explained in depth. Different content-based, citation-based, collaborative-filtering based, topic-based and keywords-based recommendation systems are discussed in-depth and conclusions are drawn.

Chapter 3 includes the proposed methodologies and its logical development.

Chapter 4 includes the implementation details and the tech stacks which we are utilizing to implement the project.

Chapter 5 presents our conclusions drawn and guides the reader to possible future works that can be built upon the proposed system.

Chapter 2

Literature Survey

The overall workflow for framing a concept of dealing with structured data like research publications follows the path from forming the graphical networks. Then, ranking the suggestions made by each, using some comparative algorithms or ensembling the results, can also serve the purpose. Several ways have been employed by various researchers to analyze how to form a system that efficiently recommends the publications. Prior works have been discussed here, along with their advantages and disadvantages to conclude these implemented methodologies. Further, some existing research paper recommendation search engines and their adopted approaches have been explained to get insight into such systems' practicality. After that, different methods deployed to measure the similarity between two natural language text entities and previous efforts made to analyze the similarity between documents have been described.

2.1 Research Paper Recommendation Approaches

Various researchers in the above-explained domain have already explored several techniques. Some prominent ones out of those are as described below.

2.1.1 Citation-Based

The proposed method tries to derive meanings from the latent relationships that co-exist between any publication and its citations. These associations help in devising a system regardless of the domains explained through the article and authors' specializations as well. It generally uses bibliographic coupling [1] or co-citation analysis [2] to find publications similar to the given input. Scholastic search engines like Google Scholar work on the conventional text mining approach and citation counts.

2.1.2 Content-Based

This system recommends an item to a user based on the description of the item and the profile of the user's interests. It matches the user's interests and profile with the content object and then recommends new items. Since we have to examine all the items in a set to correlate it with the user's interests, it takes a vast amount of item set, which is a considerable disadvantage [3].

2.1.3 Collaborative Filtering-Based

As stated by Seikyung Jung, Juntae Kim, and J. L. Herlocker [4], Collaborative Filtering (CF) identifies the need of the user before making recommendations. Hence the search becomes more specific rather than generic as it used to be in content-based recommendation systems. Information sources get rated in the first step resulting in a model that remains saved in the recommendation system. Whenever a query has been raised, the ratings get analyzed as per the user's information needs, and the output is generated. Emphasis has been laid on the fact that information need ratings have been utilized in contrast to the classical collaborative filtering methods wherein users' ratings were considered.

2.1.4 Topic-Based

Instead of using some labeled or pre-classified data, the cardinal topics of a publication can also be incorporated as a tool for the recommendation system. Chenguang Pan and Wenxin Li [5] suggest that topic modeling principles can be used to study the thematic similarity between the topics. Recommendation results will be deduced by comparing the similarity parameter.

2.1.5 Keywords-Based

Content-based recommenders typically build attribute vector representations of contents and user preferences and generate recommendations according to the degree of similarity between user interests and items, as explained in [6]. Basically, two types of networks are possible [7], [8]: one is a Co-word network in which an undirected graph has been generated by analyzing the co-occurrence nature of the keywords, and the other one is a Keyword-Citation-Keyword (KCK) network, it is a directed network in which the direction specifies the citing link between two publications and then the co-occurrence keyword property gets utilized [7], [9], the more links point to keywords for a particular paper signifies the importance of keywords thus providing more information as compared to the Co-word network. It is worth mentioning that no link exists between the same keywords in a KCK. In [9], the KCK approach has been demonstrated through its application to nano-related Environmental, Health, and Safety (EHS) risk literature.

2.1.6 Meta-Data-Based

Recommendation systems are susceptible to the problem of cold-start. The user cold-start problem refers to the task of recommending items to a new user, whose previous item preferences are not present in the system [10]. To resolve this issue, metadata, like demographic information of [11], can be used. But the problem with this approach is that metadata may not always be available and may not always be right. For example, the sensitive demographic information of users might not always be known as being used in [11].

2.1.7 Conclusion of Prior Works

All the methods, as discussed above, have both issues and advantages. As mentioned in [7], the co-word analysis only focuses on the count of keywords rather than their semantic relationships, making it challenging to derive context based information. Topic-based approaches and summarization techniques involve huge data processing and storage limitations due to the size of the data under consideration ranging to the extent of whole research paper data in summarization approaches. Only the listed keywords do not provide the main ideas of the entire literature; hence semantics play an essential role in overall development. The author in [5] has emphasized the need for textual information of any literature work and proposed a topic analysis approach by considering thematic similarity. Homographs identification becomes insignificant in citation analysis [12]. It sometimes results in assigning a publication to the wrong author. Despite having comparatively successful results using collaborative-filtering, it encounters the first-rater problem, which makes it compulsory that the items involved must be rated by at least one of its neighbors as discussed in [13]. Moreover, CF also faces sparsity problem due to the unlikeliness of users in rating all the available items.

2.2 Research Paper Recommendation Systems and their Flow of Operation

There are various open-source search engines for providing quick access to scholarly articles. They use different Natural Language Processing, semantic analysis, and ranking algorithms to serve the purpose. Some of these are as discussed below.

2.2.1 JournalFinder

Powered by the Elsevier Fingerprint Engine, JournalFinder uses semantic search technology and field-of-research specific vocabularies to match your abstract to relevant Elsevier journals. The Elsevier Fingerprint Engine applies various Natural Language Processing techniques to mine the text you enter into the JournalFinder for mentions of key concepts spanning all the major scientific disciplines and creates a structured index of weighted terms that defines the text, known as a Fingerprint. JournalFinder then compares the Fingerprint of your abstract with the Fingerprints of all journal articles in Scopus and recommends up to 50 of the most relevant journals for you to consider [https://www.elsevier.com/solutions/elsevier-fingerprint-engine].

2.2.2 Google Scholar

Google Scholar provides a simple way to search for scholarly literature broadly. From one place, users can search across many disciplines and sources: articles, theses, books, abstracts, and court opinions. Google Scholar aims to rank documents the way researchers do, weighing the full text of each document, where it was published, who it was written by, as well as how often and how recently it has been cited in other scholarly literature.

2.2.3 Springer

Springer uses semantic technology to help the user quickly choose the right journal for the paper user wants. Users can refine the results based on requirements for the Impact Factor or publishing model, including an option to match to journals that are fully open access or have open access options.

2.3 Keyword Extraction Techniques

It is cumbersome to deal with the massive volume of data or content available over any network. The same applies to the publications' content, thus raising a need to summarise it. Keywords are a far shorter summary, as stated in [14]. There are various methods available to extract keywords from available text documents. Some prominent ones are noted here.

2.3.1 Simple Statistical Approaches

These approaches don't require training data to extract the most important keywords in a text. Since they rely on stats, they may overlook relevant words or phrases mentioned once but are still considered appropriate. Different Statistical approaches:

2.3.1.1 Word Frequency

Word frequency consists of listing the words and phrases that most commonly appear. Word frequency approach considers documents as a mere bag of words leaving aside crucial aspects to the meaning, structure, grammar, and sequence of words.

2.3.1.2 TF-IDF

TF-IDF measures how important a word is to a document in a collection of documents. This calculates the number of times a word appears in a text and compares it with the inverse document frequency. Multiplying these two quantities provides the TF-IDF score of a word in a document. The higher the score is, the more relevant the word is to the document.

2.3.1.3 RAKE

RAKE is a well-known extraction method that uses a list of stop words and phrase delimiters to detect the most relevant words or phrases in a piece of text. The first thing this method does is splitting the text into a list of words and remove stop words from that list. Then, the algorithm splits the text at phrase delimiters and stopwords to create candidate expressions. Once the text has been split, the algorithm creates a matrix of word co-occurrences. After that matrix is built, and words are given a score.

2.3.2 Linguistic Approaches

Linguistic Approaches use the linguistic properties of the words, sentences, and documents. Lexical, syntactic, semantic, and discourse analysis are some of the most commonly examined properties.

2.3.3 Machine Learning Approaches

Machine Learning approaches consider supervised or unsupervised learning from the examples, but related work on keyword extraction prefers the supervised approach. Supervised machine learning approaches induce a model that is trained on a set of keywords. They require manual annotations of the learning dataset, which is too tedious and inconsistent. Thus, supervised methods require training data and are often dependent on the domain. A system needs to re-learn and establish the model whenever a domain changes [15].

2.4 Similarity Measures

This section discusses the various approaches to estimate the similarity between documents.

2.4.1 Cosine Similarity

Cosine Similarity is the most common similarity measure used when documents are represented as term vectors. Here, the cosine of the angle between the term vectors in vector space corresponds to the correlation between these vectors [16]. The closer the cosine similarity value is to 1, the smaller the angle between the vectors, and hence, the greater the similarity.

Given two documents vectors, d1 and d2, over the term set T, the cosine similarity between them is calculated as:

$$CosineSimilarity(d1,d2) = \frac{d1 \cdot d2}{|d1| \cdot |d2|} \tag{2.1}$$

2.4.2 Bibliographic Coupling

Bibliographic Coupling is a link based similarity measure that uses citations to assert similarity between two documents. Two documents are said to be bibliographically coupled if they reference one or more common works in their bibliography. Thus, instead of the contents of the documents, only citations are compared to determine similarity [17].

Given two documents, D1 and D2, the set of out-citations of D1 and D2 are defined as OC1 and OC2, respectively. The bibliographic coupling between the two given documents is then defined as the number of common out-citations, or:

$$B(D1,D2) = OC1 \cap OC2 \tag{2.2}$$

The two documents are determined as similar if their bibliographic coupling is greater than or equal to the threshold value, δ .

2.4.3 Co-Citation

Co-citation is similar to Bibliographic Coupling because it uses citations to assert similarity, but differs in the sense that while bibliographic coupling links source document citations, co-citation links the number of times, the two given documents are cited together. Thus, if document D cites two documents, D1 and D2, together, D1 and D2 are said to be co-cited. The more co-cited the two documents are, the higher the likelihood that they are semantically similar [18].

Given two documents, D1 and D2, the set of in-citations of D1 and D2 are defined as IC1 and IC2, respectively. The co-citation between the two given documents is then defined as the number of common in-citations, or:

$$C(D1, D2) = IC1 \cap IC2 \tag{2.3}$$

The two documents are determined as similar if their co-citation is greater than or equal to the threshold value, δ.

2.4.4 Keyword Co-Occurrence

Keyword Co-Occurrence is similar to Co-citation. The only difference is that it makes use of keywords rather than citations. Thus, if document D had two keywords K1 and K2 together, K1 and K2 are said to be co-occurred [19].

Given two documents, D1 and D2, the set of keywords of D1 and D2 are defined as K1 and K2, respectively. The co-occurrence of keywords between the two given documents is then defined as the number of common keywords, or:

$$C(D1, D2) = K1 \cap K2 \tag{2.4}$$

The two documents are determined as similar if their co-occurrence of keyword is greater than or equal to the threshold value, δ.

Chapter 3

Proposed Work

The proposed project aims to produce a deliverable that can be utilized among research institutions and individuals to serve their purpose of an extensive search of related publications. Moreover, the storage space and search-space complexity would also be dealt with by reducing the need for entire publication content for recommendation to some structured parts. At later stages, a ranking algorithm will be used to collaborate the results produced during intermediate stages. Context-based factor and semantic analysis are also considered for graphical-network creation of the available dataset. The whole project workflow will be as shown in **Figure 3.1** and **Figure 3.2**.

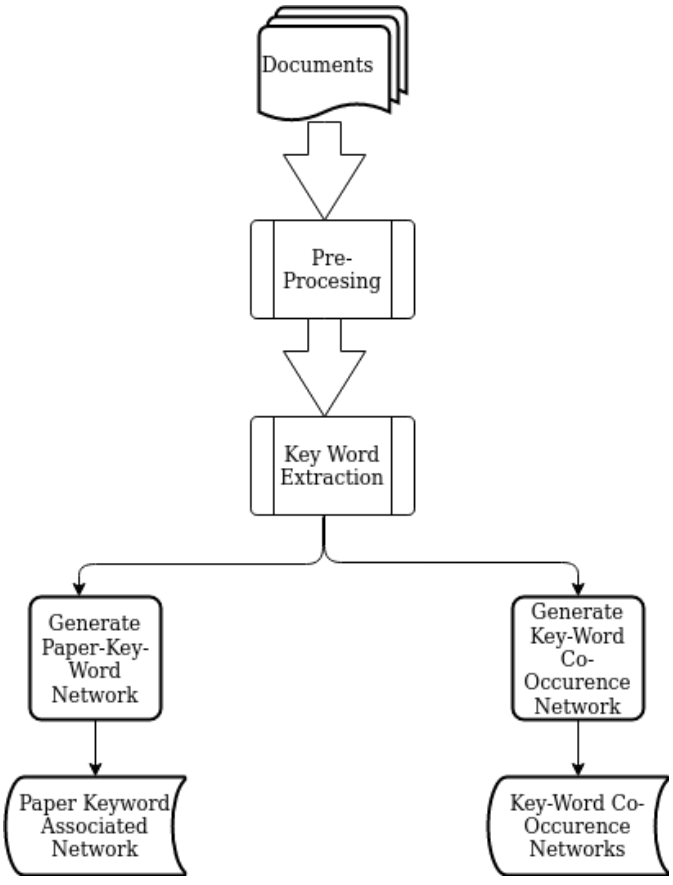

Figure 3.1: Block Diagram depicting the preprocessing and network creation stages

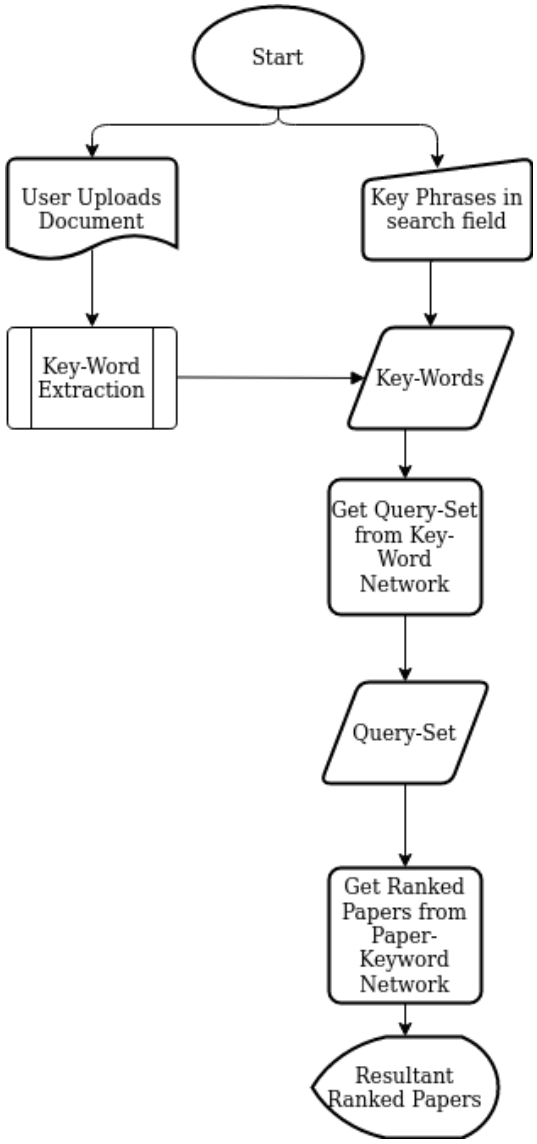

Figure 3.2: Block Diagram depicting the user interaction with the system

3.1 How does the user interact with the system?

The proposed approach holds two ways by which the user can interact with our system:

3.1.1 By Uploading a Research Paper

Here the user can directly upload a research paper and, as an output, get all the research papers that are related to it. Internally, the key phrases will be extracted from the document, and the keyword-based network will be queried based on these keywords to get the desired results.

3.1.2 By Providing Key Phrase

Here the user can provide key phrases. Based on these key phrases the keyword-based network will be queried, and as an output user will get the related papers.

3.2 How does the system work?

As shown in **Figure 3.1** and **Figure 3.2**, the workflow of the proposed system can be explained in the following manner:

- From the user's query, a set of all possible keywords will be generated (either provided by the user directly or extracted from the provided content) and fed to the system during the initial stages, as explained in the previous section.
- Now, the model prepares a query set by getting relatable keywords obtained for the input set of keywords or keyphrases using context-information and semantic and thematic relationships.
- This query set is further utilized to get a set of related and interlinked documents by analyzing the matrix or map pairs of publications and keywords using both quantitative weights and semantic similarities.
- From this collection of related papers, top n-ranked papers will be recommended to the user as an output.

Chapter 4

Simulation and Results

Under implementation, a bit of pre-processing using stemming, lemmatization, and tokenization has been carried out. Two types of graph-based networks (using keywords and paper id as nodes) are also generated as initial product development stages. Some keyword extraction techniques are also explored during the research and implementation tasks.

4.1 Pre-processing

Data-Preprocessing techniques in Data Mining are used to pre-process the raw data in a structured, well-defined, and clean format. Some of the pre-processing methods are as described below:

Stemming: Stemming is a technique used to extract the base form of the words by removing affixes from them. It is just like cutting down the branches of a tree to its stems. Python's *LancasterStemmer* class of Natural Language Toolkit (NLTK) module can be utilized for the same.

Lemmatisation: Lemmatization is similar to stemming. The process generates a 'lemma' as an output. Lemma is a root word rather than root stem, the output of stemming. After lemmatization, we will be getting a valid word that means the same thing. Python's *WordNetLemmatizer* class of NLTK module can be utilized for the same.

Tokenization: Tokenisation is splitting up a larger body of text into smaller lines, words, or even creating words for a non-English language. The various tokenization functions in-built into the NLTK module of Python itself can be utilized for the same.

4.2 Keyword Extraction

Keyword Extraction is an automated process of extracting words, group of words, or expressions from the given text. RAKE algorithm of Python's *rake-nltk* module performs the operation by analyzing the frequency of words' appearance and their correlations with other words.

4.3 Network 1 (Paper IDs as nodes)

A list of publication IDs for all the publications present in the dataset is generated in this system. Each element of the list acts as a node for an undirected graph. The weight of an edge is calculated by calculating the total number of common keywords between the two research papers depicted by the two nodes corresponding to that edge (see **Figure 4.1**).

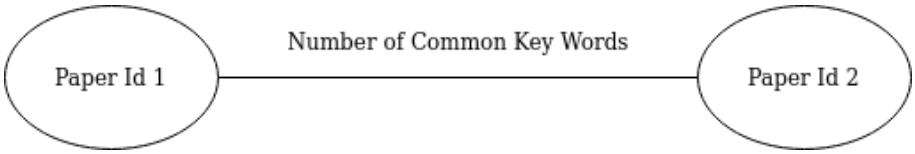

Figure 4.1: Paper-Keyword Network

4.4 Network 2 (Keywords as nodes)

A list of common keywords extracted from all the publications present in the dataset is generated in this system. Each element of the list acts as a node for an undirected graph. The weight of an edge is calculated by calculating the co-occurrence frequency between the two nodes corresponding to that edge (see **Figure 4.2**).

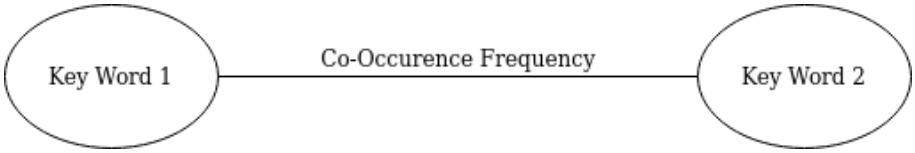

Figure 4.2: Keyword Co-Occurrence Network

Chapter 5

Conclusion and Future Work

As it has been witnessed that the system for an easy recommendation of scholarly articles is of great significance today due to the various quoted reasons, this project work would henceforth provide a probable solution to this demand. The main focus here is to develop a keyword-based recommendation system that also takes semantic relationships into consideration during the model development and utilization phases. Another worth noting fact is the capability of the model to reduce the search-space.

Further advancements in the project encompass the extension of its domain by combining multiple article recommendation approaches, namely CF, content-based, and the semantic keyword approach. This whole flow aims to develop a personalized recommendation system that also considers users' data collected from their access log with the system.

References

[1] O. Hanif, Z. Donghua, W. Xuefeng, and M. S. Nawaz, "Refining the measurement of topic similarities through bibliographic coupling and lda," *IEEE Access*, vol. 7, pp. 179 997–180 011, 2019.

[2] B. Gipp and J. Beel, "Citation proximity analysis (cpa) : A new approach for identifying related work based on co-citation analysis," in *Proceedings of the 12th International Conference on Scientometrics and Informetrics, vol. 1*, B. Larsen, Ed. São Paulo: BIREME/PANO/WHO, 2009, pp. 571–575. [Online]. Available: http://www.sciplore.org/wp-content/papercite-data/pdf/gipp09a.pdf

[3] P. Lops, M. de Gemmis, and G. Semeraro, *Content-based Recommender Systems: State of the Art and Trends*, 01 2011, pp. 73–105.

[4] Seikyung Jung, Juntae Kim, and J. L. Herlocker, "Applying collaborative filtering for efficient document search," in *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, 2004, pp. 640–643.

[5] Chenguang Pan and Wenxin Li, "Research paper recommendation with topic analysis," in *2010 International Conference On Computer Design and Applications*, vol. 4, 2010, pp. V4–264–V4–268.

[6] D. De Nart and C. Tasso, "A personalized concept-driven recommender system for scientific libraries," *Procedia Computer Science*, vol. 38, pp. 84 – 91, 2014, 10th Italian Research Conference on Digital Libraries, IRCDL 2014.

[7] Q. Cheng, J. Wang, W. Lu, Y. Huang, and Y. Bu, "Keyword-citation-keyword network: a new perspective of discipline knowledge structure analysis," *Scientometrics*, vol. 124, no. 3, pp. 1923–1943, September 2020.

[8] H. Li, H. An, Y. Wang, J. Huang, and X. Gao, "Evolutionary features of academic articles co-keyword network and keywords co-occurrence network: Based on two-mode affiliation network," *Physica A: Statistical Mechanics and its Applications*, vol. 450, pp. 657 – 669, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037843711600025X

[9] S. Radhakrishnan, S. Erbis, J. A. Isaacs, and S. Kamarthi, "Correction: Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature," *PLOS ONE*, vol. 12, no. 9, pp. 1–1, 09 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0185771

[10] H. Bharadhwaj, "Meta-learning for user cold-start recommendation," 04 2019.

[11] A. K. Pandey and D. S. Rajpoot, "Resolving cold start problem in recommendation system using demographic approach," in *2016 International Conference on Signal Processing and Communication (ICSC)*, Dec 2016, pp. 213–218.

[12] B. Gipp, J. Beel, and C. Hentschel, "Scienstein: A research paper recommender system," 01 2009.

[13] R. Torres, S. M. McNee, M. Abel, J. A. Konstan, and J. Riedl, "Enhancing digital libraries with techlens," in *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004.*, 2004, pp. 228–236.

[14] Y. HaCohen-Kerner, "Automatic extraction of keywords from abstracts," 09 2003, pp. 843–849.

[15] S. Beliga, A. Meštrović, and S. Martincic-Ipsic, "An overview of graph-based keyword extraction methods and approaches," *Journal of Information and Organizational Sciences*, vol. 39, pp. 1–20, 07 2015.

[16] A. Huang, "Similarity measures for text document clustering," *Proceedings of the 6th New Zealand Computer Science Research Student Conference*, 01 2008.

[17] B. Jarneving, "Bibliographic coupling and its application to research-front and other core documents," *Journal of Informetrics*, vol. 1, no. 4, pp. 287 – 307, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1751157707000594

[18] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, pp. 265 – 269, 07 1973.

[19] S. Radhakrishnan, S. Erbis, J. A. Isaacs, and S. Kamarthi, "Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature," *PLOS ONE*, vol. 12, no. 3, pp. 1–16, 03 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0172778

Acknowledgement

We, the final year undergraduate students of Sardar Vallabhbhai National Institute of Technology, Surat, are overwhelmed in all humbleness to acknowledge our deep gratitude to all those who have helped us to put our ideas to perfection and have assigned various tasks, well above the level of simplicity and into something concrete and unique.

We wholeheartedly thank Dr. Rupa G. Mehta, Associate Professor, Computer Engineering Department, SVNIT Surat, for having faith in us, selecting us to be a part of this worthwhile project, and constantly motivating us to do better. Her insight and knowledge of the subject matter steered us through the research. We are also very thankful to Mr. Mayur Makwana, Ph.D. Scholar, Computer Engineering Department, SVNIT Surat for his valuable suggestions and critical advice. With their brilliant guidance and encouragement, we were able to complete all the tasks assigned to us within the given time frame. We got a chance to see the stronger side of our technical and non-technical aspects during the process.

We would also like to thank Dr. Mukesh A. Zaveri, Head of Department, Computer Engineering Department.

Last but not the least, many thanks to SVNIT, Surat, and its staff for providing an enriching platform necessary in acquiring quality and sufficient knowledge to accomplish the goals perceived.