Social Network Analysis in Scientific Communities

Master Thesis

A thesis submitted in fulfilment of the requirements
for the Masters Degree

in the

Knowledge-Based Systems Group,
Department of Computer Science

Technische Universität Kaiserslautern

June 2016

Author:

Akansha Bhardwaj

Supervisors:

Prof. Dr. Prof. h. c. Andreas Dengel

Dr. Sheraz Ahmed

Declaration of Authorship

I, Akansha Bhardwaj, declare that this thesis titled, "Social Network Analysis in Scientific Communities" and the work presented in it are my own. I confirm that:

- This work was done mainly while in candidature for a masters degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signature: ______________________________

Date: ______________________________

“When you do things from your soul, you feel a river moving in you, a joy.”

Rumi

Abstract

This work makes a two-fold contribution towards social network analysis in scientific communities. Firstly, it presents an end-to-end system to analyze a scientific community by identifying hidden patterns and structures within a community. As most scientific communities are driven by the colloquial *'publish or perish'* mindset, the quantity of publications put out every year is increasing and traditional productivity metrics which focus only on quantitative data i.e., number of times a paper is cited, are becoming rapidly insufficient. In addition, currently there is no tool available to analyze the trends in a community in terms of citations and collaborations. The presented system serves as an extensive analysis and visualization tool to quantitatively and qualitatively analyze the contributions of researchers towards a community. The presented system is based on *text mining* and *social network analysis*, where the important information (title, authors, abstract, keywords and references) is extracted from scientific publications. The extracted data is then used to build the community network. Specifically, different performance indicators are computed by analyzing the scientific community network. This helps in identifying hidden patterns in the community e.g., number of connected groups, cliques, pattern of citations, etc. Secondly, this thesis provides an interactive visualization tool that offers a multitude of options to explore a scientific community. To demonstrate the feasibility of the presented framework, a case study is performed on International Conference on Document Analysis and Recognition (ICDAR) scientific community, which is the largest academic document conference. The evaluation results show that the systems is able to find a significant amount of hidden patterns in the community, e.g., number of connected groups, cliques, pattern of citations, etc.

Acknowledgements

I would like to thank Prof. Dr. Andreas Dengel for introducing the topic and providing master thesis opportunity at Knowledge-Based Systems Group, Technische Universität, Kaiserslautern and Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI).

Furthermore, I am extremely grateful to my supervisor Dr. Sheraz Ahmed for his constant guidance and motivation throughout the thesis work.

Finally, I would like to thank my friends Vinay and Benjamin, for being my sounding board for all ideas and keeping me motivated till the successful completion of this work.

I will be grateful to you all, for all your care and support.

List of Figures

List of Tables

Abbreviations

OECD	**O**rganisation for **E**conomic **C**o-operation and **D**evelopment
ICDAR	**I**nternational **C**onference on **D**ocument **A**nalysis and **R**ecognition
PDF	**P**ortable **D**ocument **F**ile
NER	**N**amed **E**ntity **R**ecognition
SNA	**S**ocial **N**etwork **A**nalysis
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
HMM	**H**idden **M**arkov **M**odel
SVM	**S**upport **V**ector **M**achines
CRF	**C**onditional **R**andom **F**ields
XML	**E**xtensible **M**arkup **L**anguage
NLP	**N**atural **L**anguage **P**rocessing
POS	**P**art **O**f **S**peech
SRL	**S**emantic **R**ole **L**abelling
PSG	**P**hrase **S**tructure **G**rammar
CSV	**C**omma **S**eparated **V**alue
SNA	**S**ocial **N**etwork **A**nalysis

This thesis is dedicated to my friends and family. Your existence is a blessing…

Contents

Chapter 1

Introduction

For every organisation, it is imperative to demonstrate the significance of its mission, and its success in achieving that mission. This demonstration becomes more important in a scientific community where stakes are high. In an atmosphere motivated by the colloquial *'publish or perish'* mindset, numerous research articles are published every year.

With more and more publications, it becomes important to develop 'productivity' indicators. Traditionally, it was sufficient to collect data on numbers of scientific articles and publications, classified by authors and/or institutions. External peer review is another traditional measure for evaluating scientific research but its scope is often limited due to the substantial costs involved. For this reason, fields like 'bibliometrics' and 'scientometrics' are being used to supplement this process. These fields introduce more multidimensional methods to evaluate a scientific work.

Bibliometrics and scientometrics are two closely related approaches to measuring scientific publications and science in general, respectively. According to Organisation for Economic Co-operation and Development (OECD) Glossary of Statistical Terms[1], "Bibliometrics is a statistical analysis of books, articles, or other publications".

Citation analysis is one of the most widely used methods of 'bibliometrics'. Citation analysis is the examination of patterns and frequency of citation in a scientific work. Some graph-based methods can also be used to perform citation analysis. This data can convey a lot of information about scholarly communication, networks of scholars and development of areas of knowledge over time.

[1] https://stats.oecd.org/glossary/

1.1 Motivation

Though there is an availability of various citation analysis measures like bibliographic coupling, co-citaion proximity analysis, there is no end-to-end pipeline that can develop visualisations and help evaluate a scientific community simply from a collection of articles. It is an interesting research question to study a scientific community through measures related to citation trends among authors. Since the co-authorship information is also available it is also interesting to check if there is correlation between frequent co-authors and their citation patterns. This might help to get insights into how much a scientific community is well-knit and how does a key player's citation patterns differ from others.

As the articles from scientific conferences are in Portable Document File (PDF) format, as a pre-requisite it was important to develop a text mining tool that extracts the relevant information from these articles and transforms it into a structured dataset. This data would be later used to draw insights into community through visualisations. Visualisations are important as they are quite often more informative than numbers; they help to create hypothesis and trigger further research.

Developing a tool that aids this research on a scientific community along with a method that can quantify the scientific impact of an author was the motivation behind this work.

International Conference on Document Analysis and Recognition(ICDAR) is an international scientific conference about character and symbol recognition, printed/handwritten text recognition, graphics analysis and recognition, document analysis, document understanding,historical documents and digital libraries, document based forensics, camera and video based scene text analysis. Since 1991, it is held every second year. The data from ICDAR 1993 – 2015 has been used to show the feasibility of the framework presented.

1.2 Contributions

This thesis makes a two-fold contribution to the '*bibliometrics* community. First, it presents an end-to-end pipeline to visualise the structure of a scientific community from a collection of articles in PDF format. Some novel approaches/contribution of this pipeline include:

- An open-source module that can convert PDF to a XML file with metadata tags of title, author, abstract, keywords (if present) and references. References can be extracted from scholarly Portable Document Files(PDF) without the use of regular expressions or label information, if needed.

- Using Named Entity Recognition (NER) to extract author names accurately.
- A novel approach using Girvan-Newman clustering to find collaboration communities.

The second contribution is a citation analysis tool which offers a multitude of options to interactively explore a scientific community. Some of the new metrices used here are:

- An overlap index graph that represents a collaboration, citation counts as bargraph for an author.
- A word cloud from keywords of an author's domain.

1.3 Thesis structure

This thesis is structured as follows. Chapter 1 gives an introduction to the topic and explains the motivation and the contributions of this work. Chapter 2 contains background information about the work. This Chapter tries to build a base for readers who are not well-acquainted with this field. Chapter 3 is titled "Related works" and describes the work of other researchers in this area. Chapter 4 deals with the methodology adopted to solve the current problem statement. Chapter 5 describes the evaluations done on the dataset and Chapter 6 discusses the conclusions and future work of this research.

Chapter 2

Background

This chapter aims to familiarize the reader with the essential context needed to understand the research problem and its significance. After reading, it is hoped that the reader can better associate with the topic of this study. Each section contains a detailed background information on the tasks adopted in this research work.

This chapter is structured into the following sections. Section 2.1 discusses the concepts related to mining of the textual data from PDF documents. Section 2.2 discusses the basics related to the field of Social Network Analysis (SNA). In last section of this chapter, Section 2.3 various data visualisation techniques are briefly discussed.

2.1 Information Extraction

Text mining techniques have been employed to extract textual data from scientific publications. In the following sub-sections, there is a brief introduction to all the technical terms and concepts which will be used throughout. It is important to mention here that '*References*' section refers to the section containing the list of the works which are cited in an article. Each individual work in this list is also known as a '*citation string*'. Usually, each *citation string* contains information on authors, title of the work cited, year of publication. The purpose of the text mining tasks is to extract the author(s) names present in each citation string. If an author A mentions the work of another author B, it can also be stated as '*A cites B*'.

2.1.1 From unstructured data to structured Information

Unstructured data is a generic label for the description of data which is not contained in a data structure or a database. This data can be 'textual' when present in sources like email messages, presentations, text documents or, 'non-textual' when present in media like images, audio or video files. Structured data is manageable because of its storage size and also because it is easily searchable and semantically more significant. There is a need across several domains to convert this unstructured data into semantically significant data.

In this work, the data was present in PDF documents and had to be extracted with PDF to text extraction techniques before other text mining operations.

2.1.2 PDF to Text

PDF is the most commonly used file format for scientific publications. It is very important to have effective means of extracting text from these PDFs in a layout-aware mechanism as scholarly articles can be in multiple column formats depending on the field of research and the respective journal or conference. Though there are several open-source projects available for this purpose [LaPDF[1], PdfToText[2], PdfBox[3], DocearPDFInspector[4]], in this work Docear's PDF Inspector has been preferred for processing the PDF documents as it is completely independent, 100% complaint to the PDF standard and open-source.

2.1.3 Text Mining

After converting the data from a non-textual to a textual format, the next step is to extract relevant information from it. This process is referred to as '*Text Mining*'. '*Text Analytics*' is the application of text-mining techniques to solve business problems. Most often, text mining involves using NLP, statistical modelling and other machine learning techniques. This can be a challenging task as natural language is often inconsistent with its syntax and semantics.

As the task at hand, required to extract names of authors that are owners of the work and whom they choose to cite in their work, it was preferred to choose a NER system. NER is a subtask of information extraction which classifies elements in text into pre-defined

[1] https://www.force11.org/node/4665
[2] http://linux.die.net/man/1/pdftotext
[3] https://pdfbox.apache.org/
[4] https://www.docear.org/software/add-ons/docears-pdf-inspector/

categories such as the names of persons, organizations, location. The NER library used in this work was based on a neural network and was developed by Collobert and Weston [15]. It has shown state-of-the-art results, is written in C, is fast and self-contained. Thus, it was a good choice for the current task.

The next step after dataset creation was to make data more organised and to extract more information from it.

2.1.3.1 Data Cleaning

This step deals with detecting and removing errors and inconsistencies from data in order to improve the quality. It is often possible to have some erroneous information with numerous false positives and inconsistencies which need to be removed from the data.

Though trivial and appearing to be irrelevant, this is the most important step for the removal of ambiguities present in names. This problem plagues such a data quite often.

2.2 Social Network Analysis

SNA focuses on the investigation into a network of a set of social entities, e.g, persons, organizations, nations, web sites, et cetera. This investigation often takes place with a graph theoretical approach, where a social network is conceptualized as a graph. There is a set of vertices that represent social actors and a set of edges representing one or more social relations between them.

A social network, however, is more than a graph because it contains additional information on the vertices and edges. In the context of this work, wherever mentioned, '*citation strength'* refers to the number of times an author cites another author.

2.2.1 Centrality

There is a long history of the concept of 'centrality' in social networks, mainly attributed to Alex Bavelas [3]. When a social network is regarded as a channel for the exchange of information, being central in this exchange is often related to influence and position.

There are several centrality measures which help to find the most important node in a network. These can be different with respect to 'kinds of trajectory' and 'methods of speed' as explained by Borgatti [10] in his work.

2.2.2 Clustering

Clustering of dataset works on the principle of ‘homophily’, *love of the same.* Homophily is the tendency of individuals to associate with similar others. This property has been very useful across various fields of network studies and provides the basic reasoning behind clustering techniques.

Hierarchical clustering is a type of clustering which seeks to build a hierarchy of clusters. This can be agglomerative or divisive. While agglomerative hierarchical clustering is a bottom-up approach, divisive approach starts with assuming the whole data as one cluster and further dividing it into various smaller clusters iteratively until each data entity has its own unique cluster.

For the purpose of finding clusters of authors that publish together and cite each other’s work more often in a scientific community, Girvan-Newman Clustering [24] has been used. This algorithm detects communities by progressively removing edges of high betweenness centrality from the original network. The steps in this algorithm for community detection are:

1. Calculation of betweenness centrality for all edges existing in the network.
2. Removal of edge(s) with highest betweenness centrality.
3. Recalculation of betweenness centrality of all remaining edges.
4. Repeating Steps 2 and 3 until no edges remain.

The intuition behind using this algorithm on collaboration network and influence network is to find subset of communities who co-author and cite each other more often than with the rest of community.

2.3 Visualisations

Since the early days of bibliometric research, the idea of visualising bibliometric networks has received a lot of attention. This has often been referred to as ‘*science mapping’*.

As a dataset grows in size, it becomes imperative to answer several questions about it. In the field of network science, these questions relate to the visualisation of complex communities, extracting dense representations from a large amount of data, realizing multiscale navigations, and visualing data to detect new patterns. It is important to supplement analysis with visualisations as human mind is better at processing visuals than numbers[5]. Also, there are some patterns and insights in a data that can be observed only with the help of visualisations.

Visualising networks of scientific research has been a topic of interest and there have been several works in the past [6, 12, 39]. Generally three basic types of networks can be created from a set of scientific publications[6], co-authorship network, publication citation networks and semantic network.

Co-authorship networks show the collaboration of authors (and their affiliated institutions) in a scientific research. Generally, the nodes represent the author, the node size conveys the number of publications by the author and the edge label denotes the number of times two authors have co-authored a paper as in the work by [6].

Publication citation networks show the relationship among scientific articles based on their citations. It could be a direct citation network, a bibliographic coupling network where an edge is drawn between two publications if both cite the same previous publication(s) or a co-citation network where an edge between two publications exist if they both cite same subsequent publication(s).

Semantic networks are formed on the basis of occurrence of a keyword in a set of publications. This is the most common form of network, where nodes represent words and edges represent the co-occurence of these words in one article.

In this work, a hybrid visualisation of co-authorship network and publication citation network displays both information. Girvan Newman Clustering, described in Section 2.2.2 is applied on the co-authorship network. The intuition behind this step is to cluster groups of author who collaborate very often. These results are represented as same node colors. Edges of this network represent authors that cite each other quite often in their research work. For example, in Figure 2.1, this cluster represents a group of people who cite each other at least 11 times.

It might also be possible to color nodes representing authors based on their affiliations or locations as done by Belter [6]. It is also important to mention that affiliation/institution resolution generally plagues all of these networks.

This is a novel way of visualing a scientific community which no previous research group in my knowledge has adopted as a solution so far.

[5] http://policyviz.com/wp-content/uploads/2015/10/power-of-visual-communication.pdf

[6] http://www.infotoday.com/online/may12/Belter-Visualizing-Networks-of-Scientific-Research.shtml

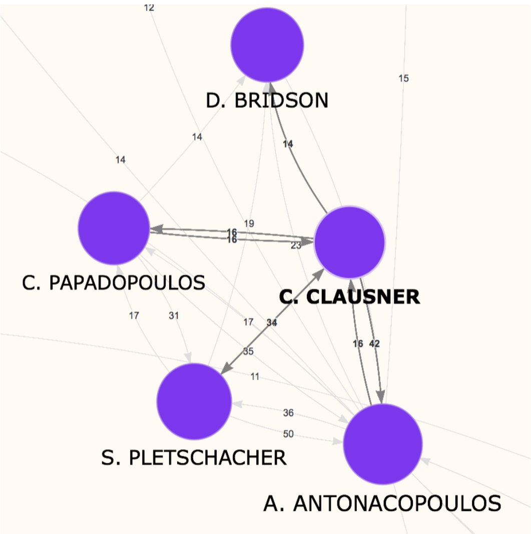


Figure 2.1: This figure shows a cluster from ICDAR community where each edge represents a citation strength of at least 11. The same color of nodes represents a strong collaboration.

"If I have seen further than others, it is by standing upon the shoulders of giants."

Issac Newton

Chapter 3

Related Work

The work done on the study of scientific communities in this thesis is different from the previous research works in a way that this work contributed towards an end-to-end pipeline for creating visualisations straight from conference PDFs rather than externally available dataset sources. This task of developing a complete pipeline was a combination of various sub-tasks which needed to be solved at each stage of this pipeline. These sub-tasks can be briefly summarized as:

1) Information extraction from PDF with meta-data
2) Text Segmentation and Dataset Transformation
3) Analysis and Visualisation

In the following sections, the related work done on these tasks has been further explained.

3.1 Information Extraction from PDF with Metadata

Text extraction from PDF is the first and most crucial step in this work. After the processing of a PDF file and extracting text, this textual data needs to be further processed and labeled with the meta-data information. The idea here is to design a complete module which processes the given PDF and generates a Extensible markup language (XML) document with informative tags. In this work, the information we required is title, header, keyword, abstract, references. This approach of dividing a document on the basis of sections is similar to the one adopted in [7] for *full text analysis* part of a reference linking application.

Though few applications use layout information for information extraction from PDF documents [7, 30, 38], most of the articles are not vey clear on how the references are located and separated from other textual data. While some works implicitly assume that references section is received as text [9, 40], others preferred to use a commercial tool for pre-processing [20]. Some state-of-the-art tools like ParsCit [17] mention the use of regular expressions and a set of heuristics like finding '*References*', '*Bibliography*', '*References and Notes*', or common variations of those strings. Though this technique works for most of the articles, sometimes there are typing mistakes or authors might use a unique name like in [37].

Generally, PDFBox[1] has been used by tools such as SciPlore Xtract [4] or ParsCit [17] for processing PDFs. In this work, Docear's PDF-Inspector [5] has been used as PDFBox sometimes has problems extracting text because of not being fully compliant to the PDF standard. Docear's PDF Inspector [5] is based on jPod[2], which is more tolerant.

A module has been written on top of Docear's PDF-inspector in Java which does not utilise the 'regular expressions' or 'label' information but depends on a unique way to identify reference section on the basis of the way a reference section is structed.This is explained in further detail in Chapter 4.

3.2 Text Segmentation

The next important step is to transform the semi-structured citation data into structured citations. It is difficult for a computer to automatically parse citations because of varying citation formats. Though we know that a citation always includes author, title and publication information, the metadata order may be different as well as their attributes. Several digital library institutes rely on various automatic parsing techniques.

In this section, a review of all the prominent works in the field of reference segmentation is presented. This study broadly resulted in four ways of accomplishing text segmentation and each of these methods is being described in the following four sub-sections:

3.2.1 Template Matching Methods

A template matching approach takes an input citation and matches its syntactic pattern against known templates. The template with the best fit to the input is then used to extract metadata about the particular citation and further label the citations tokens as fields. ParaTools(short for Paracite toolkit) [7, 8, 26, 29] and Basic Local Alignment Search Tool (BLAST) [27] used a template based reference parser.

[1] https://pdfbox.apache.org/

[2] http://java-source.net/open-source/pdf-libraries/jpod

Earlier, BLAST was being used to compare sequences of a human genome [1]. Later, authors used this well-developed protein sequence matching program to identify citations. The pre-processing included using a form translation program to translate citations into a form that is easy to process.

BLAST achieved a better alignment as compared to ParaTools and had a unique approach to template parsing. A further weakness of ParaTools was that it tagged ambiguous fields as '*Any*', equivalent to not tagging the token at all. Huang et al., 2004 [27] reported ParaTool's precision as approximately 30 percent. This level of performance and lack of portability make the approach unsuitable for high volume data processing.

3.2.2 Supervised Machine Learning Models

Supervised classification approaches involve training a network on an annotated data set. This is currently a dominant approach to solve the problem of segmenting text to extract data values in it. The complete dataset is divided into '*training*' and '*test*' dataset in an approximate 7 : 3 ratio, respectively. After the training phase is complete using the '*training*' set, the trained network is then tested on the '*test'* data to evaluate architecture performance.

For citation parsing tasks, all the supervised classification approaches have been broadly classified under two categories:
1) Hidden Markov Model (HMM) and Support Vector Machines (SVM) Based Approaches
2) Conditional Random Fields(CRF) Based Approaches

3.2.2.1 HMM and SVM Classifiers Based Approaches

Seymore et al., 1999 did some initial work on reference parsing using HMM based approaches to build a reference string sequence labeler [42]. This work led to the creation of the core dataset. They used a manually constructed model that contains multiple states per extraction field. This work primarily focused on learning the model structure from data and making the best use of labeled and unlabeled data.

The second technique involved using trained SVM-classifiers [25]. These SVM classifiers could handle many non-independent features. It is important to mention here that the work on this task was done in two stages: *first*, by classifying each line independently to assign it label; *later*, adjusting these labels based on an additional classifier that examines larger windows of labels.

3.2.2.2 CRF Based Approaches

CRFs are undirected graphical models trained to maximize a conditional probability. They were introduced in 2001 [31] but the initial work on exploring CRF techniques for information extraction started in 2003 with tasks such as NER [32], table extraction [36] and shallow parsing [43]. Another major step in the use of CRF based approaches to extract information from research papers was done in 2004 [35] where they described a large collection of experimental results on two traditional benchmark data sets and obtained dramatic improvements over SVM and HMM based results.

This was followed by ParsCit [17] in 2008 which was delivered as an open-source CRF reference string parsing package. This was also compared to three distinct reference string datasets and it compared well to other previous works. ParsCit is currently the state-of-the-art in the field of reference parsing. The package comes with utilities to run it as a web service or as standalone utility.

3.2.3 Unsupervised Classification

An unsupervised classification approach relies on a knowledge base to label citation strings. It is better than machine learning approaches in the sense that it can save resources spent on the training phase.

Flux-CIM [16] implements a knowledge-base approach. It is unsupervised and does not rely on a learning method that sometimes requires a very expensive training phase. This approach depends on a knowledge base and has shown better results as compared to Parscit [17] in 2008 and uses four stages namely, *blocking, matching, binding*, and *joining*. This approach was different from other approaches as here knowledge base is constructed automatically from an existing set of sample metadata records for a given field.

Another way to approach the problem of bibliographic reference structure is based on part-of-speech tagging [9]. RefParse[3] is another generic approach to bibliographic reference parsing and is independent of any specific reference style. Its core feature is an inference mechanism that exploits the regularities inherent in any list of references to deduce its format [40].

[3] http://plazi.org/developers/source-code/refparse/

3.2.4 Web-based look up

Web-based look up techniques are also a good option as they can save the time to construct a manually labeled training data to learn an extraction model. Experiments have shown that a combination of knowledge base, heuristics and statistical methods can automate the extraction process and achieve good performance [28].

3.3 Analysis and Visualisation

Several approaches have been taken in the past for the purpose of studying a scientific community and the impact of the authors who are a part of it. While some research groups have focused on developing innovative measures based on Pagerank and its variants [34, 47], others have focused on improving traditional indices [19, 21].

There has also been a study on the superstar phenomena/Mathew effect in scientific communities [41]. Scientific communities have also been studied as a multilayer network [18] and citation networks have been transitively reduced [13] to obtain the most relevant citations from a particular paper or to observe the general trend in which a community cites.

Visualisation of a community has mainly focused on the networks of co-authoring researchers or the networks of keywords co-occuring in publications. Several techniques have been developed for analyzing and visualising co-citation and bibliographic coupling networks. Several tools are available to help with these tasks as well [14, 44–46]. Table 3.1 shows all the available tools for the analysis of various forms of bibliometric networks.

The visualisations from the work in this thesis were developed in javascript with the help of vis.js[4] library and are interactive. Also, they show a multi-data network where same color of the node represented strong collaboration among authors. Edges with their edge labels represented the *citation strength* and relations of authors.

[4] http://visjs.org/

Table 3.1: Table of the various visualisation tools with their URL sources

Tools	**URLs**
CitNetExplorer	http://www.citnetexplorer.nl
Gephi	https://gephi.org
HistCite	http://www.histcite.com
Pajek	http://pajek.imfm.si
Sci	https://sci2.cns.iu.edu
VOSviewer	http://www.vosviewer.com

"When everything is connected to everything else, for better or for worse: everything matters."

Bruce Mau

Chapter 4

SNA in Scientific Communities: Proposed Approach

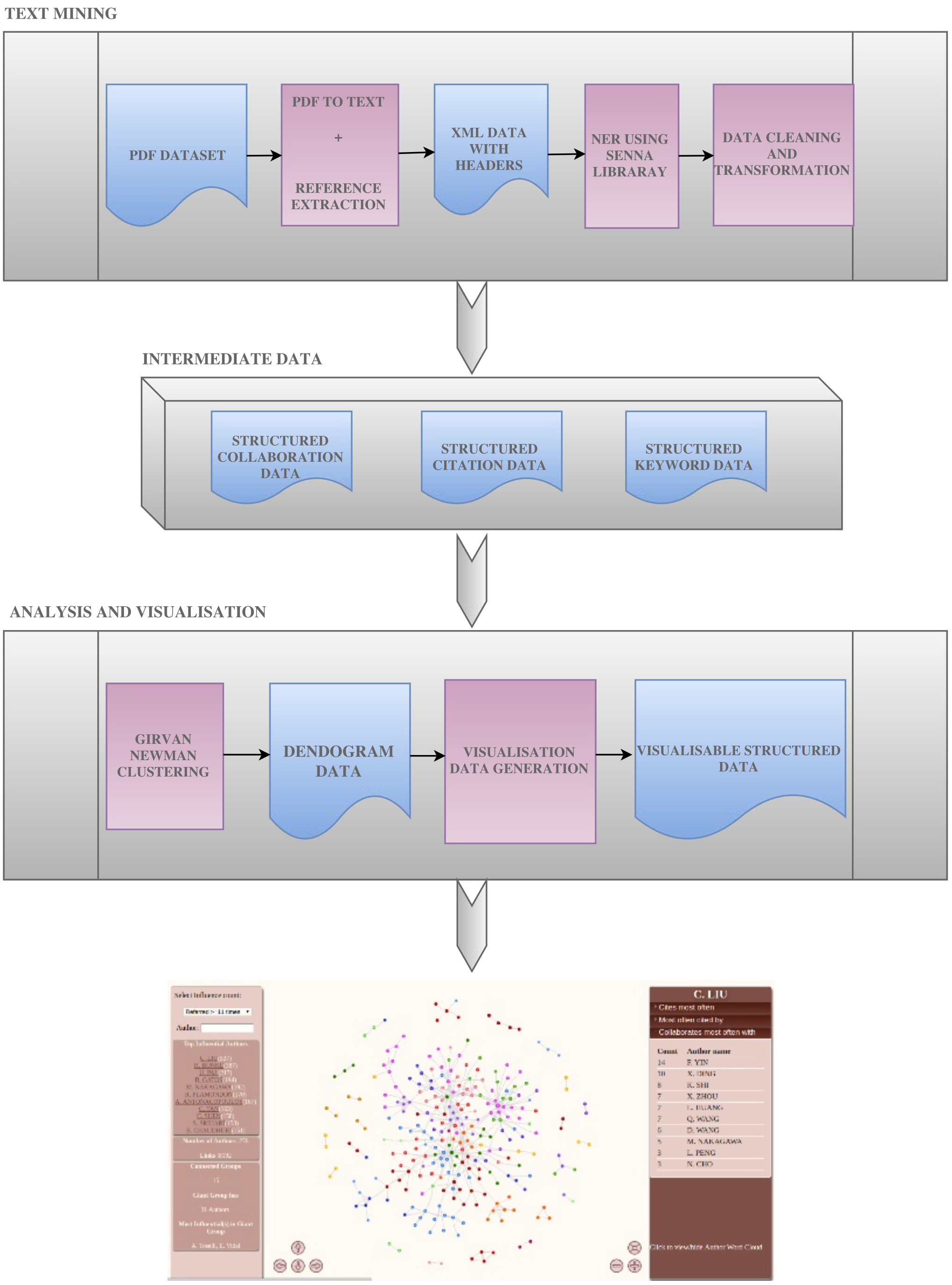
This chapter provides a detailed description of the approach adopted. This work comprises of various sub-tasks. Section 4.1 describes the *Text-Mining* sub-task that includes processing PDF documents and text segmentation. Section 4.2 describes the *Network Analysis* approach followed by development of the visualisation tool described in Section 4.3.

Please refer to Figure 4.1 for a complete workflow of the presented approach.

4.1 Text Mining

The most important step for assembling a social network is data collection. In this research work, the initial dataset was in a completely unstructured format in the form of a collection of PDF documents from ICDAR conferences. The foremost step for obtaining this data in a structured format was to process the relevant information present as text in a PDF file. This task was further divided into four steps:

1. PDF to text
2. Metadata and reference extraction
3. Extracting author names using NER
4. Data pre-processing



Count	Author name
14	F. YIN
10	X. DING
8	K. SHI
7	X. ZHOU
7	L. HUANG
7	Q. WANG
6	D. WANG
5	M. NAKAGAWA
3	L. PENG
3	N. CHO

FIGURE 4.1: This figure gives an overview of the complete approach adopted for analysis and visualisation.

4.1.1 PDF to text

For the processing of the PDF in order to generate a dataset, it is of prime importance to extract textual information from PDF before it is further processed. Several tools like pdftotext[1], pdfBox[2], Docear's PDF inspector[3] et cetera, are available for this purpose but there are few things to take care of while processing a scholarly PDF article:

1. The information might be present in single or multiple column format.
2. There might be tables or images which the tool should be able to identify and convert to text/omit respectively.
3. The OCR should be unicode character compatible.
4. Speed is also an important factor.

Considering the above requirements, Docear's PDF Inspector has been used. It is a Java library that extracts title and the full-text from a PDF file. Though other tools such as SciPlore Xtract [4] or ParsCit [17] use PDFBox for processing the PDF documents, Docear's PDF Inspector claims that being based on jPod it is better compliant to the PDF-standard, more tolerant, and thus more useful.

4.1.2 Metadata and reference extraction

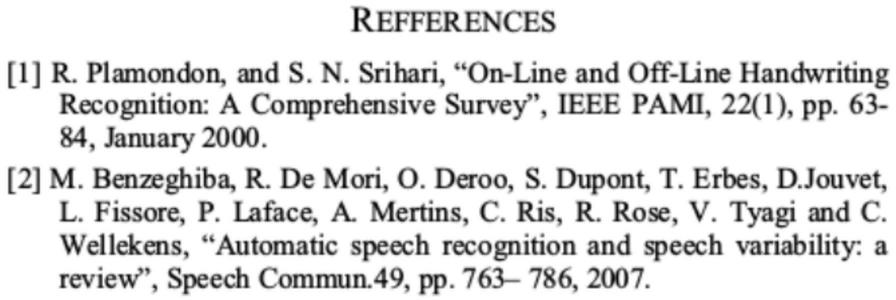
A module has been written on top of Docear's PDF-Inspector. This module segments the textual data into a broad category of title, header, keywords, abstract and references.

Though other XML tags are self-explanatory, header tag contains all the information present after title of the article and before the abstract in a scholarly article. This section contains the information about the authors of the article, most often associated with their affiliation and their contribution details. For sake of ease in this initial work, all authors under the header tag are given equal weightage.

While the title has been extracted from the library, keywords were found under the label '*Keywords*' or '*Index-terms*', based on the format of ICDAR conference papers. Abstract was also similarly extracted based on '*Abstract*' label. For extracting the references section, the code was more robust which relied upon the citations being together and a frequency of occurrence of new line in a text file. It was important to rely on something

[1] http://linux.die.net/man/1/pdftotext
[2] https://pdfbox.apache.org/
[3] https://www.docear.org/software/add-ons/docears-pdf-inspector/

more than label information here as sometimes authors can use a unique name or there might be spelling mistakes.



REFFERENCES

[1] R. Plamondon, and S. N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey", IEEE PAMI, 22(1), pp. 63-84, January 2000.

[2] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D.Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi and C. Wellekens, "Automatic speech recognition and speech variability: a review", Speech Commun.49, pp. 763– 786, 2007.

Figure 4.2: This figure shows a sample references section explaining why relying on label might result in inconsistent data. Please note the spelling mistake in the label of references section.

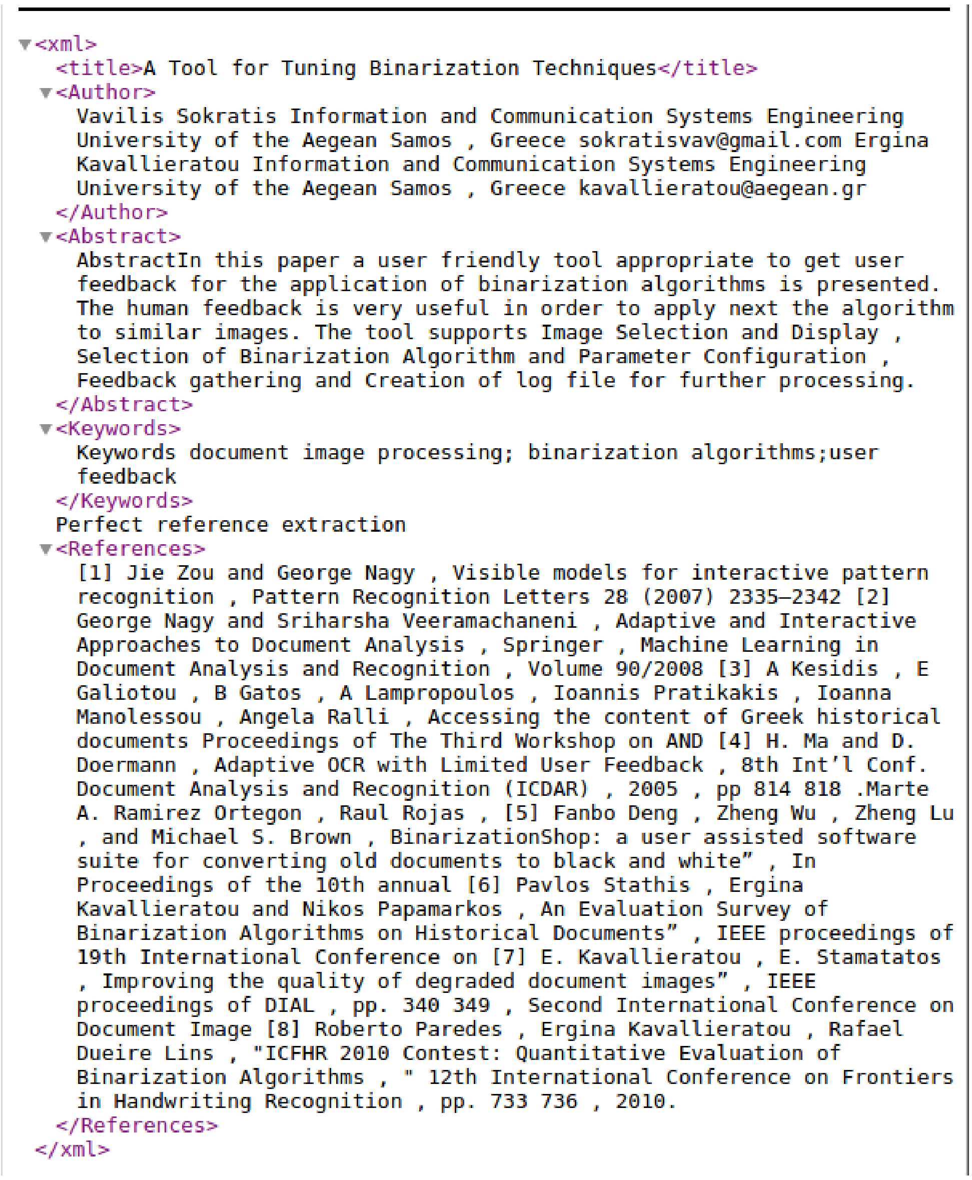
This step reduces the size of the text file considerably and makes the text more informative categorised under various XML tags using the above segmented information. It also reduces the effort for NER-extractor described in the next section Section 4.1.3, thus making the pipeline more fast.

To sum it up, this module combined with Docear's PDF-Inspector resulted in an XML file with the format as shown in a sample XML file in Figure 4.3

4.1.3 Extracting Author Names using NER

After the generation of XML files from PDF dataset as described in Section 4.1.2, the next step was to prepare structured data. This step required extraction of names of author who write the article, present in 'header' and authors who are cited, from each *citation string* present in 'references' xml tag. Keywords in the article were extracted from 'keywords' XML tag. As this work primarily focused on author relationships, extracting author names with high accuracy was of prime importance.

Though CRF based models like Parscit [17] have proved to be very good at parsing citations from a scholarly article, for this research, extraction of only author names from the 'header' and the 'references' section was required. For this reason, Senna NLP library has been used for NER task. Senna has given state-of-the-art results, it is faster and has a high precision rate[15]. It outputs a host of NLP predictions: Part of speech (POS), chunking (CHK), Named Entity Recognition (NER), Semantic Role Labelling (SRL) and Phrase Structure Grammar (PSG). In this work, only NER was used. Also, Senna is favourable as compared to other approaches because Senna identifies names as Begin-Person (B-PER), Intermediate-Person (I-PER), End-Person (E-PER) for first, middle



```
▼<xml>
  <title>A Tool for Tuning Binarization Techniques</title>
 ▼<Author>
    Vavilis Sokratis Information and Communication Systems Engineering
    University of the Aegean Samos , Greece sokratisvav@gmail.com Ergina
    Kavallieratou Information and Communication Systems Engineering
    University of the Aegean Samos , Greece kavallieratou@aegean.gr
  </Author>
 ▼<Abstract>
    AbstractIn this paper a user friendly tool appropriate to get user
    feedback for the application of binarization algorithms is presented.
    The human feedback is very useful in order to apply next the algorithm
    to similar images. The tool supports Image Selection and Display ,
    Selection of Binarization Algorithm and Parameter Configuration ,
    Feedback gathering and Creation of log file for further processing.
  </Abstract>
 ▼<Keywords>
    Keywords document image processing; binarization algorithms;user
    feedback
  </Keywords>
  Perfect reference extraction
 ▼<References>
    [1] Jie Zou and George Nagy , Visible models for interactive pattern
    recognition , Pattern Recognition Letters 28 (2007) 2335–2342 [2]
    George Nagy and Sriharsha Veeramachaneni , Adaptive and Interactive
    Approaches to Document Analysis , Springer , Machine Learning in
    Document Analysis and Recognition , Volume 90/2008 [3] A Kesidis , E
    Galiotou , B Gatos , A Lampropoulos , Ioannis Pratikakis , Ioanna
    Manolessou , Angela Ralli , Accessing the content of Greek historical
    documents Proceedings of The Third Workshop on AND [4] H. Ma and D.
    Doermann , Adaptive OCR with Limited User Feedback , 8th Int’l Conf.
    Document Analysis and Recognition (ICDAR) , 2005 , pp 814 818 .Marte
    A. Ramirez Ortegon , Raul Rojas , [5] Fanbo Deng , Zheng Wu , Zheng Lu
    , and Michael S. Brown , BinarizationShop: a user assisted software
    suite for converting old documents to black and white” , In
    Proceedings of the 10th annual [6] Pavlos Stathis , Ergina
    Kavallieratou and Nikos Papamarkos , An Evaluation Survey of
    Binarization Algorithms on Historical Documents” , IEEE proceedings of
    19th International Conference on [7] E. Kavallieratou , E. Stamatatos
    , Improving the quality of degraded document images” , IEEE
    proceedings of DIAL , pp. 340 349 , Second International Conference on
    Document Image [8] Roberto Paredes , Ergina Kavallieratou , Rafael
    Dueire Lins , "ICFHR 2010 Contest: Quantitative Evaluation of
    Binarization Algorithms , " 12th International Conference on Frontiers
    in Handwriting Recognition , pp. 733 736 , 2010.
  </References>
</xml>
```

Figure 4.3: This figure shows a sample xml file generated for a conference article.

and last part of a name respectively. This becomes important later for the purpose of name referencing which is explained in detail in Section 4.1.4.2.

4.1.4 Data Pre-processing

Preprocessing of this intermediate data, like any other, is the most important step. Real-world data is generally incomplete, noisy and inconsistent. The two major parts of this step are detailed in the following sub-sections.

4.1.4.1 Cleaning

Conventionally, data cleaning step typically includes filling in missing values, and resolving inconsistencies. In this case, NER extracted university affiliations and put them in the category of author names. These university names to which the authors are affiliated were removed from the dataset using regular expressions. If a single author name occurred twice for an article, it was removed.

4.1.4.2 Normalization

As this work focused on ICDAR dataset, it was important to transform the data and reduce its size by removing erroneous information. In this case, it was the presence of those authors in the dataset who never published a paper in ICDAR conference. This was resolved by removing each link where an author was citing an author who never published in ICDAR conference. This reduces the size of dataset considerably.

The other issue that plagued this data was of name referencing. Because of the absence of a proper format for citing an author, articles cite authors in varied formats. It is quite important to accurately reference all these formats to a common entity. In this case, a solution was adopted to reduce all names to a specific format, where the first alphabet of first name was followed by a full stop and further followed by last name of the author. If author name consisted of a middle name, it was omitted.

For instance, if an author's name existed as Breuel,T. or Breuel, T. M. or T. M. Breuel or T. Breuel or Thomas Breuel;all such appearances were reduced to T. Breuel.

4.1.5 Structured data after text mining

After all the above steps, three types of intermediate data was obtained:

1. Citation Data information: This included all the information about authors citing another author over all the years in ICDAR conference from year 1993 to 2015. It is a simple Comma Separated Value (CSV) file with two columns, the first one containing the authors who publish an article and the second column containing the authors being 'cited' or 'referred'.

2. Collaboration Data information: This included all the information about authors publishing an article together over all the years in ICDAR conference from year 1993 to 2015. It is a collection of arrays of names, names which are together in one array represent co-authorship.

3. Author, Keyword information: This contains information about all authors associated with the keywords from the fields in which they publish articles in ICDAR conference from year 1993 to 2015.

   Keyword data was also extracted from the IEEE Bibtex which has a Keyword field. This was the only information extracted using crawlers. It was important to do so in the absence of a proper clustering algorithm for keywords. IEEE keywords are often referenced and well-clustered.

4.2 Social network analysis (SNA)

SNA is a young interdisciplinary field that deals with the analysis of networks among specific groups of people and has its roots in the field of sociology. The methods for this analysis are based on the mathematical field of graph theory, where the individuals are represented by nodes and their relations by edges. Metrics exist to measure an individual node or the complete network, for example, the centrality of a node, density of the network *et cetera*.

Community detection in graphs is important as it helps to identify modules and their boundaries and allows for a classification of vertices, according to their structural position in the modules.

After the conversion of unstructured data into a structured format, the next important step was to analyse the data. In this section, all the important measures taken to make the data more informative have been described.

These measures include first clustering the data, and later calculation of few graph statistics and various centralities of the nodes in a network.

4.2.1 Clustering

Clustering works on the basic principle of 'homophily', *love of the same.* This basic property of individuals to group together on the basis of the interests they share has helped scientists detect functions of some unknown proteins, classify documents and even solving pattern recognition applications.

Keeping the goal in mind, network visualisations needed to represent two kinds of information, co-authorship and co-citation. While edges are most suited to depict citation pattern because they can represent direction with edge label, collaboration behavior is best depicted by node-color.

Generally, in similar visualisations, node color has been allocated on the basis of groups and affiliations [6]. Though intuitively it seems that finding max-clique in a collaboration graph is the appropriate approach to assign groups but it results in several 1-cliques which are of no use.

Running a Bron-Kerbosch algorithm[11] to find max-cliques resulted in approximately 600 cliques for a total of 1200 authors on ICDAR 2011-2015 data. So, to solve this problem, a novel approach was adopted to assign node colors to authors by applying the Girvan-Newman clustering algorithm [24]. The data for this was a co-authorship network where nodes represented authors and an edge existed among authors if they published a paper together. There could be multiple edges among authors if they are co-authors in multiple publications. 7 Girvan-Newman clustering is a hierarchical clustering approach which removes edges on the basis of their high betweenness centrality. The result is a dendrogram. A suitable iteration was used to cluster the groups of authors based on the results obtained from experimental verifications.

The results were tested and some of the nodes which share the same color have been shown in Figure 4.4. They were quite accurate when compared to real-world affiliations.

4.2.2 Network Analysis Metrices

In this subsection, a detailed description of all the metrices calculated for the network have been shown. These included various centrality measures for the node and information on various other graph statistics.

4.2.2.1 Centrality measures

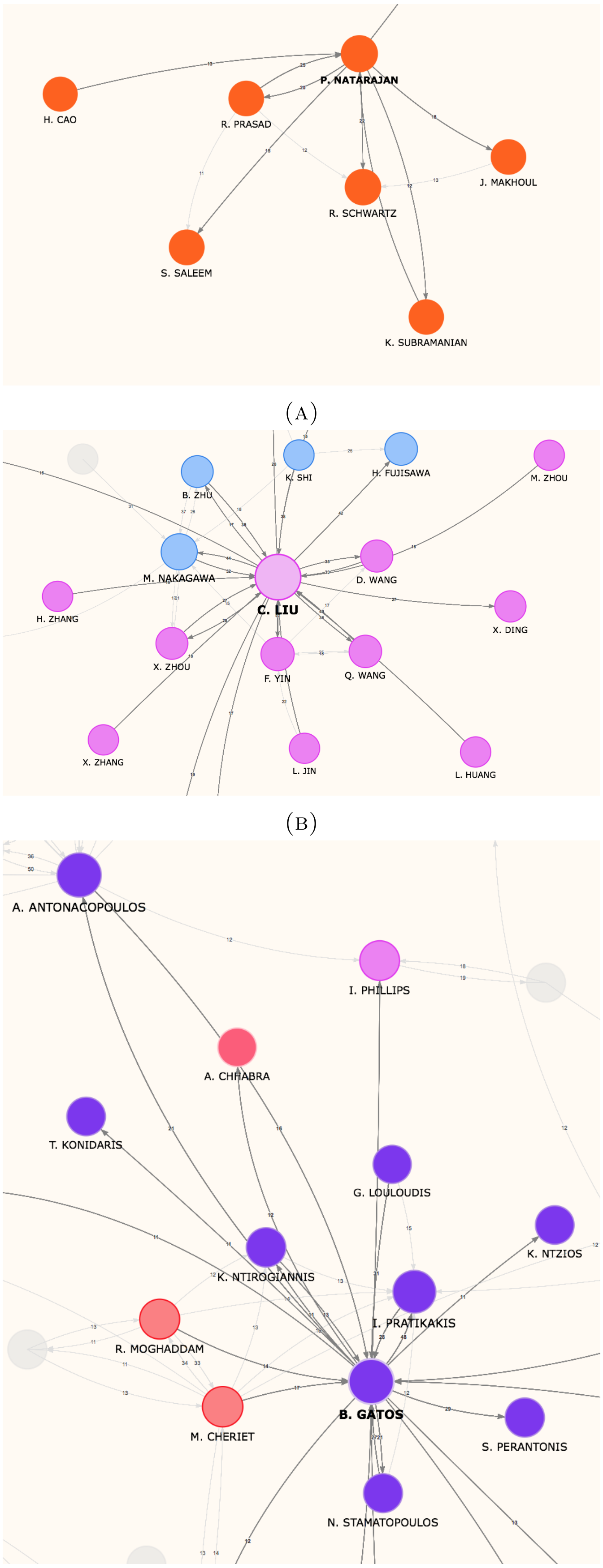
Some of the important centrality measures which have been calculated are:

1. Betweenness centrality: This centrality measure is an indicator of a node's centrality in a network. It is equal to the number of shortest paths from all vertices to all others that pass through a node [22].

   The betweenness centrality of a node v is given by the expression:

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

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



(A)

(B)

(C)

Figure 4.4: Same color nodes represent a strong collaboration among authors in a community as shown in (A), (B) and (C).

In this network, a node with high betweenness centrality represents the degree to which an author plays a role in the cohesiveness of the community. Figure 4.7 shows the centrality measures of the author with highest betweenness centrality in this network.

2. Degree centrality: This centrality measure is an indicator of number of links which are connected to a node. It is a measure of local influence in a network [23].

   The degree centrality of a vertex v, for a given graph G:=(V,E) with $|V|$ vertices and $|E|$ edges, is defined as:

$$C_D(v) = deg(v)$$

3. Eigenvector centrality: This centrality is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes [33].

   The eigenvector centrality for node i is,

$$Ax = \lambda x$$

   where A is the adjacency matrix of the graph G with eigenvalue λ .

4. Indegree centrality: This measure is a variant of *'degree centrality'* for directed networks where only the incident edges are taken into consideration.

5. Outdegree centrality: This measure is a variant of *'degree centrality'* for directed networks where only the outgoing edges are taken into consideration.

6. Closeness centrality: Closeness centraliy of a node u is the reciprocal of the sum of the shortest path distances from u to all $n-1$ other nodes. This value is normalized by the sum of minimum possible distances $n-1$.

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}$$

   Classically, the spread of information is modeled by the use of shortest paths. In this network, closeness centrality represents how well connected an author is with the rest of community.

These measures have been further normalized in the range of 0 to 1 for all the authors. They provide a good overview to assess an author's influence in a community.

4.2.2.2 Overlap graph

An overlap graph of an author is a bar-graph that denotes the quantitative overlap of an author with another author based on three categories, ‘cites’, ‘is cited by’ and ‘collaboration’ in a respectively decreasing fashion.

Figure 4.10 shows the overlap graph for author C. Liu for the top-10 authors in ICDAR community who cite him, whom he cites and with whom he collaborates. A clear overlap of 5 authors is visible for all three categories and an overlap of 11 is visible for at least two categories out of a total of 13 authors (excluding the author himself) present on y-axis in freffig:overlapgraph.

4.2.2.3 Keyword cloud

An information from the author’s domain keywords is collected and sorted according to the frequency count of keyword. This information is then later visualised in the form of a keyword cloud.

Figure 4.11 is an example and shows word-cloud for A. Dengel. This keyword cloud is based on ICDAR 2011-2015 data.

4.3 Visualisations

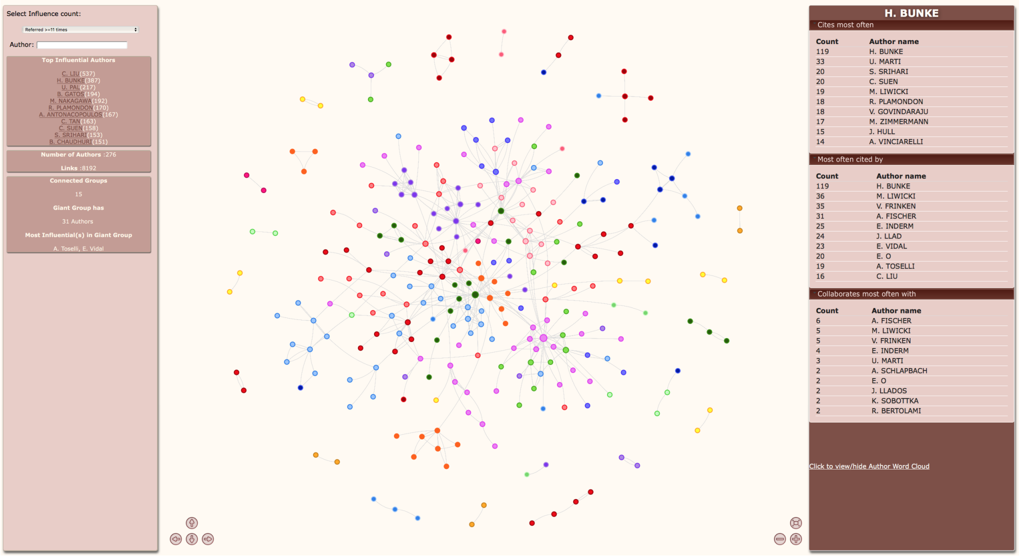
The role of visualisations today in understanding data and in knowledge advancement is quite relevant. They help to steer the complete process of a scientific work from initial to final stages. In the initial stages of research, good visualisations help create meaningful hypotheses and also to interactively explore the data. In intermediary stages of analysis, visualisations can control and steer partially automated analysis; and in final stages, visualisations provide summaries of the results that improve one’s understanding of a domain.

From field of pure sciences, like computational fluid dynamics, computational physics, computational chemistry, or molecular biology to application based scientific work, like high-dimensional data, spatio-temporal data, or time-series data; visualisations are very relevant.

In this section, the visualisation tool developed for the purpose of visualising author network in ICDAR community has been described.

4.3.1 Features of tool

The visualisation tool developed for this data offers features to visualise ICDAR author community at present. Figure 4.5 provides an overview of the complete interactive interface.



Select Influence count:

Referred >=11 times

Author:

Top Influential Authors

C. LIU(537)
H. BUNKE(387)
U. PAL(217)
B. GATOS(194)
M. NAKAGAWA(192)
R. PLAMONDON(170)
A. ANTONACOPOULOS(167)
C. TAN(163)
C. SUEN(158)
S. SRIHARI(153)
B. CHAUDHURI(151)

Number of Authors :276

Links :8192

Connected Groups

15

Giant Group has

31 Authors

Most Influential(s) in Giant Group

A. Toselli, E. Vidal

H. BUNKE

Cites most often

Count	Author name
119	H. BUNKE
33	U. MARTI
20	S. SRIHARI
20	C. SUEN
19	M. LIWICKI
18	R. PLAMONDON
18	V. GOVINDARAJU
17	M. ZIMMERMANN
15	J. HULL
14	A. VINCIARELLI

Most often cited by

Count	Author name
119	H. BUNKE
36	M. LIWICKI
35	V. FRINKEN
31	A. FISCHER
25	E. INDERM
24	J. LLAD
23	E. VIDAL
20	E. O
19	A. TOSELLI
16	C. LIU

Collaborates most often with

Count	Author name
6	A. FISCHER
5	M. LIWICKI
5	V. FRINKEN
4	E. INDERM
3	U. MARTI
2	A. SCHLAPBACH
2	E. O
2	J. LLADOS
2	K. SOBOTTKA
2	R. BERTOLAMI

Click to view/hide Author Word Cloud

FIGURE 4.5: Visualisation interface

The complete interface can be divided into three panels:

1) Network visualisation panel: This interface helps to view visualisations of ICDAR community structure. The authors have been represented as nodes and '*A cites B'* is represented as an edge. Edge label denotes the number of times an author cites another author.

Figure 4.6 represents the network visualisation. By default, this represents a graph denoting edges with citation strength of 10 or more.

Navigation buttons provide a zoom in, zoom out, fit to screen, up, down, left, right functionality to view the network of community. Hovering mouse over a node displays important centrality measures for the node and highlights connected links [refer Figure 4.7].

2) Network Information panel: Figure 4.8 represents the graph statistics in the network information panel. This panel can be used to choose a graph based on citation strength. This citation strength can be chosen and options vary from 6 to 29. A search input box helps the user to search for an author.

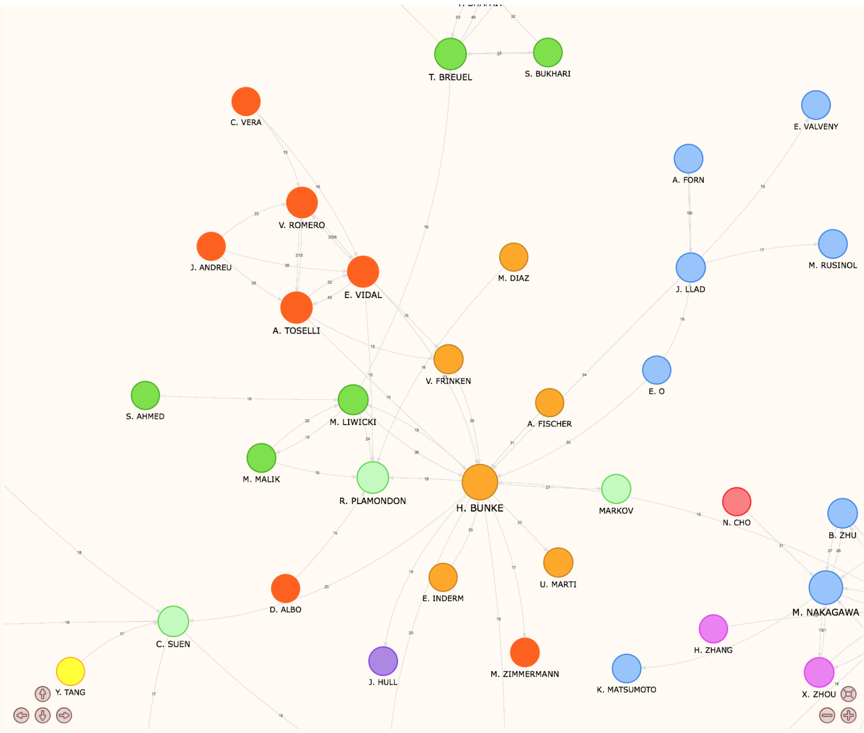
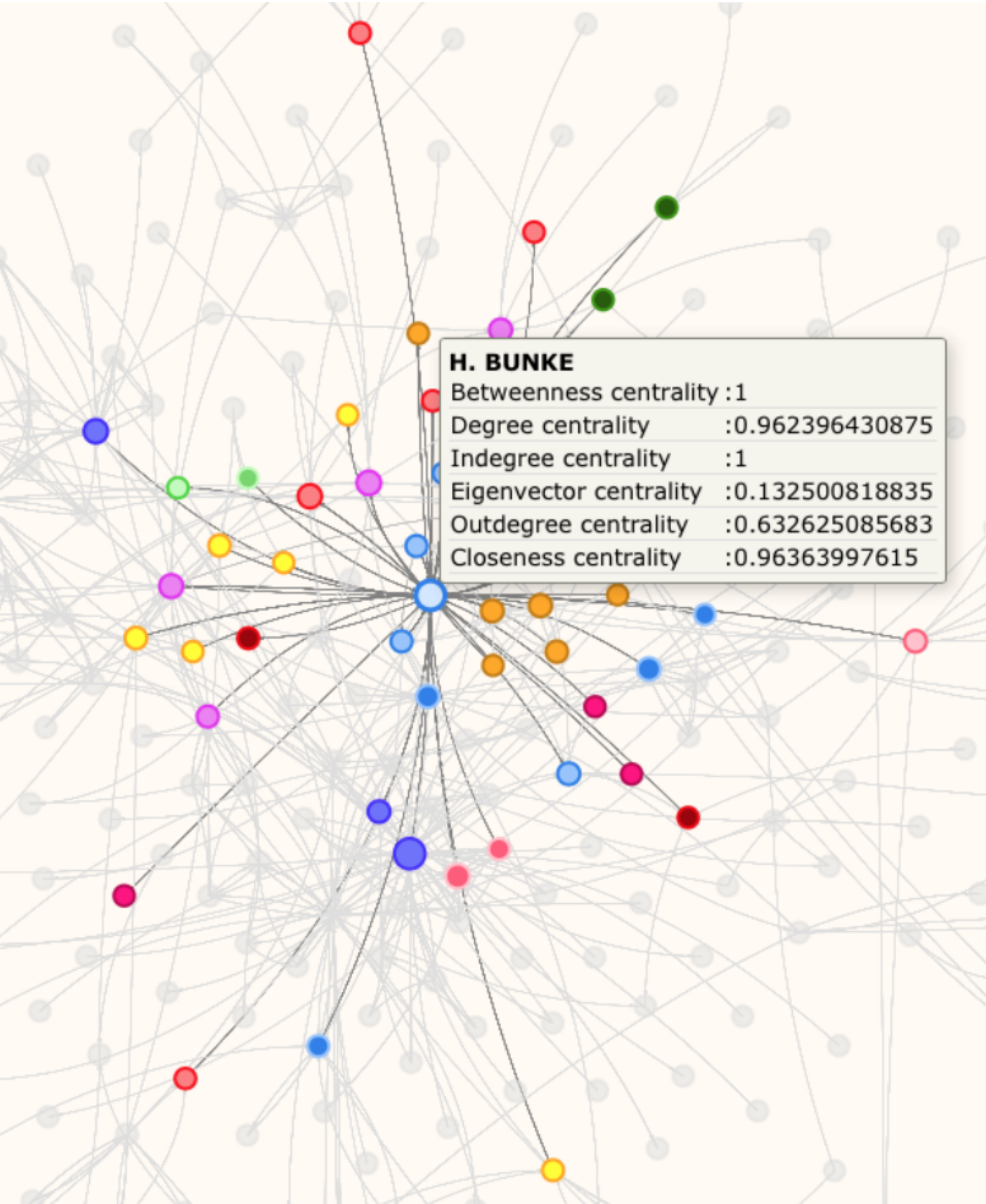


FIGURE 4.6: Network Visualisation panel

Graph statistics include information on top 10 most influential people, number of authors and links in a graph, number of connected groups, number of authors in the largest connected group and the most influential author(s) in the largest connected group. These graph statistics are displayed in this panel.

3) Author information panel: Figure 4.9 represents the author information panel. This provides information on the citation and the collaboration behavior supplemented with an overlap graph for the selected author. An overlap graph for an author *A'* is a bar chart representing the top-10 authors whom the author *a* cites, is cited by and collaborates with. These authors are present on the y-axis and can range from 10 − 30 depending upon the overlap of authors. Figure 4.10 shows an overlap graph for the author, C. Liu. This graph shows that the top-10 people whom the author cites are the ones with whom the author collaborates. Such findings can be very informative.

There is a also a possibility to view an author's word cloud of his research domain, Figure 4.11 is an example.

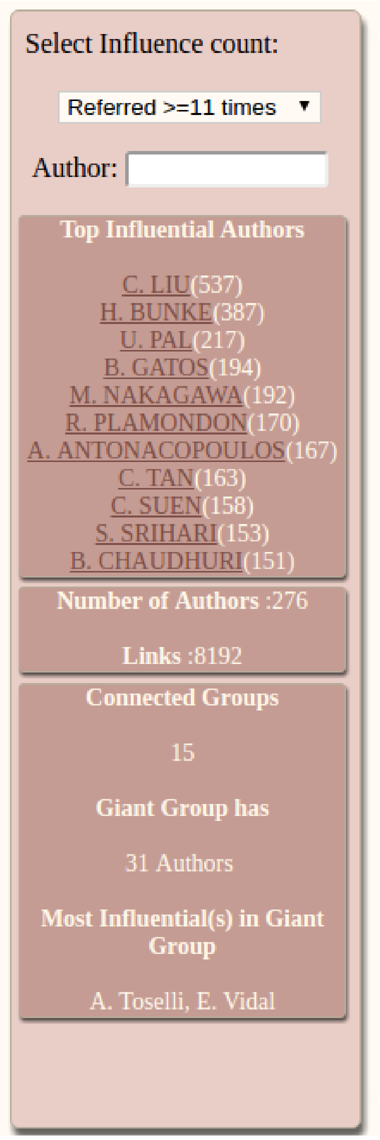


H. BUNKE
Betweenness centrality :1
Degree centrality :0.962396430875
Indegree centrality :1
Eigenvector centrality :0.132500818835
Outdegree centrality :0.632625085683
Closeness centrality :0.96363997615

Figure 4.7: Centrality values display

4.4 Summary

In this Chapter, a complete methodology of approaching the research problem has been presented. Section 4.1 describes the text mining techniques adopted to extract textual data from PDF documents and Section 4.2 explains how the data was clustered into groups using Girvan-Newman clustering. Further, this chapter concludes with Section 4.3 describing the visualisation tool and its features.



Select Influence count:

Referred >=11 times

Author:

Top Influential Authors

C. LIU(537)
H. BUNKE(387)
U. PAL(217)
B. GATOS(194)
M. NAKAGAWA(192)
R. PLAMONDON(170)
A. ANTONACOPOULOS(167)
C. TAN(163)
C. SUEN(158)
S. SRIHARI(153)
B. CHAUDHURI(151)

Number of Authors :276

Links :8192

Connected Groups

15

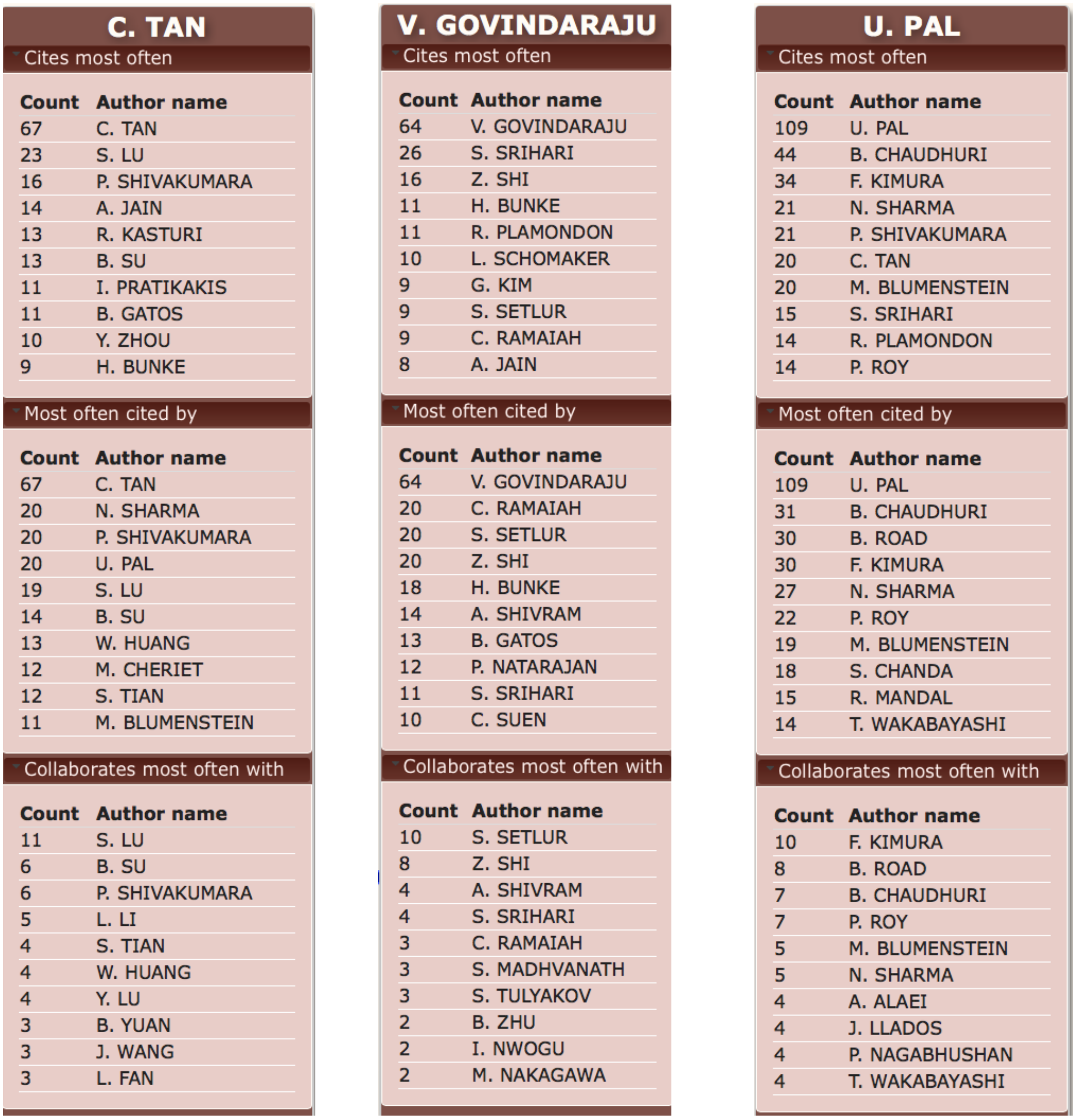
Giant Group has

31 Authors

Most Influential(s) in Giant Group

A. Toselli, E. Vidal

Figure 4.8: Network information panel



(A)

C. TAN

Cites most often

Count	Author name
67	C. TAN
23	S. LU
16	P. SHIVAKUMARA
14	A. JAIN
13	R. KASTURI
13	B. SU
11	I. PRATIKAKIS
11	B. GATOS
10	Y. ZHOU
9	H. BUNKE

Most often cited by

Count	Author name
67	C. TAN
20	N. SHARMA
20	P. SHIVAKUMARA
20	U. PAL
19	S. LU
14	B. SU
13	W. HUANG
12	M. CHERIET
12	S. TIAN
11	M. BLUMENSTEIN

Collaborates most often with

Count	Author name
11	S. LU
6	B. SU
6	P. SHIVAKUMARA
5	L. LI
4	S. TIAN
4	W. HUANG
4	Y. LU
3	B. YUAN
3	J. WANG
3	L. FAN

(B)

V. GOVINDARAJU

Cites most often

Count	Author name
64	V. GOVINDARAJU
26	S. SRIHARI
16	Z. SHI
11	H. BUNKE
11	R. PLAMONDON
10	L. SCHOMAKER
9	G. KIM
9	S. SETLUR
9	C. RAMAIAH
8	A. JAIN

Most often cited by

Count	Author name
64	V. GOVINDARAJU
20	C. RAMAIAH
20	S. SETLUR
20	Z. SHI
18	H. BUNKE
14	A. SHIVRAM
13	B. GATOS
12	P. NATARAJAN
11	S. SRIHARI
10	C. SUEN

Collaborates most often with

Count	Author name
10	S. SETLUR
8	Z. SHI
4	A. SHIVRAM
4	S. SRIHARI
3	C. RAMAIAH
3	S. MADHVANATH
3	S. TULYAKOV
2	B. ZHU
2	I. NWOGU
2	M. NAKAGAWA

(C)

U. PAL

Cites most often

Count	Author name
109	U. PAL
44	B. CHAUDHURI
34	F. KIMURA
21	N. SHARMA
21	P. SHIVAKUMARA
20	C. TAN
20	M. BLUMENSTEIN
15	S. SRIHARI
14	R. PLAMONDON
14	P. ROY

Most often cited by

Count	Author name
109	U. PAL
31	B. CHAUDHURI
30	B. ROAD
30	F. KIMURA
27	N. SHARMA
22	P. ROY
19	M. BLUMENSTEIN
18	S. CHANDA
15	R. MANDAL
14	T. WAKABAYASHI

Collaborates most often with

Count	Author name
10	F. KIMURA
8	B. ROAD
7	B. CHAUDHURI
7	P. ROY
5	M. BLUMENSTEIN
5	N. SHARMA
4	A. ALAEI
4	J. LLADOS
4	P. NAGABHUSHAN
4	T. WAKABAYASHI

FIGURE 4.9: Author information panel for author C. Tan, V. Govindaraju and U. Pal in (A), (B) and (C) respectively.

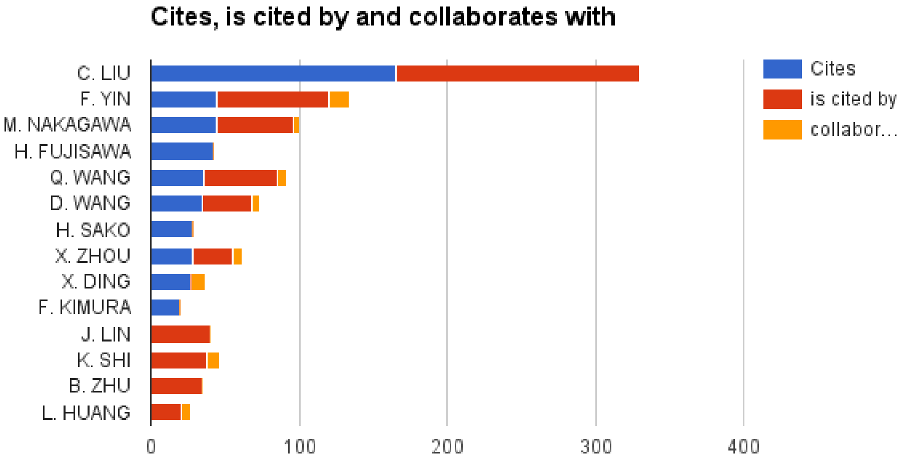
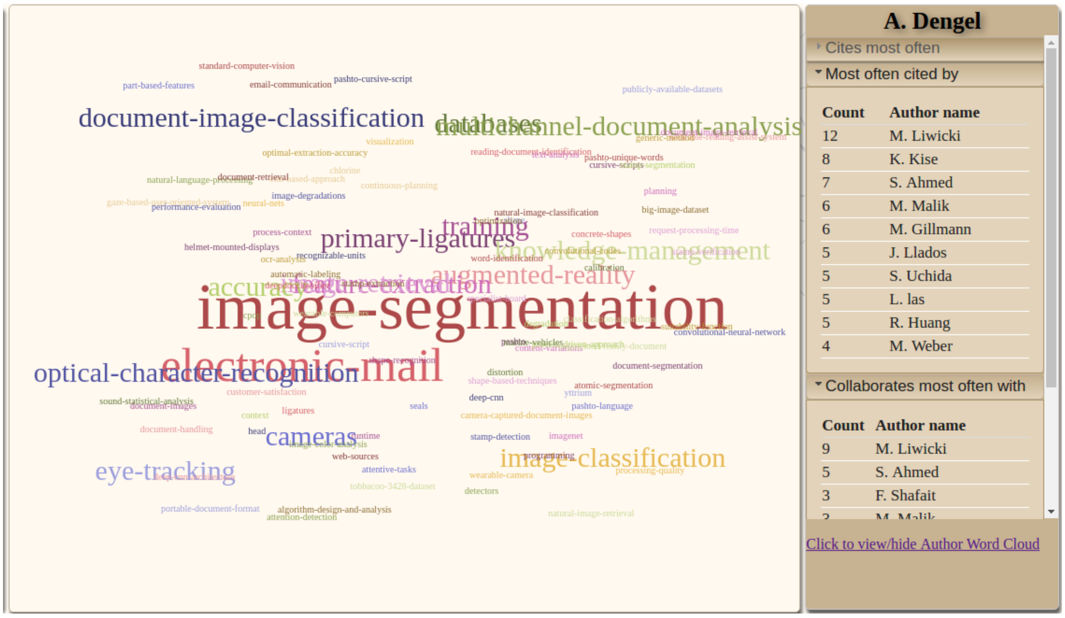


FIGURE 4.10: Author overlap graph



A. Dengel

› Cites most often

˅ Most often cited by

Count	Author name
12	M. Liwicki
8	K. Kise
7	S. Ahmed
6	M. Malik
6	M. Gillmann
5	J. Llados
5	S. Uchida
5	L. las
5	R. Huang
4	M. Weber

˅ Collaborates most often with

Count	Author name
9	M. Liwicki
5	S. Ahmed
3	F. Shafait
3	M. Malik

Click to view/hide Author Word Cloud

FIGURE 4.11: Example of an author's word cloud

Chapter 5

Evaluation

In this chapter, details of the experiments conducted to evaluate the performance of the method presented in Chapter 4 have been described. The first section describes the dataset with its content and context. The second section presents some evaluations on the information extracted.

5.1 Dataset description

It is important to ensure that the data can be accessed, understood and used over time. For this purpose, in this section a description of the dataset is being provided.

The dataset used for the proof-of-concept is a collection of ICDAR articles in PDF format from 1993 to 2015.

5.1.1 Intermediate data

The aim of this section is to provide sufficient information that enables others to understand the content and context in which the data was created and used.

The methodology adopted has been similar to as mentioned in Section 4.1. After all the data-preprocessing steps, three types of intermediate data was obtained:

1) Citation Data: This includes all the information about citations in ICDAR conference from year 1993 to 2015. It is a simple CSV file with two columns, the first one containing the authors who publish an article and the second column containing the authors being 'cited' or 'referrred'.

Before pre-processing, citation dataset contained $299,225$ rows representing citations. After pre-processing (explained in Section 4.1.4), this dataset was reduced to $133,377$ rows in it. This data includes the 6934 rows of self-citations as well.

2) Collaboration Data: This includes all the information about co-authorship in ICDAR conference from year 1993 to 2015. It is a collection of arrays of names, names which are together in one array represent co-authorship.

Before pre-processing, collaboration dataset contained 2763 arrays which were reduced to 2751 after pre-processing step. The total number of collaborations from these 2751 rows was 11097. This data was used for Girvan-Newman clustering to find communities of co-authorship in a network(explained in Section 4.2.1). There were a total of 3783 unique authors publishing in ICDAR from 1993-2015.

3) Author, Keyword Data: This includes information about all authors associated with the keywords from the fields in which they publish articles in ICDAR conference from year 2011 to 2015. It was observed that there were very few papers with 'Keyword' section in papers published before 2011 and the keywords were not well-clustered.

To present this result as a proof-of-concept, keyword data for $2011-2015$ was extracted from the IEEE Bibtex files. It was important to do so in the absence of a proper clustering algorithm for keywords. IEEE keywords are often referenced and well-clustered.

4) Girvan-Newman Grouped Data: The Girvan-Newman hierarchichal clustering algorithm (explained in Section 4.2.1) was applied to the collaboration data of 11097 rows representing co-authorship. Groups at the 40^{th} iteration were used for this task. This was obtained by experimentally verifying with few known groups who co-author frequently. This dataset consisted of arrays of author groups. Authors in one array represent same group.

Apart from the above mentioned primary datasets, few more files need to be generated for the visualisation purposes.

5.1.2 Ground truth

As there is no pre-existing ground truth for ICDAR dataset, a random subset of 100 and 50 documents was created for a quantitative analysis and qualitative analysis of author names respectively.

The random subset of 50 documents contains 795 citation strings in total and 157 author names in the header section and 1976 authors in the reference section. The total number of relevant names are 2133.

Table 5.1: Quantitative Analysis of Reference Extraction: Set 1

Quantitative Analysis: Reference Extraction	Count	Percentage
True positive	99	-
False positive	0	-
True negative	1	-
Precision	1.0	100%
Recall	0.99	99%

The evaluations performed on this random subset are presented in the next section, Section 5.2

5.2 Evaluation Tasks

In this section, various sub-tasks from the pipeline have been evaluated. As the ICDAR dataset has no ground truth, evaluations were done on a random subset of data. In the later sub-sections, a publication and citation meta-analysis is presented along with the preferential attachment properties of the structured data collected from ICDAR dataset.

5.2.1 Quantitative Analysis of Reference Extraction

In this section, two sets of quantitative analysis of reference extraction are performed.

1. This is done on a random subset of 100 PDF documents from ICDAR 1993-2015. To present a justified result, all the PDF documents which have been encoded are not a part of this subset.

   For all the 100 files, there was at least 1 reference extracted for 99 files. In 1 out of the 100 files, the reference section could not be extracted.

   This analysis resulted in a precision of 100% and a recall rate of 99%, refer Table 5.1.

2. In another set of quantitative analysis experiment, a random subset of 50 files was chosen from ICDAR 1993-2015. For each file, the ground truth here was the number of citation strings present in each file. The number of extracted citation strings were compared against the ground truth.

   For a total number of 795 citation strings present in the PDF documents, 766 citation strings were extracted. This results in an accuracy of 96.4%, refer Table 5.2.

Table 5.2: Quantitative Analysis of Reference Extraction: Set 2

Quantitative Analysis: Reference Extraction	Count	Percentage
True positive	766	-
False positive	0	-
True negative	29	-
Precision	1.0	100%
Recall	0.96	96.4%

5.2.2 Qualitative Analysis of Reference Extraction

In this section, a qualitative analysis of reference extraction is done. This is done on a random subset of 50 PDF documents collected from ICDAR 1993 to 2015. To present a justified result, all the PDF documents which have been encoded are not a part of this subset.

As ICDAR dataset has no ground truth, for qualitative analysis of reference extraction, the ground truth for the 50 PDF documents is manually annotated. Qualitative analysis deals with the number of reference strings which have been extracted from total reference strings present in the Reference section.

For the same set of 50 documents used for the quantitative analysis of reference extraction, a ground truth containing the author name from the reference column and the 'header' section in the XML file is created.

Author names are extracted from these 50 files and are compared with the manually annotated ground truth. The results are being presented for authors and author cited differently.

For the set of 50 documents, the recall for the names present in the reference section is 96%, refer Table 5.4. Recall percentage of the names present in the header section is only 77%, refer Table 5.3. The total recall value of author name extraction is 94.6%, refer Table 5.5. These results show that by using bibtex of an article, results would be much better.

It is important to mention here that precision rate is not being reported as these results are before pre-processing [Section 4.1.4] of data. Pre-processing takes place at various stages, where false positives are removed. In this case, it would not have been fair to report precision rate as the code has been modified for retrieving data with a high recall.

Table 5.3: Qualitative Analysis: Author name extraction from reference section

Author name extraction from header section	Count	Percentage
True positive	121	-
False positive	24	-
False negative	36	-
Precision	0.83	83.43%
Recall	0.77	77.07%

Table 5.4: Author name extraction from reference section

Author name extraction from reference section	Count	Percentage
True positive	2018	-
False positive	264	-
True negative	115	-
Precision	0.88	88.43%
Recall	0.95	94.6%

Table 5.5: Overall author name extraction

Overall Author name extraction	Count	Percentage
True positive	2139	-
False positive	288	-
True negative	151	-
Precision	0.88	88.13%
Recall	0.93	93.4%

5.2.3 Comparison of publication and citation in ICDAR 1993-2015

After the results from the quantitative and qualitative analysis of reference extraction methods, it was interesting to observe how the number of publications and citations have increased over all the years.

The number of publications was minimum in 1993 and maximum in 2013. While the number of publications has increased steadily from 186 to 283, the total number of citations has increased from 8844 in 1993 to 57,559 in 2015.

Figure 5.1 shows the graph showing the variation in publication and citation counts in ICDAR community from 1993-2015. The number have been reduced to a log scale.

5.2.4 Preferential attachment in authors

The term *preferential attachment* refers to the observation that in networks that grow over time, the probability that an edge is added to a node with d neighbours is proportional to d. This finding lies at the heart of scale-free network model [2]. This has

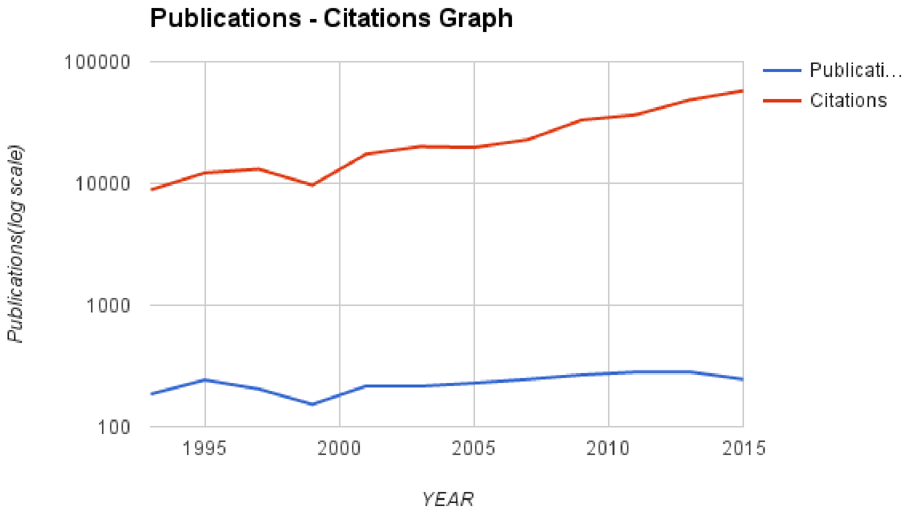


Figure 5.1: Publication and citation count in ICDAR community from 1993-2015.

been reported as a basic property of social networks where a node which is very well connected is likely to receive more new links.

Figure 5.2 and Figure 5.3 shows the preferential attachment property in author citation data according to their indegrees and degrees respectively. There are very few authors who are referred highest number of times, denoting the existence of superstar-phenomenon[41] in ICDAR community as well.

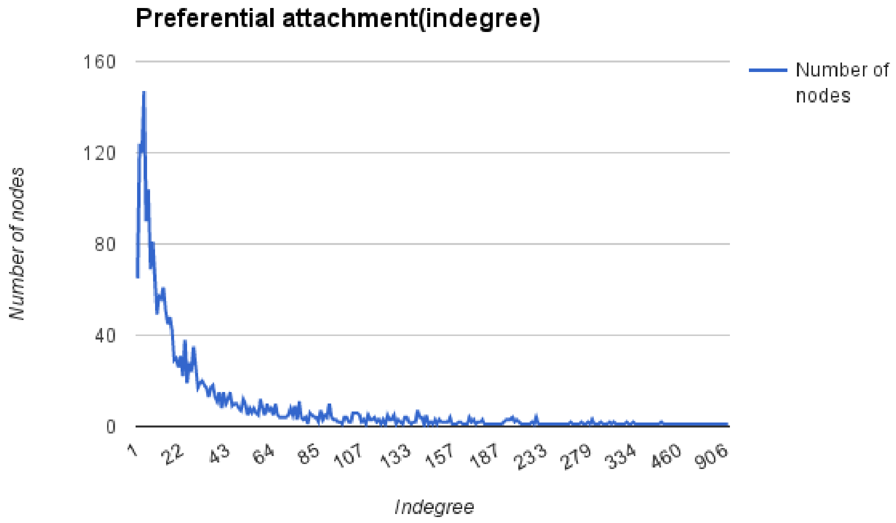


Figure 5.2: Preferential attachment in ICDAR dataset based on indegree of nodes.

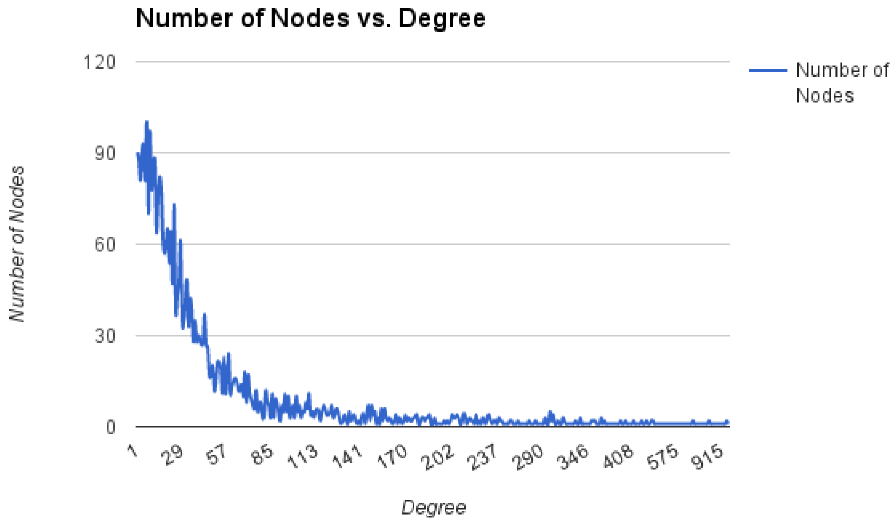


FIGURE 5.3: Preferential attachment in ICDAR dataset based on degree of nodes.

5.3 Insights From The Community

Visualisations help to create a hypothesis and steer the research. There are few initial hypotheses on which this work started and some new which have been found. All these little insights are being described in the following section:

There are various levels of graph according to increasing *citation strength*. It is important to mention that authors who end up in the highest level of graph are not necessarily influential. It just denotes that they have strong ties with other authors of the community. These levels are only for visualisation purposes. Centrality at any level of the graph represents the centrality values as in the real network.

Some of the observations from this dataset are:

- Co-authors often refer each other more often than they refer other authors of the community.
- Often, there are cliques of authors who refer each other much more than they refer others in the community. Such cliques start separating themselves from the central community at a very small threshold of citation strength.
- For an author, there was always an overlap among top authors whom he/she cites most often, authors who cite him/her most often and co-authors. This is more clear in an overlap index graph

- There is a lot of self-citation in the community.
- Most influential author in the community, C. Liu is often referred by others who are not related to each other. This is different than a typical clique for any other influential author. It conveys that the author probably works on several unique fields which are not connected to each other.

Chapter 6

Conclusions and Future Work

This work builds up a foundation towards analysing scientific communities. Section 6.1 concludes the thesis and presents a summary of proposed approach and findings, Section 6.2 proposes directions in which the existing work can be extended.

6.1 Conclusions

This work focused primarily on the network analysis of ICDAR community. Data was extracted through various text mining techniques and a network analysis tool has been developed. There are some initial insights that provide information on the dynamics in the community and supplement the evaluation process.

Chapter 1 and Chapter 2 provided the foundational concepts of this thesis. Chapter 3 mentions the contribution of other authors in this field and Chapter 4 discusses the adopted approach for this work. Further, Chapter 5 discussed the dataset and few evaluation measures on this dataset.

6.2 Future Work

There are several ways in which the initial work done in this research can be extended. The aim of this thesis and the proposed directions is to improve the presented framework.

To improve the current pipeline, it is a good idea to use machine-learning techniques to extract author names or to have a list of authors who publish in ICDAR to reference names in a better way rather than the current approach of shortening first name and appending to last name. Also, work on the presented overlap graph can be extended towards an overlap index as a single measure based on the varios overlap factors discussed in this work.

In a citation network, studying research collaborations across country might help to study collaborations and citations with respect to geographical boundaries. It might be interesting to study a popular author's citation network as compared to others and if it depends on other factors than quality publications.

While this work has been limited to citation networks, another idea would be to build up a semantic network of authors based on the keywords they are working on. It can give a brief time-based overview on main research topics and their relation with other topics in a scientific community. This time-based overview can also be used to study the development of a scientific domain over-time.

Bibliography

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[3] A. Bavelas. Communication patterns in task-oriented groups. *Journal of the acoustical society of America*, 1950.

[4] J. Beel, B. Gipp, A. Shaker, and N. Friedrich. *SciPlore Xtract: extracting titles from scientific PDF documents by analyzing style information (Font Size)*. Springer, 2010.

[5] J. Beel, S. Langer, M. Genzmehr, and C. Müller. Docear's pdf inspector: title extraction from pdf files. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 443–444. ACM, 2013.

[6] C. Belter. Visualizing networks of scientific research. *Online-Medford*, 36(3):14, 2012.

[7] D. Bergmark. Automatic extraction of reference linking information from online-documents. Technical report, Cornell University, 2000.

[8] D. Bergmark and C. Lagoze. An architecture for automatic reference linking. In *Research and Advanced Technology for Digital Libraries*, pages 115–126. Springer, 2001.

[9] D. Besagni, A. Belaïd, and N. Benet. A segmentation method for bibliographic references by contextual tagging of fields. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pages 384–388. IEEE, 2003.

[10] S. P. Borgatti. Centrality and network flow. *Social networks*, 27(1):55–71, 2005.

[11] C. Bron and J. Kerbosch. Finding all cliques of an undirected graph (algorithm 457). *Commun. ACM*, 16(9):575–576, 1973.

[12] C. Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, 57(3):359–377, 2006.

[13] J. R. Clough, J. Gollings, T. V. Loach, and T. S. Evans. Transitive reduction of citation networks. *Journal of Complex Networks*, 3(2):189–203, 2015.

[14] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera. Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, 62(7):1382–1402, 2011.

[15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

[16] E. Cortez, A. S. da Silva, M. A. Gonçalves, F. Mesquita, and E. S. de Moura. Flux-cim: flexible unsupervised extraction of citation metadata. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 215–224. ACM, 2007.

[17] I. G. Councill, C. L. Giles, and M. yen Kan. Parscit: An open-source crf reference string parsing package. In *INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION*. European Language Resources Association, 2008.

[18] J. Cui, F. Wang, and J. Zhai. Citation networks as a multi-layer graph: Link prediction and importance ranking. Technical report, Stanford University, 2010.

[19] E. Delgado López-Cózar, N. Robinson-García, and D. Torres-Salinas. The google scholar experiment: how to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, 65(3): 446–454, 2014.

[20] H. H. N. Do, M. K. Chandrasekaran, P. S. Cho, and M. Y. Kan. Extracting and matching authors and affiliations in scholarly documents. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 219–228. ACM, 2013.

[21] L. Egghe. An improvement of the h-index: The g-index. *ISSI newsletter*, 2(1):8–9, 2006.

[22] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[23] L. C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[24] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[25] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *Digital Libraries, 2003. Proceedings. 2003 Joint Conference on*, pages 37–48. IEEE, 2003.

[26] S. Harnad and L. Carr. Integrating, navigating and analyzing eprint archives through open citation linking (the opcit project). *Current science*, 79(5):629–638, 2000.

[27] I.-A. Huang, J.-M. Ho, H.-Y. Kao, and W.-C. Lin. Extracting citation metadata from online publication lists using blast. In *Advances in Knowledge Discovery and Data Mining*, pages 539–548. Springer, 2004.

[28] D. T. Huynh and W. Hua. Self-supervised learning approach for extracting citation information on the web. In *Web Technologies and Applications*, pages 719–726. Springer, 2012.

[29] M. Jewell. Paratools reference parsing toolkit-version 1.0 released. *D-lib Magazine*, 9(2), 2003.

[30] R. Kern and S. Kampfl. Extraction of references using layout and formatting information from scientific articles. *D-Lib Magazine*, 19(9):2, 2013.

[31] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL http://dl.acm.org/citation.cfm?id=645530.655813.

[32] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics, 2003.

[33] M. E. Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2(2008):1–12, 2008.

[34] A. Pal and S. Ruj. Citex: A new citation index to measure the relative importance of authors and papers in scientific publications. In *Communications (ICC), 2015 IEEE International Conference on*, pages 1256–1261. IEEE, 2015.

[35] F. Peng and A. McCallum. Information extraction from research papers using conditional random fields. *Information processing & management*, 42(4):963–979, 2006.

[36] D. Pinto, A. McCallum, X. Wei, and W. B. Croft. Table extraction using conditional random fields. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 235–242. ACM, 2003.

[37] B. Powley and R. Dale. High accuracy citation extraction and named entity recognition for a heterogeneous corpus of academic papers. In *Natural Language Processing and Knowledge Engineering, 2007. NLP-KE 2007. International Conference on*, pages 119–124. IEEE, 2007.

[38] C. Ramakrishnan, A. Patnia, E. Hovy, and G. A. Burns. Layout-aware text extraction from full-text pdf of scientific articles. *Source code for biology and medicine*, 7 (1):1, 2012.

[39] I. Samoylenko, T.-C. Chao, W.-C. Liu, and C.-M. Chen. Visualizing the scientific world and its evolution. *Journal of the American Society for Information Science and Technology*, 57(11):1461–1469, 2006.

[40] G. Sautter and K. Böhm. Improved bibliographic reference parsing based on repeated patterns. *International Journal on Digital Libraries*, 14(1-2):59–80, 2014.

[41] A. Serenko, R. A. Cox, N. Bontis, and L. D. Booker. The superstar phenomenon in the knowledge management and intellectual capital academic discipline. *Journal of Informetrics*, 5(3):333–345, 2011.

[42] K. Seymore, A. McCallum, and R. Rosenfeld. Learning hidden markov model structure for information extraction. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, pages 37–42, 1999.

[43] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics, 2003.

[44] H. Small. Visualizing science by citation mapping. *Journal of the Association for Information Science and Technology*, 50(9):799, 1999.

[45] N. van Eck and L. Waltman. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538, 2009.

[46] N. J. van Eck and L. Waltman. Visualizing bibliometric networks. In *Measuring scholarly impact*, pages 285–320. Springer, 2014.

[47] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles. Co-ranking authors and documents in a heterogeneous network. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 739–744. IEEE, 2007.